Incineration Technologies

Alfons Buekens
Vrije Universiteit Brussel, VUB, Brussels, Belgium
Zhejiang University, Hangzhou, China

Article Outline

Glossary

Air equivalence ratio Also, air ratio or air factor, (λ or k), is ratio of actual air supply to the theoretical (stoichiometric) requirements for complete combustion.

Combustion residues Ash remaining after combustion and consisting of bottom-ash or clinker, and of fly ash, entrained by flue gas and eventually separated. Chemical neutralization of flue gas also yields salts, by reaction of acid gas components with basic additives.

Emissions Output of pollutants through the stack (= guided emissions), to a minor extent also as diffuse emission, e.g., from waste pit, evaporation of spills, spreading of fly ash, and outgoing leaks.

Gasification Partial combustion generating flammable gas and conducted with deficiency of air in various reactor types.

Higher heating value (HHV) Amount of heat produced by complete combustion of a specific unit amount of fuel in oxygen.

Immission Added atmospheric concentrations attributed to specific sources, e.g., an incinerator plant, and markedly varying with atmospheric conditions. Immissions are modeled on a basis of (a) emissions, (b) their dispersion, and (c) according to variable atmospheric conditions (wind direction and speed, atmospheric stability).

Municipal solid waste (MSW) Waste produced in a city and collected by the municipality.

Pyrolysis Thermochemical decomposition of organic material in the absence of oxygen, yielding gaseous (pyrolysis gas), condensable (tar), and solid products (char).

Refuse-derived fuel (RDF) Fuel from waste, produced by mechanical processing, (possibly biological), drying, and possibly densification.

Waste-to-energy (WtE) Incineration process in which solid waste is converted into thermal energy to generate steam that drives turbines for electricity generators (http://www.businessdictionary.com/definition/waste-to-energy.html).

Definition of the Subject

Waste incineration is the art of completely combusting waste, while maintaining or reducing emission levels below current emission standards and, when possible, recovering energy, as well as eventual combustion residues. Essential features are as follows: achieving a deep reduction in waste volume; obtaining a compact and sterile residue, yet treating a voluminous flow of flue gas while deeply eliminating a wide array of pollutants.

Robert A. Meyers (ed.), *Encyclopedia of Sustainability Science and Technology*, DOI 10.1007/978-1-4419-0851-3,

Destruction by fire is almost as old as humanity. Incineration was systematically applied at some locations, both in England and the USA, from the second half of the nineteenth century [1–4]. Furnaces widely differed in conception, yet were still poked and deashed manually. A successful furnace design was the cell furnace, composed of a series of juxtaposed combustion cells with a fixed grate, or also with two superposed retractable grates [4–6]. In 1895, the first large continental incinerator was mounted in Hamburg [7] after traditional export to the countryside of *municipal solid waste* (MSW) was jeopardized by an outbreak of cholera.

The technology was strongly inspired by that of coal firing: *mechanical grate* stokers developed from the 1920s and 1930s were continuously improved to suit the special requirements of firing waste and distributing primary air, while cooling the grate bars [4, 8]. After World War II, *fluidized bed* techniques were introduced mainly in the Nordic countries, where MSW was cofired together with forest products and residues from pulp and paper industry, and also in Japan, where the suitability of fluidized bed combustors for one- or two-shift operation was valued [9–11]. *Slagging operation*, with tapping of molten residue, remained unusual until the end of the twentieth century; then it became mandatory in Japan to melt fly ash and destroy its organic contents, while either volatilizing or immobilizing its heavy metal content by conversion into a glassy state (vitrification) [12]. A search on "melting" yields more than 130 different processes, as proposed by numerous Japanese corporations [13].

Gasification of waste, a partial combustion conducted with deficiency of air, yields flammable gas, suitable as cleaned gaseous fuel or even for driving engines or turbines [9, 13–16]. This thermal conversion method is mainly apt for high-calorific waste, the complete combustion of which is difficult to control otherwise. Wood waste has been proposed as a decentralized source of heat and power [17, 18].

Pyrolysis, or thermal decomposition of waste [9, 13], may be suitable for specific waste, such as plastics [19], rubber, sewage sludge [20], or wood. These different thermal processes are not to be recommended for general waste, since their process complexity is higher and their availability, hence, lower, whereas most of the advantages claimed often failed to realize [21, 22]. Selecting unproven technology is probably about the worst possible decision in waste management.

Waste varies erratically in *composition* and *properties* and these greatly influence the selection of incinerator furnaces, heat recovery, and flue gas cleaning. Important *waste characteristics* are those determined by *proximate analysis* (moisture, ash, combustibles, subdivided further into fixed carbon and volatile matter) and *elementary analysis* (C, H, N, S, Cl. . . + O, by difference from 100%) of the combustible fraction. Moist refuse is difficult to ignite. Ash content confines the reduction in weight achievable and determines the burden of residue extraction; important is the composition of ash and its softening and melting behavior at high temperature. *Volatile matter* will lead to flaming combustion and *fixed carbon* to glowing combustion, each of these two modes showing their specific demands. Data from these analyses also allow establishing the necessary material balances, as well as estimating the *higher heating value* (HHV).

Combustion of solid waste proceeds in successive steps as schematically represented in Table 1.

In an incinerator furnace, these successive steps may well proceed in parallel and overlap partly. A combination of chemical reactor engineering and of computer fluid dynamics (CFD) may be used in modeling both physical (flow, mass, and heat transfer) and chemical phenomena (combustion, a complex chemical process, proceeding over numerous reactions involving a large number of intermediates, such as free radicals and ions). Cfr. Furnaces, Their Duties, Peripherals, Operation, Design and Control.

Combustion is never *entirely* complete, even though – thermodynamically – equilibrium approach

Incineration Technologies. Table 1 Successive steps in the combustion of waste

Step	Drying	Pyrolysis	Gasification	Combustion
Evolving to the gas phase	Water vapor	Volatile matter	Carbon monoxide, hydrogen, methane	Carbon dioxide, water vapor
Residue	Dry waste	Char and ash	Ash	Ash

could come close to unity. In practice, when most combustibles are burned, the rate of heat generation drops, temperature falls and combustion slows down and eventually stops. The reason for further completing combustion is strictly environmental: *Products of incomplete combustion* (PICs) are major atmospheric pollutants and responsible for reduced visibility, photochemical smog, as well as *soot* or *black carbon* formation. PIC's scientific and health aspects are at the heart of dedicated biannual PIC conferences [23]. Cfr. Post-combustion, Dioxins.

Coarse *combustion residues* (USA: *bottom ash* or slag; UK: *clinker*) are the principal residues of MSW incineration. After removal of unburned material and metals, these may be weathered, graded, and recycled as an aggregate material in sub-road construction and embankments [24–26] Cfr. Residues.

Fly ash is separated by flue gas dust filtration. It is considered *hazardous*, because it accumulates volatilized heavy metals (e.g., Hg, Cd, Pb, and Zn), as well as PICs, some of which are semi-volatile, such as *polycyclic aromatic hydrocarbons* (PAHs) and *dioxins* i.e., polychlorinated dibenzo-p-dioxins (PCDD) and dibenzofurans (PCDF), listed and targeted for removal and destruction by the *Stockholm Convention* on *principal organic pollutants* (POPs). Thermal processes have been applied to detoxify such residues [27–29], yet this treatment is expensive.

Incinerator furnaces. The selection of *furnace types* mainly depends on the characteristics of the waste and the strategies followed to feed the waste to be fired, to contact it with combustion air and to extract the combustion residues from the furnace. Construction of furnaces has evolved mainly empirically, with trial and error as the main method. Tremendous progress made in combustion sciences has started to see some more applications in incinerator design and operation.

Incineration requires sufficient *combustion air*, as well as suitable levels of the three T's, i.e., *T*emperature, residence *T*ime, and *T*urbulence. Turbulence is required to sustain the required macro- and micro-scale mixing to bring together combustibles and air oxygen. Conditions during incineration vary according to the technology employed and the characteristics of the waste fired. Some combustors feature active heat and mass transfer so that combustion takes place much faster, e.g., in vortex or fluidized bed burning. These require, however, size-reduced waste, i.e., preliminary shredding and grading, so that the residence time provided allows either complete burnout even of the largest particles or their recycling after separation.

Temperature ranges from as low as about 750°C (bed temperature of fluidized bed combustion) to more than 1,200°C (destruction of hazardous waste, such as PCBs, slagging operation). High temperatures are only moderately beneficial, for de-mixing of fuel, and oxygen controls combustion rates. *Pressure* is often slightly below atmospheric, to restrict the emanation of combustion products, smoke, and grit. *Residence time* at high temperature is only few seconds (generally 2–3 s) for flue gas. Solid waste and its combustion residue have a much longer residence time, from about a minute in fluidized bed combustion (time required to dry, heat and burnout the ash) to typically half an hour on a mechanical grate; yet, much depends on the time required for drying and heating. After ignition, combustion of volatile matter proceeds rapidly, but burnout of fixed carbon may take time in case of diffusion-controlled combustion, e.g., of ash-occluded carbon.

Some codes prescribe minimum values for temperature and time (e.g., 850°C for 2 s), or they limit the amount of products of incomplete combustion in flue gas (*carbon monoxide*, CO; *total organic carbon*, TOC) and carbon in residues.

Combustion air is supplied to the furnace with several purposes: *primary air* activates the fire bringing oxygen to the reaction surroundings, whereas *secondary air* (also termed over-fire air) is injected at high speed (typically 100 m/s) to induce mixing, as far as its momentum reaches. Increasing primary airflow accelerates combustion until a point at which higher cooling supersedes this stimulating effect. Air may also be used for cooling furnace walls and mechanical grates. Since several decades, water-cooled grates are also in use.

Incinerators are thermal units: liberating more combustion heat also requires supplemental combustion air. A simple rule of thumb states that this amount is directly proportional to the higher heating value, whatever the fuel fired (gas, oil, coal, or garbage of any kind).

In order to obtain complete combustion, it is essential that an adequate amount of air oxygen is supplied. The *air equivalence ratio* indicates the actual air supply, compared to the theoretical, stoichiometric

requirements for complete combustion. The difference, the excess air, merely cools the flame and inflates the volume of gas to be cleaned. Better mixing of fuel and air allow operating at lower air equivalence ratios.

One Mg (metric tonne) of MSW typically generates some 5,000–6,000 m^3 of flue gas! Flue gas flow varies proportionally with both the higher heating value and with the amount of excess air. In numerous plants, the uncontrolled entrance of air leaking into furnace and flues seriously inflates the volume of gas to be cleaned.

During waste combustion, the spatial distribution of flames (formed by combustion of volatile matter) is unpredictable and hence results in erratically active combustion zones, showing oxygen deficiency and less active zones, where oxygen requirements are much less and oxygen plentiful. This results in a complex pattern of oxygen-rich and oxygen-deficient strands that should be mixed intimately in order to reach complete combustion.

Combustion air may also be replaced by oxygen-enriched air, or even by pure oxygen, in order to improve and accelerate combustion. Such practice markedly reduces the volume of flue gas, yet it considerably adds to operating expense and is limited to exceptional cases, such as gasification by means of oxygen/steam mixtures to convert waste into *synthesis gas*.

Municipal solid waste incineration evolved into a complex plant, as represented in Table 2.

MSW storage generally takes place in a deep pit, made of impervious concrete. Storage bridges the gaps between the schedules of collection rounds and continuous firing. A traveling crane allows mixing waste of different origins, stacking waste against the bunker wall, and feeding it into the hopper on top of the load shaft. In the USA, storage floors are in widespread use.

Boiler plant. The heat from flue gas is transferred to the water, boiling in vertical pipe panels, constituting the boiler and organized around the combustion chamber (for an integrated boiler) and in successive vertical passes of the flue gas. An alternative is to suspend boiler tube panels in a horizontal flue gas channel. The resulting medium pressure steam (at 2–4.5 MPa) is superheated in case the steam is used for power generation. At lower temperature, the flue gas preheats the boiler feedwater in an economizer, and possibly the combustion air in a flue gas/air heat exchanger.

Flue Gas Cleaning

Incineration was once a source of smoke and grit. These have been mastered by improved combustion conditions and deep removal of fine dust: once reduced to below 100 mg/Nm^3 by an electrostatic precipitator, the flue gas becomes invisible, a feature that still satisfied the public in the 1950s and 1960s.

Incineration Technologies. Table 2 Composition of current municipal solid waste incinerator plant

Unit	Function	Potential problems
Storage	Bridging the gaps between delivery and firing of MSW	Dust, smells, fires
Crane	Traveling crane to mix and load MSW into a hopper	Mechanical
Hopper	Receiving mixed MSW from the storage bunker	Bridging
Valve	Sliding valve to close the furnace	Mechanical
Shaft	Junction with the combustion chamber	Air infiltration
Furnace	Combustion chamber	Refractory spalling or slagging
Grate	Mechanical grate, supporting, conveying, and poking MSW	Wear, clogging
Burner	Start up combustion, maintain temperature if required	
Boiler	Recovers the heat of combustion from flue gas	Fouling, corrosion, erosion
Dust collection	Separate the bulk of the dust from flue gas	
Scrubber	Acid gas neutralization	Corrosion, erosion, deposits

The German emission code *TA-Luft* (Technische Anleitung zur Reinhaltung der Luft), already in its first version (1974) specified emission levels requiring acid gas levels to be reduced. Since, cleaning the flue gas from waste incineration has steadily become more complex and comprehensive, throughout the 1980s and 1990s. Tables 3 and 4 show both the extent of this gas cleaning duty and the frenetic evolution of these codes in time. The European Union also promulgated successive directives on waste incineration (last directive – Directive 2000/76/EC of the European Parliament and of the Council of 4 December 2000 on the incineration of waste) and prepared codes of good practice (BREF reports: BREF stands for *BAT Reference* Document; BAT = *B*est *A*vailable *T*echnology). Other countries (the USA, Japan, and China) use distinct sets of emission Codes and reporting procedures.

For a good understanding of emission limit values, it is of interest to look at the ratio:

$$\text{Reduction ratio} = (\text{Input value})/(\text{Output value})$$

The reduction efficiencies required (in Table 3) of 95% and 99.9% respectively convert into a reduction ratio of 20 and 1,000, respectively. The first two numbers seem deceptively nearby, separated only by 4.9%; the second, the reduction ratios, come closer to the efforts really required in flue gas cleaning, which differ by a factor of 50!

The present emission values are monitored and registered continuously. Some parameters (O_2, CO_2, H_2O) remain rather constant; others are more variable (HCl) or are marked by a continuous value, spiked by peaks (CO, TOC). Dioxins cannot be monitored continuously, yet may be sampled continuously and checked on a weekly or biweekly basis.

Incineration Technologies. Table 3 Raw gas concentration, emissions, and required separation rate of flue gas cleaning devices (Adapted from [30])

	Raw gas concentration (mg/Nm^3, dry)	Emission limit value (mg/Nm^3, dry)	Required reduction rate(%)
Dust	2,000–10,000[a]	10	99.9
HCl	400–1,500	10	>99
HF	2–20	1	95
SO_2	200–800	50	94
NO_x (as NO_2)	200–400	200	50
Hg	0.3–0.8	0.05	88
Cd, Tl	3–12	0.05	>99.5
Dioxins and furans	1–10 (in ng I-TEQ/Nm^3)	0.1 (in ng I-TEQ/Nm^3)	99

[a]For fluid bed plant these figures are typically 10,000–50,000 mg/Nm^3, dry

Dust Collection

Traditionally, cyclones or *electrostatic precipitators* (ESPs) featuring 2, 3, or 4 consecutive fields were arresting the evolving grit and dust, with an efficiency approaching unity according to an exponential curve. As a consequence, it is increasingly difficult to collect the last particles. Important parameters are the size and electric resistivity of the particles to be collected, as well as their behavior (cake severance or re-entrainment) at the moment of rapping the collection electrodes. Moreover, ESPs operating at temperatures substantially above 200°C were found to generate considerable amounts of dioxins.

Current codes require retention also of the small particles around a micrometer in diameter: even though they correspond to only minor amounts when expressed in mass units (mg/Nm^3), they represent relatively large numbers of particles, strongly enriched in pollutants. *Baghouse filters* (BHFs) are capable of efficiently collecting these particles; moreover, they accumulate a layer of basic substances (injected lime, fly ash) that react with acid gases, such as HCl, SO_2, and HF, from the flue gas and adsorb some semi-volatiles.

Neutralization of Acid Gases

Historically, several solutions have been developed to the acid gas problem: wet scrubbing, dry scrubbing, semi-wet scrubbing, and semi-dry scrubbing. Generally, to neutralize these acid gases, hydrated lime is injected into the flue gas (*dry, semi-dry*, i.e., after further moistening the flue gas, and *semi-wet scrubbing*, using water slurries of lime). Wet scrubbing is even more efficient, since the principal acid gas, HCl, is eminently water soluble; yet it is also more complex and capital intensive because of the necessity of

Incineration Technologies. Table 4 Some milestones in the evolution of emission limit values (Germany and the European Union)

Compound	TA-Luft Germany, 1974	EU directive 89/369	17. BImSchV[b,c] Germany, 1990	Unit
Dust	100	30	10 (30)	mg/Nm^3
HCl	100	50	10 (60)	mg/Nm^3
HF	5	2	1 (4)	mg/Nm^3
SO_2	–	300	50 (200)	mg/Nm^3
NO_x	–	–	200 (400)	mg/Nm^3
TOC			10 (20)	mg/Nm^3
CO			50 (100)	mg/Nm^3
Heavy metals,[a]				
• Class I	20	0.2	0.5	mg/Nm^3
• Class I + II	50	0.2	0.05	mg/Nm^3
• Class I + II + III	75			mg/Nm^3
Dioxins and furans	–	0.1	0.1	ng TE/Nm^3

[a]The comparison is distorted by changes in the definition of various classes
[b]17. BImSchV Ausfertigungsdatum: 23.11.1990. Complete citation: "Verordnung über die Verbrennung und die Mitverbrennung von Abfällen in der Fassung der Bekanntmachung vom 14. August 2003 (BGBl. I S. 1633), die durch Artikel 2 der Verordnung vom 27. Januar 2009 (BGBl. I S. 129) geändert worden ist" Cfr.: http://www.gesetze-im-internet.de/bundesrecht/bimschv_17/gesamt.pdf
[c]The 17th BundesImmissionsSchutzVerordnung gives values for a daily average, as well as for a 30-min average, the latter in parentheses

maintaining a water circuit and treating the resulting wastewater, removing organic compounds as well as sludge and heavy metals. Moreover, wet scrubbing generally is conducted in two steps: in the first, acid scrubbing (pH 0–2) the bulk of HCl is removed and SO_2 follows in the second step, conducted under mild acid or basic conditions (pH 6–8). However, unless the scrubbed flue gas is reheated, wet scrubbing generates a visible plume of condensing water droplets, with its concomitant negative psychological impact. Still, deeper cooling of scrubber liquors also deepens the removal of virtually all pollutants, including mercury and the various PICs (Table 5).

Incineration Technologies. Table 5 Typical stoichiometric factors applied in flue gas cleaning (acid gas neutralization) [30]

Flue gas cleaning	Semi-dry	Semi-wet	Wet
Range (as cited)	2.4 to >3	2.2–3.0	1.1–1.4

Products of Incomplete Combustion – Organic Semi-volatile Micropollutants

In principle, ensuring steady, high-quality combustion and avoiding all combustion upsets should control *products of incomplete combustion* or *PICs*. The latter relate to large masses burning together rapidly and to poor mixing of the intrinsically heterogeneous input.

Much attention has been given to organic *semi-volatile micropollutants* (PAHs, dioxins) that occur in only minute amounts (µg/$Nm^3$and even ng/Nm^3), yet are persistent and bio-accumulating. These compounds are largely removed (>99%) by baghouse filters, after their adsorption onto fine activated carbon particles (typically injected at a dosage of 50–200 mg/Nm^3) or else provided as a fixed adsorption bed.

As an alternative, they are oxidized by means of suitable DeNOx-catalysts, active already at a very low temperature (200°C). A number of preventive measures also allow reducing the formation of PAHs and dioxins (cfr. Dioxins).

Nitrogen Oxides

Nitrogen oxides are formed during combustion, by means of complex free radical and even ionic mechanisms. NO is formed at high temperature and eventually emitted into the atmosphere. In air, slow oxidation of NO takes place, forming strongly oxidizing NO_2. Together with NO, this NO_2 forms an atmospheric oxidizing-reducing system, responsible for the formation of photochemical smog (smog = smoke + fog) and haze. Nitrogen oxides are hence termed "NOx" (NO + NO_2) and generally expressed as their NO_2 equivalent.

NOx in flue gas derives from mainly two sources: the incineration of organic N-compounds (fuel NOx) and incineration at high temperature, e.g., in cement kilns or during slagging operation (thermal NOx and also prompt NOx).

When desirable or required by codes, such NOx can be thermally (*selective non-catalytic reduction*, SNCR) or catalytically reduced (*selective catalytic reduction*, SCR) by means of suitable reducing agents, such as ammonia, urea, amines (N-compounds), hydrocarbons (reburning), and others. Thermal reduction is only possible in a high temperature window, of 760–1,000°C. Catalytic reduction is active already at much lower temperatures, typically 250–450°C.

Another nitrogen oxide is known as nitrous oxide (N_2O), or laughing gas. It forms preferentially at medium-low combustion temperature, such as the fluidized bed combustion of sewage sludge, and during reduction of the conventional NOx. It is a naturally occurring regulator of stratospheric ozone and a major greenhouse gas and air pollutant.

Heat Recovery

Heat recovery has always been central in incineration, and at times waste was regarded as free fuel, yet heat recovery is generally uneconomic in small plants. Some plants incorporate captive uses for the heat produced, e.g., by being linked to district heating systems (Denmark, Sweden) or integrated with civic centers, featuring swimming pools, sauna, and hot baths (Japan), yet generally it is difficult to market the heat produced, so that power generation emerges as a last resort, albeit at limited efficiency. Moreover, the presence of boiler and turbo-generator inflates plant downtime.

Sensible heat is difficult to recover from flue gas, since it is both fouling and corrosive. These limit the possible operating pressure of a *waste heat boiler* (consecutive to the incinerator furnace) or of an integrated boiler, with the furnace fully integrated into its boiler structure (used for highly calorific waste only). Low boiler pressure limits the possible conversion efficiency of steam energy into power. Typically, such conversion efficiency into power is only 16–24%, based on the HHV of waste compared to better than 40% for large fossil fuel–fired thermal power plants (cfr. Heat Recovery).

Co-firing. Waste can also be co-fired in non-dedicated thermal units, such as thermal power plants, cement or limekilns, and in large industrial boilers. Not all waste is suitable, though, because of both combustion and gas cleaning considerations. Table 6 lists the

Incineration Technologies. Table 6 Some specifications for RDF to be fired in cement kilns [31]

Element	Typical value (ppm)	Limit value (ppm)	Hazardous waste* (ppm)
As	9	20	300
Be	0.4	2	50
Cd	3	5	(+ Tl) 90
Co	8	15	300
Cr	40	120	3,000
Cu	100	150	3,000
Hg	0.6	1	5
Mn	50	150	2,500
Ni	50	100	2,000
Pb	50	100	2,000
Sb	25	60	150
Se	5	10	80
Sn	10	40	1,500
Te	5	20	80
Tl	1	2	
V	10	20	1,500
Zn	n.a.	n.a.	15,000

Source: Reference [31]

typical requirements for co-firing in cement kilns (cfr. Co-firing of Waste or of RDF, Thermal Power Plants, Cement and Lime Kilns).

Cost and Plant Availability

Incineration is a technically complex and expensive operation. In the European Union, an all-in cost factor for MSW incineration is ca. 100 €/Mg (1 Mg = 1 metric tonne). In Japan, this cost is about three times higher. Internal comparison is difficult, because of highly variable cost factors corresponding to buildings, other infrastructure and, in Japan, land.

Plant availability typically ranges from 84% to 92%, the latter catering for an annual shutdown, the former accounting for repeated and unscheduled stops. Availability heavily depends on the quality of plant management and maintenance.

Public Acceptance

For a variety of reasons, environmentalists have fought incineration as a waste management option: it is not natural (like composting), destroys recyclables, and generates toxic compounds. This opposition is often termed the *not in my backyard* (NIMBY) syndrome and is sometimes counterproductive to the development of adequate solutions on a sound technical and economic basis. (cfr. Public image of Incineration).

Whatever the quality or foundation of the arguments against incineration, the design and operating standards have been much further improved over recent years and today's incinerator emission standards are probably the toughest in industry.

Introduction

This introduction situates the position of waste incineration in a wider scope of waste management. Traditional waste management was limited to the three options: landfill, composting, and incineration. Landfill was suitable for reclaiming low-value lowlands or restoring the landscape affected by mines and quarries (sand, gravel, clay). Some large cities (e.g., London!) used MSW to fill lowlands, as well as empty quarries of sand, gravel, or clay, to build artificial islands (Tokyo), or even dumped MSW into the sea (New York, Istanbul). Lack of preliminary hydrogeological study and of adequate barriers to contain the leachate has led at times to serious contamination of groundwater. Moreover, landfills are responsible for important high greenhouse gas emissions (methane, carbon dioxide). Composting is still applied nowadays on selectively collected organic fractions; raw MSW yields an unacceptable quality of compost, due to the presence of heavy metals. Incineration has been widely practiced in densely populated regions, where land is at a premium (large municipalities, Japan, Switzerland) and volume reduction primordial.

The 1970s introduced numerous new concepts into waste management, such as the concept of special (Germany), poisonous (England), toxic (Belgium), chemical (the Netherlands) or otherwise hazardous waste (USA, OECD), producer responsibility, the Polluter Pays principle, and mandatory recycling. In the early 1970s, the European Union declared itself competent in environmental matters and the first Framework Directive on Waste (1975) specified the necessity of appointing authorities responsible for waste management, granting licenses, and inspecting waste processing premises. A number of waste streams received particular attention, e.g., hazardous waste, PCBs, waste oil, and packaging. Industrialized countries were repeatedly confronted with waste scandals; industrial and hazardous waste infrastructure was set up step by step and became a booming business. The lowest possible cost disposal was gradually replaced by high-tech, high-cost options. This transition was smoothened through subsidies supporting the options preferred by government and through levies penalizing low-cost landfill. Waste management was borne by the public sector, the private sector, or by public-private initiatives.

According to the Ladder of Lansink (after Dr. Ad Lansink who is a Dutch politician famous for proposing this waste management hierarchy in the Tweede Kamer [Dutch Parliament] in 1979), the generation of waste should in the first place be either prevented or reduced. Next options are reuse and recycle. Lower-ranking options are incineration (preferably with heat recovery), and landfill. Waste management is a legislation-driven business. In several EU countries and in Switzerland, the landfill option is increasingly restricted, so that combustible waste can no longer be landfilled.

Developing countries are often confronted with fast urbanization, so that public services cannot follow demand. Moreover, waste is rich in organics and barely combustible. Large Chinese cities at present are entirely surrounded by a girdle of landfills, polluting groundwater and generating hazardous fermentation gas. Incineration makes rapid progresses, using imported as well as adapted self-developed technology. The severe acute respiratory syndrome (SARS) was material in promoting incineration, in particular for hospital waste. As in numerous developing countries, Chinese MSW is still barely combustible, without resorting to auxiliary fuel!

Evaluation of Waste Incineration

In brief, waste incineration can be summarized as follows.

Advantages

- It eliminates objectionable and hazardous properties, such as being flammable, infectious, explosive, toxic, or persistent.
- Putrescible matter is sterilized and destroyed. Pathogen count becomes low and generally negligible, except in cases of deficient operation.
- It thermally treats solids while realizing a large reduction in volume, for MSW often by a factor of 10 or more.
- It destroys gaseous and liquid waste streams leaving little or no residues, except for those linked to flue gas neutralization and treatment.
- The heat of combustion generated may be put to good use.

Disadvantages

- Incineration is technically a complex process, requiring huge investment and operating cost as well as good technical skill in maintenance and plant operation, in order to conform to modern standards.
- Heat recovery takes place under adverse conditions (boiler fouling, erosion, corrosion) and is often costly and inefficient.
- Incineration generates an amount of pollutants which are not easy to control.
- Complete burnout of flue gas and residues needs to be ensured.
- As emission codes become more stringent, operating costs rise and the volume of secondary waste streams requiring further disposal increases (in decreasing order with dry, semi-wet, or wet gas scrubbers).

Some types of waste are banned from incinerator plants, unless they are specifically equipped to cope with such waste, e.g.:

- Volatile metal (i.e., principally mercury, thallium, and cadmium) bearing waste.
- PCB-containing waste, which requires special incinerators with unusually high destruction efficiency.
- Radioactive waste. The absence of such waste is now routinely checked in MSW, due to widespread use of medical radioactive preparations for either diagnostic or treatment purposes. Radioactive waste can be incinerated like other waste, with (a) volume reduction and (b) immobilization of radionuclides in ash as major aims; yet, containment is essential. Incineration may hence be conducted under slagging conditions. Dust filters should substantially retain all dust.

Waste Incineration

Waste Incineration can be described as "the controlled burning of solid, liquid or gaseous combustible wastes so as to produce gases and residues containing little or no combustible material" (Ph. Patrick, 1980. Past president of the Waste Management Institute (UK)). The technique is now considered from various viewpoints:

- Waste streams of interest
- Phenomena in waste incineration
- Stoichiometry
- Mass balances
- Incineration products
- Residues
- Thermal aspects
- Furnace capacity
- Safety aspects
- Incinerator furnaces – principles – operations – fields of application
- Post-combustion

- Heat recovery
- Corrosion problems
- Flue gas composition and cleaning
- Dioxins

Next, the major types of incinerator furnaces and the conversion of waste into refuse-derived fuel are discussed.

Waste Streams of Interest

Incineration generally addresses combustible waste, whether it is gaseous, liquid, sludge, paste like, melting or solid. Particular streams are municipal solid waste (MSW); commercial, industrial, and hazardous waste; sewage sludge; and hospital waste. Waste that fails being auto-combustible can still be incinerated by means of auxiliary fuel.

Municipal solid waste (MSW) has been routinely analyzed by manual sorting (and sieving of fines) in the Netherlands even on an annual basis and for different types of residential areas (TNO). Argus, Berlin, produced a very much detailed analysis in the early 1980s [32, 33]. Each major sorting fraction (fines, vegetal matter, paper and board, plastics, etc.) was analyzed for its pollutant contents (elementary composition, heavy metals, and dioxins).

Industrial process streams can be very diverse, e.g., gaseous, aqueous, and organic effluents from the most diverse industrial processes, sludge and dust from treating such effluents, waste oil and solvents, and, finally, solid waste. Process waste with stable characteristics is often disposed in-plant, in boilers, or furnaces. Occasional waste and small arising is stored in empty drums, bags, or barrels, grouped and sent to waste disposal centers. Some large factories, such as *BASF* (Ludwigshafen) or *Ford* (Cologne), have operated their own centers since the 1960s. The community operates some comprehensive centers (Denmark, Bavaria); private or public/private entrepreneurs manage others.

Green waste (branches, brush, and logs) may be collected separately for shredding and/or composting or for use in waste-to-energy (WtE) schemes.

Sewage sludge is also a generally occurring municipal waste, mainly consisting of water, so that mechanical dewatering and drying yield tremendous reduction in volume. Co-firing has been practiced many different ways, in mass burning, power plant, etc. Dedicated furnaces are mainly fluidized bed, multiple hearth, and rotary kiln.

Bulky waste or bulky refuse relates to waste types too large to be accepted by the regular waste collection, such as discarded furniture, large household appliances, and plumbing fixtures. The tendency to incinerate such items directly has declined: bulky waste is diverted increasingly for reuse and recycling; what remains is shredded before incineration. Some plants for bulky loads were operated on a full-day burning, nighttime cooling cycle. For fuel economy and especially for environmental reasons, such practices are no longer recommended. Dismantling for recycling and shredding of nonrecyclables is a better option.

Automotive shredder residue (ASR) often contains hazardous substances such as lead, cadmium, and PCBs. Some countries have classified ASR as hazardous waste and have established legislative controls.

Hospital waste is another stream often earmarked for incineration. Its composition varies with local systems for segregated collection. Specific compounds are sharps and disposables and infectious waste. Hospital waste is often incinerated in a two-step process, first partial oxidation then high-temperature post-combustion of fumes, derived from the pyrolyzer. There is a tendency to concentrate incineration in centralized units rather than in scattered and ill-managed small local plants.

Hazardous waste as a rule loses its hazardous properties during incineration. The hazards are more relevant during collection, storage, and pretreatment than during incineration proper. One category of waste stands out: *chlorinated waste* can best be fired to eliminate its persistent, lypophilic, and bioaccumulating properties. Alternatives, such as dehalogenation exist, yet are an order of magnitude more expensive. Particular aspects of chlorinated waste incineration are:

- Very high combustion efficiency ($\eta_{Comb} > 0.9999\ldots$) is required.
- Hence, a minimum combustion temperature of 1,200°C is stipulated.
- The Deacon reaction (forming chlorine gas) is avoided by operating with minimal excess of air, possibly addition of steam (both to steer the

equilibrium), and fast cooling or even water quenching (to freeze the high temperature composition).
- The formation of phosgene ($COCl_2$) is also avoided, by reducing Cl_2 formation and striving for complete conversion of CO into CO_2.

Dedicated thermal units are developed for *recovery* and cleaning of metal or metal parts, contaminated with paint, lacquers, or polymers.

Important Properties of Waste

Important properties of waste are related to:

- Storage behavior and potential hazards during storage (cfr. Safety Aspects)
- Form and size of individual particles and their distribution, physical and bulk density, specific surface, angle of repose
- Flammability and putrescence of *solid wastes*
- Bulk and physical density, viscosity, heat conductivity, reactivity, explosion limits, flash point, ignition temperature, vapor pressure, boiling point, gas evolution or decomposition during preheating, corrosiveness, toxicity, possibility of auto-oxidation, spontaneous polymerization or other incontrollable, exothermic or dangerous reactions of *liquid wastes*
- Density, explosion limits, toxicity, and corrosiveness of *gaseous wastes*.

Waste Gases – Liquids - Solids

The *heat of combustion* of pure chemical compounds is simply derived as the difference between the heat of formation of products and reactants.

The *higher heating value* (HHV) of fuel is derived by burning a known amount with oxygen in a bomb calorimeter and monitoring the amount of heat liberated that is largely transferred to the water mass surrounding the combustion chamber. The resulting temperature rise is proportional to the heat liberated; heat losses to the surroundings are corrected for by calibration. Several empirical formulas were developed to estimate the heat of combustion of a fuel, from its elementary composition, e.g., the Dulong equation (originally developed for coal and later modified to accommodate a variety of fuels, including municipal solid waste).

The heat of combustion of the combustible fraction of refuse is given by:

$$327.81(\mathrm{C}) + 1504.1(\mathrm{H} - \mathrm{O}/8) + 92.59\mathrm{S} + 49.69\mathrm{O} + 24.36\,\mathrm{N} \quad \mathrm{kJkg}^{-1} \tag{1}$$

In this formula C, H, O, S, and N stand for the mass percent of each of these elements. These are expressed on a moisture and ash-free (maf) basis. More formulas are cited in Niessen [35].

The *lower heating value* (LHV), also termed *calorific value*, is lower, because from the HHV value one must subtract the *latent heat of condensation* of water vapor present in the flue gas, but which generally is lost with the flue gas in the plume.

The *proximate analysis* establishes the moisture and ash content (wt.%) and – by difference – the combustible part of the waste (wt.%). Thus the proximate analysis defines the amount of moisture to be evaporated prior to combustion and the required dimension of the ash handling equipment. Moist wastes, such as garbage, sewage sludge, and aqueous solutions, burn only after evaporation of most of the moisture contained. Hence, adequate measures should be taken to ensure fast and complete drying.

The *elementary* or *chemical analysis of the combustibles* should be known in order to estimate the composition of the flue gas at a given excess of air and to determine whether wet or other scrubbing of the flue gas is required.

The other properties are helpful to select and specify the waste storage, handling and feeding facilities and the required safety provisions. Information is also required on the frequency and timing of the deliveries, the kind of containers and packaging, etc.

Individual gaseous combustible compounds are characterized by means of their chemical formula and structure and molecular mass (often termed molecular weight). Density is proportional to molecular mass, which is easily derived as the sum of all atomic masses. Denser gas requires proportionally more combustion air and hence a larger supply of air to the burner. The HHV is roughly proportional to the mass of fired gas. Gases are also often characterized by their Wobbe-index, a factor combining HHV and density.

Important properties for *liquid fuels* or waste are viscosity, density, flash point, surface tension, sooting

tendency, etc. These affect oil atomization and combustion, as well as burner construction, operation, and maintenance (Table 7).

Solid fuels or waste vary in chemical composition and thermal behavior. Coal consists of highly condensed aromatic structures capable of thermal softening, melting, and decomposing. Depending on its rank, coal generates combustible gas and volatiles during combustion, giving rise to flaming combustion and leaving a carbonizing residue. Biomass predominantly consists of cellulose structures, bounded by lignin. Worldwide, it is still an important fuel; yet, it loses much of its importance in terms of industrial use and trade.

Phenomena in Waste Incineration

Combustion science has evolved enormously, with respect to both theoretical concepts and experimental study. Some relevant references as well as past and ongoing conferences are cited in the general bibliography. Incineration is much more an empirical engineering science [35–38]. The last reference provides a state-of-the art review, composed on the basis of European experience.

Combustion of *flammable gas* follows two distinct modes: fast combustion in premixed flames (*mixing is burning*) and diffusion-controlled flames, those relevant in this context. Since waste flammable gases are difficult to store in oil refineries or petrochemical plants, they are commonly disposed of by either elevated or ground *flares*. Severe sooting may occur during an emergency, when large flows need to be flared. Sooting is reduced by addition of steam through ejectors located in nozzles that draws in ambient air. Smaller, better controllable gas streams are often burned in available boilers or furnaces. Where necessary, they are combusted either thermally in a dedicated yet simple combustion chamber, or catalytically on a fixed catalytic bed.

Combustible liquid wastes are generally fired, dispersed into fine droplets, each of which burns as a small entity, composed of evaporating liquid and diffusion flames around the periphery.

Solid fuels first dry, and then thermally decompose while heating, with evolving volatiles sustaining flaming combustion and the charring residue much slower glowing combustion. Converting fuel or waste into volatiles and fixed carbon is an essential step (pyrolysis) in their combustion. Mimicking this process is an essential test; for coal this test was standardized differently in each industrial country, yet 950°C is a typical temperature for defining the split between volatiles and fixed carbon. Heating rate applied and test duration also influence this split (Fig. 1).

In practice, these steps proceed partly in parallel, rather than in a strict sequence.

Drying is a gradual process: moisture can be absorbed quite loosely, e.g., by plastics, or firmly, physicochemically bound to its substrate.

Incineration Technologies. Table 7 Some models for combustion of liquid and solid particles

Model	Hypotheses				
	Number of compounds	Surface temperature compared to gas temperature	Heat exchange	Oil thermal conductivity	Diffusion in droplet
The d^2 law	One	Lower	Radiation	High	–
Scale model	One	Comparable		Nil	–
Homogeneous temperature	One	Comparable	Radiation + losses	High	–
Diffusion control in droplets	Several	From enthalpy balances	Rate laws	Rate laws	Species balances
Direct simulation	Several	From enthalpy balances	Rate laws	Rate laws	Species balances

Source: After Görner K [34]

Incineration Technologies. Figure 1
Flaming combustion of solids [39] (By courtesy of Wikipedia)

All organic materials decompose upon heating, generating generally smaller and simpler molecules. The emerging volatiles contain inorganic (CO, CO_2, H_2O, H_2, etc.) as well as aliphatic and aromatic organic compounds; their product distribution depends on numerous factors, such as raw materials, temperature, residence time of volatiles and solid fraction and – not in the least – catalytic effects exerted by ash, bed material, or furnace walls.

Primary pyrolysis products show structures close to those of the molecules pyrolyzed. The longer the residence times, the more these structures evolve toward thermally more stable molecules. Ultimately, mainly carbon, hydrogen, and water vapor remain when pyrolysis is concluded in the absence of air.

Cellulosic compounds, such as paper or wood, decompose already at ca. 250°C according to quite complex mechanisms that thermally soon become self-sustaining. Some plastics, conversely, follow simple-looking *unzipping* mechanisms, yielding monomer or oligomer (low polymers, such as di-mer, tri-mer, etc.) as a product. This is the case for, e.g., polymethylmetacrylate (PMMA) and polystyrene (PS).

Vinyl compounds (polyvinylchloride, polyvinyl alcohol, polyvinyl acetate) decompose at unusually low temperatures, releasing hydrochloric acid HCl, water, and acetic acid (CH_3COOH), respectively.

PVC also decomposes in two steps. HCl evolves almost quantitatively from PVC between 225°C and 275°C. This step also produces some benzene. The second step yields further, mainly aromatic compounds, by internal cyclization [40].

Polyolefins, such as polyethylene and polypropylene pyrolysis attains a maximum rate of decomposition at ca. 450°C [41]. Primary products are polyolefinic and paraffinic chain fragments, following a Gaussian molecular weight distribution: higher temperature generates in average shorter product molecules. Secondary products from polyethylene, as well as primary products from polypropylene, show more branched chain products.

Solid waste incinerators generally feature a *mechanical grate* that supports, conveys, and pokes the waste, while primary combustion air activates the fire and cools the grate. Traveling grates, roller grates, and reciprocating grates show *plug flow* characteristics, i.e., an almost even residence time for the different refuse parcels that move through the furnace. This leads to successive zones of drying, heating, ignition, and flaming combustion of waste, and residue burnout. *Reverse-reciprocating grates* create back-mixing, by pushing the burning waste upstream, underneath the incoming fresh refuse.

Incinerators burn highly flammable plastics, side by side with wet vegetal waste. Once heated high enough (>400°C) for fast pyrolysis to occur, plastics decompose swiftly and hence burn rapidly, creating oxygen-deficient flames and leaving craters in the original refuse layer. On the other hand, wet waste is slow to ignite, for first it must be superficially dried before it can start rising in temperature, generating flammable vapors, and eventually catching fire. Even then, large lumps of moist vegetal matter may remain wet internally and survive incineration. Also, massive wood, or a thick book, takes time to burn, the carbonized material thermally insulating the flammable core.

Thus, burning refuse is heterogeneous and produces strands of oxygen-deficient hot gas as well as

other gases, still prior to ignition and composed of moist air, charged with smelly products, arising in drying and heating. Unless hot oxygen-deficient and cold oxygen-rich flue gas strands are thoroughly mixed by blowing in secondary air at high speed, products of incomplete combustion will likely leave the furnace unconverted (Fig. 2).

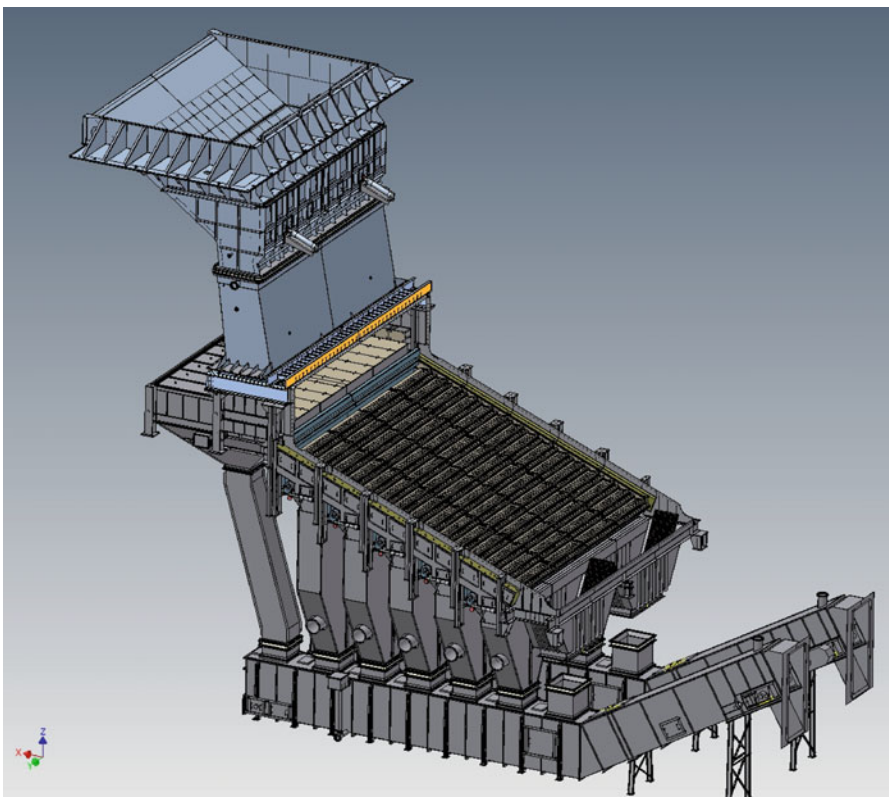
Draft is the most important physical factor determining incinerator capacity. Only the smallest units operate on natural draft, as generated by the chimney. Fans (forced draft) blow in primary and secondary air; a much larger fan in front of the stack provides induced draft. The stepwise extension of flue gas cleaning, necessitated by past progression of the cleaning levels, has inflated the head losses and increased the required capacity and the power consumption of induced draft fans.

The *residence time* of gaseous and liquid wastes in an incinerator amounts to only few seconds. The residence time required for complete combustion of solids is generally about half an hour.

Hence, incinerator feed should always be made as *homogeneous* and *constant* as possible, e.g., by mixing, blending, and for municipal solid waste (MSW) ageing, to provoke moisture transfer and to account for a wide difference of flammability between easily igniting plastics on the one hand and moist vegetal waste on the other.

Stoichiometry

Gaseous and liquid waste can be completely combusted using *low excess of air* (5–15%) as far as their composition is sufficiently predictable and constant and mixing of air and fuel well organized. In principle, much more excess is required when firing solid waste, except in incinerator types featuring first-rate air/solids contact, e.g., fluidized bed or vortex units. Lower airflow



Incineration Technologies. Figure 2
Representation of a mechanical grate incinerator (By courtesy of Keppel-Seghers)

also has other advantages: it elevates the combustion temperature, extends the residence time in a given furnace volume, and reduces entrainment of fly ash with flue gases, as well as thermal losses with flue gas in the stack.

The amount of *combustion air* provided markedly exceeds stoichiometric requirements, following from formal reaction formulas, such as:

$$C + O_2 \Rightarrow CO_2 (12g + 32g = 44g) \quad (2)$$

$$2H + \frac{1}{2}O_2 \Rightarrow H_2O (2g + 16g = 18g) \quad (3)$$

$$S + O_2 \Rightarrow SO_2 (32g + 32g = 64g) \quad (4)$$

Combustion equations are normally marked in atomic or mol units; the corresponding weight amounts are marked in grams. Under standard conditions 1 mol of gas has a volume of 22.4 l or dm^3.

Mass Balances

The *Law of Mass Conservation* also applies to incineration. Hence, the sum of all *input streams* equals the sum of all *output streams*, whether

- In *total* mass flows (kg/h)
- *Any individual element* entering and leaving the plant, and expressed either in mass units (kg/h) or in number of moles (mol/h). In combustible waste, the main elements (symbol, atomic mass) are carbon (C, 12), hydrogen (H, 1), oxygen (O, 16), sulfur (S, 32), nitrogen (N, 14), and chlorine (Cl, 35.5).

Input streams are typically (1) waste, (2) auxiliary fuel (when needed), and (3) primary and secondary combustion air and also uncontrolled air entering through leaks. The latter can be estimated along the flue gas path, simply by measuring the rising oxygen or the declining carbon dioxide content of the flue gas.

During *flue gas cleaning*, additional compounds may be added, such as basic additives (hydrated lime $Ca(OH)_2$, lime CaO, or even – at high temperature – ground limestone $CaCO_3$, and also sodium bicarbonate $NaHCO_3$ or hydroxide NaOH), ammonia or urea (DeNOx), and activated carbon, as an adsorbent for principal organic pollutants (cfr. Flue Gas Treatment).

Typical *output streams* are grate siftings, bottom ash, boiler slag, fly ash, flue gas neutralization residues, and cleaned flue gas. In some plants, the different flue gas treatment residues are extracted as a mixture, in others separately.

Mass balances directly relate input streams to output streams.

One Mg (tonne) of MSW requires 6.5–7.8 Mg (5,000–6,000 Nm^3) of combustion air. Typically, mechanical grate incineration generates (EU conditions):

- 250–350 kg of bottom ash
- 5–15 kg of boiler slag
- 20–40 kg of fly ash
- 5–15 kg of neutralization salts
- 7–8.6 Mg of flue gas

Incineration Products

On the basis of aforementioned balances, the amount and identity of the incineration products can now be derived.

The *proximate analysis* splits the waste to be fired up into three parts: (1) moisture W, (2) ash A, and (3) combustibles C. The first reports to the flue gas, the second forms the residues, whereas the third is converted into combustion products also reporting to the flue gas. Starting from W% + A% + C% = 100, Tanner represented waste composition in a triangular diagram, in which a zone of auto-combustible MSW is identified (Fig. 3).

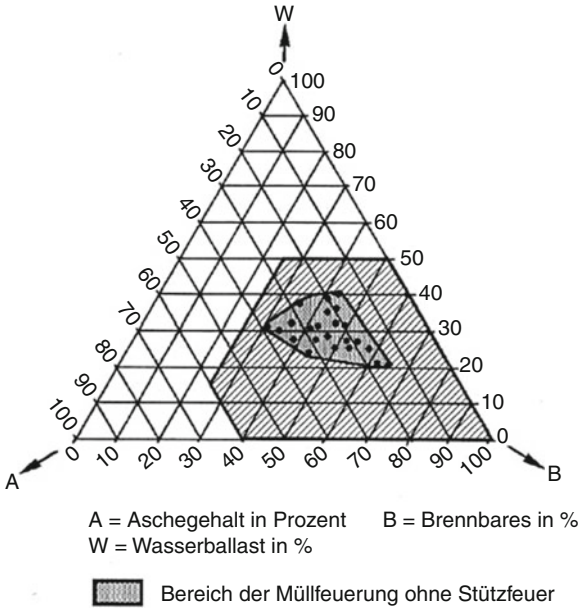
Slight disparity occurs between ashes, as originally present in waste, the "real" ash resulting during the proximate analysis test and that formed during incineration proper. Depending on the ash minerals on the one hand and the combustion conditions on the other, the original ash may differ from actual incinerator ash, because of occurrence of various thermal reactions, such as,

$$Ca(OH)_2 \Rightarrow CaO + H_2O \quad (5)$$

$$CaCO_3 \Rightarrow CaO + CO_2 \quad (6)$$

$$CaSO_4 + C \Rightarrow CaO + SO_2 + \frac{1}{2}CO \quad (7)$$

as well as many others that can only be identified by a detailed study of the ash minerals through methods such as X-ray fluorescence (XRF) or scanning electron microscopy (SEM), their thermodynamic stability, potential reactions, and state of conversion.



Incineration Technologies. Figure 3
The Tanner diagram [37]

Generic classes of such reactions are: dehydration, decarbonatation, sulfate decomposition, and decomposition of higher oxides into lower oxides. Another reason for disparity is the occurrence of volatilization at flame temperature; such volatilization depends on temperature, presence of oxidizing or reducing conditions and speciation [124]. Halogenides (chloride, bromide) are much more volatile than oxides or sulfides, carbonates, and sulfates.

Coarse or sintered ash materials report to the bottom ash, fines are at risk to be entrained. Bottom ash consists of coarse objects, such as stones, glass, or cans, and of ash proper. Low burnout temperatures preserve the original ash structures; high temperatures first cause sintering, generating larger and more solid sintered structures, and eventually more and more fusion.

The major gaseous incineration products are carbon dioxide and, to a variable extent, water vapor and, of course, a large amount of air nitrogen. *Carbon dioxide* generation is directly proportional to the amount of carbon burned, since the background carbon dioxide content in combustion air is insignificant (0.03 vol.%), compared to carbon dioxide in flue gas. This carbon dioxide concentration varies widely, from few vol.% to about 12 vol.%, depending on waste composition and on the excess air used.

Incomplete combustion leads to the formation of *carbon monoxide (CO)*, *total organic carbon (TOC)*, and *black carbon (BC)* or *soot*. The amount of carbon monoxide formed is highly variable, with generally a stable background value, spiked by rare or more frequent peaks (from less than 1 ppm to peaks of some 10,000 ppm, or 1 vol.%, occurring only during combustion upsets), yet only rarely influences the carbon dioxide content (Fig. 4).

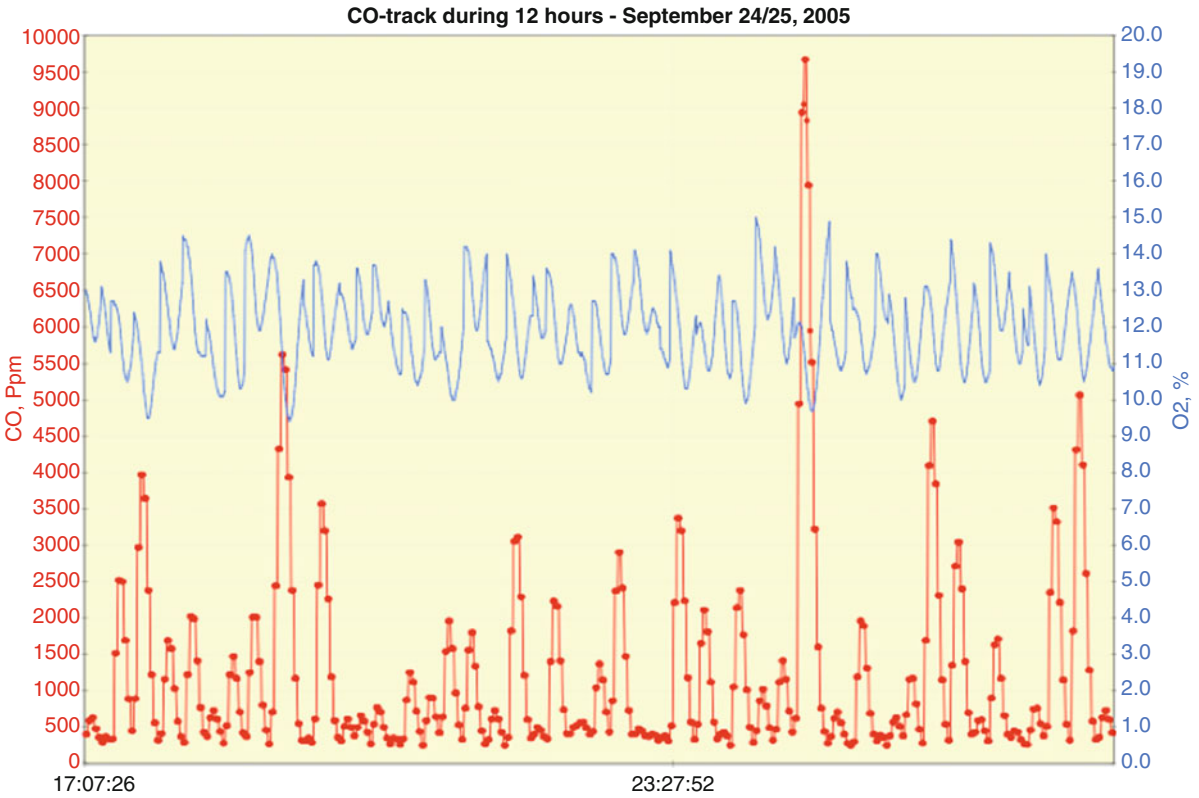
The *moisture content* of flue gas is composed of:

- The original *moisture content* of the fuel fired, which upon drying reports to the flue gas. This is normally negligible for oil and gas fuels, but it reaches several percentages for powdered coal, and is quite substantial for peat, lignite, most forms of biomass, such as sewage sludge or green wood, and for municipal solid waste (MSW).
- *Moisture* contained in *combustion air*, varying markedly with both temperature and relative humidity.
- *Chemically formed water*, derived from the hydrogen content of fuel. The amount can easily be derived by simple stoichiometric computations, based on reaction equations such as:

$$[C_6H_{10}O_5]_n + 6nO_2 \Rightarrow 6nCO_2 + 5nH_2O \qquad (8)$$

with, e.g., five volumes of water vapor formed per anhydro-cellulose unit $C_6H_{10}O_5$ (the cellulose monomer) fired, or in mass units 90 g of water vapor formed per 162 g of solid fuel.

- Water added and evaporated during *quenching* of flue gas by water injection, a usual practice in small incinerators and in the incineration of chlorinated waste.
- *Pre-conditioning* of flue gas, prior to scrubbing, to enhance the elimination of fine dust, HCl, and SO_2. The first become denser, the acid gases are absorbed more easily by hydrated lime in the presence of water vapor.
- Water evaporated in wet scrubbers, used for scrubbing out acidic gases. This treatment saturates the flue gas with water vapor; the resulting temperature is typically 65°C. Sometimes, the scrubbing water is



Incineration Technologies. Figure 4
Time evolution of carbon monoxide as a function of time [42]

cooled by heat exchangers to obtain a deeper separation of various pollutants (e.g., mercury, soluble gases, and organic vapors).

Oxygen, present in the fuel, reduces the amount of combustion air required, but does not contribute to the heating value. Heteroatoms, such as sulfur, nitrogen, chlorine, and other halogens may contribute to air pollution, since they are converted largely (sulfur, chlorine) or partly (nitrogen) into pollutants. Still, flue gas cleaning will eliminate the resulting pollutants, down to the limit values specified (cfr. Tables 3 and 4).

Formation of Pollutants

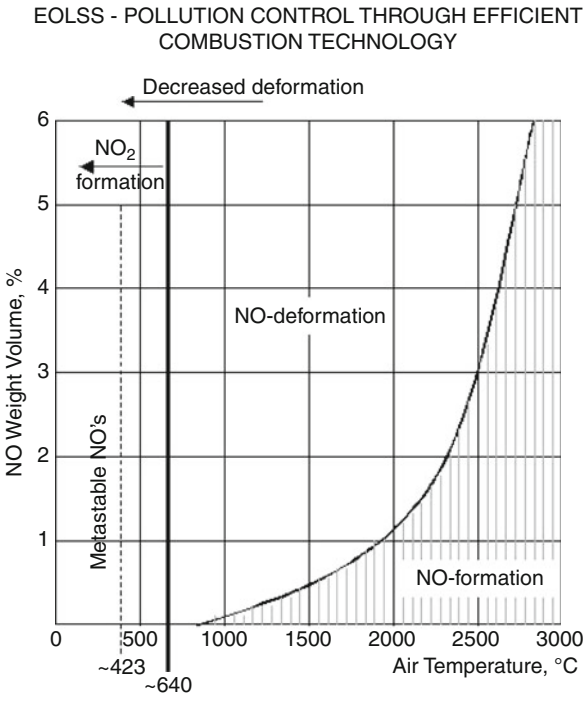
Combustion converts the S-, Cl-, and N-content into SO_2, HCl, and NO, at least as a first approximation. When the resulting flue gas is cooled down slowly and in the presence of catalytic fly ash (transition metal oxides, including iron, manganese, or vanadium oxides are catalysts), a fraction of SO_2 can oxidize further to SO_3, and HCl to Cl_2.

At high temperature (1,000°C), SO_2 and HCl are the most stable compounds; yet, below 500°C SO_3 and Cl_2 become the more stable. On the other hand, a fraction of SO_2/SO_3 and HCl/Cl_2 is removed by adsorption and neutralization by basic fly ash components, e.g., CaO.

Thus *elementary analysis* of fuel allows predicting the major combustion products:

$$\begin{aligned} &C_aH_bO_cS_xN_yCl_z + [a + 0.5(b - c) + x]O_2 \\ &\quad \Rightarrow aCO_2 + (b - 0.5z)H_2O + xSO_2 \\ &\quad + y[\alpha NO + 0.5(1 - \alpha)N_2] + zHCl \end{aligned} \tag{9}$$

Nitrogen oxide (NO) forms from *fuel*-N (i.e., the organic nitrogen, e.g., from proteins, in sewage sludge, hair or leather, or from polyamides) and also from combustion air, yet mainly at elevated temperatures, as *thermal* NO. Such NO formation is lower when the



Incineration Technologies. Figure 5
NO as a function of combustion temperature [37]

flame is cooled, e.g., by radiant heat losses or by the presence of water vapor, and also when combustion is conducted in two steps: the first under reducing conditions, the second oxidizing, yet at low temperature.

To cater for this uncertainty, fuel NO formation is given a proportion α ($0 < \alpha < 1$), the balance being reduced or decomposed to molecular nitrogen ($1 - \alpha$). The formation of thermal NO is neglected in Eq. 9 (Fig. 5).

Chlorinated Compounds

Most waste contains chlorides and also chlorinated organic compounds.

The Bundesweite Hausmüllanalyse (comprehensive analysis of refuse and its sorting fractions in the German Federal Republic) established the amount of, e.g., heavy metals, PAH, and dioxins in MSW for fractions such as fines, vegetal, synthetic, paper, and board. All these sorting fractions are contaminated with all kinds of pollutants [32, 33].

During incineration, organic compounds are destroyed and their chlorine content is converted to HCl. Typically 50% of the Cl-content comes from PVC [38]. During combustion, PVC, as well as a vast majority of organic and inorganic chlorinated compounds, is partly or completely converted into HCl. PVC liberates HCl very easily. Such release is also likely to be complete, unless some other compounds, e.g., $CaCO_3$, capture it; the latter is plausible in numerous applications featuring fillers of precipitated calcium carbonate or ground dolomite/calcite.

The presence of NaCl is ubiquitous, especially in marine surroundings. At high temperatures, NaCl reacts with steam, yet its conversion into NaOH and HCl is limited by thermodynamic equilibrium. It shifts largely to the right, however, in case NaOH is itself converted into silicates, aluminates, or other composite compounds [43, 44], e.g.:

$$NaCl + H_2O \Rightarrow NaOH + HCl \quad (10)$$

$$2xNaOH + y\,SiO_2 \Rightarrow xNa_2O\,y\,SiO_2 + H_2O \quad (11)$$

$$2xNaOH + z\,Al_2O_3 \Rightarrow xNa_2O\,z\,Al_2O_3 + H_2O \quad (12)$$

Thermodynamically, the formation of chlorine gas from hydrogen chloride is described by the industrially important Deacon equilibrium:

$$4HCl + O_2 = 2Cl_2 + 2H_2O + Heat \quad (13)$$

At combustion temperatures, HCl is by far the main Cl-compound yet – below 500°C – equilibrium conditions reverse and elemental chlorine gains ground. Chlorine is much more reactive and corrosive; moreover, it is slower to dissolve in water and thus difficult to remove. Fortunately, this reaction also becomes slower and slower, so that there is little progress towards equilibrium during the few seconds while flue gas moves from furnace to stack. The Deacon reaction also shows an effect of oxygen partial pressure, an even stronger effect of water vapor, as well as an effect of total pressure.

Other halogens follow similar equilibriums, with the elementary amount rising in a sequence: $F_2 < Cl_2 < Br_2 < I_2$. The Deacon reaction is a potential source of both corrosion and dioxin. No doubt, chlorine is only rarely produced in significant quantities and only in the presence of oxidants, such as iron ore (Fe_2O_3) or manganese ore (MnO_2).

In industry, the Deacon process is of paramount importance: chlorine is a potent reactant required in

organic chemistry and synthesis. Its use leaves HCl as a useless by-product. However, reaction (Eq. 13) allows recovering chlorine from HCl. Typical reaction conditions are: fixed or fluid bed, 450°C, $CuCl_2$ catalyst, and in dry air or pure oxygen.

Residues

In principle, incinerator residues are inert and sterile. Often, the major components in ash are silica (SiO_2), alumina (Al_2O_3), and lime (CaO), which are also the main components of the earth crust; yet virtually all elements are represented and ash composition may differ greatly from that of the earth, especially in industrial waste. Numerous studies have been devoted to the physical nature and the minerals of bottom ash and fly ash [24–26, 45, 46]. The International Ash Working Group merged worldwide experience in characterization, treatment, and leaching tests for evaluation of eventual environmental impacts of incinerator residues. Fly ash, a by-product of fossil fuel firing (coal, lignite, peat) is the subject of a site of Kentucky University and of periodic conferences published there.

Chemical analysis of the mineral ash gives information on the softening and melting behavior of the ash and hence about its tackiness and possible attack on refractory. As a rule, Na- and K-compounds decrease the melting point, in particular when present as persulfates, vanadates, borates, etc. The same holds for fluxing elements, such as boron, vanadium, or fluor. The presence of volatile compounds, such as mercury, thallium, cadmium, arsenic, antimony, and other volatile heavy metals makes the related wastes improper for incineration in conventional units. In numerous cases, stable mineral forms are different at the conditions of high-temperature combustion and at room temperature, e.g., volatile chlorides, stable at combustion temperature, tend to convert into sulfates once they condense on boiler tubes.

Thermal Aspects

During incineration, the heat content of waste, in particular its higher heating value (HHV), is liberated almost entirely. The only exceptions are the unburned materials in bottom ash, fly ash, and flue gas. Combustion efficiency η_{Comb} addresses these chemical losses by:

$$\eta_{Comb} = 1 - \text{Ash}\,C_{ash}HHV_C - (\text{Fly Ash})C_{fly\ ash}HHV_C - (\text{TOC})HHV_{TOC} \quad (14)$$

An incinerator plant is a thermal plant and should be operated as evenly and constantly as possible, close to the setpoint in its operating diagram (Fig. 6). Capacity is expressed either as (nominal) thermal load (GJ/h), or as weight throughput (Mg/h).

The *operating temperature* of an incinerator combustion chamber can be estimated from a heat balance and depends on:

- The higher heating value of waste
- The *excess air* applied
- The cooling of furnace walls (e.g., by tubes of an integrated boiler or by heat losses to the environment)
- The initial temperature of air and waste streams

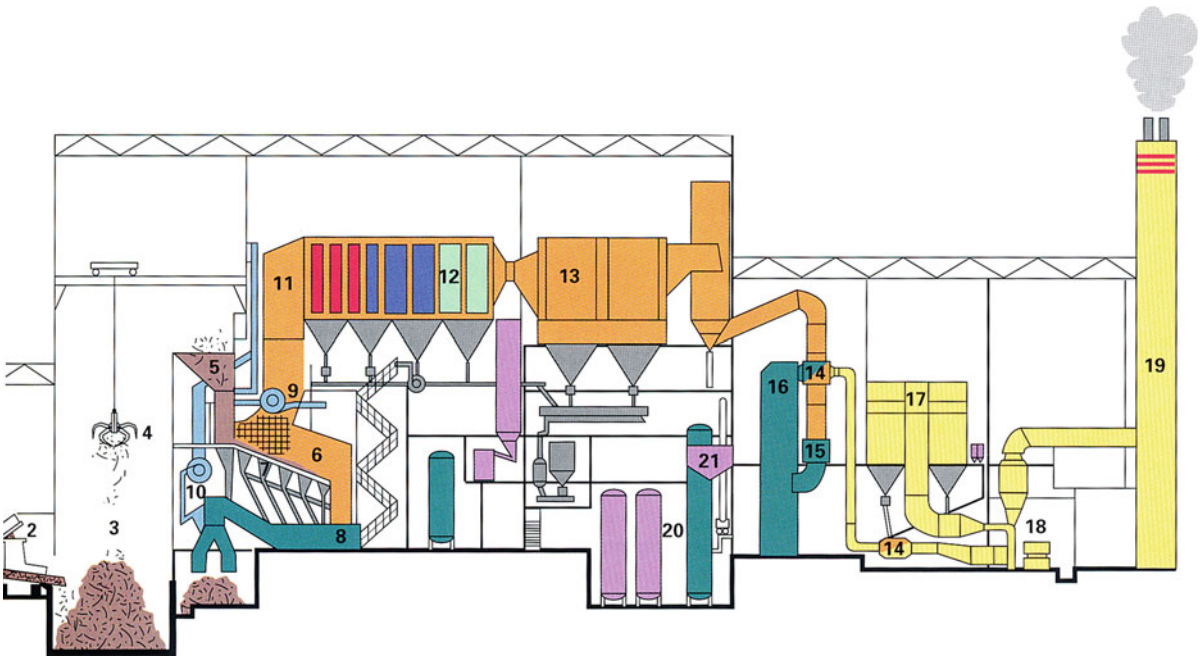
A theoretical *flame temperature* (°C) can be derived simply by plotting the heat content of flue gas ($MJ/Nm^3 \times Nm^3/kg$ waste) as a function of temperature: the flue gas reaches the theoretical flame temperature when its sensible heat equals the liberated heat of combustion (MJ/kg waste). A more complete *heat balance* over the furnace, the boiler, and all downstream equipment gives:

$$Q_{fuel} + H_{fuel} + H_{air} = (H.H.V.)F_{fuel} = Q_{heat\ duties} + Q_{wall\ losses} + Q_{sensible\ heat} \quad (15)$$

The first three terms contain the chemical energy (Q_{fuel}) liberated by combustion, augmented by the enthalpies of fuel (H_{fuel}) and air (H_{air}) when entering the furnace. After combustion, the energy entering the furnace is eventually redistributed as:

- *Useful energy* ($Q_{heat\ duties}$), taken up by the various heat duties, generally the boiler, the economizer, and the air preheater
- *Wall losses* ($Q_{wall\ losses}$) by convection and radiation
- *Sensible heat* ($Q_{sensible\ heat}$) and *latent heat* (water vapor) contained in the flue gas at the stack, i.e., the stack losses

I



1 Hall de déchargement
2 Cisaille pour déchets encombrants
3 Fosse à ordures
4 Grappin du pont roulant
5 Trémie d'alimentation du four
6 Chambre de combustion
7 Trémies sous grille
8 Canal à mâchefers
9 Brûleur d'allumage
10 Ventilateur d'air comburant
11 Chambre de post-combustion
12 Chaudière de récupération
13 Filtre dépoussiéreur
14 Echangeurs de chaleur
15 Injection d'eau (Quench)
16 Tour de lavage des fumées
17 Filtre à manches
18 Ventilateur de tirage
19 Cheminée
20 Bacs de réactifs chimiques
21 Systèmes d'épuration d'eau

Incineration Technologies. Figure 6
Operating diagram of a mechanical grate incinerator (After [37])

Thermal efficiency η_{Therm} addresses these wall losses and stack losses by:

$$\eta_{\text{Therm}} = 1 - Q_{\text{wall losses}} + Q_{\text{sensible heat}} \qquad (16)$$

It indicates the fraction of the energy entering that is recovered for useful purposes. Typical values are 0.6–0.85, or 60–85%. It can be used for district heating, water desalination, or industrial purposes. Since all these applications are site dependent and not generally available, the heat recovered as steam can be converted into electric power, by means of a turbo-alternator.

Finally, there is one more important ratio, indicating the yield of electric power derived from the initial energy in MSW or other waste incinerated. Typical values are 0.16–0.24, or 16–24%.

Air Preheating

Primary air preheating facilitates ignition, increases the flame and combustion temperature, and improves the thermal balance of the process by recovering more heat from flue gas. Combustion air is often preheated, either by flue gas/air or by steam/air heat exchangers, to assist in drying and ignition. Such heat exchangers are relatively voluminous (gas/gas heat transfer is slow) and hence capital intensive. Combustion air may also be replaced by oxygen-enriched air, or even by pure oxygen, in order to improve and accelerate combustion. This is an unusual option, since combustion air is free of charge and pure oxygen is expensive.

Low operating temperatures lead to slower and less complete combustion; excessive temperatures may

render combustion control more difficult and cause severe slagging of ash and fly ash. Tacky ash gradually builds up onto the furnace walls, the deposits eventually limiting the throughput of the furnace. Similarly, clogging problems may occur in the convection sections of the boiler, when excessive approach velocities are practiced or insufficient tube clearance is provided.

Some plants operate under *slagging conditions*, at temperatures at which the ash is molten and tapped in that state. It is important to ensure steady slag flow by:

- Carefully controlling the composition of the ash, at or close to a suitable eutectic composition; iron silicates and glass are two examples of compositions with accessible melting point
- Providing auxiliary burners and, when required, adding fluxes such as fluorspar, to enhance slag fluidity

Furnace Capacity

Nominal capacity is often expressed as the *throughput* or weight capacity (Mg/h) at which the incinerator was designed. The *load factor* of the incinerator is the ratio of the actual operating rate (Mg/h) to the nominal one (Mg/h).

Incinerator furnaces are characterized best by a minimum and a maximum *thermal capacity* (MW). Below its minimum value, the heat generation rate is so low that the furnace no longer reaches adequate temperatures to ensure smooth drying, heating, and ignition, and eventually complete combustion. When the flue gas temperatures descend below 850°C, European Union Codes stipulate that auxiliary burners must ignite and heat the combusting gases, to ensure their sufficient burnout. Excessive combustion temperatures are also undesirable, because fly ash becomes too tacky, creating deposits on furnace walls and boiler tubes. Ash similarly starts slagging; the resulting deposits on the furnace walls become ampler and ampler, eventually even restricting the movement of waste on a grate.

Furnaces feature also a minimum weight capacity (Mg/h), dictated by the necessity to maintain some minimum coverage of the grate for protecting it against furnace radiation and atmosphere. Maximum related to bed density. Finally, the relation between thermal and weight capacity is also bounded, by the necessity of producing sufficient heat for heating furnace and waste; the ratio represents the heat of combustion (MJ/kg). These different boundary conditions are represented in thermal capacity vs. weight capacity diagrams, indicating the area of smooth operation of the plant. The latter is possibly extended toward low heating values by preheating combustion air or toward high-calorific waste by cooling the combustion chamber. Thus, there are links between furnace requirements and waste characteristics.

The *volumetric heat release rate* (MJ/Nm^3, s) of a given furnace is mainly determined by the quality of contact with combustion air and by fuel reactivity, which generally decreases with larger size, higher moisture content, and lower HHV. Since combustion intensity is often unevenly distributed over the furnace, the method to consider furnace volume should be carefully defined, when citing values for volumetric heat release rates. In some cases this volume has been defined as the furnace volume at temperatures exceeding 850°C, rather than as a physical geometric volume of the combustion chamber. Dead zones at lower temperature indeed may consume a sizeable fraction of furnace volume, thus reducing real residence times and combustion efficiency η_{Comb}. Conversely, the first flue of a waste heat boiler may operate above 850°C and thus become eligible as supplemental furnace volume.

The operating domain and the limits of furnace operation may be dictated by various considerations, e.g.:

- *Heat balances*, and the concomitant higher and lower *temperature* limit (°C)
- Excessive, adequate, or insufficient *thermal load* (GJ/h)
- Adequate coverage of a mechanical grates, and hence maximum and minimum *feeding rate* (Mg/h)
- Provision of sufficient combustion air

During reception tests, the operators were supposed to deliver the proof of capacity of a given incinerator furnace over a time period of 24 h. Realizing their presumable failure, they started overcharging the furnace, bringing in more and more MSW. Due to the excessive thermal load, the furnace interior evolved from orange-red to orange, then to yellow, then turning whiter and whiter as the furnace temperature rose. Still, at that moment, more and more unburned materials

appeared among the residue: remarkably, a telephone book had crossed this furnace without even starting to convert into char!

Hazardous Waste

Hazardous waste can be identified either on the basis of inclusive lists, as proposed by the European Union [47], or on the basis of hazardous properties, an approach followed by the US EPA. In the USA, hazardous waste is waste that poses substantial or potential threats to public health or the environment. There are four factors that determine whether or not a substance is hazardous [48]:

- Ignitability (i.e., flammable)
- Reactivity
- Corrosivity
- Toxicity

The US Resource Conservation and Recovery Act (RCRA) additionally describes "hazardous waste" as waste that has the potential to [48]:

- Cause, or significantly contribute to, an increase in mortality (death) or an increase in serious
- Irreversible, or incapacitating reversible, illness
- Pose a substantial (present or potential) hazard to human health or the environment when improperly treated, stored, transported, or disposed of, or otherwise managed

Most of these hazards are entirely eliminated by incineration. Hence, HW may be incinerated at high temperature. Many cement kilns burn hazardous wastes like used oils or solvents. A more detailed discussion is to be found in various books listed at the end of this entry and in [49]. Hazardous waste poses much more problems at the levels of collection, bulking up (i.e., grouping similar waste in the same container or vessel), transportation, and intermediate or final storage than at that of incineration. Obviously, flue gas cleaning must take into account the chemical composition of the hazardous waste concerned.

Safety Aspects

Swiss Re provided a systematic discussion of some safety problems and accidents in incinerator plants. At the times of construction and annual maintenance of incinerators, lots of unusual activities take place onsite, bringing various hazards with them. During normal operation, these hazards reduce to more normal proportions, yet, numerous safety problems may occur around incinerator plants; just to name a few [50]:

- Waste bunker fires
- Explosions during the shredding of waste
- Flame flashback into the system of feeding locks
- Explosive combustion by simultaneous ignition of a large mass of waste, bringing the furnace under overpressure, with flames sorting out
- Hydrogen explosions following decomposition of water in contact with hot metal in a wet ash extractor
- Pressure vessels (boiler)
- Low levels of boiler feed water
- Boiler corrosion and tube failure
- Accidents connected to chemicals on-site, e.g., boiler feedwater treatment acids and bases and ammonia for DeNOx operation
- Rotary and moving equipment
- Transformer fires
- Fires in the wet scrubber, during shutdown

An even larger array of accidents may take place in plants treating hazardous waste, as a consequence of chemical reactivity, flammability, and corrosivity. During collection and storage it is usual practice bulking up liquid waste of similar composition and origin. Mixing distinct waste streams often leads to undesirable events; to avoid such happenings it is desirable to consult compatibility charts and data, such as [51–56], and also to mix small amounts in a test tube and then observe carefully any heating, gas evolution, precipitate formation, or other processes taking place.

Pool burning, *boiling liquid expanding vapor explosions* (BLEVEs) and *vapor cloud explosions* (VCE) are relevant concepts in industrial safety techniques [57]. Even comprehensive waste treatment centers do not necessarily reach the scale of operations or storage required to resort under COMAH eligibility conditions, although specific risk derives from the multitude and variability of waste streams potentially handled. Fires at chemical storage sites are generally impressive

and the storage, blending, and feed preparation facilities upfront a chemical waste incinerator are exposed to such occurrences.

Incinerator Furnaces and Boilers

Furnaces, Their Duties, Peripherals, Operation, Design, and Control

Most problems with incinerator plant proper are basically mechanical and arise mainly at two levels: (a) the introduction of waste into the furnace and (b) the extraction of the various combustion residues. Both should proceed without undesirable and uncontrolled entrance of ambient air.

Duties

Basically, a furnace is a heat-resistant enclosed space that should fulfill several duties simultaneously:

- Limiting the heat losses to the surroundings (heat losses ⇒ flame cooling ⇒ incomplete combustion).
- Ensuring controlled entries to primary and secondary combustion air, and exclude any notable uncontrolled entries, e.g., through the feeding or the ash removal system.
- Ensuring sufficient combustion + post-combustion time to both flue gas and solid phase (fuel, ash) to allow for their thorough and controlled burnout. This implies avoidance of short-circuiting, as well as creation of dead corners.
- Providing peripheral facilities for feeding the various waste streams to be incinerated and (when required) ash removal facilities.

Feeding Equipment

Fuel feeding peripherals strongly depend on fuel characteristics, such as the state of aggregation of the waste to be fired in a primary combustion chamber. Examples are a conventional or more specialized burner for firing gas, liquid, or pulverized, coal, in case of flammable waste gases, pumpable waste liquids, molten solids, and finely divided, free flowing solids. Burners for liquid waste may be based on centrifugal dispersion (rotary cup burners) or on pressure or auxiliary medium (steam, high pressure air) dispersion.

Chlorinated waste has been fired using the dispersion provided by a patented small auxiliary burner situated inside the main burner: the liquid chlorinated waste is supplied through apertures in a duct, leading the combustion productions from the auxiliary burner into the main combustion chamber (Vicarb technology). Some burners are even built to receive several types of wastes simultaneously, such as waste oil, emulsions, suspensions, as well as auxiliary fuel, to sustain combustion.

Solid waste can be fired by means of:

- Gravity feeding from a fuel hopper, separated from the furnace by means of a lock, composed of two sliding doors, a rotary valve, or even a pile of waste locking out the ambient air.
- Spreader stokers [59]
- Screw or piston feeders
- Mechanical or traveling grate stokers
- Pneumatic feeding of free-flowing fuel, e.g., to cyclonic or fluidized bed combustors

Cooling and extinguishing provisions may be required for preventing backfire in feeding systems, or excessive thermal decomposition in feed lines. Another frequent issue is the presence of oversized materials, metal pieces, etc., that create problems during feeding and/or residue extraction: waste containers seem to exert a fatal attraction to all kinds of extraneous matter that can block or even destroy the most sophisticated mechanical feeding or residue extraction equipment. Operators should scrutinize incinerator feed for items such as pressurized gas bottles, ammunition, or oversized concrete or metal parts.

Ash Extraction

Dry or wet ash extraction equipment is generally installed at the bottom of an ash pit or of a sequence of these ash pits, located below successive sections of a grate. It may be based on drag conveyors with suspended flights, screw conveyors, inclined vibrating conveyors, or even pneumatic conveying. These systems must be designed as a function of flow rate and the handling characteristics of fuel and ash. Failing feeding or extraction mechanisms can cause undesirable, expensive downtime (1 day of a commercial incinerator line typically costs US $20,000–$50,000).

Dry extraction plant is somewhat simpler to maintain, yet tends to be a source of persistent dust in and around the basement, where it is located. Dry extractors create considerable chimney effects and – as a consequence – they may turn into an unwelcome source of uncontrolled air in the furnace.

Wet extraction has the merit of quenching the residue and at the same time it brings in some water vapor at the level of the discharge point. Discharging hot metal may decompose water, forming potentially explosive hydrogen.

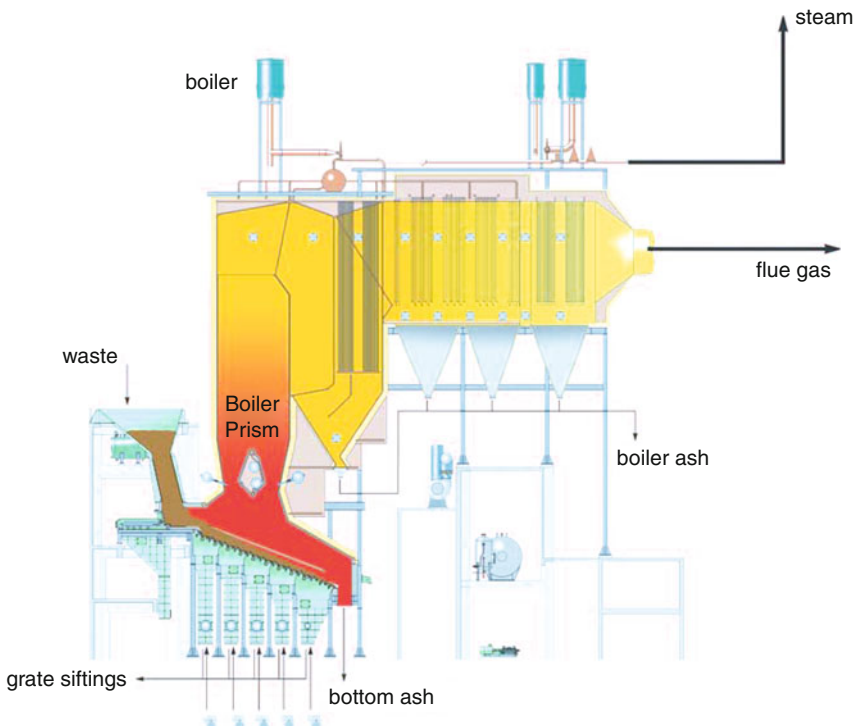
Air Supply

The combustion chamber provides suitable plenum chambers for *primary* and entrance ports for *secondary air*, supplied at possibly substantial overpressure. Primary air activates the fire, burns out combustion residues, and cools the mechanical grate, if existing. Secondary air is injected at a high speed (typically 80–150 m/s), providing the required momentum for thorough mixing of flue gas and completing their burnout. As capacity is scaled up, the available momentum declines relative to the dimensions of the furnace. Some furnace suppliers also bring in secondary air through hollow beams, situated at the level of the furnace outlet: the secondary air is split into four parts, some supplied though nozzles situated in the side walls, the remaining from the hollow beam in the middle of the furnace exit (Fig. 7).

Flow Patterns

The flow patterns in a combustion chamber are rather complex, determined by the momentum of all inputs (burners, primary and secondary air) and outputs (extraction of combusting gas), as well as by buoyancy effects caused by flames and the hot combusting gas generated.

Whatever the geometry, there is strong tendency toward short-circuiting between, on the one hand, the point(s) of entry and, on the other hand, the point(s) of



Incineration Technologies. Figure 7
Secondary air distribution beam in the middle of the exit from a combustion chamber (Courtesy of Keppel-Seghers, Willebroek [Belgium])

exit. Short-circuiting is minimal in a perfect plug flow furnace. It becomes important in the case of a voluminous combustion chamber with single entry and single exit, strong short-circuiting between entry and exit, and inactive zones in between furnace walls and short-circuit flows (Fig. 8).

A short-circuiting combustion chamber is inefficient: the short-circuiting threads show a very low, reduced residence time, the short-circuited volumes unduly long residence times, albeit at low combustion rates and temperatures. Hence, both are inefficient.

Flow patterns can be influenced by combustion chamber geometry, positioning of input and exit locations, selection of input momentum, and influencing the combusting gas pathways, e.g., by provision of baffles and changes in direction.

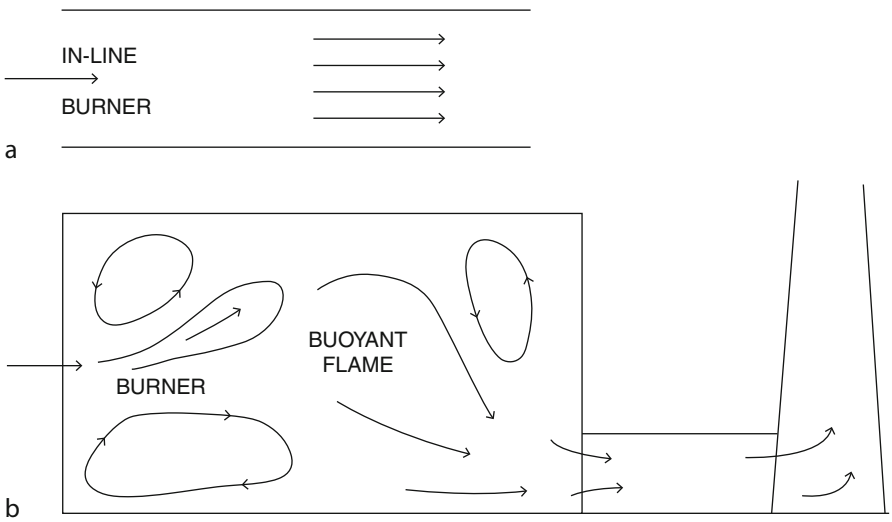
Design Aspects

Thirty years ago, only an empirical approach was practicable when designing incinerators. Tanner devised a triangle diagram to represent MSW as a ternary mixture, indicating zones with auto-combustible waste and others where auxiliary fuel was needed; Hämmerli proposed different nomograms for comparing and assessing grate loadings for mechanical stokers and rotary kilns. Today, *computer fluid dynamics* (CFD) easily derives the flow and mixing characteristics, the rates of heat generation, and the temperature and flow fields [58, 60].

Moreover, the trajectories of particles of various sizes can be predicted stochastically. Swithenbank et al. modeled the various zones (drying, pyrolysis, gasification, incineration) of a mechanical grate incinerator, using CFD, as well as the results of experimental testing at different scales [61–63]. A representative list of recent SUWIC work is given in [64]. Other important sources of solid waste incineration test work are due to ForschungsZentrum Karlsruhe, with experimental research on units such as TAMARA (small mechanical grate incinerator unit) and THERESA (rotary kiln incinerator unit).

Computer Fluid Dynamics (CFD)

Computer fluid dynamics are based on subdividing the volume of interest, i.e., the combustion chamber (or other parts of the plant) into a grid of elementary volumes. The relevant equations of conservation (mass, momentum, energy) are then applied to each of those elements, after defining all inputs, outputs, and boundary conditions. The resulting system is integrated from start to finish, after introducing momentum, mass, and heat transfer (adapted from the Laws of



Incineration Technologies. Figure 8
Plug flow versus plug flow with dead zones

Newton, Fick, Fourier, and Stefan-Boltzmann), taking into account dimensional analysis, turbulent flow, and the state functions of relevant compounds, as well as chemical kinetic reaction systems of variable complexity [60].

CFD thus allows visualizing some cardinal aspects of the combustion chamber, i.e., the fluid flow field (flow vectors, indicating flow direction, and rate in each point), temperature and pressure field, and combustion rate field and – depending on nature and composition of the reaction models – fields for any other chemical compounds of interest (PICs, specific pollutants). Modeling thermal behavior of specific compounds or waste can be conducted at a milligram or even a lower scale [65, 66] (Fig. 9).

Draft Considerations

An incinerator plant usually operates under balanced draft: a balance is struck between forced draft (blowing in combustion air) and induced draft (ID, drawing out flue gas through the stack). ID arises by means of chimney draft, supplemented by the ID-fan, so that the furnace operates steadily with a combustion chamber at a slight subatmospheric pressure, of the order of say 10 or 15 cm water column (1 atm equals more than 10 m w.c.).

Chimney draft follows from the *Law of Archimedes*: the stack is filled with light hot gas, taking the place of an equivalent physical volume of much denser ambient air. Hence, the hot stack gas aspires being replaced by the latter, which enters the furnace by all controlled inlet ports, as well as by those uncontrolled, such as a dry ash extractor or non-tight junctions between distinct parts of the plant and non-tight plant parts, e.g., a fly ash discharge valve.

Very small plants (such as a big stove) may rely on *natural draft*, controlled by means of variable obstructions regulating at the supply side or at the chimney. Medium and large plants use both *forced* and *induced draft fans*. These are major consumers of electric power. Due to the gradual extension of heat recovery and pollution control, these draft requirements have steadily risen over time. For example, power consumption in mechanical grate plant was typically 40–80 kWh/Mg of MSW around 1970. Today, it is more like 160–240 kWh/Mg of MSW.

Mechanical Drives

Until the 1920s, loading the furnace, poking the fire, and extracting ashes was largely manual, somewhat aided by gravity and appropriate tools. Mechanical grates, fans and blowers, and the use of mechanical and later hydraulic drives were first introduced to alleviate the hard labor of the stokers. Today, these tasks are largely automated and sensors monitor every operating detail.

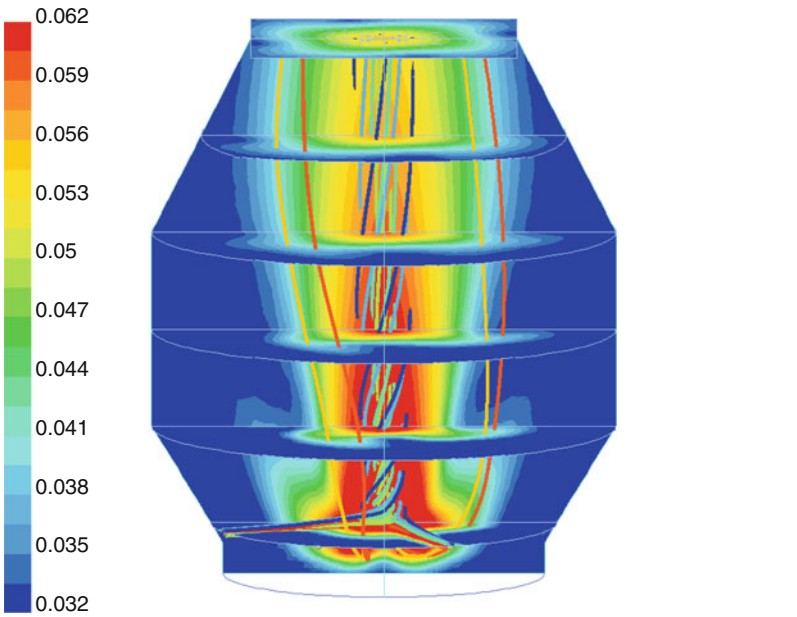
Regulation and Controls

Almost all operating parameters (action of drives, position of valves, temperatures, pressures, flows, levels) are registered continuously, for every part of the plant, as well as all relevant emission parameters (O_2, CO_2, CO, H_2O, SO_2, HCl, NOx, TOC, dust, etc.) so that all incidents can be carefully analyzed, even months post factum. Computer screens synoptically present information on storage and feeding, and on the operation of furnace, boiler, boiler feedwater treatment, steam turbo-alternator and condensers, residues extraction, air pollution control techniques, and forced and induced draft fans. Control systems are quite sophisticated and directly influence draft, furnace temperatures, and the position of the fire.

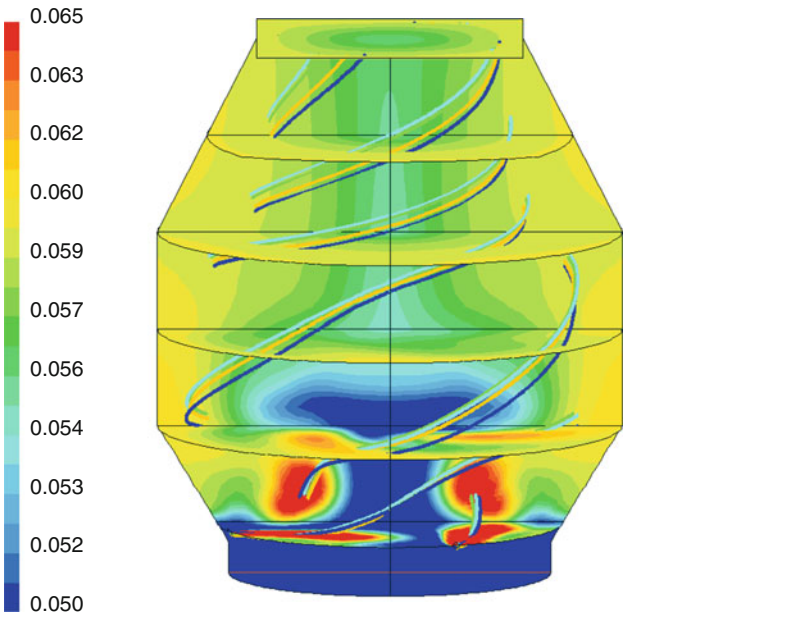
Combustion control follows complex algorithms, developed to ensure the right operating conditions, regarding temperature, pressure, airflows, etc.

Conclusions

A furnace is to achieve adequate control of air supply and draft, and thus of all major combustion conditions (temperature, time, turbulence) and emissions. Typically, combustion is conducted at more than 850°C, a residence time of combustion products in the gas phase of at least 2–3 s at these 850 C (or higher), and adequate turbulence to render these reasonably homogeneous. A minimum level of oxygen (e.g., 6 vol.% in MSW incineration) may also be specified, either by legal codes or by good practice. Ideally, combustion proceeds at a pressure slightly below atmospheric, so that combustion products do not spread to the surroundings, through the inevitable leaks that occur in between the main parts of the incinerator plant, as well as at its appendages.



a



b

Incineration Technologies. Figure 9
Computer fluid dynamics (CFD) representation of a combustion chamber (By courtesy of Prof. J. Swithenbank [SUWIC])

Post-combustion

The average *residence time* (s) in the combustion chamber is given by the ratio of the physical volume of this combustion chamber (m^3) to the volumetric flow (m^3/s) at the furnace conditions (temperature, pressure) prevailing, and determined, e.g., at the combustion chamber exit. The real *residence time* follows a distribution determined by internal flow conditions, including short-circuiting and dead zones. Such distributions are rarely established, whether by computer fluid dynamics, or by tracer experiments, as described in [67].

Combusting gas leaving the primary combustion chamber is still at about the average temperature of this chamber, i.e., typically >850°C, yet its burnout must still be completed. There are several physical and chemical reasons for this, e.g.:

- *Residence times* in the (primary) combustion chamber are rather short, to make the best use of intense combustion and the concomitant high temperatures. Too large combustion chambers operate at too low combustion temperatures, causing incomplete combustion. Conversely, too small combustion chambers operate at too high combustion temperatures, causing severe slagging of refractory walls, unless these are adequately cooled, as well as thermal NOx. Cooling of such furnace walls is technically possible by integrating the combustion chamber into the boiler, or by blowing part of the secondary air through channels prepared in the refractory walls.
- Part of the combusting gas *short-circuits* parts of the (primary) combustion chamber, so that its *real* residence time is only a fraction of the *average* residence time. Hence, it is advantageous to promote plug flow by judicious selection of furnace dimensions, make use of any constructive features promoting plug flow and avoiding dead zones, and testing the resulting furnace designs by CFD.
- In zones of intense combustion, local or even general *deficiencies of oxygen* are likely to occur, either permanently, or only in case of fast, flaming combustion of unusually large lumps of waste. As a consequence, incinerator furnaces should be fed steadily, yet in small unit doses.
- Very high combustion temperatures lead to the partial dissociation of major combustion products, such as

$$CO_2 \Rightarrow CO + \frac{1}{2}O_2 \tag{17}$$

$$H_2O \Rightarrow H_2 + \frac{1}{2}O_2 \tag{18}$$

From chemical reaction theory it follows that the best results are obtained under *plug flow* conditions. Theoretically, these can be approached by a sufficiently large number of combustion chambers. In practice, such an ideal situation can be strived for by:

- Separating the combustion chamber into a main, primary chamber, followed by a secondary and possibly third chamber. This secondary chamber is in general use when incinerating, e.g., hospital waste in the sequence (1) primary partial oxidation chamber, yielding incomplete combusted fumes, and (2) secondary post-combustion chamber fitted with an auxiliary burner for raising the temperature and provisions for generating swirl and thus promoting complete combustion.
- Conventional combustion chamber (e.g., featuring a mechanical grate stoker), followed by a zone of highly turbulent mixing, produced by the injection of high-speed secondary air.

Total organic carbon is a lump parameter of flue gas organics, measured off-line by means of flame ionization detectors and expressed as mg CH_4-equivalent per Nm^3. Detailed identification is both seldom conducted and tedious, yet of possible interest in a larger environmental debate, or for dedicated monitoring of POHC (principal organic hazardous constituent) during test burns of hazardous waste [68, 69], e.g., at the Incineration Research Facility (IRF). US EPA monitored the environmental performance of hazardous waste incinerators by ordering test burns to be conducted. The legal framework is described in [70].

Under controlled laboratory conditions Dellinger et al. applied the gas-phase thermal stability method to rank the incinerability of 20 hazardous organic compounds, selected on the basis of frequency of occurrence in hazardous waste samples, apparent prevalence in stack effluents, and representativeness among

hazardous organic waste materials. Their major findings were [71]:

- Gas-phase thermal stability is effective in ranking the incinerability of hazardous compounds in waste.
- Numerous PICs were formed during thermal decomposition of most of the compounds tested.
- A destruction efficiency of 99.99% is achieved after 2 s mean residence time in flowing air at 600–950°C (Table 8).

Conclusions

Post-combustion is essential because primary combustion chambers are too limited in residence time and in mixing and homogenization capabilities to ensure steady burnout reliably and permanently. Post-combustion is preceded by a zone of intense mixing, to homogenize oxygen-rich with oxygen-lean strands; it proceeds as long as temperature remains above, say, 500°C. As temperature decreases, all reaction rates tend to fall.

Below 500°C, oxidation may proceed further in case the remaining PICs can be adsorbed and converted catalytically.

The advent of selective catalytic reduction (SCR) paved the way for organized oxidation of PICs, the semiconductor catalysts used being capable of (first) NO reduction and (second) semi-volatile PICs (PAHs, dioxins) oxidation, even at temperatures of only 200°C.

Incineration Technologies. Table 8 Processes influencing upon the formation of products of incomplete combustion in mechanical grate municipal solid waste incinerators, factors of influence, possible remedial action, and influence of the 850°C, 2 s Rule

Nr	Process	Factors of influence	Possible positive action	Influence of the 850°C, 2 s Rule
1	Drying	Heat radiation	Mix dry and wet waste	May be mildly positive, without exerting much direct influence
		Early ignition of high-calorific materials	Preheat air	
			Use a reverse reciprocating grate (mixing)	
2	Heating and Ignition	Radiating Heat	Noncritical process	Almost none
		Ignition of adjacent materials		
3	Thermal decomposition	Material Type	Premixing refuse	None
		Temperature	Poking and mixing action of the grate	
		Heat supply rate		
4	Flaming combustion	Rate of thermal decomposition	Adapt air distribution along the grate	May be mildly negative, by requiring a hot furnace operation
		Supply of air	Enrich with oxygen	
5	Mixing the gases	Furnace geometry position and diameter of air injection nozzles	Improve the design to increase turbulence	None
			Injection of more high velocity secondary air	
6	Post-combustion	Contact time	Apply the 850°C, 2 s Rule	Important
		Temperature		
7	Avoidance of soot formation	Correlated with (3), (4), and (5)	As for (3), (4), and (5)	None

Heat Recovery

The sensible heat contained in flue gas can largely (thermal efficiency η_{Therm} typically 75–85%) be recovered in waste heat boilers. Normally, medium-pressure (1.5–4.5 MPa) boiler operation is favored, to avoid high-temperature super-heater corrosion problems. Fly ash is often tacky above 600°C; hence the contact surfaces are preceded by radiant cooling surfaces. These are specially designed for:

- Limiting adherence and deposition of hot, tacky particles
- Convenient cleaning (rapping of boiler tube panels, soot blowing, shot cleaning of tube banks)
- Easy inspection

During a furnace standstill, it is advisable to keep the boiler tubes hot, by means of imported steam, in order to avoid corrosion by hygroscopic acidic deposits, such as chlorides. The same holds for flue gas cleaning plants.

Plants Without Heat Recovery

In small or batch-operated plants, flue gas is cooled by injecting quench water in a cooling tower surmounting or following the furnace, or by admixing cooling air [9]. These methods increase the gas flow at standard temperature and pressure typically by 30–50% for water injection and by 300–400% for admixing air, which quite considerably inflate investment and operating costs of the gas cleaning plant.

In large-scale incinerators, *heat recovery* using either waste heat or integrated boilers is the most appropriate for cooling the flue gas prior to its cleaning, provided that the steam generated can be used for in-plant or other useful purposes, such as power generation, district heating (winter) and cooling (summer), water desalination, sludge drying, vacuum generation, etc. Still, such heat recovery proceeds under adverse conditions (corrosive and fouling flue gas), requiring considerable investment and diminishing plant availability.

Generated revenues and avoiding the extra cost of requiring much larger gas cleaning plant may offset these disadvantages. Moreover, since heat recovery is a more sustainable option, recovery may be mandatory, even regardless of economic factors.

Boiler Design

The design of a boiler mainly depends on steam quality (boiler pressure + superheat temperatures), water circulation requirements (MSWI boilers feature natural convection), and flue-gas characteristics (corrosion, erosion, and fouling potential). When selecting steam parameters for waste fired boilers, a compromise is searched between yield of power generation and superheater lifetime: an operating pressure of ca. 40 bar (4 MPa) and 400°C are common choices when power is generated [9].

Corrosion becomes more severe, as steam temperature increases. Steam superheaters are especially vulnerable: since they operate at the highest temperatures of the steam circuit they are located at the high temperature side of flue gas and boiler. Moreover, their internal cooling is of low grade (medium pressure steam, stead of boiling water). Corrosion-resistant materials and coatings are key in increased conversion efficiency and reduced maintenance in waste-to-energy (WTE) plants. Another possibility is to heat the steam superheater in a separate natural gas or oil-fired furnace, an option first tested at Moerdijk, the Netherlands.

During the 1960s, boilers were designed according to conventional rules: compact construction and a high rate of heat transfer, sustained by relatively high linear gas velocities. This design was at the source of failures: some superheaters, designed for 20,000 operating hours, barely reached 3–4,000 h. Linear gas velocities selected for high heat transfer rates also create conditions leading to rapid fouling or even complete clogging of entire tube banks and to rapid corrosion [72–74]!

From the 1970s, some simple rule-of-thumbs emerged that led to the design of large-volume, less efficient boilers, however, without the operating problems cited afore:

- Convection surfaces in the boiler passes are placed only after 1, 2, or even 3 empty boiler passes, so that the flue gas temperature is lower than 600°C or at most 650°C. In this temperature range, fly ash is no longer too tacky thus less fouling.
- The clearance between superheater tubes is wide and the approach velocity is low (only few ms^{-1}) limiting inertial fly ash deposition.

- Deposited fly ash is periodically removed using steam jets or dropping shot onto tube banks.

Chlorides, chlorine, and hydrogen chloride play an important role in some forms of corrosion. Yet, also other factors play a synergetic and decisive role, often related to the creation of electrochemical cells with on one side tube metal, on the other the tube deposits. Rate controlling is the electric conductivity of the deposition layer, not the amount of chlorine in the system. Basically, the presence of molten phases on the tubes must be avoided. Rasch studied the thermodynamics of the formation of these phases in some detail [75].

Corrosion Problems

Most gases attack plain steel. Combustion of MSW generates a highly corrosive environment composed of combustion gases and ash and laden with HCl, SO_2, chlorides, and (subsequently) sulfates. Corrosion rates rise with temperature and – depending on metal structure and composition – diminish by formation of protective layers. Coherent consideration of corrosion processes is difficult, as physical, chemical, operational, metallurgical, and crystallographic parameters interact and the precise origins of corrosion vary from case to case, are multiple, and generally difficult to identify. Thermodynamically speaking, some extent of corrosion is unavoidable. Countermeasures may help to reduce corrosion damage to acceptable levels. These require both constructive and operational countermeasures. Low steam parameters in the boiler system, long residence and reaction times (for preliminary sulfatation of chlorides) before entering in contact with convective heat surfaces, lowering the flue-gas speed, and leveling of the speed profile may all be successful. Protective shells, tooling, stamping, and deflectors can also be used to protect and safeguard heated surfaces. A compromise must be found in determining the boiler cleaning intensity between best possible heat transfer (metallic pipe surface) and optimal corrosion protection [76–79].

Currently, corrosion phenomena are observed on superheater tubes particularly. The key role of formation of a molten phase is obviously associated with ash composition and flue gas temperature. The deposit morphology is related to the flue gas flow pattern, to the mechanisms of corrosion and corrosion rates. A theoretical analysis and enumeration of corrosion's numerous forms and appearances are given in the EU Reference Document on the Best Available Techniques for Waste Incineration [38].

In the 1950s and 1960s, Germany built numerous large MSWI plants. Refuse was assimilated to fuel free of charge and the first generation of plants was designed to squeeze maximum power from this resource. Soon, severe corrosions were encountered and their sources were analyzed; several major areas of concern were identified [38, 72–79]:

- Severe corrosion occurred in integrated boilers, affecting mainly the lower half of the boiler tubes surrounding the combustion chamber. This form of corrosion derives from alternating oxidizing and reducing conditions, which prevent protective and coherent oxide films to form. It proceeds through formation of $FeCl_2$ in an oxygen-deficient flue-gas atmosphere, e.g., below oxide films, tube contaminations, or fireproof material especially in the furnace area. $FeCl_2$ is sufficiently volatile at these temperatures to be mobilized. An indicator for such conditions is the periodic appearance of CO. Corrosion products appear in flakey layers. Today, this part of the boiler is clad with protective refractory, often thermally conductive silicon carbide.
- High-temperature superheater corrosion. Corrosion occurs in synergy with other factors, such as inapt boiler design and the accumulation of tacky deposits on the superheater tube banks. Hydrogen chloride and chlorine play a major role in an electrochemical system constituted by boiler and especially superheater tube deposits: hydrogen chloride is released by conversion of alkaline chlorides into sulfates, and attacks iron. Corrosion is observed in MSW incinerators with flue-gas temperatures $>700°C$ and at pipe wall temperatures above 400°C. The corrosion products are black, firmly bonded, and include red hygroscopic $FeCl_3$.
- Molten salt corrosion. Flue-gas contains alkali salts, which form low-melting persulfates (Na- and $K_2S_2O_7$) and various eutectics. Such molten systems are highly reactive and cause severe corrosion or even react with the refractory lining and destroy it mechanically.

I

- Standstill corrosion creates problems mainly after a shutdown, whether scheduled or accidental. $CaCl_2$ deposits are hygroscopic and show deliquescence, whereas some heavy metal chlorides may even hydrolyze, liberating free HCl. Electric tracing is required to keep such deposits dry during standstill periods.
- Dewpoint corrosion is associated with acid gases that condense at the cold, rear end of the boiler. Temperatures below 110°C may suffer from HCl condensation; sulfuric acid may even condense below 160°C.
- Superheaters may suffer damage from erosion due to excessive flue gas approach velocities and/or excessively strong soot blowing. Such soot blowers are difficult to adjust: if the jets blow too hard they cause erosion, if too soft, soot blowing is useless. Specialized services now blast deposits by appropriate use of explosive charges.

Sulfatation

Salts and metal chlorides sublimate at furnace temperature, leaving bottom ash as a cleaner residue [75]. In the first boiler passes the temperature remains still above 650°C and fly ash is still tacky. Below 600°C, flue gas may come into contact with tube banks, without excessive risk of fouling these rapidly. Nevertheless, tube deposits still form by separation of nonsticky particles, by inertia and interception. These deposits also collect chloride salts that de-sublimate and condense. Thermodynamically, most chloride salts are no longer stable, as they were at furnace temperature. Upon contact with SO_2 they gradually convert into sulfates by generic reactions such as:

$$MeCl_2 + SO_2 + \frac{1}{2}O_2 = MeSO_4 + 2HCl \quad (Me = \text{bivalent metal}) \tag{19}$$

Such reactions also consolidate and harden deposits. Moreover, while liberating HCl they contribute to corrosion processes: HCl slowly oxidizes to Cl_2 that diffuses to the tube metal and attacks it; after it is reduced to HCl the same corrosive cycle starts over. From this viewpoint, it is favorable that the flue gas is rich in SO_2 and that sulfatation proceeds before the salt-laden fly ash deposits on the tubes.

Flue Gas Composition

Flue gas composition is determined by several factors of influence. The most important one is waste composition: all entering elements will also leave the plant, whether as flue gas or as solid residue. Mass balances, together with waste composition data, allow estimating the flue gas and the residue composition, even though some assumptions are needed regarding the distribution of the relevant elements over the various output streams. A second factor is the technology used: mass burning of MSW yields much more bottom-ash (typically 20–30 wt.% of MSW) than fly ash (2–3 wt.% of MSW). Fluid bed incineration of the same MSW will turn this relation in favor of fly ash, which may reach, e.g., 10–12 wt.% of MSW. As a consequence, the coarse fraction of fly ash will be less contaminated, following an effect of dilution by bed material and other fines. A third factor is related to the operating conditions used: lower flow and velocity of primary air reduces the entrainment of fly ash and also leads to higher bed temperatures and hence to more sintering of ash and to more volatilization of various heavy metals, e.g., Cd, Cu, Pb, and Zn, that eventually de-sublimate onto the fly ash.

Flue gas composition is also influenced by the excess air amounts practiced: primary air activates the fire in the combustion zone, yet cools the furnace in the drying and burnout zones; excess secondary air merely dilutes the flue gas. To avoid willful dilution with ambient air (to make concentration figures look lower), analytical data are generally expressed at some standard concentration of either oxygen (e.g., 11 vol.% O_2) or carbon dioxide (e.g., 6 or 8 vol.% O_2). Similarly, the concentration of obnoxious compounds is generally expressed on a dry gas basis.

In modern plants, numerous parameters are monitored continuously, e.g.:

Oxygen, on the basis of its paramagnetic properties, or using semiconductors reacting to the oxygen concentration.

Carbon dioxide, water vapor, sulfur dioxide by Fourier-transformed infrared (FTIR) absorption

Hydrogen chloride and fluoride

Nitrogen oxides

The residue composition also depends on the partition between bottom ash, boiler slag (only

a small amount), fly ash, neutralization residues, and fine dust and aerosols that escape uncollected. Numerous studies have considered such issues.

Dioxins

More than a century ago dioxins first drew the attention, while their synthesis afflicted laboratory workers with chloracne. The same happened after isolated incidents in chemical industry, e.g., Monsanto at Nitro, BASF at Ludwigshafen, or Philips-Duphar at Amsterdam. A much more spectacular accident occurred at Seveso (N. of Milan): after a run-away in a herbicide synthesis reactor, its contents were vented all over Seveso, causing trees to lose their leaves, death to various animals, as well as the evacuation of 10,000 inhabitants (1976). People exposed to dioxins are still being monitored today, to detect any eventual symptoms or mortality. Epidemiological investigations show the appearance of rare, soft tissue cancers and neurological afflictions, yet no net increase in mortality (cfr. Public Image).

Dioxins were discovered on MSW incinerator fly ash in 1977 [80]; it took some 15 years more to recognize as major sources several processes in iron and steel industry, as well as in the melting of metal scrap. Dioxins have been at the center of enormous efforts, first to develop, standardize, apply, and ameliorate analytical methods and determine potential dioxin sources as well as possible pathways to formation, then to try and meet the extremely low emission limit values during everyday operation [81–86].

Details of the mechanisms forming dioxins still today remain controversial [87–89]. Theories started with the trace chemistries of flame (Dow Chemicals Co.), continuing with various precursor theories (many researchers) and culminating with the de novo theories, worked out in most detail at Forschungs-Zentrum Karlsruhe. In the first theory, dioxins are inseparable from any combustion process [90]. Precursor theories focus on chemical, often catalytic conversion of dioxin-related structures [91–93], such as phenoxy radicals, chlorophenols, chlorobenzenes, polychlorinated biphenyls (PCBs), and also polycyclic aromatic hydrocarbons and related structures, converting into dioxins. Finally, de novo theory is based on a low-temperature catalytic conversion of almost any carbonaceous structure, such as soot or its various precursors, into dioxins and furans, or PCDD/F [94–98].

Several pathways lead to dioxins [109], yet their relative importance, as well as the precise nature of the catalysis at work will always remain elusive in each particular reactive system. Moreover, there is no mutual exclusion between pathways. Much attention was also given to metal catalysis in dioxins formation [99–102]. Other work related to the prevention of dioxins formation [103–105] or its destruction in fly ash [106–108]. Early and current abatement of dioxins from flue gas is covered in [109–111].

Dioxins in Incineration

During several decades, incinerators have formed the major source of dioxins emissions.

Strangely enough, they were also destroying dioxins, namely, those entering the furnace together with the MSW [112]. Dioxin balances have been established several times in the 1980s, showing that the input and output of dioxins in the plant was similar, yet not necessarily the dioxins fingerprint, i.e., the distribution of various isomer groups and congeners.

Although dioxins are considered to be extremely environmentally stable, they do not survive the combustion process. So, more than 99% of the dioxins entering are destroyed. At the entrance of the furnace and even after the practical end of active post-combustion, no dioxins can be found; at most their basic structures are present [112–115].

Rapid dioxin formation occurs once the flue gas attains a window between 400 and 250°C. A rate maximum of formation occurs at 300–350°C [96].

Explanations differ, yet it seems accepted that the formation is a catalytic process, so that particles play a role, whether suspended in flue gas or deposited from it. Oxygen is required, probably to reactivate the catalyst, after it is reduced while chlorinating aromatic and aliphatic structures.

Salient Factors in Dioxins Formation

Dioxins formation is affected by quite a large number of significant factors, subdivided into two groups: first, operational factors, second, related to chemical,

composition, and catalytic factors, such as catalysis, carbon, oxygen, water vapor, and chlorine [116]. Each of these has several impacts, often with various mutual interactions and it is unlikely that their ranking and relative importance under varied conditions in diverse systems will ever be established once and for all. An intrinsic difficulty in studying dioxins formation is a matter of timescale: the occurrence of a combustion setup, start-up, or shutdown has a certain timescale [117], yet that of dioxins may follow hours, days, or even weeks later (memory effects) [85, 118]. Several factors explain such memory effects: dioxins form from fly ash deposits slowly, and even slower in lower deposit temperatures. In some cases, there may be chromatographic effects, semi-volatiles such as dioxins getting adsorbed and desorbing again later. Wet scrubbers made of plastic dissolve dioxins during upsets and start-ups that desorb again into clean gas later [119].

Incinerator operating factors are of paramount importance. Poor combustion conditions may result from "bad" waste, i.e., either too poor (low temperature) or too rich (excess evolution of volatile matter). These "bad" operating conditions not only lead to more PICs and PAHs (a small fraction of which converts into dioxins), but also to a prolonged increase in dioxins (memory effects: PICs adsorb on boiler deposits and continue generating dioxins afterward). Poor combustion conditions result often from feeding too much at a time, without adequate premixing wastes of different origins and quality. Combustion upsets are notable by a development of peaks of carbon monoxide accompanied by total organic carbon (TOC), a measure for the amount of PICs present. Combustion conditions may be improved by both technology (grates, furnace geometry) and operating skills (mixing and feeding waste, providing primary and secondary air). Nevertheless, firing fuels such as MSW always bring in a factor of chance. With respect to dioxins, the following factors may help:

- Firing well-mixed, homogenized waste only. Humidity transfer from moist vegetal waste to paper and board and dispersion and mixing of high-calorific waste (plastics and rubber) in the bulk of MSW are positive factors, i.e., prolonged storage and periodic mixing of the bunker's content, or mixing moist garden waste with high-calorific commercial waste.
- Using low rates of primary air. This reduces the amount of excess oxygen in the flue gas, as well as the entrainment of dust particles, which is a source of dust deposits and of boiler fouling and corrosion.
- Steady combustion conditions. No large packs of high-calorific waste taking fire together.
- High-quality mixing of gases at the furnace exit.
- Ample post-combustion chamber volumes, at adequately high temperatures and mixing levels.
- Designing post-combustion volumes by means of computer fluid dynamics, for good mixing and avoiding short-circuiting as well as dead zones.
- Limiting residence times in a temperature window ranging from 500°C down to 200°C.
- Operating electrostatic precipitators, at low temperature, not more than 200°C, by extending waste heat boiler surfaces and limiting boiler fouling.
- Avoiding building up and extending deposits on boiler tubes, collection plates in electrostatic precipitators, in flues, etc., by limiting the approach velocity.

The quality of operation can be judged by the permanent absence of CO- and TOC-peaks.

Ideally, their frequency should be nil on a daily basis. Should such peaks still occur, they can be termed "very serious" (CO = 10^3–10^4 mg/Nm^3), "serious" (10^2–10^3 mg/Nm^3), or "benign" (10–10^2 mg/Nm^3). TOC-peaks concur with CO-peaks, yet their height and width differ. The reason for such short-lived peaks is either overfeeding (too much at a time) of fluid beds, or inadequate mixing of MSW fed to mechanical grate units [42].

Complete combustion, mixing of flue gas by blowing in secondary air at high speed, and absence of setups are all primordial operating factors; definitely less dioxin is formed in case excess oxygen is limited.

Another important operational domain is related to the cooling of flue gas: fast and deep cooling limits dioxins formation. Slow cooling of flue gas, in contact with deposited dust, has an opposite effect. For small plants, e.g., metal foundries, quenching off-gas is a suitable prevention measure.

Dust removal takes dioxins away, since these semi-volatiles report to fly ash, especially at low temperatures. Baghouse filters are designed to clean gas down to the very low dust levels required to reach the level of 0.1 ng TE/Nm^3. Any imperfections should be observed by means of tribo-electric sensors, opacity measurement, providing immediate warning in case of dust breaking through.

Chemical and Catalytic Factors

A cardinal chemical factor is related to the presence of transition metals providing the catalytic effects required to fix chlorine on carbon structures and also to oxidize the latter so that dioxins are liberated, together with scores of other surrogate and precursor compounds [47]. Catalytic metals are likely to be associated with particulate, in particular its finest fraction. The latter absorbs the de-sublimating metal salts (Zn, Pb, Cu, Cd, etc.) condensing after having been volatilized at flame temperature [67]. Copper is obviously a premium catalyst; it is often better represented in fly ash from fluidized bed units than in that from mechanical grate units [42]. This could be due to erosion effects, affecting copper wire. In China, fly ash is much leaner in heavy metals than in the EU. Another catalytic substance is iron oxide.

Mixing fly ash with inert materials and carbon creates de novo, dioxin-generating activity. Matrix effects and its particulate carrier are important [120], so is the supply of oxygen to the system: after chlorinating carbon or oxidizing carbon structures, the catalyst is in its reduced form. Oxygen restores a higher valence, required for reactivity. The relations between carbon structure and dioxin formation are still all but elucidated. The presence of the element chlorine is essential in dioxin formation, yet chlorine is ubiquitous in incineration. Factors of influence are numerous and their effects are manifestly complex, interdependent and difficult to pinpoint! Dioxins formation has been studied at full plant level [112], at pilot scale [113–115], and at laboratory level [121, 122]; it was simulated by CFD [123]. Thus, the discovery of dioxins eventually has prompted enormous research efforts, with the fortunate result that incineration became a much more controlled technical process and that the cleaning of flue gas became much deeper (cfr. Tables 3 and 4).

Flue Gas Cleaning

In MSWI flue gas a deep cleaning is essential. Public and political pressures have been so powerful that MSW incineration is at present the most regulated and best controlled form of combustion. Flue gas cleaning addresses successively [30, 37]:

- Particulates and dust, including the associated heavy metals
- Acid gases, such as HCl, HF, and SO_2
- Nitrogen oxides such as NO, NO_2, and N_2O
- Semi-volatile organic compounds, such as PAHs, PCDD/Fs (dioxins), and PCBs

Yet, the precise composition of the flue gas cleaning train depends on numerous options that can be combined in a large variety of flue gas cleaning schemes. Most existing plant during the 1980s and 1990s were forced to revamp this train at least once or even several times, leading to redundancy in the ways these various duties are addressed, e.g.,

- Baghouse filters were often added at the tail of the plant, to complete the preliminary separation by a preexisting electrostatic precipitator; in other plants the ESP was scrapped, because of redundancy and the formation of dioxins at high ESP operating temperatures.
- Dry acid gas scrubbing was supplemented at times by semi-wet or wet units.
- Activated carbon adsorption retains semi-volatile organics that eventually would be destroyed during selective catalytic reduction of NOx.

A survey of best practicable options is given in [37].

HCl is an acid, irritating gas, yet it is eminently soluble in water and thus easily scrubbed out from flue gas (together with HF and HBr, both present at about 100 times lower concentration levels). The resulting diluted solution can be distilled to yield a commercial concentration. Yet, HCl is not in high demand and sales may require removal of trace organics as well as iron. An alternative is using it as a leaching agent, to remove heavy metals from fly ash. In case such recovery options are not followed, yet, the acid needs to be neutralized, e.g., by means of lime.

Selection of Incinerator Furnaces

Selection Criteria

The *selection* of a particular type of *furnace* mainly depends not only on the type(s) of waste to be incinerated (which also determines the possible feeding methods), but also on numerous other factors, such as plant capacity, the operating schedule required, heat recovery, the amount of ash to be handled, and also its physicochemical nature and softening point, etc.

Off-gases and *liquids* are relatively easy to handle using an adapted burner in a simple, tailored combustion chamber, but the incineration of solids, sludge, and paste... may take place under a wide range of combustion conditions and in different types of furnaces.

Furnace types can be classified, according to:

- The contact of waste with combustion air (i.e., in co-current, counter-current, or cross-current relative flow; mechanical and pneumatic agitation, etc.)
- The degree of filling the combustion chamber with solid material
- The choice made between *dry* ash and slag melting conditions (so-called wet-bottom furnaces (not to be mistaken for dry or wet (in water) extraction of combustion residues).

Possible *plant capacity* may be limited by either construction methods, or experience factors; e.g., for mechanical grate at typical capacity 2–20 Mg of MSW/h or rotary kiln furnaces (typically 0.5–5 Mg of waste/h there is only limited experience available once a given size is exceeded. Higher capacity is achieved by providing parallel lines of generally identical capacity and make. Spreading capacity over two or more lines also allows more flexibility, in case of shutdown of one train or of variable supply of waste. Extrapolating existing units to an untested scale may lead to unexpected problems in thermal units. Such was the case in the 1970s for Monsanto's Landgard partial oxidation plant at Baltimore, the Andco-Torrax gasification plant at Leudelange, and the Occidental Petroleum Garrett Pyrolysis plant at El Cajon, Ca. [9, 21, 22].

Heat recovery often features a separate *waste heat boiler*, consecutive to the combustion chamber. Waste with high HHV may also be fired in a furnace, integrated into the boiler structure (*integrated boiler*). The ceiling and sidewalls of the combustion chamber are structurally formed from vertical and inclined boiler tube panels constituted from parallel finned tubes welded together. The tubes are covered by studs sustaining refractory mass, rammed onto the tubes so as to protect them from fouling and corrosion [9].

Pollutant control at times may decide upon the type of furnace to be used or on its operating conditions. Sulfur dioxide (SO_2) is easily captured in a fluidized bed combustor, operating at 850°C, which is the optimal temperature for reacting SO_2 with lime or limestone. Similarly, thermal NOx can largely be avoided at that temperature. Nevertheless, there is always a negative correlation between two types of pollutants: NOx on the one hand and CO + TOC (or PICs) on the other. In case fuel-NOx problems are expected the technique of *staged combustion* is used, which is composed of two steps:

1. Combustion conducted with a deficiency of air (first step, at high temperature), thermally reducing fuel-NOx
2. Post-combustion with ample air and at low temperature

In most cases, this technique will alleviate the problem. Combustion conditions also fix two important factors: (1) ash tends to sinter, soften, and eventually melt, as temperature rises, and (2) the distribution between fly ash and bottom ash also evolves with temperature. Other cardinal factors are the presence of oxidizing or reducing conditions and of halogens, sulfur, etc. [124].

Small-scale incinerators (capacity <2 Mg waste/h) were often operated in a one- or two-shift schedule, but today continuous operation is always to be preferred, since it enhances useful capacity and reduces auxiliary fuel requirements during start-up, thermal wear on refractory, and plant emissions.

Start-up and *shutdown periods* are much more polluting [85], and there is a strong tendency to allow only pure auxiliary fuel to be burnt during these periods. Waste firing can only start once the operating temperature is reached.

Simple, Small-Scale Forms of Incineration

Burning in the open, e.g., in a dedicated open pit, a barrel, or the foot of an old stack, is both highly

polluting and difficult to master technically. An open pit burner was developed by DuPont to incinerate waste [35]. Wigwam or tepee conical burners were used for burning trash, mostly in remote communities [35], in a more controlled manner than is feasible in the open. American apartment buildings used at times to be equipped with chute fed incinerators [125, 126].

These practices should be banned from any densely populated area. Still, in remote areas, e.g., in parts of the USA and Canada, it is often considered the only option practicable and these units are still largely advertized in the USA, even though their use would probably be forbidden in the EU and Japan. Open burning is obviously a major source of pollutants [127–129].

Stationary Furnaces

Summary. Simple, stationary furnaces are in general use for firing gaseous and liquid fuels or even solid waste, on fixed or rotary grates.

Principles. A furnace combines several essential functions, namely:

- Limiting the cooling of the flames and sustaining an adequate furnace temperature
- Providing adequate retention time in the combustion chamber
- Preventing the uncontrolled entrance of air
- Organizing the flows of incoming primary and secondary air and outgoing flue gas, without undesirable dead corners, entries of false (uncontrolled) air, or diffuse spreading of fumes in case of a temporary rise in furnace pressure

Avoiding *smoke* spreading around requires operating at slightly subatmospheric pressure, since incinerators are always somewhat leaky, a consequence of the heating and cooling cycles inflicted upon refractory and casing. For this reason, furnaces formed from welded membrane steel or boiler tube panels are very popular, ever since their first introduction ca. 1970. The selection of waste burners, their position and capacity, flame orifice, air supplies, mixing, and thermal buoyancy characteristics are prime factors determining performance. The mixing characteristics of the furnace are enhanced by appropriate injection of secondary air, enhanced back-mixing of flue gas, created by reducing the cross section of the outlet and by providing periodic changes in the direction of flue gas flow.

Heat release rates are high when burning high-calorific gases or atomized hydrocarbon liquids; they are much lower when burning sludge or wastewater. Where required, a separate post-combustion chamber is used to control PICs, soot, or smells, with its temperature controlled by an auxiliary burner. An alternative is to provide a catalytic post-combustor [130].

Construction and operation. Stationary furnaces refer to a plain combustion chamber, either horizontal or vertical, of a cylindrical shape or box-type, and fitted with the required start-up and auxiliary burners. Horizontal tubular furnaces are most common, possibly aligned with equally tubular waste heat firetube boilers. Box furnaces exhibit dead corners and were used less than half a century ago.

Vertical furnaces occupy less floor space, gradually narrowing to form the stack and deriving draft from this geometrical design. Nowadays, this arrangement becomes less common, since incinerator flue gas generally requires stepwise and multistage cleaning.

Applications. The stationary furnace is used for burning gaseous and liquid waste flows, including off-gases, solvents, oils, wastewater, pumpable sludge, and meltable and paste-like waste streams. Plastics proper are difficult to fire through a burner, for liquid burners will spin threads of molten plastics. Special burner designs fire several streams simultaneously, e.g., auxiliary fuel, waste oil, wastewater, and pumpable sludge. Alternatively, various wastes may be injected either into a stable flame or tangentially to it. Wastewater may be largely evaporated in a forced circulation evaporator and then radially blown into the flame of an auxiliary oil burner.

Advantages and disadvantages. The main technical limitation of an empty combustion chamber is the lack of provisions for eliminating ash or other residues. Ideally, the ash is fine and high melting and blown out of the furnace, and then separated by the air pollution control devices. Residual ash can be eliminated according to different schedules, such as:

- Operating on a daily shutdown schedule for manual or mechanical cleaning
- Periodic or continuous elimination of ash using suitable mechanical means, such as drag conveyers, augers, retractable grates, rotary grates, etc.

Larger units may incorporate rugged, resilient, yet flexible mechanical provisions to convey ash outward. Since it is undesirable that air leaks in through the ash removal system, a wet or dry sealing system is necessary.

Mechanical Grate Incinerators

Summary. Mechanical grate stokers were originally developed for coal [8, 59], yet since the 1930s they have increasingly been used for MSW.

Principles. Traveling grates support the fuel, while conveying it through the furnace, from the front feeding to the ash-discharging side. Staircase grates provide some tumbling action, when fuel drops from one section to the next. *Reciprocating grates* feature individual grate bars, mounted on alternating moving and fixed frames or sledges; moving the sledge conveys the overlaying fuel and – upon its retreat – turns it over that resting on bars from fixed frames. Several arrangements are possible, e.g., with alternating fixed and mobile steps, or with alternating fixed and mobile staircases juxtaposed. A survey of patent literature reveals a richness of ideas to move waste and separate ash [8].

Construction and operation. Most mechanical grates are subdivided conceptually or physically into successive and distinct drying, combustion, and burnout sections, sometimes separated by small walls, where waste tumbles from one level to the next. The position of the fire is somewhat controlled by the mechanical action of the grate, which supports, conveys, and stirs the refuse during drying, combustion, and burnout. The most common types of grate are reciprocating, reverse reciprocating, roller, rocking, and traveling grates. Proprietary, patented devices provide controlled motion, poking, mixing, and sifting ash between individual grate bars.

Primary combustion air is supplied under the various grate sections to cool the grate and accelerate the burnout of the residue. Today less primary air is used, reducing dust entrainment and the flow of flue gas per unit, and improving thermal efficiency.

Air requirements for drying refuse or for burning out clinker residue are quite low, but supply is ill-adapted to real requirements when active combustion takes place. The vapor and gases, resulting from drying and heating the refuse, are rich in oxygen; combustion products evolve as hot, oxygen-deficient strands. Both should be thoroughly mixed by means of powerful jets of secondary air, blown in through high-velocity nozzles, located at the exit of the combustion chamber. After completing further combustion the flue gas is cooled by a waste heat boiler or – in small plant – by injection of water into a cooling tower. Finally the flue gas is cleaned.

Typical combustion conditions are 850–1,050°C, excess air of 80–200%, but there is a strong tendency to limit it to 6–9 vol.% of oxygen in the flue gas. Some operating conditions are specified by codes, e.g., the EU Directive 2000/76/EC:

- Minimum operating temperatures of 850°C and minimum residence time of 2 s at this temperature
- Minimum level of 6 vol.% of oxygen

Applications. The basic application of mechanical grate stokers used to be for firing calibrated coal. Calibration ensures that all lumps or particles burn out after the same residence time, i.e., by the time the coal arrives at the end of the grate. MSW, however, is all but homogeneous. Deviations from uniformity are catered for by providing a feed that has been homogenized and aged (moisture transfer) in the MSW pit and by specific grate action. A number of options are available for co-firing sewage sludge, waste oil, plastic-rich fractions, etc.

Advantages and disadvantages. Mechanical grate operation has been evaluated a number of times, and in the Western society it can boast decisive advantages with respect to the numerous alternatives tested over more than a century.

Its major limitations are:

- Limited to waste that is supported by a grate. Powders, sludge, and liquid or melting waste are excluded, except for marginal amounts. Reporting to grate siftings impairs their quality.
- Less suitable for waste with extremely low or very high HHV, unless both are well mixed. Fluid bed units are much more flexible in this respect.

In southern climates and developing countries, MSW largely consists of putrescible organics and may be too moist to sustain combustion without auxiliary fuel.

Shaft Furnaces

Summary. Shaft furnaces were rather extensively used a century ago, yet they are currently unusual in waste management [9]. Their fields of potential application are briefly discussed.

Principles. The charge is always fed on top, and slowly descends to the hearth by gravity. The air rises generally from the bottom of the unit, activating the fire in the hearth (countercurrent operation). Unless the feed is carefully calibrated, the gas preferentially rises along bigger channels, reducing the quality of contact with air, as well as volumetric capacity. For this reason, shaft furnaces were unsuccessful in tackling raw municipal solid waste; preliminary shredding markedly improved their performance.

Vertical shaft furnaces have been operated in co-current, crosscurrent, or countercurrent (Fig. 10). Usually the last option is selected, since it easily materializes and ensures heat economy, the incoming charge being dried and preheated by the outgoing gas. As a consequence, any moisture and volatile matter evolving from the charge reports to the gas stream, charging it with organics, tars, and odors. Co-current operation hence has been applied in some gasifiers, with the purpose of cracking tars. Crosscurrent operation was applied by WSL/Foster Wheeler in an unsuccessful rubber tire pyrolysis process.

Construction and operation. The shaft furnace consists of a vertical, cylindrical shell protected by inner refractory and thermally insulating lining. Top feeding features a suitable lock for exclusion of air and possibly a distributor for equal distribution of the feed over the entire cross section. Ash extraction proceeds mostly either by means of a rotary grate for ash extraction, or by periodic molten slag tapping.

Applications. Already in Roman times shaft furnaces were used, for calcining limestone. Traditionally, they have been used in the iron and steel industry (blast furnaces), foundries (cupola furnaces for melting metals), and for wood and coal gasifiers.

Shaft furnaces appeared in some ancient incinerator systems (1880–1930), either as combustor or as ash burnout element, the inherent heat exchange assisting in burning low calorific waste with combustion air preheated by hot ash. The rising gas is heavily charged with thermal decomposition products from waste and hence it requires post-combustion. Early Dörr, Didier, Stockholm furnaces are discussed in some detail by Reimann [4].

Advantages and disadvantages. Major advantages are countercurrent heat exchange and a relatively low load of dust. Major areas of potential operating problems with shaft furnaces are [9, 22]:

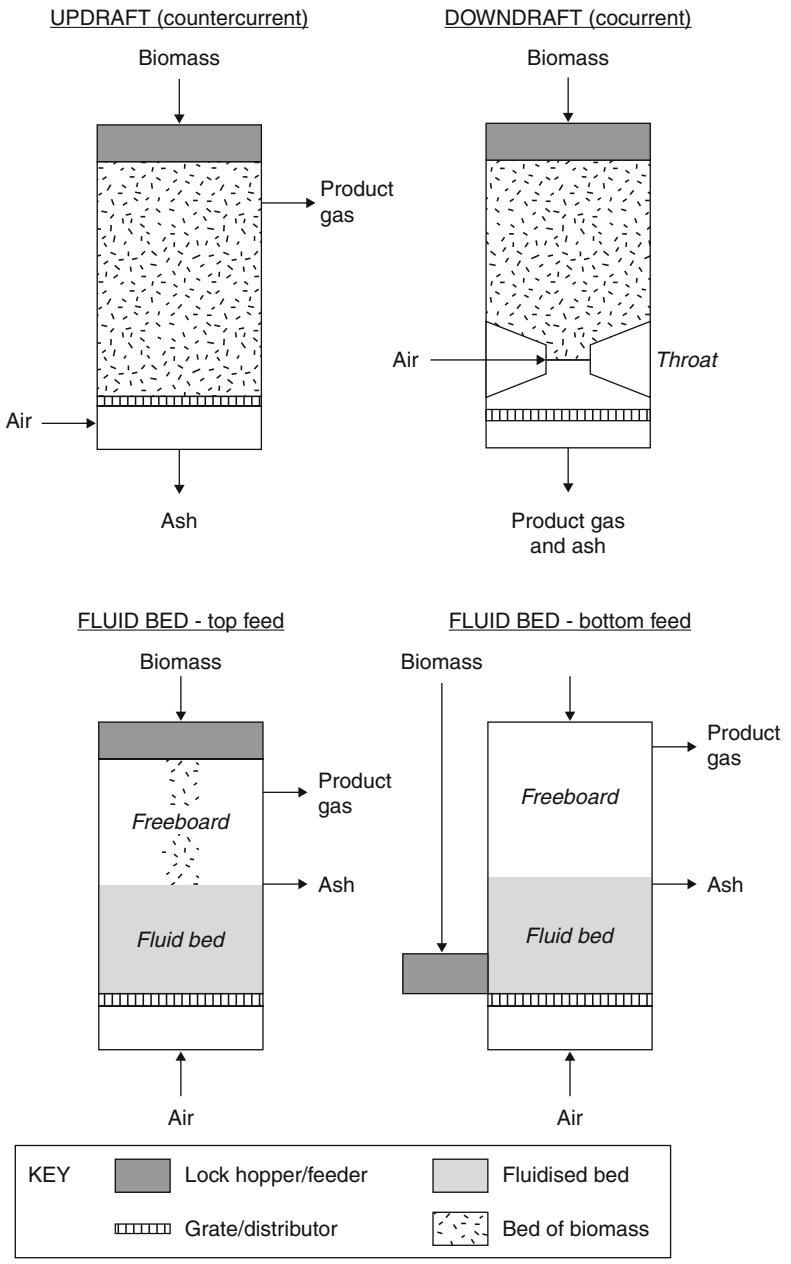
- Volatiles and moisture emanating from the charge report to the gas, requiring either post-combustion or adequate treatment of these compounds.
- Unless special care is taken to homogenize size and shape of the feed, the rising gas will be channeling through preferential pathways in the charge and along the wall, due to both irregular bed porosity and chimney effects.
- Peripheral fires occur, due to the larger voidage of the charge close to the walls.
- Failing possibilities for controlling temperature, gas flow, and oxygen distribution throughout the charge.
- Controlling ash extraction, whether as cinders or as molten slag.

Rotary Kiln Incinerators

Summary. Rotary kiln incinerators since the 1960s are state of the art in the incineration of industrial waste, including commercial and hazardous waste. These kilns operate either in countercurrent (cfr. long cement kilns), or in co-current (in short kilns, usual in industrial waste incineration). Cross-flow is possible only when using special constructions, e.g., involving telescopic kilns, or in mid-kiln feeding over a complex fixed feeding/rotary kiln device.

Principles. More or less as the shaft furnace, it can operate in countercurrent, co-current, or – with some more difficulty – in crosscurrent mode. Incinerator operation is generally in co-current, with either solid or liquid ash discharge. Tumbling the waste renews the furnace. Since there are no provisions for mixing gas phase strands, a post-combustion chamber is always required.

Construction and operation. A rotary kiln incinerator is typically composed of a stationary feeding system, a rotary kiln with slightly inclined cylindrical shell, a stationary ash discharge system, and



Incineration Technologies. Figure 10
Co-current, crosscurrent, and countercurrent operation of vertical shaft furnaces

a post-combustion chamber, followed by a waste heat boiler (or quench cooler), and air pollution control units [42].

The stationary feeding system consists of a feed hopper, a lock, and a steeply inclined chute. The whole fixed front panel can be mounted on rails. In a patented system, a knife rides on the sides of the feed hopper, cutting off ribbons, plastic film, and textiles, preventing a flashback of the flame into lock and hopper. The lock is formed by two mechanically, hydraulically, or pneumatically operated slides, which are interlocked so that a slide only opens when the other

is closed. The lower slide and the chute are both water cooled. The feeding system can be provided with an explosion relief system, diverting a shock wave into an innocuous direction.

Several systems were tested for homogenizing feed materials and supplying them at a constant rate. Some rotary kilns were fed by a screw conveyor or hydraulically operated ram feeders. Early BASF plants used simple, strong centrifugal pumps with large clearance between rotor and housing, capable of macerating material and pumping the resulting slurry. The rotary kiln was fed from a rotating mixing and storage drum, blanketed with nitrogen.

Kiln lining. The cylindrical shell is internally lined with refractory, selected with regard to the expected operating temperature and slag melting point and reactivity. When using high-quality, dense brick, a continuous operation is necessary to avoid thermal stresses. Gradual heating up may take as much as 60 h and cooling down 24 h. A lifetime of 2 years is considered to be good in normal operation.

No general rules can be formulated regarding the best or more economic furnace lining. In some plants, inexpensive ramming compound or hard chamotte bricks were successfully used. In others, chemical attack was so extensive that lifetimes remained too short, even with expensive high-alumina or magnesia bricks. Chemical attack depends on chemical composition of both lining and ash, and on temperature. Refractory is also subject to abrasion and spalling [35].

Protecting the walls with a layer of solidified molten slag and outward cooling by water sprays have been successfully applied for lengthening lifetimes. Slag accretions can be melted away periodically by slightly elevating the temperature. Iron oxide, when burning barrels, forms low melting silicates enhancing slag fluidity. Slag reactivity and melting point has sometimes been decreased successfully by addition of suitable charges, e.g., sand.

Kiln movement. The peripheral speed of the kiln can be varied continuously using a single drive, with driving pinion and bull gear. The shell is provided with riding rings, rolling on support rollers to obtain a uniform distribution of bearing forces over the shell. In case of power failure, an auxiliary motor should drive the shell to prevent thermal deformation.

Air sealing between rotary shell and stationary loading and discharging equipment at the ends is critical. Excessive air leakage should be prevented by provision of suitable angle or segment seal rings.

In the most usual *co-current operation* both wastes and combustion air are introduced at the front end of the furnace. An auxiliary burner is installed in the fixed front panel of the kiln to enhance drying and accelerate preheating and ignition of the wastes. In the absence of such a burner, drying and preheating completely depends on radiant heat transfer from the rear part; the rate of radiant heat transfer is proportional to the fourth power of temperature (in °K), attained in the hottest part of the kiln. Operation is at 1,200–1,500°C in a slagging operating mode or below 1,000°C in the dry extraction mode.

The rotary kiln is sometimes operated in *counter-current* when relatively wet wastes with a low heating value are to be incinerated (e.g., sewage sludge). Counter-current operation is unsuitable for other waste, because of the risk of flame flashback into the charging lock.

A *partial countercurrent* operation is sometimes used in very short kilns. A good mixing pattern is obtained by using an auxiliary burner in the fixed rear panel. The burner creates a backward gas flow along the kiln axis Fig. 8.

Kiln Internals

The residence time and flow pattern in principle can be modified by installing conveying spirals to guide the materiel flow, ring-dams to retain melted or chains for granular material, or by providing an enlarged cross section near the discharge end to reduce the gas velocity and provide a soaking period at high temperature. A spiraling dentition can be provided in the refractory to retard the forward movement of wastes and enhance the contact between burning wastes and combustion air. The higher cost of the lining limits the practical use of these various patented devices, also prone to erosion and clogging.

Air Supply

Primary air is blown in through a set of nozzles, located on the fixed, front side of the furnace. No secondary air can be distributed along the kiln, unless it is composed of several sections of a different diameter in a telescopic arrangement. The excess of air is large, to make up for

sudden variations in calorific value, and often amounts to 200–300%. Typically 8.000–12.000 m^3 of flue gas is generated per ton of waste.

Combustion air is blown in tangentially at the front end and creates a whirling movement along the wall. Superposition of the two different flow patterns results in a reasonable amount of gas phase mixing, a more uniform temperature and increased kiln capacity.

Residence time and turbulence in the gas phase are both limited. Hence, combustion is to be completed in post combustion chambers providing a supplemental residence time of 2–3 s. The temperature in these chambers is often maintained above 800°C with auxiliary burners firing fuel or liquid waste (solvents, oil).

Facts and Figures

The volumetric rate of heat generation varies, depending on the combustion temperature, between ca. 400,000 and 1,000,000 MJ h^{-1} m^{-3}. Thermal efficiency of the rotary kiln plant is low, limited typically to 55–60%, due to the large excess of air and the various heat losses.

Rotary kilns are built industrially with diameters from 1 to 4.5 m and a length typically from 3 to 15 m. The largest kilns have a capacity of 60 GJ h^{-1}. Scaling-up problems arise, because kiln volume is proportional to the square of the inner diameter, the available exposed surface of waste only to the inner diameter D_i. Hence, multiple kilns are preferred over a single, large diameter one.

Applications

The concept of rotary kiln incinerators was developed at BASF-Ludwigshafen, probably inspired by the much longer units used for producing cement clinker, for calcining limestone or for roasting pyrite and sulfide ores. Yet, the short rotary kilns used in incineration retain the tumbling action rather than countercurrent operation and intrinsic heat exchange.

Dedicated rotary kiln incinerators are capable of eliminating almost any type of industrial wastes, e.g., plastics, oil contaminated sludge, waste paint, solvents, pesticides, spent chemicals, and even explosives (in small amounts). Explosive combustion is relatively harmless, the combustion chamber being spacious and followed by a post-combustion chamber. The rotary kiln is not highly regarded as an incinerator of municipal refuse because of excessive wear of the lining and the absence of possibilities for longitudinal air distribution.

Solid and paste-like wastes, sometimes even complete barrels filled with waste are introduced into a hopper, with a lock system and a chute, located at the stationary upper end of a slowly rotating, slightly inclined cylindrical furnace. The wastes slowly slide and tumble by the rotary movement of the kiln; this provides for mixing and a periodic surface renewal of the burning charge. On their way from the higher feeding side to the lower discharge end, the wastes are rapidly dried, heated, and ignited under the action of radiant heat from the furnace walls. The kiln is generally filled up to 10–20% of its volume. The *residence time* of solid and paste-like waste depends on the length of the kiln, its speed of rotation, the possible presence of a profile in the refractory lining, and gas velocity. Generally residence times of less than 1 h are selected. The ash is discharged into a water bath, located under the lower end of the kiln. In some plants larger pieces of residue are retained on grizzly screen bars, to protect the ash-discharging conveyor [131].

The rotary kiln is used as a drying furnace for, e.g., sewage sludge, in the roasting of pyrites and sulfide ores, the calcination of limestone and the production of cement clinker. In Great Britain, Belgium, and Germany, pulverized refuse was used as a supplemental fuel in coal-fired kilns. Later, rubber tires and hazardous waste in numerous plants became a standard supplement in cement clinker manufacturing.

Westinghouse proposed a unique combination of a rotary combustor with an integrated boiler (O'conner) [35]. In only few cases wastes have been treated or incinerated in a metal-walled rotary kiln having no refractory lining at all (the red factory, of Prayon, Engis, Belgium).

Advantages and Disadvantages

The following problem areas have been identified:

- The charging chute is exposed to heavy wear because of feed sliding and tumbling and condensation of corrosive vapors. Sometimes cracks occur along the welding.

- The kiln lining is exposed to heavy wear and chemical attack. The action of corrosive melted slag is important at the kiln end mainly.
- Air sealings are exposed to dirt, wear, and high temperatures.
- The lower part of the combustion chamber is exposed to attack by entrained droplets of melted ash.

Tilting Furnaces

There are various kinds on rotary kilns, distinguished by their shape (cylindrical, conical), their aspect ratio L/D, or even their rotary movement.

Laurent-Bouillet proposed a particular type of furnace, with a typical conical-cylindrical shape, in which MSW is subjected to an oscillating movement [132].

Multiple Hearth Furnaces (MHF)

Summary MHF have been developed in the nineteenth century for ore roasting and treatment and metallurgical applications are still leading in Europe. Especially in the USA, they have been applied for sewage sludge incineration.

Principles The MHF is a cylindrical construction, composed of a number of circular hearths mounted one above the other. Each hearth contains an air-cooled rabble arm, driven from a common central shaft. Blades, fitted to the slowly rotating rabble arm move the material forward – depending on the angle at which they suspend from the arms – either toward the center or toward the periphery, until it passes over a discharge aperture and falls onto the lower hearth.

The retention time of the charge is varied by changing the speed of rotation of the rabble arms or, rarely, by adapting their relative position to the floor.

Construction and Operation Multiple hearth furnaces (MHFs) consist of a series of superimposed hearths, solids being fed on top and descending stepwise by gravity, after describing a spiraling movement on each hearth, starting at the discharge point of the higher hearth and ending at that to the lower hearth. Gases generally mount, in countercurrent to the movement of solids, aided by buoyancy. The feed material is charged onto the upper hearth and slowly makes its way down, falling from one hearth to the next, while it is progressively dried, heated, ignited, combusted, and finally cooled by the combustion air. The latter is introduced in part or all at the bottom of the furnace, preheated by the ash on the lower hearth(s), and partly consumed on successive combustion hearths. The resulting flue gas is cooled by the incoming feed and leaves the unit toward possible post-combustion and cleaning. Auxiliary burners are used for preheating and adapting and controlling the temperature profile. The atmosphere is controlled by balanced introduction of air, recirculation, or other means [131].

Applications MHFs were originally developed for roasting sulfide ores (Nichols-Herreshoff). Later they were adapted for sewage sludge incineration and for competing with fluidized beds. They provide a controllable temperature record to the feed, generally involving sequential drying, heating, reacting, and cooling hearths. There is much contact surface with air and this surface is periodically renewed by the passing rabble arms with attached plates, plowing through the material.

The main application in waste is incinerating sewage sludge and regeneration of spent carbon or lime. The heat required for drying sludge can – when desirable – be supplied by firing pulverized refuse on lower hearths as an auxiliary fuel.

Lucas Furnace Developments, Ltd., once designed a rotary, single hearth furnace, sloping down from the periphery toward the center. It was proposed for incinerating sewage sludge, old tires (without any prior size reduction), and plastics. After preheating the furnace, waste was fed at regular intervals by means of a ram feeder. As the solid hearth slowly rotates the waste first moves along the outer periphery and gradually spirals to the central discharge point. Finally, the ash falls into a quench tank and is removed by a scraper conveyor.

In this Lucas furnace, the gas flow is organized for cyclonic combustion. High-velocity nozzles direct the combustion airflow tangentially into the furnace, cooling the walls to 850–900°C. The central temperature attains 1,450°C. The plant operates at 80–100% excess of air. Operation at reduced capacity suffers from loss of turbulence in the gas phase, a problem that can be tackled using auxiliary steam jets.

I

Advantages and Disadvantages

MHFs are a proven and traditional technology that allows a flexible adaptation of operating conditions on each hearth. This versatility is less available in rotary kilns or shaft furnaces.

Because of its complex construction this furnace is limited in its maximum capacity. Moreover, it takes long times to preheat and shutdown MHFs, since it is important to avoid thermal shocks.

Fluidized Bed Incinerators

Summary Fluidized (bubbling) bed combustors are exceptionally adaptable, allowing to burn (or gasify, if air supply is sub-stoichiometric) solid, pasty, melting, liquid, sludge, slurried, and gaseous waste, simultaneously and at unusually low temperatures. Moreover, they admirably accomplish in-bed solids mixing and heat transfer and leave a neatly polished solid residue, sinking in the bed. Fine ash is entrained, including particulate formed by attrition or erosion. Desulfurization with limestone or dolomite is possible in bed at combustion temperature and thermal NO_x-formation remains negligible. The principle limitations are the relatively important pressure drop, as well as bed agglomeration, in the presence of tacky ash or salts. Draining decanted residues requires a dedicated circulation circuit of bed materials, adding to mechanical complication: the extracted bed material is sieved and the underflow returns to the bed.

Circulating bed incinerators have been developed and used since the late 1990s by Zhejiang University and also by Tsinghua University. In China, circulating fluid bed units are unusually popular, since combustion stability can be maintained simply by adding cheap coal, instead of using expensive oil [133].

Principles of Fluidization Consider a fixed bed of granular media, such as sand, ash, or limestone, supported on a porous plate, the distributor plate. An upward current of fluid traversing this layer incurs a pressure drop Δp, which rises as the fluid flow rate increases. Meanwhile, the bed porosity (i.e., the void volume) gradually expands. At a given flow rate the pressure drop Δp even equals the pressure, exerted by bed weight. Then, the minimum velocity of fluidization u_{mf} is reached and the head loss corresponds to the weight of the entire bed per unit of cross section (+ the friction loss, omitted here):

$$\Delta p(u_{mf}) = \rho A H g \tag{20}$$

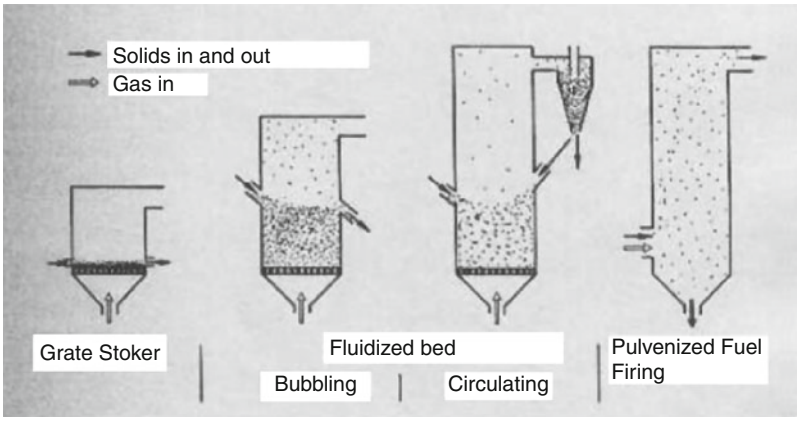
with Δp = pressure drop, m^2
u_{mf} = minimum fluidization velocity, m s^{-1}
A = bed cross-section, m^2
H = bed height, m
g = acceleration of gravity, m s^{-2}

In principle, the bed thus reaches kind of a state of levitation. However, the fluid (further termed "primary air") will not carry the bed upward, given its granular structure. Rather, excess air trickling through the bed starts forming bubbles at the orifices of the distributor. Such bubbles, after a while, leave the distributor and rise through the bed, more or less like bubbles do in boiling water. The bubbling bed thus resembles a boiling liquid, with series of bubbles rising from the bottom and bursting at the surface. This upward movement of air bubbles creates excellent mixing patterns in the bed leading to temperature homogeneity. Light materials tend to float; dense materials sink to the bottom of the bed. Fluidized beds may be used both as separator of stones or metals present in waste, at low gas velocities, and as a mixer, at velocities a multiple of u_{mf}.

The value of u_{mf} can be estimated using empirical correlations. The gas flows both as bubbles and as a "dense phase" trickling flow. Smooth fluidization is obtained only at a gas velocity of three to four times u_{mf}, together with suitable particle mixing and heat transfer characteristics.

With rising gas velocity, entrainment of fines becomes more important, depending on their terminal falling velocity. The entrainment of bed particles specifies an upper limit of gas velocity. Depending on particle size (often 0.3–0.8 mm), gas velocities typically range between 0.3 and 5 m s^{-1}. There is gradual transition from a bubbling bed to a circulating fluid bed, the latter characterized by a cycle of entrainment, particle separation, and recirculation (Figs. 9 and 11).

Principles Fluidized bed incinerators burn waste, suspended and moved around erratically in a vigorously bubbling bed of hot granular material. High heat generation rates can be obtained despite the low operating temperatures, typically 750–900°C.



Incineration Technologies. Figure 11
Evolution from a fixed bed, over a bubbling bed and a circulating fluid bed to an entrained flow combustor, with rising gas flow (Görner [34])

In bubbling fluidized beds, combustion of volatile matter mainly takes place in the freeboard zone, i.e., above the bed. This freeboard zone should be ample and well mixed, so that reducing and oxidizing strands can mix and burn out completely. Secondary air provides the swirl required for mixing. The bed material provides a thermal flywheel that effectively copes with short-term fluctuations in feed rates and quality.

Construction and Operation A fluidized bed incinerator consists (starting below and moving upward) of an empty plenum chamber; a distributor supporting the bed; the bed of granular material to be fluidized; a freeboard zone ensuring disengagement of entrained particles and post-combustion, and adapted feeding; start-up and auxiliary heating burners; waste heat boiler; and pollution abatement equipment. The *distributor* supports the bed material and evenly distributes the primary combustion air over the entire cross section. Even distribution requires the pressure drop over the distributor to be at least 0.1–0.2 times the pressure drop over the bed. The distributor should also prevent weeping of bed material into the plenum chamber, or its cycling between both zones, with erosion as a consequence. The distributor is made of heat-resistant alloy or forms an arch of refractory material. In principle, there is a wide range of possible designs; a bubble cap distributor is selected most often.

The *bed material* consists of a graded fraction of clean heat-resistant material, such as sand, more seldom alumina, limestone, dolomite, or ash. Pollutants, such as SO_2, can be removed in situ by simply feeding limestone or dolomite into the bed. The bed is preheated to operating temperature by generating hot air in a separate furnace or using a start-up burner directed toward the surface of the bed and then fluidizing gently.

Combustible waste and *auxiliary fuel* are generally fed into the bed to ensure that the heat of combustion is largely generated inside, rather than above the bed, in the freeboard zone. Yet, much of the combustion of the volatile matter will burn above the bed. Gas is fed through independent bubble caps, liquid fuel, slurries and pumpable sludge, and lances, and solids by means of a screw or a pneumatic feeder.

Low-calorific wastes can be dropped onto the top of the bed by a chute fed by a conveyor belt, or sprayed over the bed by means of suitable nozzles. Mechanical spreaders may assist in obtaining a more uniform distribution of the feed. The falling droplets or particles are partly dried while dropping onto the bed. Once in the bed, drying, heating, ignition, and burnout proceed very rapidly. Feeding large pieces lead to the local evolution of excessive amounts of volatiles, causing total depletion of oxygen, evolution of clouds of pyrolysis and gasification products, and, eventually, of

sequences of CO and TOC peaks. For this reason the feeding rate should be bit by bit and steady, and feed materials should not be larger than 5 or 10 cm.

Bubbling beds project particles into the freeboard. Most settle in the *freeboard zone*, but the finer ones are partly entrained. Finer particles are separated in internal or subsequent cyclones and flow back into the bed, to complete their combustion. Secondary air is injected into the freeboard zone to complete the combustion.

Applications Fluidized bed technology was first applied in the 1920s, in coal gasification (Winkler) [134]. Fluid catalytic cracking of gas-oil to gasoline followed during World War II (Massachusetts Institute of Technology, Esso) [134]. Other important industrial applications are the roasting of sulfide ores and the drying of polymer powders. The most significant applications are the incineration of wastewater sludge and black liquors from wood pulp manufacturing [135].

Fluidized bed incineration, gasification, and pyrolysis of shredded or classified refuse have been widely developed in Japan, Finland, and Scandinavia.

Advantages and Disadvantages Fluidized bed incinerators are relatively simple to build, operate, automate, and maintain. They have no moving parts at high temperatures. Yet, high heat generation rates and bed-to-wall heat transfer rates are obtained due to the high-quality gas-solids contact. Complete combustion is possible already at a low temperature (750–850°C) and a low excess of air (15–35%); hence the volume of flue gas to be cleaned and the NO_x generation rate are relatively small. Bed material can easily be added or removed (draining at the bottom or overflow), which allows adding also lime or dolomite.

Thanks to the large thermal capacity it is also possible to absorb important step changes in feeding rate and even to operate intermittently: cooling of the bed after shutting down is very slow, so that proper operating conditions can rapidly be reached after a standstill of 1 or 2 days.

On the other hand, both the power requirements for fluidization and the dust content of flue gas are quite high. Dense material may segregate and accumulate on the distributor plate, which can be avoided by using an appropriate distributor design, such as sloping distributor plates or arrays of spaced perforated tubes sparging air into the bed. A waste heat boiler and/or preheated air are required to reduce the stack heat losses.

The most serious operating problem occurs when the combustion temperature increases beyond the softening point of the ash. Rapid particle agglomeration then occurs, followed by solidification of part or all of the bed material. When this happens, the solidified material has to be excavated by pneumatic hammers after cooling of the bed.

Vortex Combustors

Summary Vortex incinerators can be used for high-rate combustion of gaseous, liquid, and finely divided solid fuels or wastes. Larger particles require longer residence times and may not burn out completely; in this case, supplemental mechanical means, such as a specific burnout grate, have to be provided for retaining the burning residue [9].

Principles Vortex firing involves a highly turbulent mode of combustion, featuring fast transfer of heat and mass and resulting in high volumetric rates of heat release.

Construction and Operation Fuel (or waste) is blown in tangentially into a conical or cylindrical furnace. The rotary movement of combustibles suspended in combustion air as a carrier creates excellent mixing conditions and hence high combustion intensities and temperature homogeneity.

Two vortices are formed: an outer one consisting of combustion air and waste and an inner one of burning gases. The cooler outer flow shields the refractory walls from overheating and it is rapidly preheated by the hot inner core, considerably stabilizing a steady combustion. A wide range of operating temperatures and a low excess of air can be used. This may lead to a slagging operation.

Applications Tangential firing involving vortex combustion has been used extensively in coal-fired utility boilers. In one design, pulverized coal is fired in a separate cylindrical vessel, somewhat inclined to the horizontal. In a second design, pulverized coal, together with combustion air, is injected from the

four corners of a vertical chamber with a square cross section. Jets are directed tangentially to an imaginary circle contained in this section.

Heat rates in such cyclonic furnaces allow liquid tapping of slag, also with medium HHV waste, such as dry straw or wood chips. Wet bottom cyclonic furnaces for firing coal with an unusually low ash melting point have been developed.

Advantages and Disadvantages Cyclonic furnaces are compact and highly productive. These advantages weigh more, in case the unit capacity is important. The principle is applied less often for small plants. A high-temperature operation may lead to considerable NOx formation, unless excess oxygen is really minimized.

Slagging Incineration

Summary *Slagging incineration* also termed "wet bottom operation", is an option whenever combustion is conducted at quite high temperatures, or when the resulting ash has an unusually low melting range. This occurs when the waste contains certain groups of chemicals in its ash, such as borates and numerous alkali salts.

Principles *Wet bottom* combustors fire fuel (coal, or waste) at temperatures exceeding the melting point of ash. Incinerators may operate in this slagging mode in case:

- Waste is sufficiently high calorific and combustion temperatures are adequately high
- The resulting slag either has adequate fluidity, or fluxes are added
- Provisions for tapping molten slag are available

Preheating combustion air, enriching it with oxygen, providing auxiliary fuel, and dissipating electric power in the charge (Ohm effect, electric arc, hot plasma) all allow to raise combustion temperature to higher levels and addition of fluxing agents (fluorspar, iron oxides, lime) may be used to lower the melting range and enhance slag fluidity. Molten slag can be tapped discontinuously and discharged to solidify to large crystalline blocks. Generally, however, it is quenched by pouring the melt into a water bath, converting it into small glassy grains. Continuous tapping is uncommon, given the small capacity. When treating metal rich waste, two phases might be formed: a light slag floating on top of molten metal.

Construction and Operation There are numerous different methods to conduct incineration or even gasification under slagging conditions. Such methods encompass, e.g.:

- Shaft furnaces, operating like a blast furnace. Examples: Lurgi pressurized moving bed oxygen/steam gasifiers. Union Carbide, Andco-Torrax and Nippon Steel gasifiers.
- Rotary kiln incinerators operating on high-calorific waste and in a slagging mode.
- Mechanical grate incinerators, fitted with a dedicated furnace to melt the residue.
- Electric arc furnaces (cfr. metallurgical Industry).
- Suspension firing. Cfr. The Vortex Furnace and Koppers-Totzek Coal Gasifiers.

Applications Already in 1934 Rummel [134] pioneered the concept of using molten material (these could be slag, metals, or salts) as heat carrier and oxygen transfer agent. Ever since these 1930s, slagging incinerators have been experimented with, in association with cokes addition, electric arc furnaces, plasma torches, with special waste with unusually low ash melting traject, or with waste streams, warranting high disposal cost, e.g., radioactive waste, or PCBs.

Recently, slagging operation in Japan became a standard, following the necessity of converting incinerator residues into glassy slag.

Nippon Steel has developed blast furnace technology, applied to MSW.

Ebara Co. has pioneered a fluidized bed gasification plant, featuring post-combustion under slagging conditions of the gas.

Advantages and Disadvantages Slagging incineration has several potential advantages, such as

- Simpler furnaces, ash flowing along an inclined floor
- Generating dense glassy slag, with low leaching rates and almost free from combustible inclusions
- Operating at low excess of air, reducing flue gas flows to be cleaned as well as stack losses.

The major problem is ensuring steady fluidity of slag, while still limiting its attacks on refractory. A thin layer of solidified slag may be maintained on the refractory lining to cover and protect the refractory (Fig. 12).

It is recommendable to separate combustion from ash melting, by providing a controlled supply of heat and possibly flux in the slag tapping area. Slagging operation exists in many variants. Heat supply is secured by the following means, alone or in combination:

- Plasma torch
- Addition of coke
- Auxiliary burners
- Combustion of gasification products

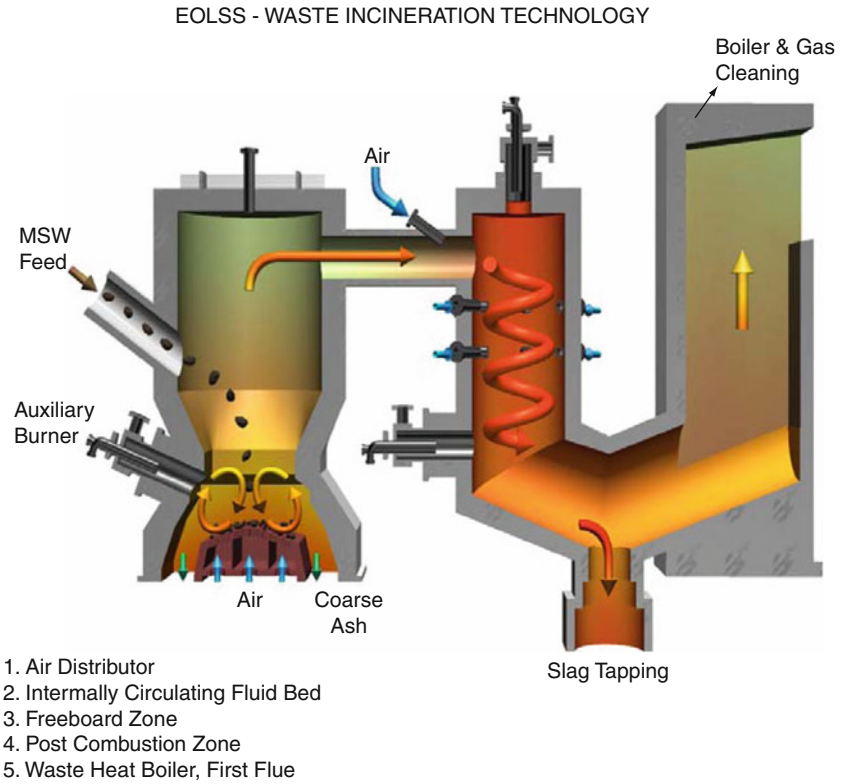
Example: Incineration in a Molten Salt Bath Incineration in a molten salt bath has also been applied when dealing with hazardous wastes, such as pesticides, explosives, etc. Combustion in a bath of molten salts (e.g., sodium carbonate, potassium carbonate,) could present the following advantages:

- Molten salt acts as a heat carrier and combustion catalyst, ensuring complete oxidation at temperatures lower than normally required.
- Carbonized residues and dust and ash particles are entrapped in the bath.
- Acidic pollutants in the off-gas react with the salt and are also retained in the bath.

A potential disadvantage is the required disposal or regeneration of spent salt. Moreover, managing volatilizing salts is problematic, since deposits of desublimated salt fumes will need to be removed periodically.

Submerged Combustion

Summary Submerged combustion combines a vertical conventional combustion chamber with immediate



Incineration Technologies. Figure 12
Fluid bed gasifier with slagging post-combustor (By Courtesy of Ebara Co., Japan)

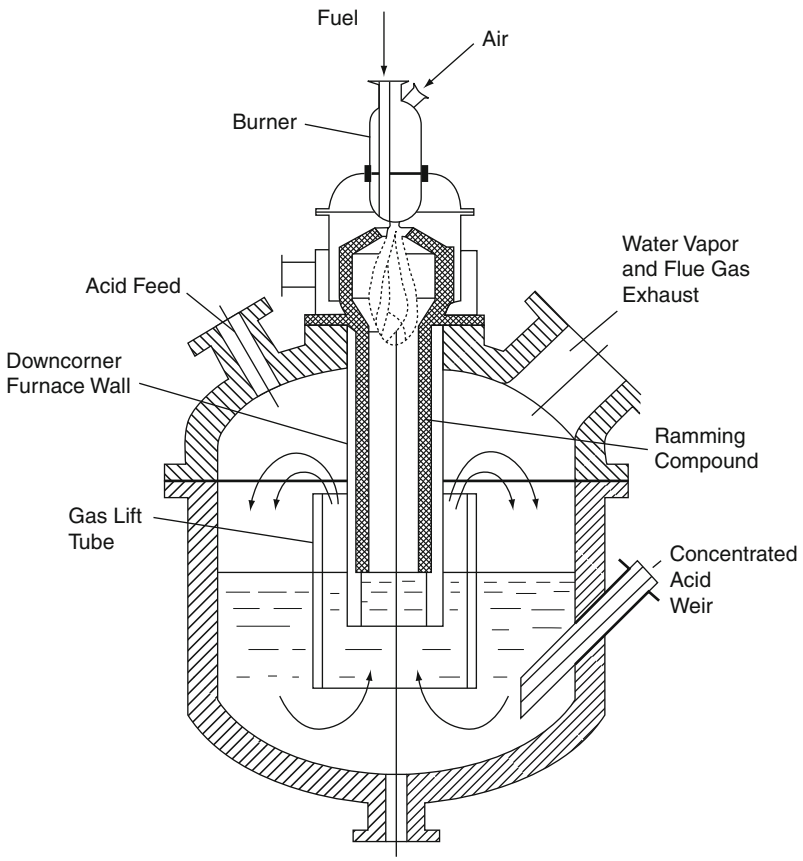
direct-contact quenching of flue gas by its immersion in aqueous liquor, brine, lye or acid, ensuring fast heat and mass transfer in a bubbling liquid mass.

Principles Submerged combustors feature a vertical, refractory lined steel shell furnace equipped at the top with a down-firing burner, the flue gas of which bubbles through a reservoir filled with quenching and scrubbing water or other aqueous liquids (Fig. 13).

Construction and Operation The plant consists of a slender, elongated, vertical furnace, plunging as a downcomer tube into a bath retained in a wider container. Frequently, a concentric tube surrounds the downcomer, forming an annular space, acting as an airlift and promoting internal mixing of the liquor contained in a quenching bath. As a result the gas is suddenly quenched, freezing undesirable reactions, such as forming chlorine according to the Deacon equilibrium of reaction (13).

Waste sulfuric acid, brine, or other corrosive solutions can thus be concentrated by direct contact between hot flue gas and the liquor to be treated. Heat and mass transfer between flue gas and liquid quench are almost instantaneous: the quenched flue gas is almost completely saturated with water vapor and leaves substantially at the temperature of the quenching bath.

Applications Submerged combustors have a long tradition, with first patents in the nineteenth century. Chemico used it as a means to concentrate dilute sulfuric acid, Nittetu Chemical Engineering to treat chlorinated waste [137]. It is mentioned in the IPPC report [136]. Submerged combustors are used to:



Incineration Technologies. Figure 13
Submerged combustor (Günther [34])

- Quench and clean flue gas, arising from the combustion of chlorinated organics [137] or CFCs [138, 139]
- Concentrate wastewaters or corrosive acids, maintained as quenching bath
- Recover a solution of inorganic salts, when firing aqueous solutions of organic or inorganic salts
- Recover dilute hydrochloric acid of acceptable concentration when firing chlorinated wastes

The liquid wastes are atomized and injected into the flame, so that they are rapidly dried, thermally decomposed, and completely combusted. Inorganic compounds are converted to tiny molten salt particles and are recovered as a salt solution or slurry in the quench vessel.

Another design (Nittetu) features a fractionating column mounted on top of the water vessel. Wastewater contaminated with volatile hydrocarbons is fed on top of the column and hydrocarbons are stripped off in contact with a rising mixture of flue gas and water vapor. After condensation of the vapors, the condensed heavier hydrocarbons are recycled to the quench vessel. The noncondensable is combusted.

Advantages and Disadvantages Submerged combustors are relatively simple, efficient equipment. Quenching and scrubbing are fast and direct-contact. Chlorine formation is suppressed when firing halogenated solvents and vapors. Also dioxins formation is suppressed.

A dedicated website is [140]; various hardware are presented in [141] (Table 9).

Refuse-Derived Fuel

Rather than firing waste as it comes, one can convert it into storable fuel, following a suitable sequence of operations, composed of primary and secondary shredding, grading, wind sifting and screening, magnetic and eddy-current separation, etc. Suitable combinations of such operations may convert municipal solid waste, packaging, wood, paper and plastics, etc., into better manageable and storable refuse-derived fuel (RDF) with more predictable characteristics and specifications, such as HHV, and proximate and ultimate analysis. RDF assumes different forms, such as fluff, powdered (after adding embrittling agents), or densified, i.e., in bales, pellets. Already in the 1970s, the National Centre for Resource Recovery (Washington) tried to standardize RDF, to improve its acceptance and access to the energy markets [9, 142].

The preparation of RDF may proceed according to very simple as well as more complex schemes, promising higher quality as well as more investment and operating cost. EcoFuel® was a powdered product, obtained by raising the temperature and adding an embrittling agent, e.g., sulfuric acid. In one case, processing started by wet pulping (Black Clawson at Franklin, Ohio); the resulting RDF was wet during processing, which is counterproductive for thermal utilization. Moreover, the only large plant in Florida suffered from odor problems and was eventually dismantled. Yet the processing generates clean fractions of metals and glass.

Today, mechanical biological treatment [143, 144] is proposed as a generic group of processes to produce a fuel fraction and sorting fractions.

The complete combustion of solids generally requires residence times of typically half an hour, except for some high-intensity incinerators, such as those firing a *refuse-derived fuel* (RDF). In principle, the residence time available for combustion is comparable now to that of flue gas, i.e., a matter of seconds only for suspension firing up to about a minute in bubbling fluidized bed plants. Circulating fluidized bed units feature cyclonic separators that collect coarse matter for recycling. In practice, the denser RDF will thus be retained by suitable aerodynamic and geometrical means (gravity or centrifugal force) that extend its residence time until the residue is so fine that it is entrained.

RDF preparation is expensive, typically in the range of 20–80 US$/Mg. These costs derive both from investment and operations and include power consumption and heavy wear on equipment. Fire and explosion hazards are notoriously present during shredding, drying, and even longer-term storage.

Co-firing of Waste or of RDF

In most cases waste is incinerated in dedicated furnaces. In some instances, it may be more attractive to

Incineration Technologies. Table 9 Typical combustion conditions

Combustion conditions	Combustion temperature (°C)	Furnace temperature (°C)	Gas velocity (m s^{-1})	Residence time (s)	Air number (–)	Thermal volumetric load (MW m^{-3})	Thermal cross-sectional load (MW m^{-2})
Power plant:							
Natural Gas	1,100–1,400	1,000–1,100	5–10	1–3	1.05–1.1	0.25–0.35	5–8
Oil	1,100–1,400	1,000–1,100	5–10	1–3	1.05–1.2	0.25–0.35	5–8
Grate stoker	1,100–1,300	1,000–1,100	4–9	1–3	1.3–2.5	0.15–0.35	0.5–2.5
Fluidized bubbling bed combustor	750–1,050	750–1,050	0.5–5	1–3	1.2–1.4	2–5	1–2
Circulating fluid bed combustor	750–950	750–950	5–8	0.5–6	1.12–1.3	8–20	2–8
Pulverized coal firing	1,100–1,500	1,050–1,250	5–10	1–3	1.13–1.3	0.06–0.3	0.6–3
Pulverized lignite firing	1,100–1,300	950–1,150	4–8	1–3	1.2–1.5	0.06–0.15	2.5–5
Pulverized coal firing (wet bottom)	1,300–1,600	1,000–1,150	5–10	1–3	1.15–1.3	0.1–0.4	4–6
Mechanical grate MSW-incinerator	1,000–1,250	1,000–1,100	3–8	3–6	1.5–2.0	0.15–0.35	1.4–1.6
Fluid bed sludge combustor	750–900				1.05–1.8	1–3	0.5–1
Stationary combustion chamber					1.2–3	0.1–0.3	0.1–1
Rotary kiln incinerator					1.6–3.5	0.15–0.2	1.5–2.5
Postcombustion chamber	1,050–1,300	1,050–1,250			>1.4	0.08–0.35	1–1.5

Source: Compiled from Görner

incinerate waste in preexisting plants, such as industrial furnaces, power plant, cement kilns, etc. The advantages are obvious:

- Investment cost is limited to the additional plant, required to prepare, store, feed, and fire the waste.
- The energy content of waste is put to good use, often at much higher efficiency than is possible in a dedicated plant. A thermal power plant typically operates at an efficiency of HHV to power of ca. 44%, against typically 16–24% for dedicated incinerators.

Theoretically, it would thus be attractive to replace dedicated incineration by usage of waste as a fuel. Yet there also disadvantages, such as below:

- Integrating distinct activities (waste elimination and heat and power generation) also means declining the operating flexibility of each individual activity.
- Incineration is subject to much more stringent emission codes, compared to those for thermal power plants, industrial furnaces, or cement kilns.

Co-firing has been denigratingly termed "solution by dilution." The EU directive on incineration has considered this problem and offered a solution featuring flexible emission limits.

- The waste composition should be confronted with that of the conventional, generally solid, fuel with respect to the pollutants S, N, Cl, and heavy metals.
- The ash arising from waste may affect and often lower product quality, e.g., in cement and especially limekilns.
- Some elements contribute not only to pollution, but also to the creation of operating problems, such as superheater fouling and corrosion (biomass or RDF co-firing) or cycling of heavy metals (cfr. Cement Kilns).

Thermal Power Plants

Co-firing RDF in thermal power plants offers solutions that hold the promise of limited investment, related to the production (possibly off-site), storage, and firing of RDF. Conversely, RDF co-firing may create serious problems at the level of boiler fouling and emissions.

Thermal power plants in general are large-capacity units (40–400 MW_{el}), typically one order of magnitude larger than the usual waste-to-energy (WtE) projects. As they stand, they are fully equipped with provisions for fuel supply, firing, and ash storage, boiler feedwater production, steam generation at high pressure, turbo-alternator, transmission lines, steam cooling, and condensation provisions. Sharing these provisions with incineration plant allows sharing all provisions related to the steam circuit and power generation.

Co-firing of RDF has been proposed consistently since the early 1970s. Ideally, the hosting power plant fires solid fuels such as coal or lignite. The RDF must be reduced in size, so that the individual particles burn out in suspension. Dense parts falling out can be collected on a dump grate for completing their combustion.

Co-firing of biomass has also been considered, at first in Denmark, to eliminate the polluting practice of field burning. Biomass is often lean in pollutants (sulfur, nitrogen, heavy metals). Unfortunately, the ash is also rich in low-melting potassium salts and hence tacky, causing extensive superheater fouling and corrosion. Pure wood is low in ash (0.5–2 wt.%), yet real biomass, such as straw, is much higher, up to 8 wt.%.

Cement and Lime Kilns

Cement (and lime) kilns are increasingly used for incinerating hazardous and also high-calorific waste. The kilns always operate in countercurrent (Fig. 12) and feature combustion temperatures of almost 2,000°C, with kiln lengths ranging from some 50 m (dry process) to about 150 m (wet process). This ensures longer residence times at temperatures above 850°C than any other furnace. Even hazardous pollutants, such as PCBs, requiring a destruction efficiency of at least six 9s (99.9999%), are completely combusted in such kilns. The waste is fired at the lower end of the kiln so that all flue gas starts at flame temperature and then remains in contact with the high temperature reaction zone in which the clinker is formed. Most ash drops out at high temperature and is incorporated into the clinker.

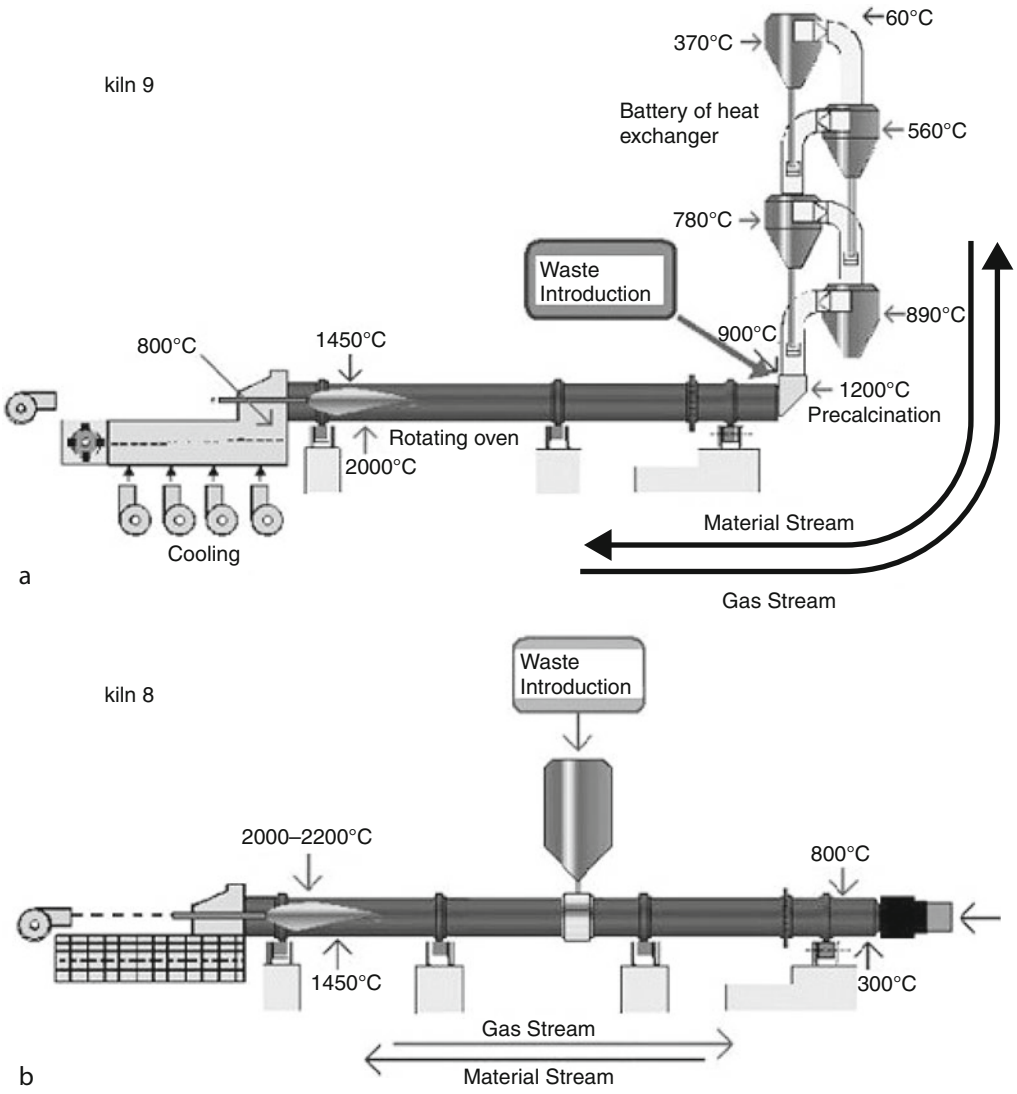
Wet kilns are even longer, since the raw materials mix is to be dried, dehydrated, decarbonated (conversion of $CaCO_3$ into lime), and eventually reacted to clinker at around 1,500°C.

Wet kilns allow separating dust from flue gas and sluicing it out, since the feed enters as a paste. Much shorter dry kilns do not have this feature, the incoming dry meal being heated in direct contact in batteries of cyclonic heat exchangers and enter the much shorter kiln already at high temperature, after preliminary decarbonatation at ca. 800°C. Because of energetic considerations, dry kilns are always preferred for cement manufacturing.

Waste serves primarily as substitute fuel. Yet, waste with appropriate mineral composition (silica, alumina, lime, and iron) may also replace natural feed materials (clay, shale, limestone), alleviating the needs for quarrying (Fig. 14).

Cement kilns are thus prime substitutes for hazardous waste incinerators, as long as the wastes are introduced as solid or liquid fuels at the lower, kiln discharge side. There is sufficient oxygen, temperature, and time to complete combustion of even the most refractory hazardous compounds, such as PCBs, even though the mixing in the gas phase tends to be weak. The solids are in better contact, due to the tumbling action of the kiln. Typical feed requirements are presented in Table 6.

When waste is introduced mid-kiln, however, or – worse – at the higher end of the kiln (in a dry plant, yet after the battery of heat exchangers),



Incineration Technologies. Figure 14

Cement kilns treating contaminated soil with feed point (**a**) after the cyclonic heat exchangers and (**b**) mid-kiln. (14**a**) Shows a battery of four cyclonic heat exchangers, featuring direct contact with hot rising flue gas. The feed entering the plant is dried and heated from 60°C to 900°C, completely calcining the limestone in the feed. In the rotary kiln, it converts into clinker. In this unit, contaminated soil is added at the kiln entrance and it is not certain how far emerging volatiles are still combusted completely. (14**b**) Shows an alternative with mid-kiln feeding, and still converts the contaminated soil into clinker, yet leaves more room for post-combustion than in the first case

a sizeable part of this high-temperature residence time is sacrificed. Feeding organics along with the raw materials, however, must be considered carefully from an environmental viewpoint, since any volatiles evolving from the feed report to the off-gas without post-combustion or cleaning.

The ash from waste is largely incorporated into the clinker. This has raised questions regarding the

eventual leaching of any heavy metals from clinker, as well as regarding the state of oxidation of chromium, i.e., Cr^{III} or Cr^{VI}. Halogens and volatile heavy metals create cycling and emission problems. Several heavy metals (Pb, Zn, and Cu) volatilize in the presence of chlorides, yet de-sublimate and deposit during cooling.

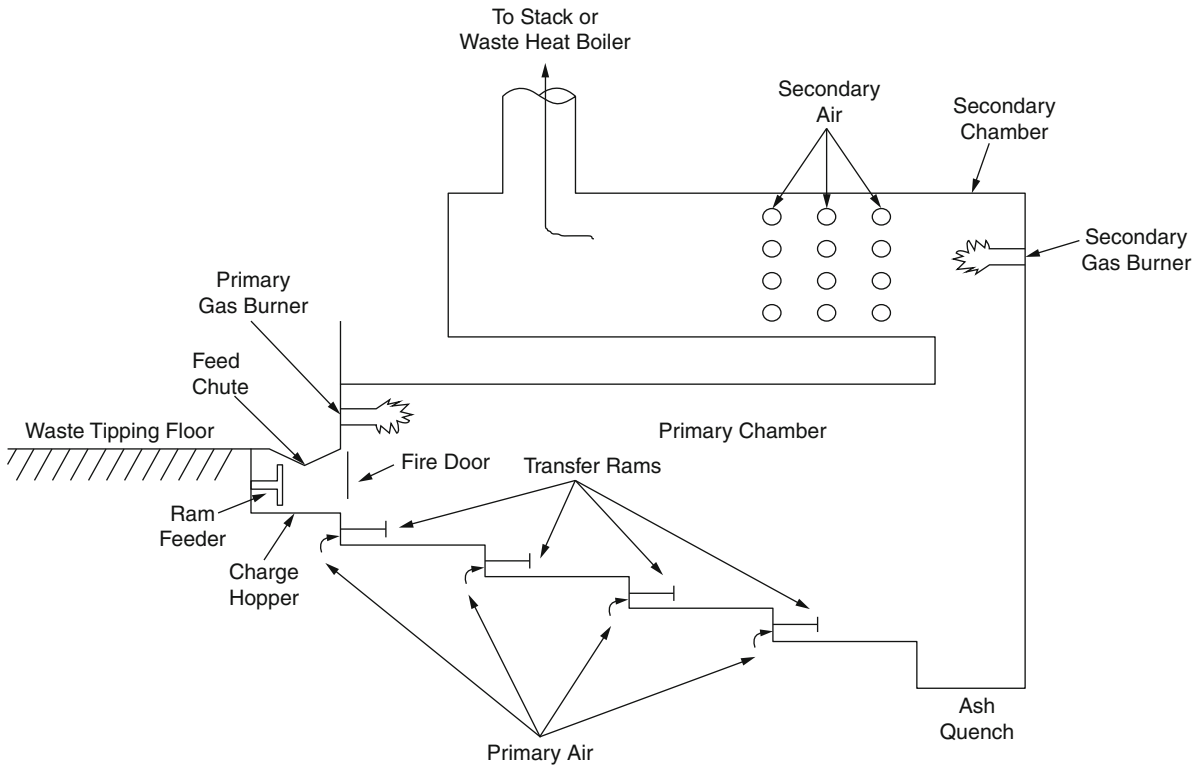
Cement kilns are important emitters of carbon dioxide, dust, and nitrogen oxides. Emission limit values are much more lenient than for dedicated incinerator plants. The cement route has hence been denigrated as "solution by dilution." Nevertheless it is clear that there is scope, worldwide, for the cement route especially in countries devoid of dedicated plant.

There is very extensive literature regarding the cement route [145–148]. Individual cement plants are well documented, relative to inputs and emissions. The lime route has been much less publicized. Obviously, any ash reporting to the product deteriorates product quality.

Public Image of Incineration

Incineration has been branded as a substantial source of environmental pollution (dioxins, heavy metals), as well as an easy way around voluntary or even mandatory recycling. Greenpeace has been quite vocal in these criticisms, attacking incineration, and also PVC as a source of dioxins [149–152]. In the meantime it became clear that dioxins in MSW incineration are formed at comparable rates whether or not PVC is present: the element chlorine is so ubiquitous that its concentration in MSW as a rule is not rate controlling [153]. Much more important factors are the steady quality of combustion, the catalytic effects of transition metals, and dioxins forming increasingly in electrostatic precipitators as their operating temperatures rise (Fig. 15).

Plume dispersion computations show that immission values of state-of-the-art incinerators are negligible, compared to background values. Such



Incineration Technologies. Figure 15
Starved air incinerator (By courtesy of)

deposition values also have been measured many times [154]. Epidemiological studies relating incinerator emissions to public health never clearly condemned the old generations of incinerators, let alone current units with much lower emission values [155, 156]. One reason for these results is that the body burden of most toxic organics relates to food uptake, rather than to inhalation [157]. Today, the main medical interest is related to minute submicron particles that readily migrate through the lung membranes. Also prenatal exposure to dioxins and POPs has been studied.

Other important aspects are related to absorption, digestion, and eventual health effects of dioxins and dioxin-like compounds, in particular polychlorinated diphenyls (PCBs). PCBs are man-made chemicals, yet they also form (with a different congener profile) in thermal processes, typically representing less than 5% of dioxins and furans, when expressed in toxicity equivalents (TEQs). Dioxin diets were established in numerous countries, since food is the major route to take up TEQs. One pathway between the incinerator stack and the food chain is as follows: emission – particle deposition – absorption by grazing cattle – secretion with the milk. In the past, cow milk has been declared unfit for consumption around both incinerator and metallurgical plants. Halting the emissions at Zaandam (the Netherlands) rapidly restored milk quality. Far higher dioxin emission levels at Gien (France) produced no undesirable effects: dioxins were emitted in the gas phase and apparently degraded in the atmosphere, rather than impairing the quality of dairy products.

Incinerators are no longer major polluters, given the extremely stringent emission codes applied today. There is a tendency at present to compare incineration to its traditional alternatives, landfill and composting, incorporating additional criteria, concepts and methods, such as the impact upon climate change and global warming. Landfill is responsible for important greenhouse gas emissions; evolving fermentation gas contains carbon dioxide and also methane, a much more potent greenhouse gas. Composting also emits greenhouse gases, yet does not get the bonus of producing green energy. Markets for compost are as precarious as those for incinerator heat.

Waste incineration is also held responsible for destroying values, available for recycling. Yet, the main bottlenecks of recycling are markets for secondary raw materials showing low-grade specifications or containing *pernicious contraries*: any imbalance between supply and demand exerts strong leverage on market prices.

Future Directions

At present, *waste incineration* has evolved to mature technology, with mechanical grate incinerators as standard in MSW incineration, rotary kiln plant for firing industrial and hazardous waste, and fluidized bed units for sewage sludge, as well as for co-firing wastes with extremely dissimilar properties. Yet, each of these units still has some technical limitations. Mechanical grate stokers must support the waste during combustion, yet have difficulty in coping with both very wet and high-calorific waste. For rotary kiln units, gas phase mixing and wear are major problem areas. Fluid bed units require steady, size-reduced feed and may experience loss of fluidization in the presence of low-melting ash and at too high temperatures.

Flue gas cleaning has evolved considerably since the early 1970s, under pressure of ever tightening limit values (Tables 3 and 4). Given the thorough cleaning generally practiced, it seems unlikely that these limit values would evolve even further. In particular cases, e.g., dioxins, emission values were promulgated on the basis of rather thin evidence and in the absence of proven technology to reach the new limit values. Yet, it cannot be excluded that still new parameters would be brought forward, such as nanoparticles and nitrous oxide (N_2O). However, it is obvious that more can be gained by cracking down on open burning of waste and other primitive and polluting forms of combustion.

The concept of *refuse-derived fuel* (RDF) production holds the promise of conducting incineration in non-dedicated units, such as thermal power plant, and cement and lime kilns. Although RDF is still produced and fired in such plants, initial promise has been mitigated by the added cost and complexity of fuel preparation, by both environmental and product quality (lime, cement) concerns, and by logistic and operational requirements.

Gasification and *pyrolysis* processes have frequently been proposed and tested at laboratory, pilot, and full-scale level. Their theoretical advantages, such as simpler

operation and lower volumetric rates of gas production, have materialized in practice in only few cases. A decisive disadvantage is poor reliability and availability. A large majority of actually constructed plants have actually been scrapped after realizing precarious operating records (e.g., Siemens at Fürth, Thermoselect at Karlsruhe). Some communities were forced to pay for an entire generation for such plants (Andco-Torrax plant in Grasse). By far the most experience has been gathered in Japan, with slagging shaft furnace operation (Nippon Steel) and fluidized bed gasification, followed by post-combustion and fly ash melting and granulating (Ebara Co.) as most successful representatives.

Slagging operation produces glassy aggregate, rather than clinker and fly ash. The question rises whether the more attractive residue can justify considerable supplemental cost, higher energy consumption, and lower availability.

Some organizations important in matters of incineration:

- Air & Waste Management Association (A&WMA)
- American Academy of Environmental Engineers (AAEE)
- American Institute of Aeronautics & Astronautics (AIAA)
- American Institute of Chemical Engineers (AIChE)
- American Society of Mechanical Engineers (ASME)
- Chartered Institution of Wastes Management, London
- Coalition for Responsible Waste Incineration (CRWI)
- Electric Power Research Institute
- Institute for Professional Environmental Practice (IPEP)
- Institute of Chemical Engineers – United Kingdom (IChemE)
- Institution of Mechanical Engineers – United Kingdom (IMechE)
- Integrated Waste Services Association
- International Solid Waste Association (ISWA)
- Japan Waste Management Association ()
- Korea Associate Council of Incineration Technology (KACIT)
- Korea Society of Waste Management (KSWM)
- National Institute for Environmental Studies ()
- National Institute of Environmental Health Sciences (NIEHS)
- Society of Chemical Engineers – Japan
- Solid Waste Association of North America
- Swedish Chemical Society – Sweden
- UK Environmental Agency – United Kingdom
- United States Department of Energy (US DOE)
- United States Environmental Protection Agency (US EPA)
- Waste-to-Energy Research and Technology Council (WtERT)

Bibliography

Primary Literature

1. Lewis H (2007) Centenary history of waste and waste managers in London and South East England, Chartered Institution of Wastes Management, London. http://www.iwm.co.uk/web/FILES/LondonandSouthernCentre/London_and_Southern_Centenary_Histroy.pdf. Accessed July 2011
2. Kleis H, Dalager S (2007) 100 years of waste incineration in Denmark – from refuse destruction plants to high-technology energy works. DTU, Copenhagen. http://www.ramboll.com/services/energy%20and%20climate//media/Files/RGR/Documents/waste%20to%20energy/100YearsLowRes.ashx. Accessed July 2011
3. Reimann DO (1991) Abfallentsorgung mit integrierter Abfallverbrennung – Verfahren von gestern und heute. In: Reimann DO (ed) Rostfeuerungen zur Abfallverbrennung. EF-Verlag für Energie und Umwelt, Berlin, pp 1–20
4. Reimann DO (1991) Die Entwicklung der Rostfeuerungstechnik für die Abfallverbrennung – Vom Zellenofen zur vollautomatischen, emissions- und leistungsgeregelten Rostfeuerung. In: Reimann DO (ed) Rostfeuerungen zur Abfallverbrennung. EF-Verlag für Energie und Umwelt, Berlin, pp 21–60
5. Reimann DO (1991) Rostfeuerungen zur Abfallverbrennung. EF-Verlag für Energie und Umwelt, Berlin
6. Tanner R (1965) Die Entwicklung der Von Roll-Müllverbrennungsanlagen. Schweizer Bauzeitung 83(16)
7. Picture of the first Hamburg incinerator (1985) http://fr.wikipedia.org/wiki/Fichier:Erste_M%C3%BCllverbrennungsanlage_Hamburg.jpeg. Accessed 29 Dec 2011
8. Schoeters J (1975) Patent study on mechanical grate development. VUB, Brussels
9. Buekens A, Schoeters J (1984) Final Report Thermal methods in waste disposal – pyrolysis, gasification – incineration – RDF-firing, Contract Number ECI 1011/B 7210/83B
10. Ebara Co. (1993) Fluidised-bed combustion of municipal solid waste in Japan. Company document
11. Buekens A (1978) Resource recovery and waste treatment in Japan. Resour Recov Conserv 3(3):275–306

12. Buekens A (2008) Schmelzverfahren – erfahrungen in Japan. In: Bilitewski B, Urban AI, Faulstich M (eds) Schriftenreihe des Fachgebietes. Abfalltechnik Universität, Kassel
13. Global Environment Centre Foundation, Japanese Advanced Environment Equipment, http://www.gec.jp/JSIM_DATA/company_index.html
14. E.U. (2009) E.U. Guideline for safe and eco-friendly biomass gasification (gasification – guide). http://www.gasification-guide.eu/. Accessed 11 July 2011
15. Buekens A, Bridgwater AV, Ferrero GL, Maniatis K (eds) (1990) Commercial and marketing aspects of gasifiers. Commission of the European Communities, Elsevier Applied Sciences, Luxembourg, pp 1–239
16. Malkow T (2004) Novel and innovative pyrolysis and gasification technologies for energy efficient and environmentally sound MSW disposal. Waste Manag 24(1):53–79
17. Buekens A, Masson H (1980) Wood waste gasification as a source of energy. Conserv Recycl 3(3–4):275–284
18. Siemons RV (2002) A development perspective for biomass-fuelled electricity generating technologies. PhD thesis, University of Amsterdam. http://www.cleanfuels.nl/Projects%20&%20publications/Siemons_PhD%20Thesis_Internet.pdf. Accessed 11 July 2011
19. Scheirs J, Kaminsky W (2006) Feedstock recycling and pyrolysis of waste plastics. Wiley, Chichester
20. Inguanzoa M, Dominguez A, Menéndez JA, Blancoa CG, Pisa JJ (2002) On the pyrolysis of sewage sludge: the influence of pyrolysis conditions on solid, liquid and gas fractions. J Analy Appl Pyrol 63(1):209–222
21. Buekens A, Schoeters J (1980) Basic principles of waste pyrolysis and review of European processes. ACS Symposium Series 130:397–421
22. Buekens A (1978) Schlussfolgerungen hinsichtlich der praktischen Anwendung der Hausmüllpyrolyse aufgrund weltweiter Erfahrungen. Müll und Abfall 12(6):184–191
23. 12th international congress on combustion by-products and their health effects: combustion engineering and global health in the 21st century – issues and Challenges, Zhejiang University in Hangzhou, China, 5–8 June 2011
24. Chandler AJ, Eighmy TT, Hartlén J, Hjelmar O, Kosson DS, Sawell SE, van der Sloot HA, Vehlow J (1997) Municipal solid waste incinerator residues. Elsevier, Amsterdam\Lausanne\New York\Oxford\Shannon\Tokyo
25. Izquierdo M, López-Soler A, Ramonich EV, Barra M, Querol X 2002) Characterisation of bottom ash from municipal solid waste incineration in Catalonia. J Chem Technol Biotechnol 77(5):576–583
26. Vehlow J (2002) Bottom ash and APC residue management. Expert meeting on power production from waste and biomass – IV, Hanasaari Cultural Center, Espoo, 8–10 Apr 2002. VTT Information Service, Espoo, pp 151–176
27. Sakai S, Hiraoka M (2000) Municipal solid waste incinerator residue recycling by thermal processes. Waste Manag 20:249–258
28. Bergfeldt B, Jay K, Seifert H, Vehlow J, Christensen TH, Baun DL, Mogensen EPB (2004) Thermal treatment of stabilized air pollution control residues in a waste incinerator pilot plant. Part 1: fate of elements and dioxins. Waste Manag Res 22:49–57
29. Baun DL, Christensen TH, Bergfeldt B, Vehlow J, Mogensen EPB (2004) Thermal treatment of stabilized air pollution control residues in a waste incinerator pilot plant. Part 2: leaching characteristics of bottom ashes. Waste Manag Res 22:58–68
30. Achternbosch M, Richers U (2002) Materials flows and investment costs of flue gas cleaning systems of municipal solid waste incinerators. Forschungszentrum Karlsruhe Wissenschaftliche Berichte (FZKA), Karlsruhe, 6726
31. CBR (2011) Personal communication
32. ARGUS – ARBEITSGRUPPE UMWELTSTATISTIK (1981) Bundesweite Hausmüllanalyse 1979/80. Umweltbundesamt, Berlin. Forschungsbericht 103 03 503.
33. ARGUS –ARBEITSGRUPPE UMWELTSTATISTIK (1986) Bundesweite Hausmüllanalyse 1983-1985;Laufende Aktualisierung des Datenmaterials. Umweltbundesamt, Berlin. Forschungsbericht 103 03 508
34. Görner K (1991) Technische verbrennungssysteme, grundlagen, modellbildung, simulation. Springer, Berlin\Heidelberg\New York, p 27
35. Niessen WR (2010) Combustion and incineration processes: applications in environmental engineering. Taylor and Francis, Baco Raton
36. Brunner CR (1996) Incineration systems handbook. Incinerator Consultants, Reston
37. Hämmerli H (1991) Grundlagen zur Berechnung von Rostfeuerungen. In: Reimann D (ed) Rostfeuerungen zur Abfallverbrennung. EF-Verlag, Hrsg
38. European Commission (2006) Integrated pollution prevention and control – reference document on the best available techniques for waste incineration
39. http://en.wikipedia.org/wiki/File:Et_baal.jpg
40. Wilkes JW, Summers CE, Daniels CA, Berard MT (2005) PVC handbook. Hanser Verlag, MŘnchen
41. Buekens A (2006) Introduction to feedstock recycling of plastics. In: Scheirs J, Kaminsky W (eds) Feedstock recycling and pyrolysis of waste plastics: Converting waste plastics into diesel and other fuels. John Wiley & Sons
42. Buekens A (2008) Solving emission problems in a fluid bed MSWI. In: 5th i-CIPEC: international conference on combustion, incineration/pyrolysis and emission control – eco-conversion of biomass and waste, Chiang Mai
43. Briner E, Roth P (1948) Recherches sur l'hydrolyse par la vapeur d'eau de chlorures alcalins seuls ou additionnés de divers adjuvants, Helv Chim Acta 31(2):1352–1360
44. Buekens A, Schoeters J (1986) PVC and waste incineration. APME, Brussels
45. Chimenos JM, Segarra M, Fernández MA, Espiell F (1999) Characterization of the bottom ash in municipal solid waste incinerator. J Hazard Mater 64(3):211–222
46. Meima JA, Comans RNJ (1997) Geochemical modeling of weathering reactions in municipal solid waste incinerator bottom ash. Environ Sci Technol 31(5):1269–1276

47. Commission Decision of 3 May 2000 replacing Decision 94/3/EC establishing a list of wastes pursuant to Article 1(a) of Council Directive 75/442/EEC on waste and Council Decision 94/904/EC establishing a list of hazardous waste pursuant to Article 1(4) of Council Directive 91/689/EEC on hazardous waste (notified under document number C (2000) 1147)
48. Wikipedia, Hazardous Waste
49. Buekens A (2011) Hazardous waste and pollution prevention, course organized by VMAC, Premier Provider of Business Intelligence, Abu Dhabi (U.A.E.)
50. Suisse de Réassurance (1995) Les usines de traitement des déchets urbains, Zurich
51. EPA's Chemical Compatibility Chart (1980) http://www.uos.harvard.edu/ehs/environmental/EPAChemicalCompatibilityChart.pdf. Accessed 11 July 2011
52. Mallinckrodt Specialty Chemicals Co–Chemical compatibility list, 5/1989 http://www.uos.harvard.edu/ehs/environmental/MallinckrodtChemicalCompatibilityList.pdf. Accessed 29 Dec 2011
53. Cole-Palmer Instrument Company-Chemical compatibility (2011) http://www.coleparmer.com/techinfo/ChemComp.asp. Accessed 29 Dec 2011
54. University of Georgia-Chemical storage plans for laboratories (2003) http://www.esd.uga.edu/chem/chemstorage.htm, http://www.esd.uga.edu/chem/pub/lsmanual.pdf, http://www.esd.uga.edu/chem/pub/hmrelocating.pdf. Accessed 29 Dec 2011
55. The University of Vermont, http://www.uvm.edu/~esf/chemicalsafety/chemicalstorage.html. Accessed 29 Dec 2011
56. Magazine Lab Manager, Chemical storage plan fundamentals. http://www.labmanager.com/?articles.view/articleNo/1161/article/8-Chemical-Storage-Plan-Fundamentals. Accessed 29 Dec 2011
57. COMAH (Control of Major Accident Hazards), http://www.hse.gov.uk/comah/
58. Ferziger JH, Peric M (2001) Computational methods for fluid dynamics, 2nd edn. Springer, Berlin, http://elib.tu-darmstadt.de/tocs/100561322.pdf
59. Reményi K (1987) Industrial firing. Akadémiai Kiado, Budapest, 496 p
60. Ferziger JH, Peric M (2001) Computational methods for fluid dynamics, 2nd edn. Springer, New York, http://elib.tu-darmstadt.de/tocs/100561322.pdf
61. Yang YB, Nasserzadeh V, Swithenbank J (2002) Mathematical modelling of MSW incineration in a travelling bed. J Waste Manag 22(4):369–380
62. Yang YB, Goodfellow J, Nasserzadeh V, Swithenbank J (2002) Parameter study on the incineration of MSW in packed beds. J Inst Energy 75(504):66–80
63. Lim CN, Nasserzadeh V, Swithenbank J (2001) The modelling of solid mixing in waste incinerator plants. J Powder Technol 114(1):89–95
64. SUWIC papers (2011) http://www.suwic.group.shef.ac.uk/Journal%20Papers.html. Accessed 29 Dec 2011
65. Buekens A, Mertens J, Schoeters J, Steen P (1979) Experimental techniques and mathematical models in the study of waste pyrolysis and gasification. Conserv Recycl 3(1):1–23
66. Moilanen A (2006) Thermogravimetric characterisations of biomass and waste for gasification processes, VTT Publications 607. 103 pp. + app. 97 pp. Espoo, Finland
67. Nasserzadeh V, Swithenbank J, Lawrence D, Garrod N, Jones B (1995) Measuring gas-residence times in large municipal incinerators, by means of a pseudo-random binary signal tracer technique. J Inst Energy 68(476):106–120
68. Gorman P, Bergman F, Oberacker D (1984) Field experience in sampling hazardous waste incinerators. US Environmental Protection Agency, Washington, DC, EPA/600/D-84/134 (NTIS PB84201573)
69. Carroll GJ (1994) Pilot scale research on the fate of trace metals in incineration. In: Hester RE (ed) Waste incineration and the environment. Royal Society of Chemistry (Great Britain), Cambridge, pp 95–121
70. http://cfr.vlex.com/vid/270-62-hazardous-waste-incinerator-permits-19820277, (2010). Accessed 29 Dec 2011
71. Dellinger B, Torres JL, Rubey WA, Hall DL, Graham JL (1984) Determination of the thermal decomposition properties of 20 selected hazardous organic compounds. Prepared for the U.S. EPA Industrial Environmental Research Laboratory. Prepared by the University of Dayton Research Institute. EPA-600/2-84-138. NTIS PB-84-232487
72. von Paczkowski K (1979) Der Kessel als Bestandteil einer Müllverbrennungsanlage. Seine Entwicklung, sein Entwurf, WÄRME 85:121–125
73. von Paczkowski K (1984) Tendenzen bei Kesseln in Müllverbrennungsanlagen. In: Thome-Kozmiensky KI (ed) Recycling international. EF-Verlag, Berlin
74. Jachimowski A (1978) Kessel für Abfallverbrennungsanlagen. Chemie-Technik 7:403–5
75. Rasch R (1976) Korrosionsvorgänge im Feuerraum. In Kumpf, Maas, Straub, Müll und Abfallbeseitigung, E. Schmidt Verlag, 39 Lfg/III, 7300
76. Vaughan DA, Krause HH, Boyd WK (1974) Study of corrosion in municipal incinerators versus refuse composition. EPA-R-800055
77. Schroer C, Konys J (2002) Rauchgasseitige hochtemperaturkorrosion in müllverbrennungsanlagen – ergebnisse und bewertung einer literaturrecherche. Forschungszentrum Karlsruhe (FZKA), Karlsruhe, 6695
78. Brossard JM, Lebel F, Rapin C, Mareche JF, Chaucherie X, Nicol F, Vilasi M (2009) Lab-scale study on fireside superheaters corrosion in MSWI Plants. In: Proceedings of the 17th annual north american waste-to-energy conference, NAWTEC17, 18–20 May 2009, Chantilly
79. Deuerling C, Maguhn J, Nordsieck H, Benker B, Zimmermann R, Warnecke R (2009) Investigation of the mechanisms of heat exchanger corrosion in a municipal waste incineration plant by analysis of the raw gas and variation of operating parameters. Heat Trans Engin 30(10–11):822–831

80. Olie K, Vermeulen PL, Hutzinger O (1977) Chlorodibenzop-dioxins and chlorodibenzofurans are trace components of fly ash of some municipal incinerators in the Netherlands. Chemosphere 6:455–459
81. Rappe C, Andersson R, Bergqvist PA, Brohede C, Hansson M, Kjeller LO, Lindström G, Marklund S, Nygren M, Swanson SE, Tysklind M, Wiberg K (1987) Overview on environmental fate of chlorinated dioxins and dibenzofurans-sources, levels and isomeric pattern in various matrices. Chemosphere 16:1603
82. Rappe C, Andersson R, Bergqvist PA, Brohede C, Hansson M, Kjeller LO, Lindström G, Marklund S, Nygren M, Swanson SE, Tysklind M, Wiberg K (1987) Sources and relative importance of PCDD and PCDF emissions. Waste Manag Res 5(3):225–237
83. Huang H, Buekens A (1995) On the mechanisms of dioxin formation in combustion processes. Chemosphere 31:4099–4117
84. Weber R, Iino F, Imagawa T, Takeuchi M, Sakurai T, Sadakata M (2001) Formation of PCDF, PCDD, PCB, and PCN in de novo synthesis from PAH: mechanistic aspects and correlation to fluidized bed incinerators. Chemosphere 44:1429–38
85. Weber R, Sakurai T, Ueno S, Nishino J (2002) Correlation of PCDD/PCDF and CO values in a MSW incinerator–indication of memory effects in the high temperature/cooling section. Chemosphere 49:127–34
86. Sakai SI, Hayakawa K, Takatsuki H, Kawakami I (2001) Dioxin-like PCBs released from waste incineration and their deposition flux. Environ Sci Technol 35:3601–7
87. McKay G (2002) Dioxin characterisation, formation and minimisation during municipal solid waste (MSW) incineration: review. Chem Engin J 86:343–368
88. Everaert K, Baeyens J (2002) The formation and emission of dioxins in large scale thermal processes. Chemosphere 46:439–448
89. Stanmore BR (2004) The formation of dioxins in combustion systems. Combust Flame 136:398–427
90. Bumb RR, Crummett WB, Cutie SS, Gledhill JR, Hummel RH, Kagel RO, Lamparski LL, Luoma EV, Miller DL, Nestrick TJ, Shadoff LA, Stehl RH, Woods JS (1980) Trace chemistries of fire: a source of chlorinated dioxins. Science 210(4468):385–90
91. Karasek FW, Dickson LC (1987) Model studies of polychlorinated dibenzo-p-dioxin formation during municipal refuse incineration. Science 237(4816):754–756
92. Gullett BK, Bruce KR, Beach LO (1990) Formation of chlorinated organics during solid waste combustion. Waste Manag Res 8:203
93. Sidhu S, Edwards P (2002) Role of phenoxy radicals in PCDD/F formation. Int J Chem Kinet 34:531
94. Vogg H, Metzger M, Stieglitz L (1987) Recent findings on the formation and decomposition of PCDD/PCDF in municipal solid waste incineration. Waste Manag Res 5(3):285–294
95. Hagenmaier H, Kraft M, Brunner H, Haag R (1987) Catalytic effects of fly ash from waste incineration facilities on the formation and decomposition of polychlorinated dibenzo-p-dioxins and polychlorinated dibenzofurans. Environ Sci Technol 21(11):1080–1084
96. Stieglitz L, Zwick G, Beck J, Roth W, Vogg H (1989) On the de-novo synthesis of PCDD/PCDF on fly ash of municipal waste incinerators. Chemosphere 18:1219–1226
97. Schwarz G, Stieglitz L (1992) Formation of organohalogen compounds in fly ash by metal-catalyzed oxidation of residual carbon. Chemosphere 25(3):277–282
98. Stieglitz L, Jay K, Hell K, Wilhelm J, Polzer J, Buekens A (2003) Investigation of the formation of polychlorodibenzodioxins /-Furans and of other organochlorine compounds in thermal industrial processes, Forschungszentrum Karlsruhe, Wissenschaftliche Berichte – FZKA 6867
99. Gullett B, Bruce K, Beach L (1990) The effect of metal catalysts on the formation of polychlorinated diobenzo-p-dioxin and polychlorinated diobenzofuran precursors. Chemosphere 20:1945–1952
100. Olie K, Addink R, Schoonenboom M (1998) Metals as catalysts during the formation and decomposition of chlorinated dioxins and furans in incineration processes. J Air Waste Manag Assoc 48:101–105
101. Kuzuhara S, Sato H, Kasai E, Nakamura T (2003) Influence of metallic chlorides on the formation of PCDD/Fs during low-temperature oxidation of carbon. Environ Sci Technol 37(11):2431–5
102. Hinton WS, Lane AM (1991) Characteristics of municipal solid waste incinerator fly ash promoting the formation of polychlorinated dioxins. Chemosphere 22:473–483
103. Tuppurainen K, Halonen I, Ruokojärvi P, Tarhanen J, Ruuskanen J (1998) Formation of PCDDs and PCDFs in municipal waste incineration and its inhibition mechanisms: a review. Chemosphere 36(7):1493–1511
104. Addink R, Paulus RHWL, Olie K (1996) Prevention of polychlorinated dibenzo-p-dioxins/dibenzofurans formation on municipal waste incinerator fly ash. Environ Sci Technol 30(7):2350–2354
105. Pandelova M, Lenoir D, Schramm K-W (2007) Inhibition of PCDD/F and PCB formation in co-combustion. J Hazard Mater 149(3):615–8
106. Vehlow J, Braun H, Horch K, Merz A, Schneider J, Stieglitz L, Vogg H (1990) Semi-technical demonstration of the 3R process. Waste Manag Res 8(6):461–472
107. Weber R, Nagai K, Nishino J, Shiraishi H, Ishida M, Takasuga T, Kondo K, Hiraoka M (2002) Effects of selected metal oxides on the dechlorination and destruction of PCDD and PCDF. Chemosphere 46:1247–1253
108. Stach J, Pekarek V, Grabic R, Lojkasek M, Pacakova V (2000) Dechlorination of polychlorinated biphenyls, dibenzo-p-dioxins and dibenzofurans on fly ash. Chemosphere 41:1881–1887
109. Alderman SL (2005) Infrared and X-ray spectroscopic studies of the copper (II) oxide mediated reactions of chlorinated aromatic precursors to PCDD/F, Ph.D. Dissertation Louisiana State University, Chapter 1. http://etd.lsu.edu/docs/available/etd-01112005-150557/unrestricted/Alderman_dis.pdf. Accessed 11 July 2011

I

110. Buekens A, Huang H (1998) Comparative evaluation of techniques for controlling the formation and emission of chlorinated dioxins/furans in municipal waste incineration. J Hazard Mater 62:1–33
111. Wielgosiński G (2010) The possibilities of reduction of polychlorinated dibenzo-p-dioxins and polychlorinated dibenzofurans emission. Int J Chem Eng. Review article 392175:11
112. Düwel U, Nottrodt A, Ballschmiter K (1990) Simultaneous sampling of PCDD/PCDF inside the combustion chamber and on four boiler levels of a waste incineration plant. Chemosphere 20(1):839–846, More papers are to be found at: http://www.nottrodt-ing.de/de/publi.htm
113. Wikström E, Ryan S, Touati A, Tabor D, Gullett BK (2004) Origin of carbon in polychlorinated dioxins and furans formed during sooting combustion. Environ Sci Technol 38(13):3778–84
114. Wikström E, Ryan S, Touati A, Gullett BK (2004) In situ formed soot deposit as a carbon source for polychlorinated dibenzo-p-dioxins and dibenzofurans. Environ Sci Technol 38(7):2097–101
115. Wikström E, Ryan S, Touati A, Tabor D, Gullett BK (2003) Key parameters for de novo formation of polychlorinated dibenzo-p-dioxins and dibenzofurans. Environ Sci Technol 37(9):1962–70
116. Addink R, Olie K (1995) Mechanisms of formation and destruction of polychlorinated dibenzo-p-dioxins and dibenzofurans in heterogeneous systems. Environ Sci Technol 29:1425–1435
117. Konduri R, Altwicker ER (1994) Analysis of time scales pertinent to dioxin/furan formation on fly ash surfaces in municipal solid waste incinerators. Chemosphere 28(1):23–45
118. Zimmermann R, Blumenstock M, Heger HJ, Schramm K-W, Kettrup A (2001) Emission of nonchlorinated and chlorinated aromatics in the flue gas of incineration plants during and after transient disturbances of combustion conditions: delayed emission effects. Environ Sci Technol 35:1019–1030
119. Kreisz S, Hunsinger H, Vogg H (1997) Technical plastics as PCDD/F absorbers. Chemosphere 34(5–7):1045–1052
120. Pekarek V, Weber R, Grabic R, Solcova O, Fiserova E, Syc M, Karban J (2007) Matrix effect on the de novo synthesis of polychlorinated dibenzo-p-dioxins, dibenzofurans, biphenyls and benzenes. Chemosphere (Eng) 68(1):51–61
121. Altwicker ER (1994) Formation of PCDD/F in municipal solid waste incinerators: laboratory and modeling studies. J Hazard Mater 47(1–3):137–161
122. Buekens A, Tsytsik P, Carleer R (2007) Methods for studying the de novo formation of dioxins at a laboratory scale. In: International conference on power engineering-2007, Hangzhou, 23–27 Oct 2007
123. Buekens A, Swithenbank J (2007) CFD modelling of industrial plant from a viewpoint of dioxins formation. In: International conference on power engineering (ICOPE-2007), Hangzhou
124. Verhulst V, Buekens AG, Spencer P, Eriksson G (1996) The thermodynamic behaviour of metal chlorides and sulfates under the conditions of incineration furnaces. Environ Sci Technol 30:50–56
125. http://www.termwiki.com/EN:chute-fed_incinerator_(Class_IIA). Accessed 29 Dec 2011
126. http://www.seas.columbia.edu/earth/wtert/sofos/nawtec/1964-National-Incinerator-Conference/1964-National-Incinerator-Conference-25.pdf. Accessed 29 Dec 2011
127. http://www.dioxinfacts.org/sources_trends/trash_burning.html. Accessed 29 Dec 2011
128. http://www.epa.gov/oaqps001/community/details/barrelburn.html. Accessed 29 Dec 2011
129. Gullett BK, Lemieux PM, Lutes CC, Winterrowd CK, Winters DL (1999) PCDD/F emissions from uncontrolled, domestic waste burning. Presented at Dioxin '99, the 19th international symposium on halogenated environmental organic pollutants and POPs, Organohalogen compounds, vol 41, Venice, 12–17 Sept 1999, pp 27–30
130. Lemieux PM, Lutes CC, Abbott JA, Aldous KM (2000) Emissions of polychlorinated dibenzo-p-dioxins and polychlorinated dibenzofurans from the open burning of household waste in Barrels. Environ Sci Technol 34:377–884
131. iran
132. http://www.sittommi.fr/fonctionnement-usine-incineration-ordures-menageres-pontivy.html. Accessed 29 Dec 2011
133. Buekens A, Yan M, Jiang XG, Li XD, Lu SY, Chi Y, Yan JH, Cen K (2010) Operation of a municipal solid waste incinerator – Pontivy. i-CIPEC
134. Winnacker
135. Saxena SC, Jotshi CK (1994) Fluidized-bed incineration of waste materials. Prog Energy Combust Sci 20(4):281–324
136. Integrated pollution prevention and control reference document on best available techniques for the waste treatments industries, August 2006
137. Santoleri JJ (1972) Chlorinated hydrocarbon waste recovery and pollution Abatement. In: Proceedings of the 1972-National-incinerator-conference, New York
138. Mizuno K (2002) Destruction Technologies for ozone depleting substances in Japan. National Institute for Resources and Environment, in UNEP: http://www.unep.fr/ozonaction/information/mmcfiles/3521-e-file2.pdf. UNON Nairobi
139. http://www.uneptie.org/ozonaction/information/mmcfiles/3521-e-file2.pdf. Accessed 29 Dec 2011
140. http://submergedcombustion.org.uk/Default.aspx. Accessed 29 Dec 2011
141. Tsukishima Kankyo Engineering (2010) http://www.tske.co.jp/english/index.html. Accessed Dec 2011
142. Buekens AG, Schoeters JG, Jackson DV, Whalley LW (1986) Status of RDF-production and utilization in Europe. Conserv Recycl 9:309–309
143. Friends of the Earth (2008) Briefing – mechanical and biological treatment (MBT)
144. Wikipedia, Mechanical and biological treatment (MBT)

145. IPPC (1999) Integrated pollution prevention and control (IPPC): reference document on best available techniques in the cement and lime manufacturing industrie. Formation and release of POPs in the cement industry, 2nd edn. European Commission, Directorate General JRC, Institute for Prospective Technological Studies, Seville
146. Ökopol (1999) Economic evaluation of dust abatement techniques in the European cement industry, Report for EC DG11, contract B4-3040/98/000725/MAR/E1; and "Economic evaluation of NOx abatement techniques in the European cement industry", Report for EC DG11, contract B4-3040/98/000232/MAR/E1. Ökopol GmbH, Hamburg
147. Rabl A (2000) Criteria for limits on the emission of dust from cement kilns that burn waste as fuel. ARMINES/Ecole des Mines de Paris, Paris
148. SINTEF (2006) Formation and release of POPs in the cement industry, second edition. Report of the World Business Council for Sustainable Industry, Cement sustainability initiative, Geneva
149. Greenpeace International (1991) http://archive.greenpeace.org/toxics/reports/gopher-reports/inciner.txt. Amsterdam. Accessed 29 Dec 2011
150. Greenpeace International (1994) http://archive.greenpeace.org/toxics/reports/azd/azd.html. Greenpeace Communications, London
151. Costner P (2001) Chlorine, Combustion and Dioxins: Does Reducing Chlorine in Wastes Decrease Dioxin Formation in Waste Incinerators? http://archive.greenpeace.org/toxics/reports/chlorineindioxinout.pdf
152. PVC WASTE AND RECYCLING. Solving a Problem or Selling a Poison? (1999) http://archive.greenpeace.org/toxics/html/content/pvc3.html#top. Accessed 29 Dec 2011
153. Buekens A, Cen KF (2011) Waste incineration, PVC, and dioxins. J Mater Cycles Waste Manag 13:190–197
154. Xu MX
155. Travis CC (1991) Municipal waste incineration risk assessment: deposition, food chain impacts, uncertainty, and research needs. Plenum Press, New York
156. Hattemer-Frey HA, Travis CC (1991) Health effects of municipal waste incineration. CRC Press, Baco Raton
157. Roberts SM, Teaf CM, Bean JA (1999) Hazardous waste incineration: evaluating the human health and environmental risks. Lewis, Boca Raton

Books and Reviews

Air Pollution Control Association, American Society of Mechanical Engineers. Research Committee on Industrial and Municipal Wastes (1988) Hazardous waste incineration: a re-source document sponsored by the ASME Research Committee on Industrial and Municipal Wastes; co-sponsored by the Air Pollution Control Association, the American Institute of Chemical Engineers, the United States Environmental Protection Agency

Alter H, Horowitz E (1975) STP 592, Resource recovery and utilization. In: Proceedings of the national materials conservation symposium. http://www.astm.org/BOOKSTORE/PUBS/STP592.htm. Accessed July 2011

Bilitewski B, Härdtle G, Marek K (2000) Abfallwirtschaft. Handbuch für Praxis und Lehre. Springer, Berlin

Bonner T, Dillon AP (1981) Hazardous waste incineration engineering, pollution technolo-gy review 88. Noyes Data Corporation, Park Ridge

Gershman, Brickner & Bratton, Inc. (1986) Small-scale municipal solid waste energy recovery systems. Van Nostrand Reinhold, New York

de Souza-Santos ML (2004) Solid Fuels combustion and gasification: modeling, simulation, and equipment operations. Marcel Dekker, New York

Freeman HM (1988) Incinerating hazardous wastes. Technomic, Lancaster

Görner K (1991) Technische Verbrennungssysteme. Springer, Berlin\Heidelberg\New York

Grover VI (2002) Recovering energy from waste: various aspects. Science, Enfield

Günther R (1974) Verbrennung und Feuerungen. Springer, Berlin \Heidelberg\New York

Hester RE (1994) Waste incineration and the environment. Royal Society of Chemistry (Great Britain), Cambridge

Institute of Electrical and Electronics Engineers (1975) Incineration and treatment of hazardous waste. In: Proceedings of the eighth annual research symposium CRE: conversion of refuse to energy, vol 1. World Environment and Resources Council, Institute of Electrical and Electronics Engineers

International conference on combustion, incineration/pyrolysis (i-CIPEC). In: Proceedings of the 1st (Seoul, Korea in 2000), 2nd (Jeju, Korea in 2002), 3rd (Hangzhou, China in 2004), 4th (Kyoto, Japan in 2006), 5th (Chiangmai, Thailand in 2008), and 6th International Conference on Combustion, Incineration/Pyrolysis (Kuala Lumpur, Malaysia, 2010)

International conference on thermal treatment technologies

National Research Council (US). Committee on Health Effects of Waste Incineration (2000) Waste incineration and public health. National Academies, Washington

National-Incinerator-Conference 1964, 1966, 1968, 1970, 1972, 1974 (visit the proceedings at the WTERT-site of Columbia University, e.g. at http://www.seas.columbia.edu/earth/wtert/sofos/nawtec/1966-National-Incinerator-Conference/). Accessed July 2011

National-Waste-Processing-Conference 1976, 1978, 1980, 1982, 1984 1986, 1988, 1990, 1992, 1994 (visit the proceedings at the WTERT-site of Columbia University, e.g., http://www.seas.columbia.edu/earth/wtert/sofos/nawtec/1980-National-Waste-Processing-Conference/). Accessed July 2011

North American Waste to Energy Conferences (NAWTEC) http://nawtec.swana.org/. Accessed July 2011

EPA (1989) Environment Canada. Proceedings of the international conference on municipal waste combustion. Hollywood, Florida

Robinson WD (1986) The solid waste handbook: a practical guide. Wiley, Chichester

I

Rogoff MJ, Screve F (2011) Waste-to-energy: technologies and project implementation. Elsevier Science, Amsterdam

Santoleri JJ, Theodore L, Reynolds J (2000) Introduction to hazardous waste incineration. Wiley-IEEE, New York

Solid Waste Association of North America (1998) Asian-North American solid waste management conference. Paper presented at the 17th biennial waste processing conference, Atlantic City (Proceedings available at the WTERT-site of Columbia University)

Theodore L, Reynolds J (1987) Introduction to hazardous waste incineration. Wiley, New York

Warnatz J, Maas U, Dibble RW (2001) Combustion – physical and chemical fundamentals, modeling and simulation, experiments, pollutant formation, 3rd edn. Springer, Berlin\ Heidelberg\New York

World Health Organization. Regional Office for Europe (1985) Solid waste management: selected topics. World Health Organization, Copenhagen

Young GC (2010) Municipal solid waste to energy conversion processes: economic, technical, and renewable comparisons. Wiley, Hoboken

Incinerator Grate Combustion Phenomena

J. Swithenbank[1], Vida N. Sharifi[1,2]
[1]Energy and Environment Engineering (EEE), Sheffield University, Sheffield, UK
[2]EEE Group, Chemical and Biological Engineering, Sheffield University, Sheffield, UK

Article Outline

Glossary

AD Anaerobic digestion
CFD Computational fluid dynamics
CHP Combined heat and power
CHP/DHC Combined heat and power/District heating and cooling
CV Calorific value
DEFRA Department of Environment Food and Rural Affairs
DHC District heating and cooling
DVC Depolymerization-vaporization-cross-linking model
EfW Energy-from-Waste
E.U. European Union
FG Functional group
FG-DVC Functional group model and a depolymerization-vaporization-cross-linking model.
GJ Gigajoule
kWh Kilowatt hours
kW_h Kilowatt heat
MBT Mechanical biological treatment
MSW Municipal solid waste
MRF Material recovery facility
MW Megawatt
NO Nitric oxide
NO_x Nitrogen oxides
PCDD/Fs Polychlorinated dibenzo-dioxins and furans
RDF Refuse-derived fuel
ROC Renewable Obligation Certificate
SRF Solid-recovered fuel
TEQ Toxic equivalent
TGA Thermogravimetric analysis
WID Waste incineration directive
WTE Waste-to-energy

Definition of the Subject

Historically, waste materials from cities were simply dumped in huge piles of polluting material. The liquid runoff usually polluted water courses, and the rotting material continued to emit greenhouse gases, methane, and carbon dioxide for 50 years. The area of land required also became a problem and most societies now accept that such waste dumps are unacceptable. However, an important fact is recognition that dumping of wastes without recovering reusable materials is unsustainable and waste should be: (a) minimized at source, (b) the recovery of reusable or recyclable materials should be optimized, and (c) the recovery of energy-from-waste (EfW) must be maximized.

Minimization of waste at source poses many problems. For example, most cultures have celebrations, such as weddings, which result in waste from gift packaging. Yet these events contribute to the quality of life, and it would be unpopular to ban them. However, packaging consists of paper or cardboard which are biofuels similar to wood, and their combustion in an energy-from-waste plant simply returns captured carbon dioxide to the atmosphere and displaces fossil fuel that would otherwise be used. The important point is that waste material that it is not viable to reuse or recycle should be used in an energy-from-waste plant.

Waste consists of a wide variety of materials such as cans and paper that are initially separate but become mixed in a crude waste-collection system. The subsequent separation or de-mixing requires considerable cost and energy, and usually results in cross-contaminated products. Thus, in accordance with the principle of entropy (or disorder), wastes should be separated at source wherever it is viable. Nevertheless, material handling machines can recover some recyclable material such as metal, paper, and some plastics. The residue, which amounts to about 50% of the raw waste, is flock or pellets known as Solid-Recovered Fuel (SRF), which can be burned or possibly gasified to generate power.

Composting raw municipal waste converts much of the material to carbon dioxide without recovering energy and leaves a semi-toxic residue, whereas methane generation by anaerobic digestion is generally more suitable for wet food wastes.

This presentation focuses on thermal energy-from-waste technology using incineration since this is now a mature and bankable technology that is delivering electricity from hundreds of plants worldwide. Nevertheless, attention is drawn to the fact that the efficiency of the electrical power generation is only about 23% due to boiler corrosion problems. Fortunately, there is an engineering solution to this situation since the remaining energy is available as hot water that can be used for district heating (or building cooling), thus raising the energy conversion efficiency to about 90%. Not only does this save fossil fuels, but because most of the waste residue is biomass, the net carbon dioxide emission is quite low, whilst the heat is generated close to the consumer and transport of waste is minimized.

A key feature of this environmentally friendly "trash-into-resource" strategy is that it helps to reduce global warming by reducing the **net** emissions of carbon dioxide to the atmosphere.

In the light of these observations, it is clear that EfW is a key technology for modern society. The aim of this article is to present; the rationale, the underpinning scientific principles, and key engineering aspects of this topical subject.

Introduction

The impending Fossil Fuel Poverty that is developing in the world largely explains why countries such as Denmark, which have no national fossil fuels, have developed efficient energy-from-waste (EfW) systems incorporating district heating that exploit low-grade heat from waste and integrate it with low-grade heat from their electricity power generation. The result is that two thirds of their buildings are already connected to district heating, and these two systems should be generally considered together for an energy-efficient city. Furthermore, the common perception that old cities cannot have district heating is false since many old cities like Vienna and Paris have successfully installed very extensive heat distribution networks.

The relation between process plant-scale and plant-capital cost generally follows the 0.6 power law, leading to the relative economy of large-scale process plants. Waste plants such as incinerators follow this rule; however, smaller plants can be operated, provided the waste

is suitably shredded. However, in many cases, the problem of large-scale waste supply is solved by the use of waste transfer stations to integrate the wastes sources from several towns.

An important point that is not usually emphasized sufficiently is the relative efficiency of different energy-from-waste technologies. Firstly, many studies assume that the low-grade heat from power generation can be ignored. However, as pointed out above, the efficiency of power generation is usually less that 25%, and this results in ignoring about 60% of the energy that is available to displace fossil fuels currently used to heat buildings. Surely, society will not be able to ignore this issue for long.

The recent development of materials recovery facility (MRF) plants that separate recyclables from residual waste are already starting to produce millions of tons per year of solid-recovered fuel (SRF or RDF). However, the SRF from a particular process will be much more consistent in quality than raw MSW. Complex process plants such as gasifiers generally require the feed to be tightly specified and, hence, will be better suited to use such wastes. It is also noteworthy that SRF can be shipped and stored more easily than raw MSW. An important factor is the need to apply the Waste Incineration Directive (WID) to emissions from all waste combustion systems. The importance of defining the status of new fuels derived from wastes therefore requires considerable attention, especially with regard to the emission of organic material and inorganics, such as heavy metals.

The importance of the moisture content of wastes must be also recognized. In the case of thermal processes, this simply represents 2 MJ/kg heat gain for every 10% reduction in moisture content. The calorific value (CV) of raw MSW is about 10 MJ/kg, hence drying results in a very significant gain in CV. The significance of reducing this moisture in order to recover energy from materials such as AD and sewage sludge requires appropriate recognition. The UK sewage industry often uses centrifuges and thermal systems to dry sewage sludge before burning it in an autothermal fluidized bed; however, the additional use of a belt press can result in a waste cake that delivers enough heat to generate power.

A similar question arises when the power-generation capability of different technologies is compared. In the case of incineration, a typical EfW plant generates electrical energy at 600 kWh/t plus 2,500 kWh/t of heat for buildings. By comparison, an anaerobic digestion plant generates electrical energy at 75–160 kWh/t of the organic fraction of the waste. In the future, new thermal gasification plants could double the electrical energy that can be generated from each ton of waste.

Brief Overview of Materials and Energy Recovery from Solid Wastes

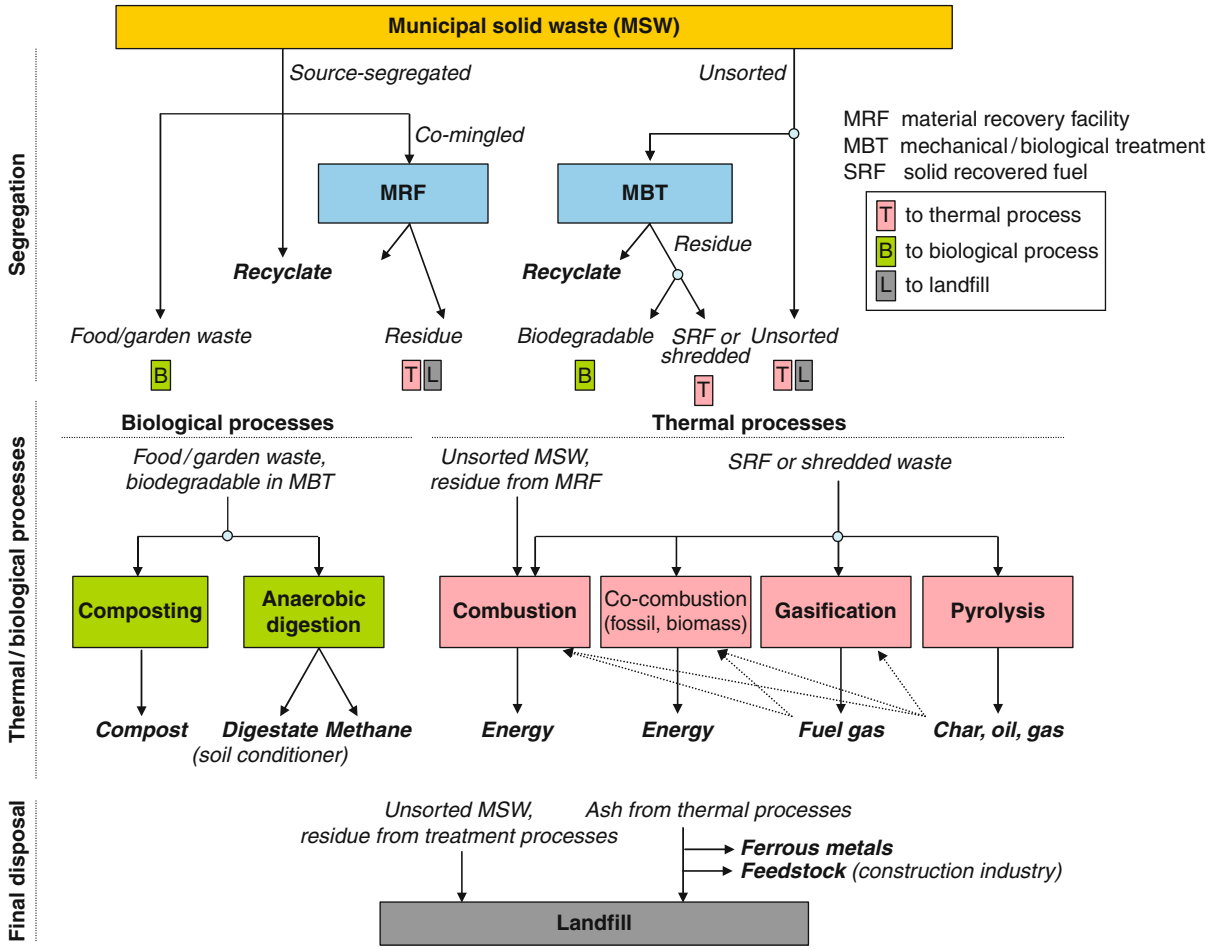
Municipal solid wastes (MSW) can be treated by various methods, as illustrated in Fig. 1. However, direct combustion, also called incineration, of post-recycling-mixed MSW is currently the main route for energy and metals recovery practiced by the global energy-from-waste industry (EfW; also called waste-to-energy or WTE).

The predominant MSW incinerators burn wastes of a wide range of calorific values on a moving grate, without any waste preprocessing. However, in some countries, these incineration systems suffer from unfavorable public perception. The main treatment alternatives to incineration are:

(a) Production of solid-recovered fuel (SRF) from mechanical/biological treatment,
(b) Pretreatment of MSW to "refuse-derived fuel" (SRF) as discussed in another section of this document.
(c) Gasification of preprocessed or as-received waste followed by either combustion or use of the gas as a fuel.

Such processes have *potentially* higher energy efficiency when compared to incineration and perhaps more flexibility in the use of primary products. However, further technology developments with appropriate regulatory drivers are still required for them to compete effectively with incineration.

Another thermal route available for treating the high calorific components of MSW is pyrolysis, in order to produce cheap and storable fuel products. Pyrolysis is normally only suitable for specific types of waste material such as plastics or waste wood, where it is used to produce charcoal, a storable fuel product.



Incinerator Grate Combustion Phenomena. Figure 1
Various methods for recovering materials and energy from municipal solid wastes

Energy Recovery by Thermal Treatment of Wastes

The energy output available from a material is conveniently specified by its calorific value in MJ/kg. Thus, if the fuel feed is 1 kg/s (i.e., 3.6 t/h or ~31,500 t of waste per year), a typical municipal waste fuel containing 10 MJ/kg can produce a maximum of 10 MJ/s, that is, 10 MW of combined heat and power. In practice, the efficiency of the appropriate electrical power generating system is only 23%, so a more realistic figure for the electrical output would be 2.3 MW_e of electricity. However, the rest of the thermal energy is available as low-grade heat that is suitable for district heating and can provide about 6.2 MW_h of district heating.

Putting these figures into perspective, the average person in the E.U. produces approximately half a ton of nonrecyclable waste per year, so the example above would correspond to a fairly large city with a population of 60,000 and 20,000 homes. Heating each home requires up to 5 kW_h, hence the number of homes that could be heated from their own MSW is 6,200/5 = 1,240, or a proportion of 20%. The electricity generated from the waste also provides power for a similar proportion of households in a community served by EfW.

Therefore, the energy available from the waste generated by a local population can make a significant contribution to fulfilling the local demand for heat and electricity, while also replacing the use of an

important amount of nonrenewable fossil fuel. Furthermore, by doing so, the local urban population assumes the responsibility for managing their waste instead of sending it to pollute a rural area. The fact that the waste volume is reduced by a factor of 10 in an energy-from-waste plant is also an important consideration in minimizing the transportation involved in waste management.

To access the energy available in the waste, it must be converted from chemical energy to heat and/or power. The direct method of recovering EfW is thus incineration, which actually involves all the complex reaction processes of pyrolysis, gasification, and oxidation. The other two processes (Gasification and Pyrolysis) produce intermediate products of combustion that can be burned for energy generation or used as a feedstock, depending on process conditions. Furthermore, a gasification or pyrolysis system used to produce heat and power energy also involves combustion of the gas or char and thus conversion of the initial feed to carbon dioxide, water, and ash. Each alternative process should therefore be considered holistically, where one starts from the raw waste and then follows the process to completion of the reaction of the fuel with air, the production of heat and power, and the disposal of the ash residue. Also, any gas cleaning operation that is required to meet the air and water environmental control regulations must be included in such a comparison of alternatives.

An important consideration is the internal use of power within the plant; systems should be compared on the basis of net power production rather than the output of the electricity generator. The internal power-consuming devices include waste preprocessing shredders, fans, material conveyers, pumps, etc.

Incineration is a mature technology that can meet all regulations and that is bankable. Why then are other technologies being considered? The reason lies in the potential efficiency of electrical power production. In the case of the mass-burn system, the combustion gases are used to raise steam for use in the turbine/generator with a Rankine power-generation cycle. Thermodynamic considerations show that the efficiency of the power generation depends on the top temperature of steam generation. This temperature is limited by the acceptable material life of the superheater steam tubes. In a conventional coal-fired power station, steam temperatures are typically about 565°C and result in a power-generation efficiency of about 35%. In the case of a conventional energy-from-waste plant, the composition of the flue gas is more corrosive and a steam temperature yielding an acceptable superheater tube life is about 400°C, resulting in power-generation efficiency of about 23–25%. Although most of the balance of the energy is available as heat for a district heating system, electricity is a more valuable form of energy than heat; hence, research workers are striving to develop a more efficient system.

In principle, if one could generate gas from waste with an efficiency of about 70% and use the gas in an internal combustion engine or a combined gas turbine/steam turbine generator with an efficiency of 45–55%, then an overall power-generation efficiency of 31.5–38.5% should be attainable. However, this target has been elusive and remains a "gleam-in-the-eye" at the present time. For this reason, the following discussion will largely focus on conventional incinerator combustion as the route to produce energy from waste. Furthermore, since the waste combustion process is particularly complex and its study is a specialty of the authors, this entry will focus on this aspect of energy-from-waste.

Energy Recovery from MSW by Incineration

Typical incineration systems are mass-burn combustion chambers that operate at high temperatures (i.e., 850°C at the furnace exit) and with excess air (typically 50–80% of stoichiometric air) in order to ensure efficient combustion of waste material and complete destruction of toxic organic pollutants. As mentioned above, mass-burn incinerators can handle wastes with a wide range of calorific value without any specific waste pretreatment, which is a key advantage of this technology.

The major public issues concerning the "incineration option" are pollution, health risks, and low energy efficiency. All combustion systems inevitably generate certain amounts of pollutants. In the case of incineration, the main flue gases are; nitrogen, oxygen, water vapor, carbon dioxide, acidic gases, NO_x, particulates, heavy metals, and dioxins. The pollutants are efficiently removed by a combination of gas cleaning technologies to a level below the emission limits set by the Waste Incineration Directive of the E.U. (Table 1). Modern

incinerators operate well below these emission limits. Dioxins and fly ash are of particular public interest. However, the dioxin level from modern incineration systems is far below the European emission limits, for example, the total UK dioxin emissions from incineration account for only about 1% of the UK national total [1]. A similar result was also reported for Germany and other nations.

Incinerator Grate Combustion Phenomena. Table 1
Daily average values of air emission limits for incineration in WID

Pollutants	Values
Total dust	10 mg/m^3
Total organic carbon	10 mg/m^3
HCl	10 mg/m^3
HF	1 mg/m^3
SO_2	50 mg/m^3
NO_x	200 mg/m^3
CO	50 mg/m^3
Hg	0.05 mg/m^3
Cd/Tl	Total 0.05 mg/m^3
Sb, As, Pb, Cr, Co, Cu, Mn, Ni, V	Total 0.05 mg/m^3
PCDD/Fs	0.1 ng I-TEQ/m^3

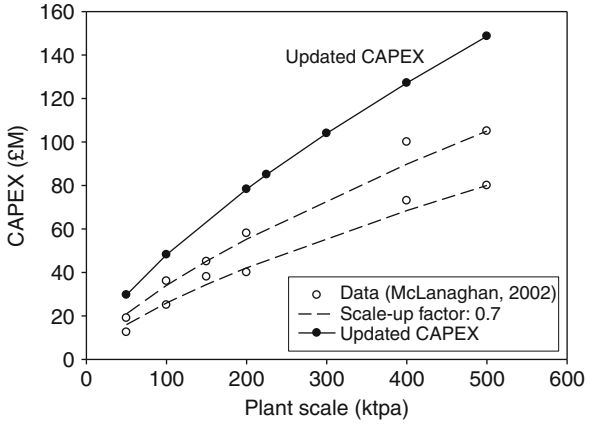
Incineration with Combined Heat and Power (CHP)

As noted above, due to corrosion problems the efficiency of electricity generation by incinerators using a steam turbine is about 25%, which is rather low compared to about 35% for coal-fired power plant. However, the use of combined heat and power systems (CHP) can dramatically increase the overall energy efficiency by exploiting the low-grade heat from the steam turbine for heating (or cooling) domestic, commercial, and industrial buildings. The overall energy efficiency of a CHP system using conventional or waste-derived fuels is as high as 85%. The best example in the UK is the Sheffield incinerator operated by Veolia Sheffield Ltd (Fig. 2). It produces up to 60 MW thermal plus 19 MW electrical energy from a waste throughput of 225 kt/year and supplies heat to commercial properties and over 5,000 residential buildings through a network of 42 km of piping.

An important factor concerning EfW plants in the UK is their eligibility for the Renewable Obligation Certificates (ROCs), which provides an incentive to promote electricity generation from renewable sources, such as the biomass content of wastes.

Incineration is capital intensive (Fig. 3). The major cost factors are equipment and building costs, land acquisition, planning, labor, maintenance, and ash

Incinerator Grate Combustion Phenomena. Figure 2
Sheffield incinerator



Incinerator Grate Combustion Phenomena. Figure 3
Curve-fitted and updated (2009) for various plant scales

disposal. The costs are very much site-dependent and are affected by a number of parameters such as:

- Scale of the plant (typically ranging from 50 to 400 kt/year)
- Requirement for flue gas treatment
- Treatment and disposal process for the bottom and fly ash, including recovery of metals from the bottom ash and their sale
- Efficiency of energy recovery

The Waste Combustion Process

Two alternative technologies are available to burn waste. These are the mass-burn grate and the fluidized bed. The former consists of a moving grate that burns waste travelling on a grate from a feed shaft to the ash pit. The latter consists of a fluidized bed of near mono-size sand particles enveloping the burning waste material. Air passing up through the bed is almost universally used as the source of oxidant.

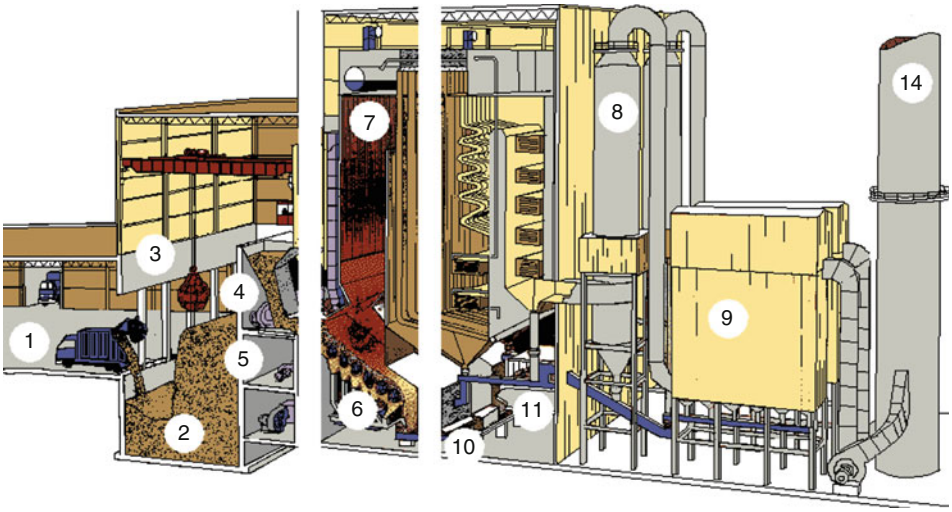
The two systems differ in the peak temperature achieved during combustion. In the case of a mass-burn grate, the maximum temperature is about 1,300°C, whereas the fluidized bed operates at about 850°C. Because the ash melting temperature lies between these two temperatures, the ash from a mass-burn grate can contain large pieces of slag. These two processes also differ in the amount of power that is used within the process. In the case of a fluidized bed, the waste must be preprocessed by shredding, which incurs a significant cost and efficiency penalty; also the air pressure needed to suspend the fluidized bed is much greater than that used with a mass-burn grate. The overall effect is that the net electrical power output of a fluidized bed system tends to be less than with a mass-burn grate. These and other considerations have led to a preference for installation of the mass-burn system in Europe, where there is more emphasis on power production, whereas in Japan the balance between the uses of each system is more nearly even.

Mathematical Modeling of Combustion in an Energy-from-Waste Plant

Travelling grate combustion systems such as that illustrated in Fig. 4 are commonly used in industry for incineration of MSW waste. In this type of reactor, waste burns on top of the moving grate with primary air supplied from below. The fresh waste fed onto the grate at one end ignites by radiation from the hot environment and the ignition front propagates into the bed, as it is slowly transported to the discharge end of the moving grate. The combustion products are released to a gas plenum above the bed, where further oxidation of combustible gases takes place. Secondary air is injected through jets into the gas in the shaft above the bed to enhance the gaseous mixing and combust the volatile gases and carbon monoxide that emanate from the burning bed of solids.

Mathematical modeling of the burning bed of waste particles ("FLIC" code) was developed at Sheffield University for thermal processes in a packed bed such as moving grate combustion, fixed bed air-steam gasification, and slow pyrolysis. After establishing the governing equations for the two-phase reacting bed, the heat transfer and reaction parameters are optimized using test results carried out in parallel in a so-called pot burner. This code is usually coupled with a gas flow modeling ("FLUENT" code) to provide comprehensive information on the thermal processes in a furnace. This novel simulation technique has been applied widely to provide a detailed understanding of the combustion process and to help optimize the design and operation of energy-from-waste plants worldwide.

To segregate recyclables, the unsorted MSW can be mechanically treated before being fed into the furnace, as shown in Fig. 5. This can be an attractive waste



Incinerator Grate Combustion Phenomena. Figure 4
Typical EfW plant [2]

disposal route for local authorities in order to meet the recycling targets by post-collection segregation and reduce landfilling. Based on a least-cost optimization of the waste management options, incineration with mechanical treatment of waste is most likely to be adopted in the near future. However, the recent trend is for more recyclables to be source-separated and collected [3] rather than for post-collection segregation of unsorted wastes, principally due to the poor quality of recyclables sorted out from mixed MSW.

Mass and Energy Balance in the Incineration Process

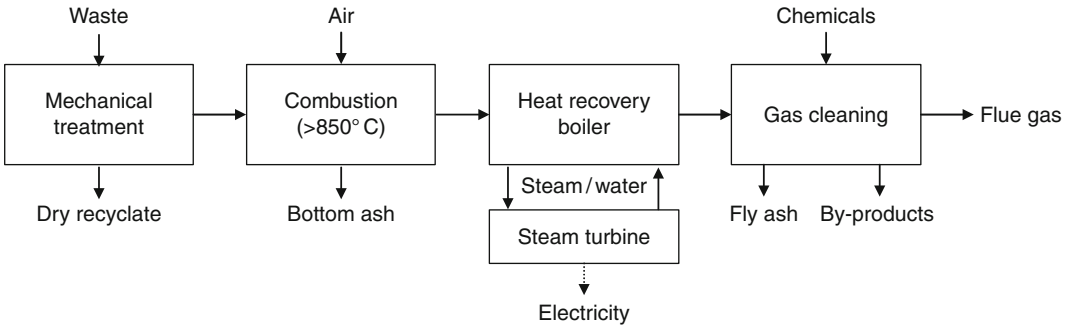
Incineration generally operates at temperatures between 850°C and 1,100°C with a large excess of oxygen and converts waste materials into ash and exhaust gas while producing thermal/electrical energy from the chemical energy of waste. The overall reaction for incineration can be written as:

$$\begin{aligned} &\textit{waste} + \textit{air}\,(\textit{products}) \\ &\quad \mathrm{w(H_2O)} \bullet \mathrm{CxHyOz} \bullet \mathrm{Ash} \\ &\quad + (1+\lambda)\mathrm{a(O_2 + 3.76N_2)} \rightarrow \mathrm{xCO_2} \\ &\quad + (\mathrm{w} + 2/\mathrm{y})\mathrm{H_2O} + \mathrm{a}\lambda\,\mathrm{O_2} + 3.76(1+\lambda)\mathrm{aN_2} + \mathrm{Ash} \end{aligned}$$

where λ: excess air ratio and $\mathrm{a} = \mathrm{x} + \mathrm{y}/4 - \mathrm{z}/2$.

The chemical composition of waste varies significantly between locations and seasons. For an indicative mass and energy balance calculation, typical ultimate and proximate analyses for MSW, such as that shown in Table 2, must be obtained. In this case, the combustible component of MSW becomes $C_{20}H_{37.3}O_{11.6}$ with 31.2% by weight of moisture and 24.2% by weight of noncombustible ash.

A simple mass and energy balance for incineration based on the above properties of waste is shown in Fig. 6. For 1,000 kg of waste, the amount of air required at 70% excess air ($\lambda = 0.7$) is 5,377 kg. The amount of solid residues is 242 kg. Most of these are bottom ash, and the balance is collected in bag filters as fly ash. The amount of exhaust gas is 6,120 kg in which the fraction of oxygen is 8.8% on a dry basis. The typical temperatures of exhaust gas in the gas cleaning process, and at the chimney are 200°C and 130°C, respectively. The electricity output is 2.15 GJ from 8.6 GJ of chemical energy stored in 1,000 kg of waste. This is based on 25% electrical energy efficiency which is typical in a large-scale incineration plant with a steam turbine. As mentioned above, for incineration with combined heat and power (CHP), the energy efficiency can be as high as 85%.



Incinerator Grate Combustion Phenomena. Figure 5
Schematic of mass flow for mechanical treatment followed by incineration

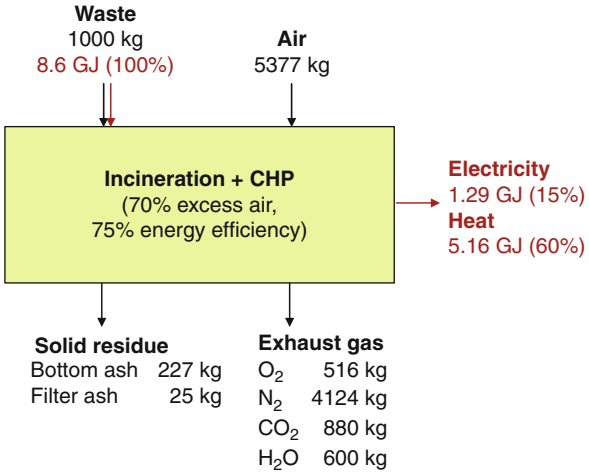
Incineration Modeling Equations

A mathematical model [4], also known as the FLIC code, is used to predict the fixed bed combustion of the waste materials. This model assumes the fuel bed to be a unidimensional bed of uniform spherical particles and gas voids and tracks the combustion in the bed as a function of time. The combustion processes of moisture evaporation, devolatilization, and char burnout for the solid phase, and also the reactions in the gas phase are simulated by various sub-models and modes of heat transfer. The details of the model assumptions, governing equations, various sub-models for reaction and heat transfer, and the numerical scheme are presented below. Application of the FLIC code to the mathematical modeling of the characteristics of a burning bed has provided very valuable results for various solid fuels in fixed or moving bed furnaces.

Early development of solid waste reaction models was largely based in previous work on coal combustion and was linked to biomass combustion due to the fact that a large proportion of municipal solid wastes is biomass or originates from it [4–6]. The devolatilization rate depends not only on fuel type but also on particle size, heating rate, and final temperature. For experimental studies, thermogravimetric analysis (TGA) has been used at low heating rates. However, the pyrolysis processes in an actual incineration furnace or packed-bed pyrolyser differ from the conditions of TGA tests in terms of sample weight and particle sizes. Furthermore, evidence [7] indicated that a bed filled with charcoal at elevated temperatures could reduce tar concentration in the exit gases by 95%; therefore, mathematical models for an actual solid-fuel pyrolysis

Incinerator Grate Combustion Phenomena. Table 2
Ultimate analysis of MSW

Material	% by weight
Water	31.2
Carbon	24.0
Hydrogen	3.2
Oxygen	15.9
Nitrogen	0.7
Sulphur	0.1
Chlorine	0.7
Ash	24.2



Incinerator Grate Combustion Phenomena. Figure 6
Typical mass and energy balance for incineration with combined heat and power

system should include the tar cracking and re-polymerization processes. In the following discussion, solid waste pyrolysis is modeled by two different methods and the effect of devolatilization rate on the combustion of MSW is investigated by numerical simulations. A specific tar cracking feature is also mathematically simulated to optimize the process operation.

Char from solid wastes constitutes an important part of the waste-to-energy conversion process as the burnout time of char can be more than 10 times the devolatilization time. In addition, chars from different parent fuels are expected to have different kinetic rates since minerals in the ash can act as catalysts and, also, the burnout rate of char also depends heavily on mass transport between the gas and solid phases through the outer ash layer. Thus, collapse of this outer ash layer or particles breaking down into smaller fragments during the char burnout processes can significantly affect the results. This effect is taken into account in the mathematical model by introducing a particle mixing coefficient in a packed-bed agitated by a moving grate; numerical predictions indicate that the overall burning rate of the solid fuel can be doubled. To investigate the effect of grate movement on the combustion processes, a transient mathematical model has also been developed which can reflect the random nature observed for the dynamic processes. However, to analyze the effect of fuel properties and most of the other operating parameters such as airflow rate, the transient mathematical model is unnecessary. Therefore, the average or "steady-state" mathematical model has been generally used.

One of the earliest models for packed-bed solid combustion was proposed by [6]. Early models were generally based on the assumption of thermally thin particles; however, in the case of MSW combustion, particle sizes may be very large in some cases. To correctly model the combustion processes of thermally thick particles, an advanced double-mesh numerical model was developed that takes into account the three-dimensional nature of the boundary conditions for an individual particle in a packed bed. The advantage of such a model is its ability to take into account the temperature gradient inside thermally thick particles, making the modeling more realistic.

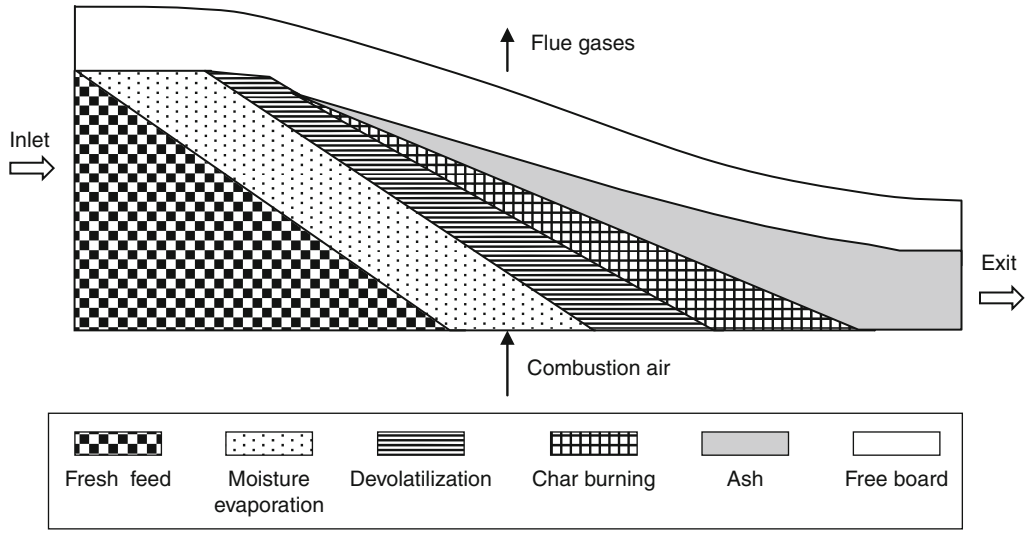
Radiation penetration is important in packed-bed waste combustion as it is the major mode of heat transfer inside packed beds of burning solid fuels and one of the major factors influencing the combustion characteristics. The absorption and scattering coefficients of the packed-particle assembly are therefore important parameters required for advanced modeling of heat transfer inside packed-beds. Detailed 3D calculations of radiation penetration in packed beds of various bed configurations were carried out by employing the Discrete Ordinates method and treating the particle surface as wall boundaries. The radiation absorption and scattering coefficients for the particle assembly were thus deduced and correlated to particle size, shape, surface emissivity, and bed porosity.

Most of the model developments have been validated against small-scale bench-top experiments for obvious economic and practical reasons. Scale-up rules allow for the developed models to be applied to large-scale industrial plants, which is the ultimate goal of all the modeling work.

Steady-State Model for Packed Bed Combustion

A schematic description of a solid-to-energy conversion bed is shown in Fig. 7. Layers of solid particles are packed on a grate and under-fire air flows upward through the bed. Solid material is supplied continuously at one end of the grate and the force of gravity and motion of the grate slowly propel the solids to the discharge end of the grate. Most of the combustible material in the solid is burned within the bed and the freeboard zone above the bed, producing CO_2, H_2O, and minor traces of CO and unburned hydrocarbons. The combustible gases are later burned in the over-bed region by adding secondary air. The solid products are bottom ash and fly ash carried in the combustion gases.

It is assumed that the major bed properties, i.e., temperatures of gas and solid phases inside the bed, gas compositions (O_2, H_2, CO, CO_2, etc.), and solid compositions (moisture, volatiles, fixed carbon and ash), can be described as continuous functions of space. It is also assumed that the bed can be treated as a porous medium, where mass and heat transfer take place between the solid and gas phases and that the shape of the particles is spherical (the surface-volume averaged diameter is used). Under such assumptions, the individual bed processes (moisture evaporation, devolatilization, and char burning) can be viewed as



Incinerator Grate Combustion Phenomena. Figure 7
Schematic of the solid combustion processes in a moving packed bed

taking place layer by layer from the bed top to the bottom. This model is suitable for thermally thin particles. However, for much larger particles, the combustion processes may not occur layer by layer from the bed top to the bottom. Instead, the burning process may be more on an individual particle basis, and the numerical model for thermally thick particles must be used.

General Transport Equations

The general governing equations for both the gas and solid phases in a moving bed include mass and momentum conservation, heat transfer, and species transport. For the gas phase, the following equations apply,

$$\text{Continuity:}\quad \partial\left(\phi\rho_{\mathrm{g}}\right)/\partial \mathrm{t} + \nabla\bullet\left(\phi\rho_{\mathrm{g}}\left(\mathrm{V_g}-\mathrm{V_B}\right)\right) = \mathrm{S_{sg}} \tag{1}$$

$$\text{Momentum:}\quad \partial\left(\phi\rho_{\mathrm{g}}\mathrm{V_g}\right)/\partial \mathrm{t} + \nabla\bullet\left(\phi\rho_{\mathrm{g}}\left(\mathrm{V_g}-\mathrm{V_B}\right)\mathrm{V_g}\right) = -\nabla\rho_{\mathrm{g}} + F(\mathbf{v}) \tag{2}$$

$$\text{Species transport:}\quad \partial\left(\phi\rho_{\mathrm{g}}Y_{\mathrm{ig}}\right)/\partial \mathrm{t} + \nabla\bullet\left(\phi\rho_{\mathrm{g}}\left(\mathrm{V_g}-\mathrm{V_B}\right)Y_{ig}\right) = \nabla\bullet\left(D_{\mathrm{ig}}\nabla\left(\rho_{\mathrm{g}}Y_{\mathrm{ig}}\right)\right) + S_{y_{\mathrm{ig}}} \tag{3}$$

$$\text{Energy:}\quad \partial\left(\phi\rho_{\mathrm{g}}H_g\right)/\partial \mathrm{t} + \nabla\bullet\left(\phi\rho_{\mathrm{g}}\left(\mathrm{V_g}-\mathrm{V_B}\right)\mathrm{H_g}\right) = \nabla\bullet\left(\lambda_{\mathrm{g}}\nabla T_{\mathrm{g}}\right) + S_{\mathrm{a}}\,h_{\mathrm{s}}'\left(T_{\mathrm{s}}-T_{\mathrm{g}}\right) + Q_{\mathrm{h}} \tag{4}$$

In the above equations, $F(\mathbf{v})$ represents resistance of solids to fluid flow in a porous medium and is calculated by Ergun's equations. The fluid dispersion coefficients, D_{ig} for mass and λ_{g} for thermal, consist of diffusion and turbulent contributions. For $\mathrm{Re} > 5$, the corresponding cross-flow and in-flow dispersion coefficients are given by the following equations [8]:

$$\mathrm{D_r} = \mathrm{E^0} + 0.1\,d_{\mathrm{p}}|\mathrm{V_g}| \text{ and } \mathrm{D_{ax}} = \mathrm{E^0} + 0.5\,d_{\mathrm{p}}|\mathbf{V_g}| \tag{5}$$

$$\lambda_{\mathrm{r}} = \lambda_{\mathrm{g}}{}^{0} + 0.1\,d_{\mathrm{p}}|\mathbf{V_g}|\rho_{\mathrm{g}}\,\mathrm{C_{pg}} \text{ and } \lambda_{\mathrm{ax}} = \lambda_{\mathrm{g}}{}^{0} + 0.5\,d_{\mathrm{p}}|\mathbf{V_g}|\rho_{\mathrm{g}}\,C_{\mathrm{pg}} \tag{6}$$

For the solid-phase processes, the equations of particle movement, species transport, and heat transfer are described below,

$$\text{Continuity:}\quad \partial\rho_{\mathrm{sb}}/\partial \mathrm{t} + \nabla\bullet\left(\rho_{\mathrm{sb}}(\mathbf{V_s}-\mathbf{V_B})\right) = \mathrm{S_s} \tag{7}$$

$$\text{Momentum:}\quad \partial\rho_{\mathrm{sb}}\mathbf{V_s}/\partial t + \nabla\bullet\left(\rho_{\mathrm{sb}}(\mathbf{V_s}-\mathbf{V_B})\mathbf{V_s}\right) = -\nabla\bullet\sigma - \nabla\bullet\tau + \rho_{\mathrm{sb}}\mathbf{g} + A \tag{8}$$

$$\text{Species: } \partial\rho_{sb}Y_{is}/\partial t + \nabla\bullet(\rho_{sb}(\mathbf{V}s-\mathbf{V}_B)Y_{is}) = \nabla\bullet(D_s\nabla(\rho_{sb}Y_{is})) + Sy_{is} \quad (9)$$

$$\text{Energy: } \partial\rho_{sb}H_s/\partial t + \nabla\bullet(\rho_{sb}(\mathbf{V}_s-\mathbf{V}_B)H_s) = \nabla\bullet(\lambda_s\nabla T_S) + \nabla\bullet q_r + Q_{sh} \quad (10)$$

The fourth term on the right-hand side of (8) is included to account for particle random movements (mixing) caused by the mechanical disturbance of the moving grate and other random sources. However, for simplicity, the particle velocities are not actually calculated from (8). Instead, predetermined values are given to the horizontal average particle velocity, and the vertical velocity is then calculated by means of (7).

The solid bed effective thermal conductivity λ_s in (10) consists of two parts: the solid material conductivity, λ_{s0}, and the thermal transport caused by random movement of the particle, λ_{sm}.

$$\lambda_g = \lambda_{s0} + \lambda_{sm} \quad (11)$$

The transport coefficients of the solid-phase, i.e., μ_s, D_s, and λ_{sm} in a packed bed have to be estimated. "Particle Prandtl Number" can be defined as Pr_s, and "Particle Schmidt Number," Sc_s, by analogy to the fluid phase; by assuming that $Pr_s/Sc_s = 1$,

$$\mu_s = \rho_{sh}D_s \quad \text{and} \quad \lambda_{sm} = \mu_s C_{ps} = \rho_{sh}C_{ps}D_s \quad (12)$$

In the above transport equations, steady-state is achieved by setting the time-differentiation term to zero.

As discussed previously, thermal radiation is the major heat transfer mode in the packed bed under combustion conditions. The radiation heat transfer in the bed is represented by a four-flux radiation model [9]:

$$dI_r^{\pm}/dr = -(k_a+k_s)I_r^{\pm} + 1/4\,k_a E_b + 1/4\,k_s\left(I_x^{+} + I_x^{-} + I_z^{+} + I_z^{-}\right) \quad (13)$$

where $r = x$ or $r = z$. The scattering coefficient, k_s, is assumed zero as the first approximation, and the absorption coefficient, k_a, is taken as [10],

$$k_a = -1/d_p\,\ln(\phi) \quad (14)$$

Evaporation of Moisture

It is assumed that the progress of drying is limited by the transport of heat inside the particle and the moisture evaporation rate is approximated by (6)

$$\dot{r}_M = \begin{cases} f_M\frac{(T_s-T_{evop})\rho_M c_{pM}}{\Delta H_M \delta t} & \text{if } T_s \geq T_{evap} \\ 0 & \text{if } T_s < T_{evap} \end{cases} \quad (15)$$

where $f_M = 1$.

To overcome potential numerical instability during the computation, the above equation is modified with $f_M = X_M/M$, where M is the initial moisture content in the fuel.

Devolatilization

There are two types of models that are generally used to simulate solid-fuel pyrolysis:

(a) Parallel reaction models
(b) The function-group, depolymerization, vaporization, cross-linking (FG-DVC) model

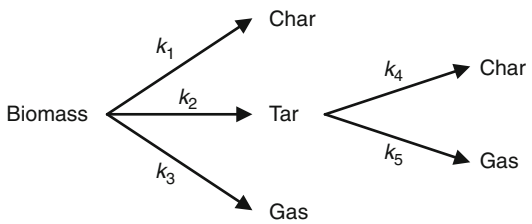
The FG-DVC model is advocated by Advanced Fuel Research, Inc. in USA [11] and was originally developed for coal pyrolysis. The model combines two previously developed models, a functional group (FG) model, and a depolymerization-vaporization-cross-linking (DVC) model. The DVC subroutine is employed to determine the amount and molecular weight of macromolecular fragments, the lightest of which evolves as tar. The FG subroutine is used to describe the gas evolution and the elemental and functional group compositions of the tar and char. The reaction rates are assumed to follow first-order kinetics with distributed activation energy of width σ. In particular, the rate constant k_i for release of the i-th functional group (also called: pool) is expressed by an Arrhenius expression of the form

$$k_i = A_i\,\exp\left(-\frac{E_i \pm \sigma}{RT}\right) \quad (16)$$

where A_i is the pre-exponential factor, E_i is the activation energy, and σ_i the width of distribution in activation energies.

Direct application of the FG-DVC model data to large particles and fuel batches may encounter difficulty due to secondary cracking of tar and re-polymerization, both inside a pyrolyzing particle and on the activated surfaces of other particles; this can significantly increase the char yield and reduce tar production. It has been found that char yield can increase as much as 30–100% in the packed-bed pyrolyser as

compared to standard TGA tests on which the FG-DVC model data are based. Parallel reaction models, on the other hand, provide a much simpler tool to predict the distribution of char, tar, and gases under more realistic conditions, but the elemental composition of each of the products are not predicted. Parallel reaction models are also known as two-stage, semi-global models [12]. In these models, the virgin material is considered as a homogeneous single species and reaction products are grouped into gas, tar, and char. The virgin fuel undergoes thermal degradation according to primary reactions giving as products gas, tar, and char. Tar may undergo secondary reactions in the gas/vapor phase within the pores of the solid matrix, related either to cracking, to give light hydrocarbons, or to depolymerization to give char, as illustrated in the following:



The Function-Group model can be reduced to the more traditional one-step global reaction model if a representative single species is employed

$$dv/dt = k_v(v_\infty - v) \quad s^{-1} \tag{17a}$$

$$\text{where} \quad k_v = A_v \exp(-E_v/RT_s) \tag{17b}$$

In the above equations, v_∞ denotes the ultimate yield of volatile, and v the remaining volatile amount at time t. A_v and E_v are parameters determining the kinetic rate.

Char Burnout

The primary products of char combustion are CO and CO_2,

$$C(s) + \alpha O_2 \rightarrow 2(1-\alpha)CO + (2\alpha - 1)CO_2$$

where the ratio of CO and CO_2 can be expressed as [14]:

$$CO/CO_2 = 2500 \exp(-6420/T) \tag{18}$$

for temperatures between 730 and 1170 K. Ratios outside of this temperature range are calculated using one of the limiting temperatures.

The char consumption rate is expressed as [9]:

$$R_C = P_{O_2}/(1/k_r + 1/k_d) \tag{19}$$

where k_r and k_d are rate constants due to chemical kinetics and diffusion, respectively.

Combustion of Volatiles in the Porous Bed

Gaseous fuels released from the devolatilization process have first to mix with the surrounding air before they can be combusted. Therefore, burning of the volatile hydrocarbon gases is limited not only by the reaction kinetics (temperature dependent), but also by the mixing-rate of the gaseous fuel with the under-fire air. The mixing rate inside the bed is assumed to be proportional to energy loss (pressure drop) through the bed and by recalling the Ergun equations can be expressed as:

$$R_{mix} = C_{mix}\,\rho_g \left\{ 150 \frac{D_g(1-\phi)^{2/3}}{{d_p}^2 \phi} + 1.75 \frac{V_g(1-\phi)^{1/3}}{d_p \phi} \right\} \min\left\{ \frac{C_{fuel}}{S_{fuel}}, \frac{C_{O_2}}{S_{O_2}} \right\} \tag{20a}$$

where C_{mix} is an empirical constant, D_g the molecular diffusivity of the combustion air, V_g the air velocity, d_p the particle diameter, ϕ the local void fraction of the bed, C the molar fractions of the gaseous reactants, and S their stoichiometric coefficients in the reaction.

In the freeboard area immediately above the bed surface, "flame tongues" form and the mixing rate of the volatile gases with surrounding air decreases with increasing distance from the bed surface. Detailed CFD calculations with simulated particle beds were conducted and have produced the following correlation between mixing rate and the distance from the bed top:

$$R_{mix} = R'_{mixo}\left(2.8e^{-0.2y^{++}} - 1.8e^{-2y^{++}}\right) \min\left\{ \frac{C_{fuel}}{S_{fuel}}, \frac{C_{O_2}}{S_{O_2}} \right\} \tag{20b}$$

$$y^{++} = y^{+}/l_p \tag{20c}$$

where R'_{mixo} is calculated from (20a) without the species concentration terms at the bed surface. y^+ denotes the physical distance from the bed top.

A representative hydrocarbon species, $C_mH_nO_l$, can be assumed in the devolatilization products. It is oxidized to produce CO and H_2:

$$C_mH_nO_l + (m/2 - l/2)O_2 \rightarrow mCO + n/2\ H_2$$

CO is then burned by further oxidization to form CO_2

$$CO + {}^1\!/_2O_2 \rightarrow CO_2$$

and H_2 is combusted to H_2O:

$$H_2 + {}^1\!/_2O_2 = H_2O$$

The kinetic rate for $C_mH_nO_l$ oxidation, $R_{C_mH_nO_l}$, is assumed to be the same as that for C_mH_n:

$$R_{C_mH_nO_l} = 59.8\ T_g\ P^{0.3}\ \exp(-12200/T_g)\ C_{C_mH_nO_l}{}^{0.5} C_{O_2} \quad (21)$$

and the kinetic rate for CO, R_{co} was shown by [15] to be

$$R_{co} = 1.3 \times 10^{11} \exp(-62700/T_g) C_{CO} C_{H_2O}{}^{0.5} C_{O_2}{}^{0.5} \quad (22)$$

The kinetic rate for H_2, R_{H2} is adopted from [17] as

$$R_{H_2} = 3.9 \times 10^{17} \exp(-20500/T_g) \times R_{H_2}{}^{0.85} C_{O_2}{}^{1.42} C_{C_mH_nO_l}{}^{-0.56} \quad (23)$$

where the original $C_{C_2H_4}$ has been replaced with $C_{C_mH_nO_l}$ as an approximation.

For tar combustion, a similar reaction rate as $C_mH_nO_l$ is adopted.

In the above equations, C_{CO}, C_{H_2O}, C_{O_2}, and $C_{C_mH_nO_l}$ represent species concentrations, and P the pressure (taken as atmospheric for the packed-bed combustion). The actual reaction rates of volatile species are taken as the minimum of the temperature-dependent kinetic rates and their mixing-rates with oxygen:

$$R = \mathrm{Min}[R_{\mathrm{kinetic}}, R_{\mathrm{mix}}] \quad (24)$$

NO Formation and Destruction

Fuel-nitrogen is assumed to be the dominant source of NO formation in the incinerator furnace. Fuel-nitrogen is released from the solids during pyrolysis either as HCN or NH_3, depending on fuel type and heating rate. In the presence of oxygen, NH_3 or HCN is further oxidized to produce NO and, at the same time, it can also act as a de-NO_x agent to reduce the already-formed NO to nitrogen molecules. In this study, NH_3 is assumed to be the only intermediate product and the well-known De-Soete model is used to calculate the rates:

$$R_{NO} = 4.0 \times 10^6\ X_{NH3}\ X_{O_2}^h \exp\left(\frac{134,400}{RT}\right)\ \ s^{-1} \quad (25)$$

and

$$R_{N_2} = 1.8 \times 10^6 X_{NH_3}\ X_{NO} \exp\left(\frac{134,400}{RT}\right)\ \ s^{-1} \quad (26)$$

where R_{NO} and R_{N_2} are the NO formation and destruction rates, respectively.

It is assumed that the nitrogen remaining in the char after devolatilization can be heterogeneously oxidized to NO at a rate proportional to the rate of char burnout by a factor related to the relative distribution of carbon and nitrogen in the char [16]. At the same time, the surface oxidation of the char carbon by NO to from N_2 can be calculated from:

$$(dNO/dt) = 4.18 \times 10^4 \exp(-34.70\ \mathrm{Kcal/RT})\ A_E P_{NO}\ \ \mathrm{moles/sec} \quad (27)$$

where A_E is the external surface area of the char in m^2/g, and P_{NO} is in atmospheres.

Nomenclature

A	Particle surface area, m^2m^{-3}
A_r	Pre-exponent factor in char burning rate, $kgm^{-2}s^{-1}$
A_v	Pre-exponent factor in devolatilization rate, s^{-1}
C	Constant; molar fractions of species
C_{fuel}	Fuel concentration, kgm^{-3}
C_{pg}	Specific heat capacity of the gas mixture, $Jkg^{-1}K^{-1}$
C_{mix}	Mixing-rate constant, 0.5
$C_{w,g}$	Moisture mass fraction in the gas phase
$C_{w,s}$	Moisture mass fraction at the solid surface
D_{ig}	Dispersion coefficients of the species Y_i, m^2s^{-1}
d_p	Particle diameter, m

E_b	Black body emission, Wm^{-2}
E_r	Activation energy in char burning rate, $Jkmol^{-1}$
E_v	Activation energy in devolatilization rate, $Jkmol^{-1}$
E^0	Effective diffusion coefficient.
H_{evp}	Evaporation heat of the solid material, Jkg^{-1}
H_g	Gas enthalpy, Jkg^{-1}
H_s	Solid-phase enthalpy, Jkg^{-1}
h_s	Convective mass transfer coefficient between solid and gas, $kgm^{-2}s^{-1}$
h_s'	Convective heat transfer coefficient between solid and gas, $Wm^{-2}K^{-1}$
I_x^+	Radiation flux in positive x direction, Wm^{-2}
I_x^-	Radiation flux in negative x direction, Wm^{-2}
k_a	Radiation absorption coefficient, m^{-1}
k_d	Rate constants of char burning due to diffusion, $kgm^{-2}s^{-1}$
k_r	Rate constants of char burning due to chemical kinetics, $kgm^{-2}s^{-1}$
k_v	Rate constant of devolatilization, s^{-1}
k_s	Radiation scattering coefficient, m^{-1}
p_g	Gas pressure, Pa
Q_h	Heat loss/gain of the gases, Wm^{-3}
Q_{sh}	Thermal source term for solid phase, Wm^{-3}
q_r	Radiative heat flux, Wm^{-2}
R	Universal gas constant; process rate, $kgm^{-3}s^{-1}$
R_{mix}	Mixing-rate of gaseous phase in the bed, $kgm^{-3}s^{-1}$
S	Stoichiometric coefficients in reactions
S_{sg}	Conversion rate from solid to gases due to evaporation, devolatilization, and char burning, $kgm^{-3}s^{-1}$
Sy_{ig}	Mass sources due to evaporation, devolatilization, and combustion, $kgm^{-3}s^{-1}$
Sy_{is}	Source term, $kgm^{-3}s^{-1}$
t	Time instant, s
T	Temperature, K
U	x velocity, ms^{-1}
V	y velocity, ms^{-1}
VM	Volatile matter in fuel, wt%
x	Coordinate in bed forward-moving direction, m
y	Coordinate in bed height direction, m
Y_{ig}	Mass fractions of individual species (e.g., H_2, H_2O, CO, CO_2, C_mH_n, . . .).
Y_{is}	Mass fractions of particle compositions (moisture, volatile, fixed carbon, and ash)
ε_s	System emissivity
σ_b	Stefan–Boltzmann constant, 5.86×10^{-8} $Wm^{-2}K^{-4}$
υ	Remaining volatile in solid at time, t
υ_∞	Ultimate yield of volatile
Φ	Void fraction in the bed
λ_g	Thermal dispersion coefficient, $Wm^{-1}K^{-1}$
λ_g^0	Effective thermal diffusion coefficient, $Wm^{-1}K^{-1}$
λ_s	Effective thermal conductivity of the solid bed, $Wm^{-1}K^{-1}$
Subscripts	
env	Environmental
g	Gas phase
i	Identifier for a component in the solid
p	Particle
s	Solid phase

Solution Technique

Except for radiation, the governing equations described above (summarized in Table 3) are generalized into a standard form:

$$\frac{\partial \rho \Phi}{\partial t} + \nabla \bullet (\rho V \Phi) = \nabla \bullet (\lambda \nabla \Phi) + S_\Phi \tag{28}$$

where ρ represents density, V velocities, Φ the parameter to be solved, λ transport coefficient, and S_Φ the source term. The whole geometrical domain of the bed is divided into a number of small cells and (28) is discretized over each cell and solved numerically using the SIMPLE algorithm [13]. For each cell, the equation becomes:

$$a_{i,j}\Phi_{i,j} + a_{i-1,j}\Phi_{i-1,j} + a_{i+1,j}\Phi_{i+1,j} + a_{i,j-1}\Phi_{i,j-1} + a_{i,j+1}\Phi_{i,j+1} = S_{i,j} \qquad (i = 1, M;\ j = 1, N) \tag{29}$$

The radiation equations are solved by the fourth-order Runge–Kutta method.

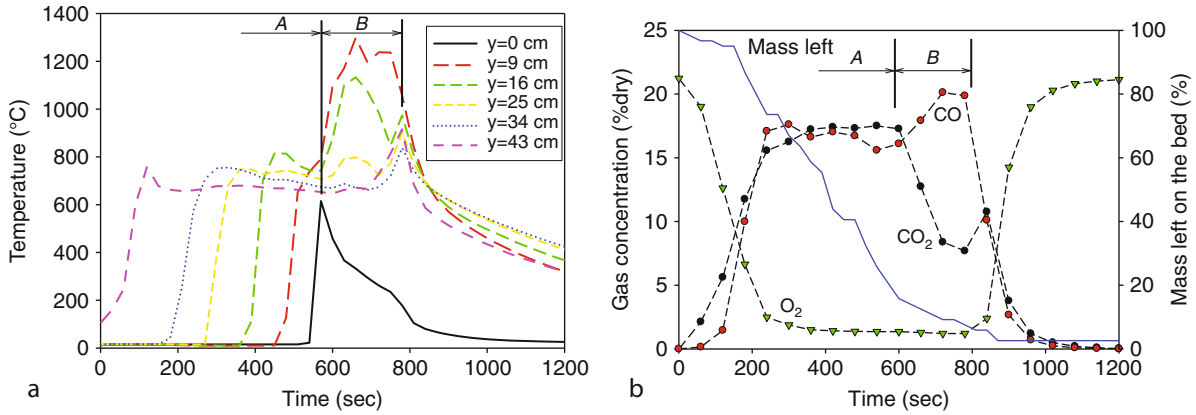
Incinerator Grate Combustion Phenomena. Table 3 Transport equations and reaction rates used in the mathematical models

Process rate equations	Moisture evaporation	$R_m = A_p h_s (C_{w,s} - C_{w,g})$ when $T_s < 100°C$
		$R_m = \frac{A_p\left[h'_s(T_g - T_s) + \varepsilon_s \sigma_b (T_{env}^4 - T_s^4)\right]}{H_{evp}}$ when $T_s = 100°C$
	Devolatilization	$R_v = (1-\phi)\rho_s k_v (v_\infty - v)$ $k_v = A_v \exp\left(-\frac{E_v}{RT_s}\right)$
		$C_m H_n + \frac{m}{2} O_2 \rightarrow mCO + \frac{n}{2} H_2$ $R_{C_mH_n} = 59.8\, T_g P^{0.3} \exp\left(\frac{-12200}{T_g}\right) C_{C_mH_n}^{0.5} C_{O_2}$
	Combustion of volatiles	$CO + \frac{1}{2} O_2 \rightarrow CO_2$ $R_{CO} = 1.3 \times 10^{11} \exp\left(\frac{-62700}{T_g}\right) C_{CO} C_{H_2O}^{0.5} C_{O_2}^{0.5}$
		$H_2 + \frac{1}{2} O_2 \rightarrow H_2O$ $R_{H_2} = 3.9 \times 10^{17} \exp\left(\frac{-20500}{T_g}\right) C_{H_2}^{0.85} C_{O_2}^{1.42}$
		$R_{mix} = C_{mix} \rho_g \left[150 \frac{D_g (1-\phi)^{2/3}}{d_p^2 \phi} + 1.75 \frac{V_g (1-\phi)^{1/3}}{d_p \phi}\right] \min\left[\frac{C_{fuel}}{S_{fuel}}, \frac{C_{O_2}}{S_{O_2}}\right]$ $R = \min[R_{kinetic},\ R_{mix}]$
	Char gasification	$C(s) + \alpha O_2 \rightarrow 2(1-\alpha) CO + (2\alpha - 1) CO_2$ $\frac{CO}{CO_2} = 2500 \exp\left(-\frac{6420}{T}\right)$
		$R_4 = A_P C_{O_2} \Big/ \left(\frac{1}{k_r} + \frac{1}{k_d}\right)$ $k_r = A_r \exp\left(-\frac{E_r}{RT_s}\right)$
Gas-phase conservation equations	Continuity	$\frac{\partial(\rho_g \phi)}{\partial t} + \frac{\partial(\rho_g U_g \phi)}{\partial x} + \frac{\partial(\rho_g V_g \phi)}{\partial y} = S_{sg}$
	x – Momentum	$\frac{\partial(\rho_g U_g \phi)}{\partial t} + \frac{\partial(\rho_g U_g U_g \phi)}{\partial x} + \frac{\partial(\rho_g U_g V_g \phi)}{\partial y} = -\frac{\partial p_g}{\partial x} + F(U_g)$
	y – Momentum	$\frac{\partial(\rho_g V_g \phi)}{\partial t} + \frac{\partial(\rho_g V_g U_g \phi)}{\partial x} + \frac{\partial(\rho_g V_g V_g \phi)}{\partial y} = -\frac{\partial p_g}{\partial y} + F(V_g)$
	Species	$\frac{\partial(\rho_g Y_{i,g} \phi)}{\partial t} + \frac{\partial(\rho_g U_g Y_{i,g} \phi)}{\partial x} + \frac{\partial(\rho_g V_g Y_{i,g} \phi)}{\partial y} = \frac{\partial}{\partial x}\left(D_{ig} \frac{\partial(\rho_g Y_{i,g} \phi)}{\partial x}\right) + \frac{\partial}{\partial y}\left(D_{ig} \frac{\partial(\rho_g Y_{i,g} \phi)}{\partial y}\right) + S_{Y_{i,g}}$
		$D_{ig} = E^0 + 0.5 d_p \lvert V_g \rvert$
	Energy	$\frac{\partial(\rho_g H_g \phi)}{\partial t} + \frac{\partial(\rho_g U_g H_g \phi)}{\partial x} + \frac{\partial(\rho_g V_g H_g \phi)}{\partial y} = \frac{\partial}{\partial x}\left(\lambda_g \frac{\partial T_g}{\partial x}\right) + \frac{\partial}{\partial y}\left(\lambda_g \frac{\partial T_g}{\partial y}\right) + Q_h$
		$\lambda_g = \lambda^0 + 0.5 d_p \lvert V_g \rvert \rho_g C_{pg}$
Solid-phase conservation equations	Continuity	$\frac{\partial((1-\phi)\rho_s)}{\partial t} + \frac{\partial((1-\phi)\rho_s U_s)}{\partial x} + \frac{\partial((1-\phi)\rho_s V_s)}{\partial y} = -S_{s_g}$
		$U_s = f(x)$, predefined
	Species	$\frac{\partial((1-\phi)\rho_s Y_{i,s})}{\partial t} + \frac{\partial((1-\phi)\rho_s U_s Y_{i,s})}{\partial x} + \frac{\partial((1-\phi)\rho_s V_s Y_{i,s})}{\partial y} = -S_{Y_{i,s}}$
	Energy	$\frac{\partial((1-\phi)\rho_s H_s)}{\partial t} + \frac{\partial((1-\phi)\rho_s U_s H_s)}{\partial x} + \frac{\partial((1-\phi)\rho_s V_s H_s)}{\partial y} = \frac{\partial}{\partial x}\left(\lambda_s \frac{\partial T_s}{\partial x}\right) + \frac{\partial}{\partial y}\left(\lambda_s \frac{\partial T_s}{\partial y}\right) + \frac{\partial q_{rx}}{\partial x} + \frac{\partial q_{ry}}{\partial y} + Q_{sh}$
Radiation heat transfer	$\frac{dI_{si}^*}{dx_i} = -(k_a + k_s) I_{xi}^{\pm} + \frac{1}{2N} k_a E_b + \frac{1}{2N} k_s$ $\left(I_{xi}^+ + I_{xi}^-\right), i = 1, N$	
	$k_s = 0$ $k_a = -\frac{1}{d_p} \ln(\phi)$	

Modeling Validation

The modeling predictions have been investigated experimentally for a number of sample materials, where the model parameters were not changed other than the fuel properties. In the simulation, the fuel bed was generally divided into 200 cells, and the time step was fixed at 2 s.

Figures 8a and 8b show the measured temperature history, gas composition, and mass left on the grate for cardboard samples at a low airflow rate (468 kg/m^{2}h).



Incinerator Grate Combustion Phenomena. Figure 8

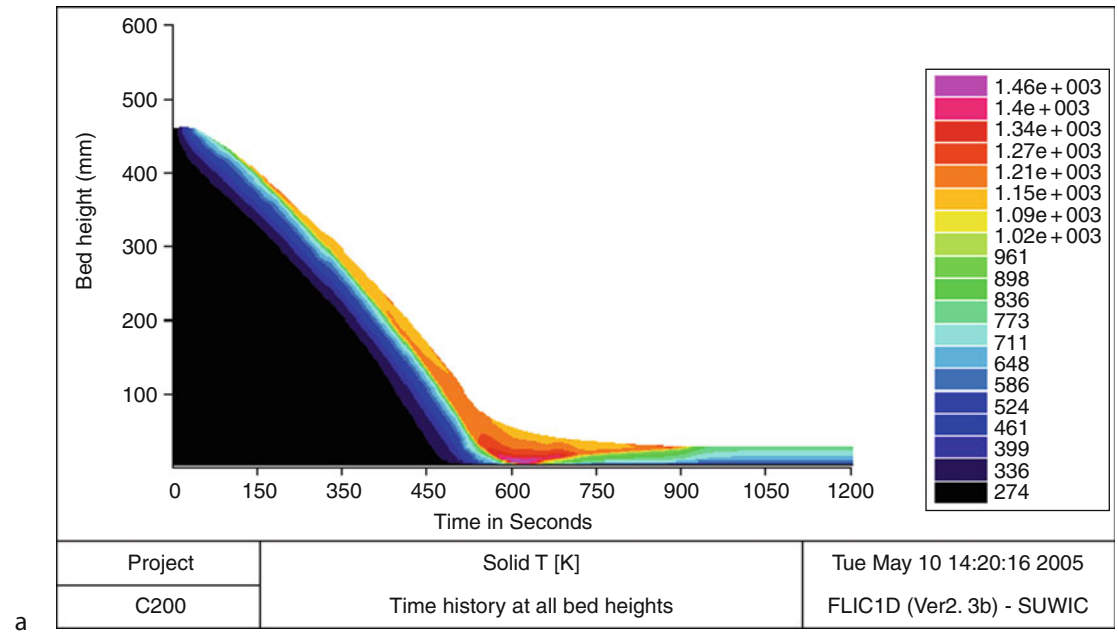
(**a**) Bed temperatures for cardboard combustion at an airflow rate of 468 kg/m^{2}h. (**b**) Gas composition and mass left on the bed for cardboard combustion at an airflow rate of 468 kg/m^{2}h

This type of temperature history is typical for fixed bed combustion as reported in the literature. The temperature plot in Fig. 8a shows two distinct stages with different combustion characteristics. In the ignition propagation stage designated as "A" in Figs. 8a and 8b, the ignition front initiated by the start-up burner propagated into the bed. As the ignition front passed one of the thermocouples in the bed, the volatiles from the burning particles ignited and formed local flames around them and above. This resulted in the temperature increases to around 800°C within a few minutes. The heat released from the reaction of volatiles and char with air transfers downward toward the fresh particles (below the ignition front). This heat is then consumed for further evaporation and heating of fresh fuel. Since oxygen is consumed first by the volatile gases, a layer of char normally accumulates above the ignition front as it propagates downward.

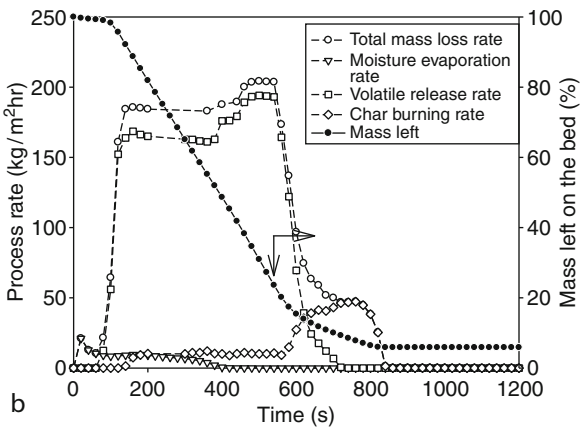
When the ignition front reaches the grate ($y = 0$ cm), the devolatilization of the particles terminated and only the gasification of the remaining char takes place. This stage is designated as "B" in Figs. 8a and 8b. Glowing char particles without flames and high bed temperatures characterize this stage of the combustion process. The temperature at $y = 0$ cm dropped immediately due to the convective cooling. At the same time, the temperatures at $y = 9$ and 16 cm increased rapidly due to active char gasification in this region. As a result, the level of CO immediately began to rise and reached a peak within 150 s as shown in Fig 8b. Note that there were about 60 s delays in the indicated gas composition due to the retention time needed for the sample gas to reach the gas analyzer. The char gasification stages for the cases at an airflow rate of 587 and 702 kg/m^{2}h did not have a CO peak and were shorter than for the above case.

Figures 9a to 9c show the results predicted by the FLIC code for cardboard and waste wood both at an airflow rate of 468 kg/m^{2}h. The thin burning layer and hot spot at 800 s (Fig. 9a) illustrates the key features of bed combustion. As shown, the predicted temperature, weight loss, and gas composition results were in good agreement with the measured data. There was some minor discrepancy between the predicted and measured results for CO and CO_2 during the char gasification stage, although the overall trend was similar. The possible reasons are underprediction of char by the CO_2 to CO reaction or the unaccounted effect of the water gas shift reaction ($CO + H_2O \leftrightarrow CO_2 + H_2$) in this model. In a later model, the char was assumed to be pure carbon, although the char contains a small amount of H element, which is released to the gas phase in the char gasification stage. There was also discrepancy in the bed height in the char gasification stage. Therefore, more information is required on the properties of char particles.

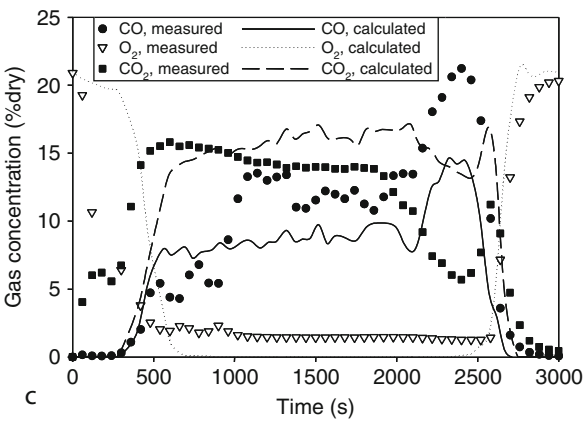
Figure 10 compares the predicted and measured mass loss, ignition rate, and burning rate as a function of airflow for samples of cardboard and waste wood. The predicted ignition rate for cardboard



a



b



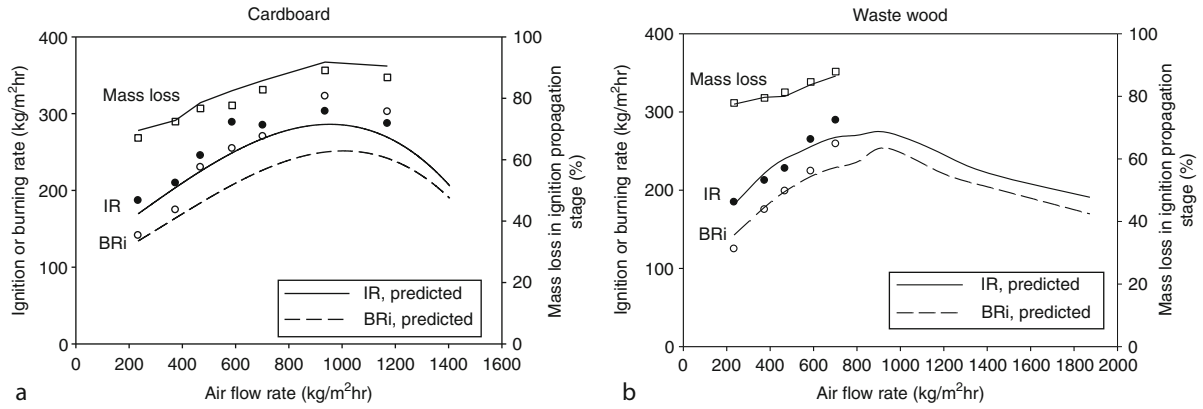
c

Incinerator Grate Combustion Phenomena. Figure 9

(**a**) Predicted solid temperature for cardboard combustion at an airflow rate of 468 kg/m^{2}h. (**b**) Predicted process rates for cardboard combustion at an airflow rate of 468 kg/m^{2}h. (**c**) Predicted and experimental gas composition for waste wood at an airflow rate of 468 kg/m^{2}h

shown in Fig. 10a was reasonably close to the measured values, but the burning rate was underpredicted. This point will now be discussed to illustrate some of the issues to be taken into account when modeling the combustion of municipal wastes:

Particle size determines the surface to volume ratio of a particle, which directly affects the heat transfer (i.e., radiation penetration through the bed and convection) and the reaction rates of char. Since the model herein assumes the fuel particles to be spherical, the particle diameter was determined to be 10 mm in order to maintain the specific surface area (surface area to volume ratio of a particle). However, the cardboard particle has internal corrugated layers, which significantly increase the area available for mass transfer and heterogeneous reactions. The ignition rate is not significantly affected by the particle size since the bulk density of the bed and the calorific value are unchanged. However, it directly increases the predicted burnout time for char, which gives slower burning rates



Incinerator Grate Combustion Phenomena. Figure 10
Comparison of measured and predicted ignition and burning rates for cardboard and waste wood (*IR*: ignition rate, *BRi*: burning rate) for the ignition propagation stage. (**a**) Cardboard. (**b**) Waste wood

than the measured values. The char gasification stage also appears at high airflow rates in the prediction. Therefore, more parametric studies for different particle sizes and corresponding model improvement would be required for cardboard. The introduction of a shape factor would be essential in the model in order to consider the actual surface area of a complex particle shape.

However, as expected, the FLIC model prediction for the waste wood shown in Fig. 10b matched well with the measurements. The predicted ignition rate and burning rate had maximum values at an airflow rate of 936 kg/m^{2}h and gradually decreased at higher airflow rates.

Future Directions

- The incineration of a bed of waste by combustion on a moving grate is a widely used and mature technology. Numerical modeling has helped engineers and scientists to understand the key features of the process and optimize plant design and operation.
- The Sheffield CHP/DHC system represents a successful demonstration of this technology. This waste management strategy is already used widely in several European cities but could be applied in many other towns and cities worldwide to reduce fossil fuel consumption and achieve socioeconomic and environmental objectives.
- As indicated above, the potential electricity output of energy-from-waste plants could be almost doubled by the development of an efficient waste gasification system operating with a gasification efficiency of about 70%. The use of this fuel gas in a modern combined cycle gas turbine power plant that has an efficiency of 55% would give and overall efficiency of electricity generation of 38.5%. Since electricity is a more valuable form of energy than low-grade heat, this is an attractive opportunity for further research and development

Acknowledgments

The contribution of the EEE team to this article is gratefully acknowledged: Q Chen, X Zhang, H Li, K M Finney, C Ryu, N Phan, A Khor and Y B Yang.

Bibliography

1. DEFRA (2004) Review of environmental and health effects of waste management: municipal solid waste and similar wastes. London. Available from http://www.defra.gov.uk/ENVIRONMENT/waste/research/health/
2. http://www.ref-fuel.com.technology.htm
3. Department of Environment, Food and Rural Affairs (DEFRA) (2007) Municipal waste management statistics. London. Available from http://www.defra.gov.uk/environment/statistics/wastets/index.htm
4. Yang YB, Goh YR, Zakaria R, Nasserzadeh V, Swithenbank J (2001) Mathematical modelling of MSW incineration, IT3-2001, Philadelphia

5. Goh YR, Siddall RG, Nasserzadeh V, Zakaria R, Swithenbank J, Lawrence D, Garrod N, Jones B (1998) Mathematical modelling of the waste incinerator burning bed. J I Energy 71:110–118
6. Peters B (2003) Thermal conversion of solid fuels. WIT Press, Southampton
7. San SH, Lu A, Stewart DF, Connor MA, Fung PYH, Ng HS (2004) Tar levels in a stratified downdraft gasifier. Science in thermal and chemical biomass conversion. Victoria, pp 1–6, 30 Aug–2 Sept 2004
8. Wakao N, Kaguei S (1982) Heat and mass transfer in packed beds. Gorden & Breach, New York
9. Smoot LD, Pratt DT (1979) Pulverized-coal combustion and gasification. Plenum Press, New York
10. Shin D, Choi S (2000) The combustion of simulated waste particles in a fixed bed. Combust Flame 121:167–180
11. http://www.afrinc.com/products/fgdvc/default.htm
12. Di Blasi C (1996) Heat, momentum and mass transport through a shrinking biomass particle exposed to thermal radiation. Chem Eng Sci 51(7):1121–1132
13. Patankar SV (1980) Numerical heat transfer and fluid flow. Hemisphere, Washington/London
14. Arther JA (1951) Reactions between carbon and oxygen. Trans Faraday Soc 47:164–178
15. Howard JB, William GC, Fine DH (1973) Kinetics of carbon monoxide oxidation in postflame gases. In: Proceedings of the 14th symposium (international) on combustion, the Combustion Institute, Pittsburgh, pp 975–986
16. Jones J, Williams A, Bridgeman T, Darvell L (2005) Modelling pyrolysis and modelling potassium release. EPSRC (UK) Supergen bio-energy consortium, Annual Researchers' Meeting. Stafford, UK, 23–25 Nov 2005
17. Hautman AN, Dryer FL, Schlug KP, Glassman I (1981) A multiple-step overall kinetic mechanism for the oxidation of hydrocarbons. Combust Sci Technol 25:219

Increasing Salinity Tolerance of Crops

Stuart J. Roy, Mark Tester
Australian Centre for Plant Functional Genomics, University of Adelaide, Adelaide, SA, Australia

Article Outline

Glossary

Exclusion The ability to maintain low concentrations of toxic ions in the plant shoot.

Ionic stress The stress imposed on a plant by the accumulation of salts to toxic concentrations in cells, particularly those of the shoot, leading to premature death.

Genetically modified plant A plant which has been transformed by artificial means with single or multiple genes from another variety or species.

Osmotic stress The stress imposed on a plant by the accumulation of high concentrations of salt around the root, which reduces plant growth.

Osmotic tolerance The ability of a plant to maintain growth under osmotic stress.

Saline soil Soils affected by excess accumulation of salts. Accumulation of sodium chloride (NaCl) on agricultural land has a severe impact on crop yield.

Salt tolerant plant A plant with the ability to grow and set seed in saline environments without significant reductions in plant biomass or yield.

Selective breeding Where two plant species with desirable phenotypes are bred together in an attempt to produce an offspring with both traits.

Tissue tolerance The ability to withstand high concentrations of toxic ions in the shoot.

Definition of the Subject

Plant growth and yield are severely affected by saline soils. High concentrations of salt in the soil make it difficult for plants to take up water, while the accumulated salts in cells, particularly the sodium (Na^+) and chloride (Cl^-) ions, are toxic to plant metabolism. These two factors result in a reduction in plant growth, an increase in the rate of leaf senescence, and a loss in crop yield. The fact that significant areas of farmland worldwide are affected by salt brings with it potentially serious implications for crop yield.

Introduction

Saline soils have been defined as areas where the electrical conductance (ECe; a means of measuring the amount of ions in the soil) is greater than 4 dS/m. It is at around 4 dS/m (approximately 40 mM NaCl) that most plants start to exhibit significant reductions in yield [1]. Over 800 million hectares of land worldwide are affected by saline soils; this accounts for more than 6% of the total land area of the world [2]. Most of this salt-affected land has arisen from natural causes, such as the weathering of rocks, which releases a variety of soluble salts including Cl^-, Na^+, calcium, magnesium, sulfates, and carbonates [3]. Other sources of salt accumulation include the deposition of salts from seawater that is transported by wind and rain, as well as from salts carried in rainwater. It has been estimated that rainwater contains 6–50 mg/kg of sodium chloride (NaCl) which, over time, results in large-scale salt depositions [1].

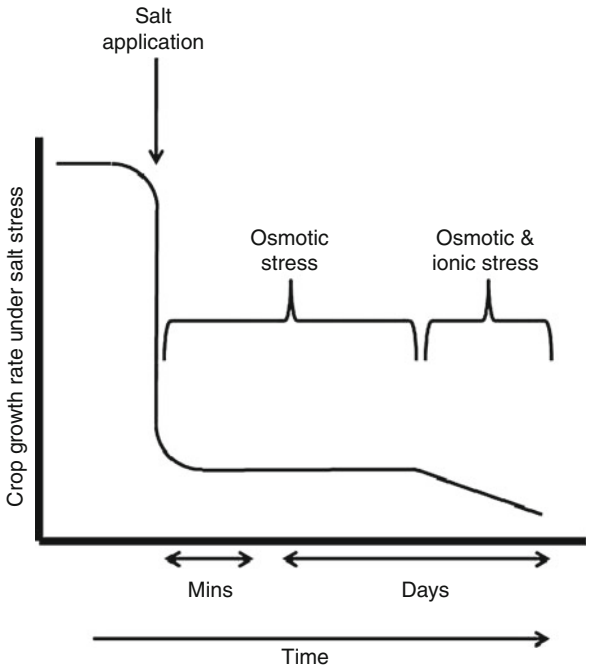
In addition to the natural processes of salinization, farmland areas are affected by secondary types of salinity which are a consequence of human activities such as land clearing and/or irrigation. This secondary form of salinity results in the raising of water tables and an increase in the concentration of salts around plant roots. Approximately 32 million hectares of the 1,500 million hectares farmed by dryland agriculture are affected by secondary salinity, while 45 million hectares of the 230 million hectares of irrigated land are salt affected [2]. Although it accounts for only 15% of the total cultivated area, irrigated land is twice as productive as dryland agriculture. Consequently, losses of yield which result from an increase in soil salinization in irrigated areas have a disproportionally large effect. Unfortunately, the areas of farmland affected by salinization are increasing and irrigated land is particularly at risk [1].

The deleterious effect of soil salinity on agricultural yields is enormous. To solve this problem will require a variety of approaches including altering farming practices to prevent soil salinization; the implementation of remediation schemes to remove salt from soils; and programs aimed at increasing the salt tolerance of crop plants, either through traditional breeding or by genetic manipulation technologies. By increasing crop salinity tolerance, plant varieties can be generated which will grow on marginal saline soils while longer term land management practices are being introduced. However, before crop salinity tolerance can be improved, an understanding is required of the two separate stresses imposed on a plant when it is grown on a saline soil: osmotic stress and ionic stress.

Effects of Salt Stress on Plant Growth

Osmotic Stress

Osmotic stress affects a plant as the salt concentration around the root reaches 4 dS/m and results in an immediate reduction in shoot growth [1, 4, 5] (Fig. 1). Osmotic stress reduces the rate at which growing leaves expand, the rate of emergence of new leaves, and the development of lateral buds. As this stage of salt stress concerns the inability of a plant to maintain water relations, the cellular and metabolic processes



Increasing Salinity Tolerance of Crops. Figure 1
The effect of salt stress on the growth rate of a crop plant. Plants experience an immediate reduction in growth rate after exposure to salt as a result of osmotic stress. Overtime, the effect of ionic stress increases as shoot Na^+ concentrations build to toxic levels (Adapted from [1])

involved are similar to those observed in drought-stressed plants [6]. In dicotyledonous crop species, such as soybean, this osmotic stress results in reductions in the size of leaves and the number of branches [1]; in the monocotyledonous cereals, such as wheat, barley, and rice, the major effect is a reduction in total leaf area and number of tillers [6, 7].

Although it is the roots that are initially exposed to the saline soil, it is actually the growth of the shoot which displays a greater sensitivity to salt; root growth recovers quickly even after exposure to high levels of NaCl [4, 8]. The reduction in leaf development has been attributed to the high salt concentration outside the roots and not to toxic levels of Na^+ or Cl^- within the tissues of the plant [9–11]. This is supported by experiments where plants demonstrate reduced shoot growth when grown in a mixture of salts which individually are at concentrations below those necessary for ionic toxicity but together cause osmotic stress [7, 12]. The mechanisms underlying this down regulation of leaf growth and shoot development remain unclear, but a decrease in shoot area is likely to reduce water use by the plant, thereby conserving soil moisture and preventing an increase in the soil salt concentration. It has been suggested that this reduction in growth rate is regulated by long distance signals in the form of plant hormones and, as the reduction in growth rate is independent of carbohydrate or water supply, is not due to nutrient deficiency [13, 14]. However, it is not just vegetative growth that can be affected by osmotic stress but also the reproductive development of a crop plant – osmotically stressed plants have been found to exhibit either early flowering and/or a reduced number of flowers [1].

Osmotic stress has other detrimental effects on crop plants. On the surfaces of their leaves, plants have stomatal pores, tiny holes through which carbon dioxide (CO_2) enters the leaf for use in photosynthesis and carbohydrate production, and water and oxygen leave the plant. Due to reduced water uptake, osmotically stressed plants close these stomatal pores [1, 15]. The consequent reduction in CO_2 assimilation results in a reduction of carbohydrate production which is detrimental to crop yield. Many plants are able to compensate partially for the reduction in the amount of CO_2 entering the leaf by producing smaller, thicker leaves with more densely packed chloroplasts although this is expensive in terms of expenditure of energy [1]. The decrease in photosynthesis caused by the closure of stomatal pores has the secondary effect of a build up of reactive oxygen species (ROS) [16, 17]. Reactive oxygen species are high energy forms of oxygen, such as superoxide and hydrogen peroxide, which can damage plant DNA and proteins. These ROS accumulate in plant leaves when the energy absorbed from sunlight by chloroplasts cannot be used to synthesize carbohydrates as there is insufficient CO_2 in the leaf to provide the carbon source. If left unchecked these ROS can cause significant damage to plants so cells must produce a range of enzymes, such as superoxide dismutase, ascorbate peroxidase, and catalase, to detoxify and convert the ROS into harmless forms [16, 17]. These detoxifying enzymes are naturally present in plants to protect leaves from sudden burst of sunlight, such as that which occurs when the sun emerges from behind a cloud, but more must be manufactured in response to salt stress, this again being an energy expensive process.

Ionic Stress

Ionic stress has a slower speed of onset than osmotic stress (Fig. 1). It occurs only when the Na^+ or Cl^- accumulation in older leaves reaches a high concentration which results in premature leaf senescence [1, 4, 6, 18]. All salts at high concentrations can affect plant growth but in saline soils it is the Na^+ and Cl^- ions which cause the most detrimental effects on growth. For some plant species, especially citrus, soybean, and grapevines, it is the accumulation of Cl^- ion in the shoot which leads to toxicity, as Na^+ is retained within the roots and the stem [18–22]. However, for most crop plants including the cereals, Na^+ reaches toxic concentrations before Cl^- and is the ion responsible for most of the damage caused to plants [1].

Na^+ and Cl^- are delivered to the shoot in the transpiration stream, that is, in the water which is being transported from the root to the shoot in the xylem of the plant. For most plants, the movement of Na^+ and Cl^- back from the shoot to the roots via the phloem is relatively small, most of the salt delivered to the shoot remaining there [1, 18]. High concentrations of Na^+ in the shoot can cause a range of metabolic and osmotic problems for plants [1, 23]. The metabolic toxicity of Na^+ is largely as a result of its ability to

compete with K^+, which is required for many essential cellular functions. Over 50 essential enzymes have been shown to be activated by K^+. Consequently, high levels of cellular Na^+, which will increase the cellular Na^+:K^+ ratio and decrease the availability of K^+, can disrupt a variety of enzymatic processes [24]. In addition, protein synthesis requires high concentrations of K^+ so that tRNA can bind to ribosomes [25]. A reduction in the amount of available cellular K^+ due to high concentrations of Na^+ will disrupt protein synthesis [26]. As older leaves have ceased expanding they cannot use additional water to dilute the salt being transported into, and this leads to an increase in the senescence of older leaves. Consequently, a failure to exclude Na^+ from the shoot over time will result in the accumulation of toxic levels of ions leading to premature senescence and leaf death [4, 27]. If the rate of leaf death is greater than leaf production, the photosynthetic capacity of the plant will be reduced, the plant will be unable to supply carbohydrates to any new leaves, and the growth rate of the plant will decrease.

There is also an osmotic component to ionic stress. During ionic stress, Na^+ and/or Cl^- remain when water from the transpiration stream evaporates and can, therefore, accumulate to high concentrations in the leaf apoplast [5, 28]. These high extracellular concentrations of ions will result in water leaving cells with a consequent severe impact on cellular function. High concentrations of Na^+ and Cl^- in the leaf also present another osmotic problem, that of maintaining cellular water potential below that of the soil, thereby facilitating water uptake for growth. Under conditions of elevated salt concentrations in the soil, plants need to accumulate solutes in order to maintain water uptake. Under such circumstances, the most readily available and energy efficient solutes are the Na^+ and Cl^- ions; however, high cellular concentrations of these ions are toxic. Although Na^+ and Cl^- can be stored within the vacuole of a cell or in the apoplastic space, plant cells have difficulty in maintaining low cytosolic Na^+ and Cl^- [29–31]. Therefore, in order to reduce the water potential within the cell, plants need to synthesize solutes which can be accumulated at high concentrations in the cytoplasm of cells without interfering with metabolism [32, 33]. The synthesis of such compatible solutes, however, is energetically expensive and can make significant demands on the energy resources of a plant.

Overall, in comparison with non-stressed plants, salt stressed plants grow more slowly and die more rapidly. It has been estimated that, due to its immediate effect on plant growth, the osmotic stress has a greater impact than ionic stress on the growth rate of a crop [1]. Ionic stress affects plants only at a later stage and has a lesser effect than osmotic stress, particularly at moderate salinity levels. Only when salt levels are high or if a plant is extremely salt sensitive will the ionic effect be greater than the osmotic.

Variation in Plant Salinity Tolerance

Plants vary widely in their response to saline soils. Many show reduced rates of growth and yield while others, such as the salt tolerant saltbush (*Atriplex amnicola*), only reach an optimal growth rate when a moderate level of salt is present [1, 34]. Depending on their sensitivity to saline soils, plants can be divided into two groups: the salt sensitive glycophytes, which are relatively easily damaged by salt; and the salt tolerant halophytes which can tolerate, and may even require, high concentrations of salt in the soil. Indeed, the halophytic saltbush has been shown to survive at concentrations of salt similar or higher than that of seawater [34]. It has been estimated that only 2% of plant species are true halophytes, while the majority of species, including most crops, are glycophytes [35]. Within the monocotyledonous cereals, rice (*Oryza sativa*) is one of the most salt sensitive [36–39], and shows a significant decrease in growth and yield when exposed to moderate levels of NaCl. By contrast, barley (*Hordeum vulgare*) is significantly more salt tolerant [1, 40]. While not as tolerant as barley, the hexaploid bread wheat (*Triticum aestivum*), which contains the genomes of three different wheat species (AABBDD), is moderately salt tolerant and is able to exclude 97–99% of Na^+ entering the shoot. The tetraploid durum wheat (*T. Turgidum* ssp *durum*), which has the genomes of two species (AABB), is more salt sensitive than bread wheat as it lacks genes for salinity tolerance found on the bread wheat D genome and can exclude only 94–95% of the Na^+ entering the root [6, 30]. In durum wheat, there is a clear deleterious relationship between the amount of Na^+ that accumulates in its shoot and the yield of the plant: the higher the Na^+ concentration, the lower the yield [41].

The variation in salinity tolerance of dicotyledonous crop species is even greater than that observed in the cereals. On a scale of salt sensitivity, sugar beet has been reported as salt tolerant, cotton and tomatoes intermediate in tolerance, and chickpea, beans, and soybean as sensitive to salt [42, 43]. Many fruit trees, such as citrus, are classified as very salt sensitive [43]. A number of legumes have been shown to be extremely salt sensitive, even more so than rice; others, such as alfalfa (*Medicago sativa*) are more salt tolerant than barley [1]. In addition to this variation in salinity tolerance between different crop species, variation also exists within species, some varieties and lines having significantly greater salinity tolerance than others [40–44].

Mechanisms of Salt Tolerance

Osmotic Tolerance

Osmotic stress immediately reduces the expansion rate of shoots and roots. It also results in the closure of stomatal pores. Plants that are more tolerant to osmotic stress will exhibit greater leaf growth and stomatal conductance. This would be desirable in irrigated farmland where water is not limiting, but may be problematic in dryland agricultural systems if the soil water content is depleted before the end of the growing season.

Although it is believed that considerable variation for osmotic tolerance may exist within crop species, until recently this was not easily measured. The estimation of growth rates requires daily measurements of leaf growth or measurements of stomatal conductance [7, 41, 45–47]. These methods are usually either time consuming or have required destructive measurements of plant material to ensure accuracy. Nondestructive imaging technologies have been developed which use digital photographs to calculate plant area and mass [48], or infrared thermography to measure leaf temperature and, thereby, stomatal conductance [49]. These technologies have been used to measure the growth rates of plants in saline environments and, hence, measurement of osmotic tolerance. Variation for osmotic tolerance has now been observed in durum wheat [45, 49] and in wild relatives of wheat, such as *T. monococcum* which is a modern day variety of the plant which donated the A genome to both durum and bread wheat [48].

Ionic Tolerance

Na^+ can accumulate in the shoots of plants to reach toxic levels at concentrations which are below those required of Cl^- for toxicity. Consequently, most studies have focused on revealing any variation in shoot Na^+ accumulation and on the transport of Na^+ within the plant. Ion concentrations in specific tissues can easily be measured at a specific developmental age, and either image analysis [48] or a meter that measures chlorophyll content can be used to measure leaf senescence.

Ionic Tolerance: Exclusion A long established mechanism for salinity tolerance in crop plants is the exclusion of ions, particularly Na^+, from the shoot. Due to the ease of experimentation, this is the mechanism perhaps most studied. A strong correlation between salt exclusion and salt tolerance has been shown for many crops, such as in durum wheat [41, 50], rice [51, 52], barley [40, 53, 54], lotus [55], and *Medicago* [56]. Na^+ enters a plant initially from the soil through the root and is then rapidly transported to the shoot in the water of the transpiration stream. Roots are able to maintain relatively constant levels of NaCl by exporting excess salt either back to the soil or to the shoot. As a result, there is a higher accumulation of Na^+ in the shoot compared with the root. If the net delivery of Na^+ to the shoot could be reduced, this may enable a plant to become more salt tolerant. There are four distinct components that can be modified in order to reduce shoot Na^+ and Cl^- concentrations, all of which occur in the root: reduction in the initial influx of ions from the soil into the root; maximization of the efflux of ions from the roots back to the soil; reduction of the efflux of ions from the inner root cells into the xylem cells which are carrying water and ions to the shoot in the transpiration stream; and maximization of ion retrieval from the transpiration stream into root cells thereby retaining Na^+ and Cl^- in the root.

Ionic Tolerance: Tissue Tolerance Tissue tolerance is the ability to accumulate Na^+ or Cl^- ions in the absence of any detrimental effects on plant health. Tolerance requires the toxic ions to be compartmentalized into areas where they can do no damage. At the cellular level, this usually involves avoiding the accumulation of Na^+ and Cl^- in the cytoplasm of the plant cell where

most of important metabolic processes occur. One strategy of tissue tolerance involves compartmentalization of ions within the vacuole, a large plant cell organelle which can be used as a storage structure. Employing such a mechanism will allow a plant to accumulate high concentrations of Na^+ and Cl^- within the shoot, while avoiding all of the toxicity effects. There already exists a large body of evidence for variation between different varieties of crops in terms of the rates of accumulation of shoot Na^+ and Cl^-, as well as for the concentrations of these ions which the different varieties can tolerate.

Generation of Salt Tolerant Crops

Salt tolerant crop plants may be generated only once there is a clear understanding of the mechanisms underlying salinity tolerance, and of the variation between plant species in effecting such mechanisms. Once identified, the benefit of introducing these salinity tolerance mechanisms into crops must be considered. For example, there would be little point in introducing into a cereal the salinity tolerance mechanisms from a slow growing highly tolerant halophyte if that mechanism involved a slow growth phenotype which would result in the cereal taking years to reach maturity for flowering. In addition, a salt tolerant crop plant must do as well as a sensitive plant when grown in the absence of salt. A high yielding salt sensitive crop which shows a 50% yield reduction under salt stress will still be of greater value to a farmer than a salt tolerant variety which displays little reduction in yield but which produces only 40% as much grain as the salt stressed sensitive variety in the first place.

Crop plants developed to have increased tolerance to both ionic and osmotic stresses would be able to grow at productive rates throughout the life cycle, and the severe losses of yield experienced for most crops growing on saline soils would be reduced. It should also be noted that it may be necessary to develop crop plants with different salinity tolerance mechanisms depending on the environment in which the plants will grow. Crops grown by dryland agriculture may benefit particularly from possessing tissue tolerance mechanisms, as the accumulation of high concentrations of ions within the vacuoles of the plant cells may assist the plant in retrieving more water from the soil. By contrast, an osmotic tolerance strategy, combined with Na^+ exclusion, may be more beneficial to crops grown under irrigation so that water availability is not an issue but the Na^+ content in that water may be high.

Two approaches may be taken to improve the salt stress tolerance of current crops: the exploitation of natural variation in salinity tolerance between different varieties and species of crops; or, the generation of transgenic plants with altered gene expression to increase salinity tolerance. Both approaches have advantages and disadvantages as discussed below.

Exploitation of Natural Variation

For many crop species there exists large natural variation in salinity tolerance mechanisms, with some lines and varieties producing significant yields under salt stressed environments. The screening of 5,000 accessions of bread wheat led to the identification of 29 accessions which produced seed when grown in 50% seawater [57], while screening of 400 Iranian wheats identified several accession with high grain yield under both salt stressed and control environments [58]. Varietal differences in yield in saline conditions have been observed in many crops such as durum wheat [41, 45, 59], barley [60, 61], soybean [62], citrus [19, 63], chickpea [42, 64], and rice [65, 66]. The selection and breeding of these salt tolerant varieties with the current elite varieties grown by farmers would be a step forward in the generation of salt tolerant plants.

The selective breeding of lines with desirable salt tolerance traits with those lines possessing desirable traits for yield is an approach for generating salinity tolerant crops that has been practiced for thousands of years. One limitation with this approach is the time and space necessary to grow offspring from these crosses, test their salinity tolerance, obtain viable seed, and then repeat the crossing with a parent to produce the next generation. Recently, new molecular technologies have been developed which have aided this approach considerably. Different varieties and species have different DNA sequence. The difference between the DNA may be subtle, such as between varieties of the same species where there may be a single nucleotide change in the coding sequence of a gene, or the differences can be extreme, such as the gene duplications or deletions observed between species. Modern molecular techniques

enable the detection of these differences between individuals, varieties, and species and allow the design of molecular markers which recognize specific differences in the DNA between two individuals. Using these molecular markers as DNA landmarks it is possible to produce a map of plant chromosomes which can be divided into regions. By finding differences in regions of DNA between two varieties of plants and then observing the phenotype of the offspring produced by breeding the two original varieties, it is possible to identify regions in DNA linked to that phenotype. These regions are often called quantitative trait loci (QTL). By identifying two different plant varieties with differences in salinity tolerance and by observing molecular markers that are different between the two parents, it is possible to identify QTL linked to salinity tolerance by screening their offspring. As salt tolerance is a complex trait, both genetically and physiologically, it is not uncommon to observe several QTL associated with tolerance.

QTL have now been identified for salinity tolerance in a number of plant species including barley [67, 68], tomato [69], rice [70], citrus [63], bread wheat [71], and durum wheat [41]. When QTL have been discovered, one approach is to then identify the gene in that region of DNA which is responsible for the salt tolerance phenotype. The *SKC1* QTL identified on chromosome 1 in rice and the *Nax1* and *Nax2* loci observed in durum wheat on chromosomes 2A and 5A, respectively, have been narrowed down to genes belonging to a family of Na^+ transporters [70, 72, 73], which have been shown previously to be important for exclusion of Na^+ from the shoot [74–79]. Once a QTL has been discovered, the plant which contains that important piece of DNA for salt tolerance can be bred with salt sensitive varieties to introduce into them the salt tolerant phenotype. As a molecular marker will be linked to the QTL, it is not necessary to screen every offspring produced from this cross with a salt sensitivity assay, rather, it is possible to identify which of the offspring have the piece of DNA important for salt tolerance by screening for the molecular marker linked to the salt tolerance QTL. While this does not necessarily speed up the length of time it takes an individual plant to reach maturity, it does reduce the necessity to screen hundreds of plants in saline conditions looking for those that are salt tolerant, so that more focus can be placed on breeding tolerance traits into crops.

While one approach is to identify a variety of a crop with good salt tolerance and then cross it to other members of that species, a second approach is to introduce salt tolerance traits from related species or near-wild relatives.

Bread wheat and durum wheat are two separate plant species, but because of their genetic background there is the possibility of breeding these two species together to exchange valuable traits. Durum wheat is a tetraploid (AABB) containing two genomes from an ancient ancestral cross, the A genome and the B genome. Bread wheat is a hexaploid (AABBDD), with the same A and B genomes as durum wheat and also a third genome, the D genome [80]. It is possible to breed a bread wheat and durum wheat together to produce a pentaploid hybrid offspring, which has an AABBD genome. During sexual reproduction, there is the possibility of the chromosomes from the different wheat backgrounds to swap DNA, a process called recombination, thereby transferring genes from bread wheat to durum wheat and vice versa. Importantly, however, only chromosomes from the same genome can recombine, i.e., durum genome A with bread genome A and not durum genome A with bread genome B. By crossing this offspring with either bread wheat or durum wheat it is possible to re-obtain tetraploid durum wheat and hexaploid bread wheat, only now containing genes from the other species. This has been done successfully to transfer disease resistance genes [80] and could be used for transferring salt resistance traits between the two species. Although difficult, because durum wheat contains no D genome, it is possible to introduce genes from the bread wheat D genome into durum wheat; however, a special wheat plant, with a mutation that affects the way in which chromosomes align in recombination, is required [81]. This technique was used to transfer the K^+/Na^+ discrimination locus *Kna1* from chromosome 4D of bread wheat to chromosome 4B of durum wheat [82]. This new durum wheat line was able to maintain a high K^+/Na^+ ratio in the leaves [82, 83], thereby increasing its salinity tolerance. However, there was no significant difference in grain yield between durum plants with the bread wheat *Kna1* and those without, perhaps due to a yield penalty imposed by having a large section of the bread wheat D genome in durum wheat. Unfortunately, no agronomically

acceptable durum variety containing the bread wheat *Kna1* locus has been released [80].

In addition to looking for variation in plant salt tolerance in current cultivars, there is the possibility of introducing salinity tolerance traits to crop from their near-wild relatives. These species may have been evolving in areas of high salinity, away from the selective pressures inflicted on domesticated crops. It is, therefore, likely these relatives have developed novel salt tolerance mechanisms which might be introduced into current crops. This approach is not new and there have been many attempts to introduce genes from salt tolerant wild relatives to current salt sensitive crops. Traits for salt tolerance have been discovered in wild relatives of tomato, [84], potato [85], rice [44], wheat [80, 86–88], and barley [40, 86] and several attempts have been made to introduce them to cultivated crops. Screening of eight wild *Hordeum* species, wild relatives of domesticated barley, revealed that seven of the eight had better Na^+ and Cl^- exclusion than domesticated barley under a variety of salt stressed environments. A number of these relatives, such as *H. spontaneum*, *H. marinum*, and *H. intercedens*, had significantly higher relative growth rates than domesticated barley when grown under high salinity stress [40, 89].

T. urartu (AA) is the modern day ancestor of the species that gave rise to the A genomes of durum and bread wheat. Both *T. urartu* and other closely related A genome species, such as *T. monococcum* spp. *monococcum* and *T. monococcum* spp. *aegilopoides*, show greater Na^+ exclusion than durum wheat [80, 90]. Lines of *T. monococcum* also show great variation in both osmotic and Na^+ tissue tolerance [48] and are, therefore, a potential source of novel genes for salinity tolerance. It is possible to cross these species with durum wheat and transfer salinity tolerance traits. One success story of breeding a salinity tolerant crop has been the introduction of a Na^+ exclusion trait into durum wheat from a near-wild relative *T. monococcum*. Screening of multiple durum wheat lines for Na^+ exclusion from the shoot identified a durum landrace, line 149, with significantly lower shoot Na^+ than cultivated durum [59]. It was discovered that Na^+ exclusion in these lines was controlled by two major genes, *Nax1* and *Nax2*, which had been introduced into durum from a cross with *T. monococcum* [91]. These two genes have now been introduced separately into the Australian durum wheat Tamaroi and have undergone field trials.

Another wild relative source for wheat is from *Aegilops tauschii*, which is the modern day version of the species that donated the D genome to bread wheat – there appears to be no modern day equivalent of the B genome [80]. Several screens of *Ae. tauschii* have identified lines with lower shoot Na^+ accumulation and enhanced K^+/Na^+ discrimination than durum wheat, although the phenotype was comparable to bread wheat [29, 71, 92]. As the natural environment of *Ae. tauschii* is dry and moderately saline [80], there exists the possibility of introducing novel salinity tolerant genes into wheat. One possibility is the re-creation of the original cross that generated bread wheat by breeding durum wheat (AABB) with *Ae. tauschii* (DD), thereby creating a synthetic wheat (AABBDD) with a genome similar in style to bread wheat. While the technique is tricky, it has been successful in the past [92]. The advantage of this technique is that the Na^+ exclusion locus found on the D is introduced which is not on the A or B genomes of wheat. Synthetic hexaploid lines with enhanced Na^+ exclusion have been created successfully to have Na^+ exclusion similar to that of the parent *Ae. tauschii* and significantly greater exclusion than that of the durum parent [29] at both high and moderate levels of salt. Indeed, some of the synthetic hybrids produced have significantly lower shoot Na^+ accumulation than bread wheat and often greater yield under salt stress conditions [80, 93]. These results indicate that the approach clearly has validity.

The use of wild relatives in breeding programs remains controversial as few salt tolerant crops are released through this approach [39]. Wheat was one of the earliest crops to be crossed with halophytic wild relatives but over 25 years have elapsed since that initial cross, and no new tolerant varieties has yet been released to farmers [39]. However, a recent report of significant yield advantage in a saline field site of durum wheat plants incorporating a Na^+-excluding locus, *Nax2*, from *T monococcum* appears to be particularly promising [94].

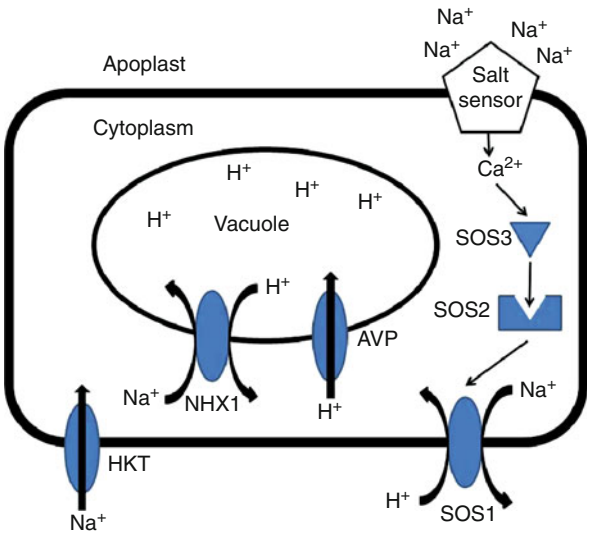
A considerable disadvantage with introducing salinity tolerance traits into crops which are already well adapted for cultivation is the introduction of

undesirable traits encoded by genes which may be physically close to the desirable gene for salinity tolerance in the plant genome [80]. This is a particular problem when breeding current crop varieties with wild relatives, as cultivated crops have been designed by breeders for thousands of years to have desirable traits such as high grain yield, appropriate height and disease resistance. When new traits are introduced into crops by breeding it is not possible to introduce only the gene responsible for that trait. The piece of DNA introduced from the wild relative can be quite large and will contain many genes, for most of which the functions are unknown. If these genes have an undesirable effect which impacts on the agricultural value of the crop leading, for example, to low yield or incorrect flowering time, the crop will be of no value to a farmer. This phenomenon is known as linkage drag [39, 80]. It is possible to reduce the size of the DNA insertion from the wild species by breeding the line with a cultivated crop as, over time, fragments of the wild species DNA will be replaced by that of the cultivated crop. This process, however, can take many generations and requires the breeder to have a molecular marker specific to the Na^+ tolerance gene that has been inserted into the genome; otherwise this gene also would be lost [80].

Transgenic Approaches to Generating a Salt Tolerant Crop

Transgenic approaches are attractive in the generation of salinity tolerant plants, as the sequences of genes known to encode proteins involved in salinity tolerance can be artificially introduced directly into the target variety, without the compounding effects of bringing in multiple, and often undesirable, genes through traditional breeding approaches. In theory, the transformation of commercially relevant crop plants directly with genes for salinity tolerance would help to reduce the time required before farmers can use these crops in the field.

There are numerous possibilities for generating transgenic crops with increased salinity tolerance, either by introducing novel genes for salinity tolerance into crops from other plant species, or by altering the expression of existing genes within the crop (see Fig. 2 for examples).



Increasing Salinity Tolerance of Crops. Figure 2 Intracellular location of proteins in a plant cell which are encoded by candidate genes for transformation into transgenic crop plants. NHX1 is a transporter which is involved in the compartmentation of Na^+ into the vacuole by swapping a cytoplasmic Na^+ ion with a vacuolar proton (H^+). AVP is a proton pump that uses the energy released from the breakdown of PPi to move protons into the vacuole. These protons can then be used by transporters such as NHX to transport Na^+ into the vacuole. HKT proteins are involved in the transport of Na^+ from the extracellular space (apoplast) into the cytoplasm. In the salt overly sensitive (SOS) pathway, high concentrations of Na^+ are detected by a membrane bound salt sensor, which results in the release of Ca^{2+} to the cytoplasm. This cytoplasmic Ca^{2+} binds to the calcium binding protein SOS3, which activates the protein kinase SOS2. Together SOS3 and SOS2 activate the Na^+ transporting ability of the SOS1, which moves Na^+ out of the cell

To date, the greatest success with the development of transgenic salinity tolerant crops has been the generation of plants which are better able to compartmentalize Na^+ in the vacuole, where Na^+ can accumulate to high levels without detrimental effects on the plant cells. Central to this process of vacuolar compartmentation is a gene encoding a vacuolar Na^+/H^+ antiporter (NHX), which transports Na^+ into the vacuole in exchange for a proton (H^+) [1, 95, 96] (Fig. 2). The activity of this transporter has been shown to increase

I

under salinity and it is expressed in a variety of different plant species including barley [97], maize [98], sunflower [99], tomato [100], cotton [101], and Arabidopsis [102]. Constitutive expression in yeast of the *NHX* gene from Arabidopsis, *AtNHX1*, had the effect of significantly increasing the salinity tolerance of the yeast [103, 104]. Transgenic plants which have been created to constitutively overexpress the same Arabidopsis *AtNHX* gene, such as Arabidopsis [105], tomato [106], *Brassica napus* [107], and cotton [108], also show increased Na^+ accumulation in the shoot and greater salinity tolerance. These plants are, therefore, Na^+ tissue tolerators. Importantly from a farmer and a consumer point of view, Na^+ accumulation only occurred in the green tissue and not in the fruit, as in the case for tomato [106]. Of particular interest is that both increasing or decreasing the expression of the Arabidopsis *AtNHX* gene has been shown to significantly affect the expression patterns of other genes involved in salinity tolerance mechanisms [96, 109, 110]. This finding has significant implications for the generation of transgenic crops as it indicates that it may not be necessary to transform a plant with multiple salinity tolerant genes but rather with one gene which can regulate others.

Several homologues of the *AtNHX1* gene have now been identified in a number of crops including wheat [111, 112], barley [113], cotton [101], *Medicago* [114], Maize [98], and rice [115, 116], and the constitutive overexpression of these gene in Arabidopsis [111], rice [115–117], wheat [118], tobacco [101], and barley [113] has also been reported to improve salinity tolerance.

Another candidate gene family for the generation of salinity tolerant transgenic crops are the vacuolar H^+ pyrophosphatase genes [1, 96, 119]. Similar to the *NHX* genes, H^+ pyrophosphatase genes, such as Arabidopsis *AVP1*, are involved in Na^+ sequestration to the vacuole. These genes do not encode proteins that are directly responsible for the transport of Na^+ into the vacuole, but rather ones that use the energy released from the breakdown of the high energy molecule inorganic pyrophosphate (PPi) to pump protons (H^+) into the vacuole (Fig. 2). PPi is produced as a by-product of a wide range of biosynthetic pathways. Use of PPi as an energy donor for the activity of the vacuolar H^+-PPase allows ATP to be conserved and improves plant cell performance under more demanding environmental conditions. Once the vacuolar H^+ pyrophosphatase proteins have transported protons into the vacuole, these protons can then be used by Na^+ transporters such as NHXs to move Na^+ into the vacuole. Analysis of plants that are growing under salt stress, such as barley and Arabidopsis, reveals that these genes are significantly upregulated [102, 113]. Arabidopsis, alfalfa (*Medicago sativa*), tobacco (*Nicotiana tabacum*), bentgrass (*Agrostis stolonifera* L.), and rice plants genetically engineered to either express *AVP1* alone, or in combination with *NHX*, have been shown to have increased salinity tolerance [111, 117, 119–122]. Transgenic alfalfa which was constitutively overexpressing *AtAVP1* maintained a greater shoot biomass than wild type alfalfa when grown on 200 mM NaCl [122]. Similarly, transgenic bentgrass expressing the *AtAVP1* gene was not greatly affected when grown on 100 mM NaCl, and was able to survive salt stress of 200 and 300 mM NaCl, levels which severely reduced the growth of wild type bentgrass [123].

In addition to the success in generating salt tolerant plants using genes involved in the mechanisms for sequestering Na^+ in the vacuole, transformation of plants with genes controlling other processes, such as exclusion of Na^+ from the plant, have also been successful. Other candidate genes for increasing the salinity tolerance of crop plants include members of the Salt Overly Sensitive (SOS) pathway.

Many aspects of plant growth, development, and responses to environmental stresses are mediated by the calcium ion (Ca^{2+}) as a secondary messenger signaling molecule. The external cue is first perceived by receptors on the plant cell membrane and this then activates a signaling cascade, using calcium, which regulates the activities of proteins and gene expression [124–127]. The SOS pathway mediates the response of a plant cell to salinity stress. The SOS pathway was so named due to the extreme salt sensitivity of plants which had mutations in key genes of this pathway [128]. Initially, three genes from these mutants, *AtSOS1*, *AtSOS2*, and *AtSOS3*, were identified in Arabidopsis as being important in salinity tolerance [129]. It should be noted, however, that the SOS name refers to a specific salt sensitive phenotype and that the genes sharing the same SOS identifier are unrelated to each other. Indeed, the proteins encoded

by these genes are quite different: AtSOS1 is a plasma membrane Na^+ transporter [130]; AtSOS2 is a protein kinase belonging to a large family of Calcineurin B-like Interacting Protein Kinases (CIPKs) [125, 127]; and AtSOS3 is a plasma membrane bound Ca^{2+} binding protein which belongs to the Calcineurin B-Like proteins (CBL) [125, 127]. However, although they have completely different functions, it is the interactions of these proteins that help a plant cell survive salt stress.

It has been shown in Arabidopsis that under salt stress Ca^{2+} is released into the plant cell cytoplasm from either internal or external cellular stores and it binds to the plasma membrane bound AtSOS3 (AtCBL4). CBL proteins have specific regions which allow them to bind to specific CIPKs, such as SOS2. When Ca^{2+} becomes bound to AtSOS3, it recruits AtSOS2 to the plasma membrane where the kinase phosphorylates the Na^+/H^+ antiporter AtSOS1, thereby activating the transporter and allowing the movement of Na^+ out of the cell [1, 124, 125, 127] (Fig. 2). Although these genes were identified initially in Arabidopsis, homologues for all three genes have now been discovered in a variety of plant species, including crops, such as *Thellungiella halophila* [131], poplar [132], and rice [133]. In all of these species, the genes involved in the SOS pathway have been shown to be significantly upregulated under salt stress, particularly in the plant roots. This would make them ideal as candidate genes for transformation into transgenic crops to increase salinity tolerance.

Arabidopsis plants that were engineered to constitutively express the *AtSOS1* gene had significantly increased salinity tolerance, showing greater biomass, increased chlorophyll retention, and reduced concentrations of Na^+ in the shoot when compared to wild type plants when grown under high saline conditions [134]. Importantly, these plants did not suffer any yield penalty when grown under non-stressed conditions. The increase in the salinity tolerance of the transgenic plants was attributed to them having a great efflux of Na^+ at the cellular level, when compared to control plants.

It is not always necessary to generate a salt tolerant plant by altering the expression level of a gene that encodes a transporter of ions. The salinity tolerance of a plant can also be increased by overexpressing genes encoding molecules that are involved in signaling or activating genes. Overexpression of the transcription factor *Alfin1* in alfalfa resulted in plants with increased root and shoot growth under both control and salt stressed conditions [135]. Enhanced expression of genes involved in signaling pathways, such as those encoding calcium binding CBL proteins and the protein kinase CIPKs, increases the salt tolerance of Arabidopsis, rice, and tobacco [136–139], presumably through enhancing the signaling response of the cell when it is under salt stress. However, the way in which some genes contribute to overall salt tolerance remains unclear. Transgenic tomato that had been transformed with the yeast gene *HAL1* showed increased salt tolerance under stressed conditions but had reduced shoot weight when grown in control conditions, significantly lower than non-transformed plants [140]. This demonstrates that there remains significantly more to understand about the timing and regulation of genes *in planta* before a transgenic salt tolerant plant can successfully be produced.

Certain genes that have been identified as important for plant salinity tolerance have nevertheless not been shown to increase the salinity tolerance of genetically modified plants when constitutively overexpressed. For example, although the *HKT* gene family has been shown to be important in salinity tolerance, the constitutive overexpression of an *HKT* gene was found to have a detrimental effect. The *HKT* gene family can be divided into those genes encoding a Na^+ transporting protein (subfamily 1) or a K^+/Na^+ transporting protein (subfamily 2) [23, 74]. Members of subfamily 2 are considered to be involved in nutrition and the uptake from the soil of ions essential to plant growth (small quantities of Na^+ can be beneficial to plant growth) [141–144], whereas members of subfamily 1 are believed to be important for plant salt tolerance [1, 23, 96]. Members of the subfamily 1 *HKT* gene family have been shown to encode proteins important for the retrieval of Na^+ from the xylem in both the root and the shoot, thereby reducing the accumulation of Na^+ in the shoot [23, 74, 75, 79, 96, 145]. The protein moves Na^+ from the transpiration stream into the cells surrounding the xylem (Fig. 2). Evidence for this function has now been found in a number of plant species in addition to Arabidopsis, such as rice [70] and wheat [72, 73]. Both naturally occurring ecotypes and mutant lines of Arabidopsis

which have reduced expression of this gene show increased shoot Na^+ accumulation [75, 102, 146, 147]. However, constitutive overexpression of this subfamily 1 *HKT* gene also results in higher concentrations of Na^+ and salt sensitive plants [77]. As HKT proteins move Na^+ into cells, the increased salt sensitiveness of constitutive overexpressing plants may be due to the fact that, when the gene is expressed throughout the plant, the protein encoded by the gene transports more Na^+ from the soil into the root, resulting in more Na^+ being transported to the shoot in the transpiration stream. Expression of this gene only in the cells surrounding the xylem would result in a plant being more efficient in retrieving Na^+ from the transpiration stream.

Plants consist of multiple tissues and multiple cells. Each tissue is adapted for a specific purpose – roots for nutrient uptake, leaves for photosynthesis, stems for support – and, therefore, will not necessarily express the same genes. Genes responsible for the maintenance of photosynthesis in the leaves will not be expressed in the roots, and genes for nutrient uptake from the soil will not be expressed in floral tissue. Similarly, not all genes in a plant are expressed all the time; many genes are activated only when required. When growing in non-stressed environments there is little point in a plant using critical energy supplies to generate and maintain proteins important for salinity tolerance. It is unsurprising, therefore, that the continuous expression throughout a plant of a gene important for salinity tolerance, such as *AtHKT1;1*, often results in detrimental effects [77]. A critical feature in the generation of crops engineered to have increased salinity tolerance is the spatial and temporal control of the transgene which has been introduced.

Recently, transgenic Arabidopsis plants have been produced with cell-specific expression of the *AtHKT1;1* gene in the root cells surrounding the xylem [148]. Unlike plants with constitutive overexpression of *AtHKT1;1*, these cell-specific plant lines showed a significant reduction in shoot Na^+ and increased salt tolerance [148].

In a different approach, rice plants designed to overexpress a gene involved in the synthesis of trehalose only when the plants experienced stress exhibited reduced shoot Na^+ concentrations and better growth in saline conditions than non-transformed plants [149]. Trehalose is a sugar involved in protecting cells from long periods of desiccation and possibly aids salinity tolerance through an ability to scavenge reactive oxygen species, thereby protecting cellular proteins [39]; however, plants with constitutive overexpression of genes for trehalose synthesis display severe stunting [150]. The use of a stress-inducible promoter is, therefore, an important control to minimize growth inhibition of transgenic plants when grown in non-stressed environments. The focus now is the identification of gene promoters (sequences of DNA which are used to activate genes) which allow the cell- and temporal-specific expressions of genes in crops.

In addition to the fine control of genes transformed into transgenic crops, there is also the need to identify gene combinations which may have the potential to increase crop salt tolerance. As has been observed, plants employ multiple salinity tolerance mechanisms to survive saline soils, all of which rely on a variety of different genes and proteins. It seems unlikely, therefore, that the generation of a successful commercial salt tolerant crop will be achieved by the constitutive overexpression of one single gene. Recent research promoting salt tolerance in plants focuses on either boosting the intracellular salt-sequestering processes, or on the Na^+ exclusion mechanisms by transferring into selected crop species genes for salinity tolerance from model organisms (such as *Arabidopsis*) or from salt tolerant plants. A complementary approach focuses on the challenging task of reducing net input of salt into plants by perturbing the function of channels and transporters involved in sodium uptake but without disturbing potassium uptake. An ideal scenario contemplates the generation of transgenic plants with an enhanced capability for vacuolar salt sequestration combined with a reduced uptake of salt. While a number of genes involved in these processes have now been identified, the challenge is to switch these genes on at the appropriate time and in the appropriate tissues where they can be most effective. In order to achieve this aim, improved knowledge is required of gene promoters that are stress-inducible and cell specific.

While it is clear that there are potentially many avenues for the generation of a genetically modified salt tolerant plant, there remain significant challenges.

Although it is now possible to generate salt tolerant plants in a laboratory, it has yet to be shown whether this relates to actual yield improvements in the field. There are cases where using genetic modification to generate a salt tolerant plant has a negative effect on yield when no stress is present. It is clear, therefore, that more information on gene promoters is required to enable the activation of salt tolerant genes in specific tissues/cell types only when plants are grown in salt. Furthermore, there are areas of the world, such as Europe, where there remains considerable resistance to the acceptance of genetically modified plants. This may well be due to the lack of availability to consumers of clear, accurate information as well as the prevalence of extremist views. A more open, transparent approach by scientists is required explaining the potential advantages and disadvantages of this technology. Only then will consumers be able to make their own informed choices about genetically modified organisms.

Future Directions

Crops growing on saline soils suffer severe reductions in yield due to both ionic and osmotic stresses. As considerable areas of farmland are currently affected by saline soils much research has been undertaken to enhance crop salinity tolerance by exploitation of natural variation in salinity tolerance or through the generation of transgenic plants expressing genes shown to be important for salt tolerance. While salinity tolerant plants have been generated by both approaches, the focus should now be on the production of viable crop plants for farmers to grow in affected areas. For this to occur, the new cultivars of tolerant plants need to be tested under rigorous field conditions and those with enhanced salt tolerance and, as equally important, no yield penalties when grown in nonsaline conditions pass to breeders for incorporation into future crops.

Approaches are still required to help speed up the generation of salinity tolerance crops through the exploitation of natural variation. Sequencing of cereal genomes will greatly speed up the identification of candidate genes underlying salt tolerance QTL, thereby enabling highly specific molecular markers that are tightly aligned to the trait. Using these markers will help reduce the effects of linkage drag bringing undesirable traits into the population.

For transgenic plants, it is clear that refined control over when and where a gene is expressed is essential. As there are now multiple candidate genes, with potential for enhancing salinity tolerance, research should now focus more on identifying the controlling elements in a plant's genome, which dictate when and where a gene is expressed, and less on the identification of candidate genes. In addition, combinations of genes which have additive effects on salinity tolerance need to be identified, thereby allowing the production of the most optimal salt tolerant plants. When these factors are known, crops can be produced which have the ability to activate multiple genes for salinity tolerance in different areas of the plant but only when saline soils are experienced.

Although further research is clearly still required, considerable progress has been made in generating salt tolerant *plants* through the exploitation of natural variations and the generation of genetically modified organisms. The next step is to deliver salt tolerant *crops* to farmers.

Acknowledgments

We thank Christina Morris for her comments on the manuscript, and the Australian Research Council and Grains Research Development Corporation for financial support.

Bibliography

Primary Literature

1. Munns R, Tester M (2008) Mechanisms of salinity tolerance. Annu Rev Plant Biol 59:651–681
2. FAO. FAO Land and Plant Nutrition Management Service (2008) Available from http://www.fao.org/ag/agl/agll/spush
3. Szabolcs I (1989) Salt-affected soils. CRC Press, Boca Raton
4. Munns R (2002) Comparative physiology of salt and water stress. Plant Cell Environ 25:239–250
5. Flowers TJ, Hajibagherp MA, Yeo AR (1991) Ion accumulation in the cell walls of rice plants growing under saline conditions: evidence for the Oertli hypothesis. Plant Cell Environ 14: 319–325
6. Munns R, James RA, Lauchli A (2006) Approaches to increasing the salt tolerance of wheat and other cereals. J Exp Bot 57:1025–1043

7. Yeo AR et al (1991) Short- and long-term effects of salinity on leaf growth in rice (*Oryza sativa L.*). J Exp Bot 42:881–889
8. Frensch J, Hsiao TC (1994) Transient responses of cell turgor and growth of maize roots as affected by changes in water potential. Plant Physiol 104:247–254
9. Hu Y et al (2007) Short-term effects of drought and salinity on mineral nutrient distribution along growing leaves of maize seedlings. Environ Exp Bot 60:268–275
10. Hu Y, Fricke W, Schmidhalter U (2005) Salinity and the growth of non-halophytic grass leaves: the role of mineral nutrient distribution. Funct Plant Biol 32:973–985
11. Fricke W (2004) Rapid and tissue-specific accumulation of solutes in the growth zone of barley leaves in response to salinity. Planta 219:515–525
12. Termaat A, Munns R (1986) Use of concentrated macronutrient solutions to separate osmotic from NaCl specific effects on plant growth. Aust J Plant Physiol 13:509–522
13. Munns R et al (2000) Leaf water status controls day-time but not daily rates of leaf expansion in salt-treated barley. Aust J Plant Physiol 27:949–957
14. Fricke W, Peters WS (2002) The biophysics of leaf growth in salt-stressed barley. A study at the cell level. Plant Physiol 129:374–388
15. Fricke W et al (2004) Rapid and tissue-specific changes in ABA and in growth rate in response to salinity in barley leaves. J Exp Bot 55:1115–1123
16. Apel K, Hirt H (2004) Reactive oxygen species: metabolism, oxidative stress and signal transduction. Annu Rev Plant Biol 55:373–399
17. Logan BA (2005) Reactive oxygen species and photosynthesis. In: Smirnoff N (ed) Antioxidants and reactive oxygen species in plants. Blackwell, Oxford, pp 250–267
18. Tester M, Davenport R (2003) Na^+ tolerance and Na^+ transport in higher plants. Ann Bot 91:503–527
19. Storey R, Walker RR (1999) Citrus and salinity. Sci Hortic 78:39–81
20. Flowers TJ, Yeo AR (1988) Ion relations of salt tolerance. In: Baker D, Halls J (eds) Solute transport in plant cells and tissues. Longman, Harlow, pp 392–416
21. Läuchli A (1984) Salt exclusion: an adaptation of legumes for crops and pastures under saline conditions. In: Staples RC (ed) Salinity tolerance in plants: strategies for crop improvement. Wiley, New York, pp 171–187
22. Teakle NL, Tyerman SD (2010) Mechanisms of Cl^- transport contributing to salt tolerance. Plant Cell Environ 33:566–589
23. Horie T, Hauser F, Schroeder JI (2009) HKT transporter-mediated salinity resistance mechanisms in Arabidopsis and monocot crop plants. Trends Plant Sci 14:660–668
24. Bhandal IS et al (1988) Potassium estimation, uptake, and its role in the physiology and metabolism of flowering plants. Int Rev Cytol 110:205–254
25. Wyn Jones RG, Brady CJ, Spears J (1979) Ionic and osmotic relations in plant cells. In: Laidman DL, Wyn Jones RG (eds) Recent advances in the biochemistry of cereals. Academic, London, pp 63–103
26. Blaha G et al (2000) Preparation of functional ribosomal complexes and effect of buffer conditions on tRNA positions observed by cryoelectron microscopy. Methods Enzymol 317:292–306
27. Munns R (1993) Physiological processes limiting plant growth in saline soils: some dogmas and hypotheses. Plant Cell Environ 16:15–24
28. Öertli JJ (1968) Extracellular salt accumulation, a possible mechanism of salt injury in plants. Agrochimica 12: 461–469
29. Gorham J (1990) Salt tolerance in the *Triticeae*: K/Na discrimination in synthetic hexaploid wheats. J Exp Bot 41:623–627
30. Dubcovsky J et al (1996) Mapping of the K^+/Na^+ discrimination locus *Kna1* in wheat. Theor Appl Genet 92:448–454
31. Maathuis FJM, Amtmann A (1999) K^+ nutrition and Na^+ toxicity: the basis of cellular K^+/Na^+ ratios. Ann Bot 84:123–133
32. Hu YC, Schnyder H, Schmidhalter U (2000) Carbohydrate deposition and partitioning in elongating leaves of wheat under saline soil conditions. Aust J Plant Physiol 27:363–370
33. Chen THH, Murata N (2002) Enhancement of tolerance of abiotic stress by metabolic engineering of betaines and other compatible solutes. Curr Opin Plant Biol 5:250–257
34. Aslam Z et al (1986) Effects of external NaCl on the growth of *Atriplex amnicola* and the ion relations and carbohydrate status of the leaves. Plant Cell Environ 9:571–580
35. Glenn EP, Brown JJ, Blumwald E (1999) Salt tolerance and crop potential of halophytes. Crit Rev Plant Sci 18:227–255
36. Aslam M, Qureshi RH, Ahmed N (1993) A rapid screening technique for salt tolerance in rice (*Oryza sativa* L.). Plant Soil 150:99–107
37. Walia H et al (2005) Comparative transcriptional profiling of two contrasting rice genotypes under salinity stress during the vegetative growth stage. Plant Physiol 139:822–835
38. Flowers TJ, Yeo AR (1981) Variability in the resistance of sodium chloride salinity within rice (*Oryza sativa* L.) varieties. New Phytol 88:363–373
39. Flowers TJ (2004) Improving crop salt tolerance. J Exp Bot 55:307–319
40. Garthwaite AJ, von Bothmer R, Colmer TD (2005) Salt tolerance in wild *Hordeum* species is associated with restricted entry of Na^+ and Cl^- into the shoots. J Exp Bot 56:2365–2378
41. Munns R, James RA (2003) Screening methods for salinity tolerance: a case study with tetraploid wheat. Plant Soil 253:201–218
42. Flowers TJ et al (2010) Salt sensitivity in chickpea. Plant Cell Environ 33:490–509
43. Greenway H, Munns R (1980) Mechanisms of salt tolerance in nonhalophytes. Ann Rev Plant Phys 31:149–190
44. Sengupta S, Majumder AL (2010) *Porteresia coarctata* (Roxb.) Tateoka, a wild rice: a potential model for studying salt-stress biology in rice. Plant Cell Environ 33:526–542
45. James RA et al (2008) Genetic variation in tolerance to the osmotic stress component of salinity stress in durum wheat. Funct Plant Biol 35:111–123

46. James RA et al (2002) Factors affecting CO_2 assimilation, leaf injury and growth in salt-stressed durum wheat. Funct Plant Biol 29:1393–1403
47. Silva C, Martínez V, Carvajal M (2008) Osmotic versus toxic effects of NaCl on pepper plants. Biol Plantarum 52:72–79
48. Rajendran K, Tester M, Roy SJ (2009) Quantifying the three main components of salinity tolerance in cereals. Plant Cell Environ 32:237–249
49. Sirault XRR, James RA, Furbank RT (2009) A new screening method for osmotic component of salinity tolerance in cereals using infrared thermography. Funct Plant Biol 36:970–977
50. Poustini K, Siosemardeh A (2004) Ion distribution in wheat cultivars in response to salinity stress. Field Crop Res 85: 125–133
51. Lee K-S et al (2003) Salinity tolerance of japonica and indica rice (*Oryza sativa* L.) at the seedling stage. Planta 216: 1043–1046
52. Zhu GY, Kinet JM, Lutts S (2001) Characterization of rice (*Oryza sativa* L.) F3 populations selected for salt resistance. I. Physiological behaviour during vegetative growth. Euphytica 121:251–263
53. Forster B (2001) Mutation genetics of salt tolerance in barley: an assessment of Golden Promise and other semi-dwarf mutants. Euphytica 120:317–328
54. Wei W et al (2003) Salinity induced differences in growth, ion distribution and partitioning in barley between the cultivar Maythorpe and its derived mutant Golden Promise. Plant Soil 250:183–191
55. Teakle N et al (2007) *Lotus tenuis* tolerates the interactive effects of salinity and waterlogging by 'excluding' Na^+ and Cl^- from the xylem. J Exp Bot 58:2169–2180
56. Sibole JV et al (2003) Ion allocation in two different salt-tolerant Mediterranean Medicago species. J Plant Physiol 160:1361–1365
57. Kingsbury RW, Epstein E (1984) Selection for salt-resistant spring wheat. Crop Sci 24:310–315
58. Jafari-Shabestari J, Corke H, Qualset C (1995) Field evaluation of tolerance to salinity stress in Iranian hexaploid wheat landrace accessions. Genet Resour Crop Evol 42:147–156
59. Munns R et al (2000) Genetic variation for improving the salt tolerance of durum wheat. Aust J Agric Res 51:69–74
60. Richards RA et al (1987) Variation in yield of grain and biomass in wheat, barley, and triticale in a salt-affected field. Field Crop Res 15:277–287
61. Slavich P, Read B, Cullis B (1990) Yield response of barley germplasm to field variation in salinity quantified using the EM-38. Aust J Exp Agr 30:551–556
62. Lee J-D et al (2009) Inheritance of Salt Tolerance in Wild Soybean (*Glycine soja* Sieb. and Zucc.) Accession PI483463. J Hered 100:798–801
63. Tozlu I, Guy CL, Moore G (1999) QTL analysis of Na^+ and Cl^- accumulation related traits in an intergeneric BC_1 progeny of *Citrus* and *Poncirus* under saline and nonsaline environments. Genome Biol 42:692–705
64. Vadez V et al (2007) Large variation in salinity tolerance in chickpea is explained by differences in sensitivity at the reproductive stage. Field Crop Res 104:123–129
65. Gregorio GB et al (2002) Progress in breeding for salinity tolerance and associated abiotic stresses in rice. Field Crop Res 76:91–101
66. Sahi C et al (2006) Salt stress response in rice: genetics, molecular biology, and comparative genomics. Funct Integr Genomics 6:263–284
67. Mano Y, Takeda K (1997) Mapping quantitative trait loci for salt tolerance at germination and the seedling stage in barley (*Hordeum vulgare* L.). Euphytica 94:263–272
68. Shavrukov Y et al (2010) *HvNax3* – a locus controlling shoot sodium exclusion derived from wild barley (*Hordeum vulgare* ssp. *spontaneum*). Funct Integr Genomics 10:277–291
69. Bretó MP, Aśins MJ, Carbonell EA (1994) Salt tolerance in *Lycopersicon* species. III. Detection of quantitative trait loci by means of molecular markers. Theor Appl Genet 88:395–401
70. Ren Z-H et al (2005) A rice quantitative trait locus for salt tolerance encodes a sodium transporter. Nat Genet 37: 1141–1146
71. Gorham J et al (1987) Chromosomal location of a K/Na discrimination character in the D genome of wheat. Theor Appl Genet 74:584–588
72. Huang S et al (2006) A Sodium Transporter (HKT7) Is a Candidate for Nax1, a Gene for Salt Tolerance in Durum Wheat. Plant Physiol 142:1718–1727
73. Byrt CS et al (2007) HKT1;5-like cation transporters linked to Na^+ exclusion loci in wheat, *Nax2* and *Kna1*. Plant Physiol 143:1918–1928
74. Platten JD et al (2006) Nomenclature for HKT transporters, key determinants of plant salinity tolerance. Trends Plant Sci 11:372–374
75. Davenport RJ et al (2007) The Na^+ transporter AtHKT1; 1 controls retrieval of Na^+ from the xylem in Arabidopsis. Plant Cell Environ 30:497–507
76. Huang S et al (2008) Comparative mapping of HKT genes in wheat, barley, and rice, key determinants of Na^+ transport, and salt tolerance. J Exp Bot 59:927–937
77. Rus A et al (2004) AtHKT1 facilitates Na^+ homeostasis and K^+ nutrition in Planta. Plant Physiol 136:2500–2511
78. Rus A et al (2001) AtHKT1 is a salt tolerance determinant that controls Na^+ entry into plant roots. Proc Natl Acad Sci USA 98:14150–14155
79. Sunarpi H et al (2005) Enhanced salt tolerance mediated by AtHKT1 transporter-induced Na^+ unloading from xylem vessels to xylem parenchyma cells. Plant J 44:928–938
80. Colmer TD, Flowers TJ, Munns R (2006) Use of wild relatives to improve salt tolerance in wheat. J Exp Bot 57:1059–1078
81. Dvořák J, Gorham J (1992) Methodology of gene transfer by homoeologous recombination into *Triticum turgidum*: transfer of K^+/Na^+ discrimination from *Triticum aestivum*. Genome Biol 35:639–646
82. Dvořak J et al (1994) Enhancement of the salt tolerance of *Triticum turgidum* L. by the Kna1 locus transferred from the

Triticum aestivum L. chromosome 4D by homoeologous recombination. Theor Appl Genet 87:872–877
83. Gorham J et al (1997) Genetic analysis and physiology of a trait for enhanced K^+/Na^+ discrimination in wheat. New Phytol 137:109–116
84. Perez-Alfocea F et al (1994) Comparative salt responses at cell and whole-plant levels of cultivated and wild tomato species and their hybrid. J Hortic Sci Biotech 69:639–644
85. Sherraf I et al (1994) Production and characterization of intergeneric somatic hybrids through protoplast electrofusion between potato (*Solanum tuberosum*) and *Lycopersicon pennellii*. Plant Cell Tissue Organ Cult 37:137–144
86. Nevo E, Chen G (2010) Drought and salt tolerances in wild relatives for wheat and barley improvement. Plant Cell Environ 33:670–685
87. Gorham J et al (1986) Salt Tolerance in the Triticeae: Solute Accumulation and Distribution in an Amphidiploid derived from *Triticum aestivum* cv. Chinese Spring and Thinopyrum bessarabicum. J Exp Bot 37:1435–1449
88. King IP et al (1997) Introgression of salt-tolerance genes from *Thinopyrum bessarabicum* into wheat. New Phytol 137:75–81
89. Yan J et al (2008) Phenotypic variation in caryopsis dormancy and seedling salt tolerance in wild barley, Hordeum spontaneum, from different habitats in Israel. Genet Resour Crop Evol 55:995–1005
90. Gorham J et al (1991) The presence of the enhanced K/Na discrimination trait in diploid Triticum species. Theor Appl Genet 82:729–736
91. James RA, Davenport RJ, Munns R (2006) Physiological characterization of two genes for Na^+ exclusion in durum wheat, *Nax1* and *Nax2*. Plant Physiol 142:1537–1547
92. Shah SH et al (1987) Salt tolerance in the Triticeae: the contribution of the D genome to cation selectivity in hexaploid wheat. J Exp Bot 38:254–269
93. Schachtman DP, Lagudah ES, Munns R (1992) The expression of salt tolerance from *Triticum tauschii* in hexaploid wheat. Theor Appl Genet 84:714–719
94. CSIRO. CSIRO develops highest yielding salt tolerant wheat (2010) Available from http://www.csiro.au/news/CSIRO-develops-highest-yielding-salt-tolerant-wheat.html
95. Blumwald E, Aharon GS, Apse MP (2000) Sodium transport in plant cells. Biochim Biophys Acta 1465:140–151
96. Plett DC, Skrumsager Møller I (2010) Na^+ transport in glycophytic plants: what we know and would like to know. Plant Cell Environ 33:612–626
97. Garbarino J, DuPont FM (1989) Rapid induction of Na^+/H^+ exchange activity in barley root tonoplast. Plant Physiol Biochem 89:1–4
98. Zörb C et al (2005) Molecular characterization of Na^+/H^+ antiporters (*ZmNHX*) of maize (*Zea mays* L.) and their expression under salt stress. J Plant Physiol 162:55–66
99. Ballesteros E et al (1997) $Na^+//H^+$ antiport activity in tonoplast vesicles isolated from sunflower roots induced by NaCl stress. Physiol Plant 99:328–334
100. Wilson C, Shannon MC (1995) Salt-induced Na^+/H^+ antiport in root plasma membrane of a glycophytic and halophytic species of tomato. Plant Sci 107:147–157
101. Wu C-A et al (2004) The cotton *GhNHX1* gene encoding a novel putative tonoplast Na^+/H^+ antiporter plays an important role in salt stress. Plant Cell Physiol 45:600–607
102. Jha D et al (2010) Variation in salinity tolerance and shoot sodium accumulation in *Arabidopsis* ecotypes linked to differences in the natural expression levels of transporters involved in sodium transport. Plant Cell Environ 33:793–804
103. Aharon GS et al (2003) Characterization of a family of vacuolar Na^+/H^+ antiporters in *Arabidopsis thaliana*. Plant Soil 253:245–256
104. Gaxiola RA et al (1999) The *Arabidopsis thaliana* proton transporters, AtNHX1 and AVP1, can function in cation detoxification in yeast. Proc Natl Acad Sci USA 96:1480–1485
105. Apse MP et al (1999) Salt tolerance conferred by overexpression of a vacuolar Na^+/H^+ antiport in Arabidopsis. Science 285:1256–1258
106. Zhang H-X, Blumwald E (2001) Transgenic salt-tolerant tomato plants accumulate salt in foliage but not in fruit. Nat Biotechnol 19:765–768
107. Zhang H-X et al (2001) Engineering salt-tolerant *Brassica* plants: characterization of yield and seed oil quality in transgenic plants with increased vacuolar sodium accumulation. Proc Natl Acad Sci USA 98:12832–12836
108. He C et al (2005) Expression of an Arabidopsis vacuolar sodium/proton antiporter gene in cotton improves photosynthetic performance under salt conditions and increases fiber yield in the field. Plant Cell Physiol 46:1848–1854
109. Sottosanto J, Gelli A, Blumwald E (2004) DNA array analyses of Arabidopsis thaliana lacking a vacuolar Na+/H+ antiporter: impact of AtNHX1 on gene expression. Plant J 40:752–771
110. Sottosanto J, Saranga Y, Blumwald E (2007) Impact of *AtNHX1*, a vacuolar Na^+/H^+ antiporter, upon gene expression during short- and long-term salt stress in *Arabidopsis thaliana*. BMC Plant Biol 7:18
111. Brini F et al (2007) Overexpression of wheat Na^+/H^+ antiporter *TNHX1* and H^+-pyrophosphatase *TVP1* improve salt- and drought-stress tolerance in *Arabidopsis thaliana* plants. J Exp Bot 58:301–308
112. Yu J et al (2007) An Na+/H+ antiporter gene from wheat plays an important role in stress tolerance. J Biosci 32:1153–1161
113. Fukuda A et al (2004) Effect of salt and osmotic stresses on the expression of genes for the vacuolar H^+-pyrophosphatase, H^+-ATPase subunit A, and Na^+/H^+ antiporter from barley. J Exp Bot 55:585–594
114. Zahran HH et al (2007) Effect of salt stress on the expression of NHX-type ion transporters in Medicago intertexta and Melilotus indicus plants. Physiol Plant 131:122–130
115. Fukuda A et al (2004) Function, intracellular localization and the importance in salt tolerance of a vacuolar Na^+/H^+ antiporter from rice. Plant Cell Physiol 45:146–159

116. Chen H et al (2007) Over-expression of a vacuolar Na^+/H^+ antiporter gene improves salt tolerance in an upland rice. Mol Breed 19:215–225
117. Zhao F-Y et al (2006) Co-expression of the *Suaeda salsa SsNHX1* and Arabidopsis *AVP1* confer greater salt tolerance to transgenic rice than the single *SsNHX1*. Mol Breed 17: 341–353
118. Xue Z-Y et al (2004) Enhanced salt tolerance of transgenic wheat (*Tritivum aestivum L.*) expressing a vacuolar Na^+/H^+ antiporter gene with improved grain yields in saline soils in the field and a reduced level of leaf Na^+. Plant Sci 167: 849–859
119. Gaxiola RA et al (2001) Drought- and salt-tolerant plants result from overexpression of the *AVP1* H^+ -pump. Proc Natl Acad Sci USA 98:11444–11449
120. Duan X-G et al (2007) Heterologous expression of vacuolar H^+-PPase enhances the electrochemical gradient across the vacuolar membrane and improves tobacco cell salt tolerance. Protoplasma 232:87–95
121. Gao F et al (2006) Cloning of an H^+-PPase gene from *Thellungiella halophila* and its heterologous expression to improve tobacco salt tolerance. J Exp Bot 57:3259–3270
122. Bao A-K et al (2009) Overexpression of the Arabidopsis H+-PPase enhanced resistance to salt and drought stress in transgenic alfalfa (*Medicago sativa* L.). Plant Sci 176:232–240
123. Li ZG et al (2010) Heterologous expression of Arabidopsis H^+-pyrophosphatase enhances salt tolerance in transgenic creeping bentgrass (*Agrostis stolonifera L.*). Plant Cell Environ 33:272–289
124. Zhu J-K (2003) Regulation of ion homeostasis under salt stress. Curr Opin Plant Biol 6:441–445
125. Luan S (2009) The CBL-CIPK network in plant calcium signaling. Trends Plant Sci 14:37–42
126. Batistic O, Kudla J (2004) Integration and channeling of calcium signaling through the CBL calcium sensor/CIPK protein kinase network. Planta 219:915–924
127. Weinl S, Kudla J (2009) The CBL-CIPK Ca^{2+}-decoding signaling network: function and perspectives. New Phytol 184:517–528
128. Wu SJ, Ding L, Zhu JK (1996) SOS1, a genetic locus essential for salt tolerance and potassium acquisition. Plant Cell 8: 617–627
129. Qiu Q-S et al (2002) Regulation of SOS1, a plasma membrane Na^+/H^+ exchanger in *Arabidopsis thaliana*, by SOS2 and SOS3. Proc Natl Acad Sci USA 99:8436–8441
130. Qiu Q-S et al (2003) Na^+/H^+ exchange activity in the plasma membrane of Arabidopsis. Plant Physiol 132:1041–1052
131. Vera-Estrella R et al (2005) Salt stress in *Thellungiella halophila* activates Na^+ transport mechanisms required for salinity tolerance. Plant Physiol 139:1507–1517
132. Wu Y et al (2007) Molecular characterization of PeSOS1: the putative Na+/H+ antiporter of *Populus euphratica*. Plant Mol Biol 65:1–11
133. Kolukisaoglu U et al (2004) Calcium sensors and their interacting protein kinases: genomics of the arabidopsis and rice CBL-CIPK signaling networks. Plant Physiol 134: 43–58
134. Shi H et al (2003) Overexpression of a plasma membrane Na^+/H^+ antiporter gene improves salt tolerance in *Arabidopsis thaliana*. Nat Biotechnol 21:81–85
135. Winicov I (2000) Alfin1 transcription factor overexpression enhances plant root growth under normal and saline conditions and improves salt tolerance in alfalfa. Planta 210: 416–422
136. Cheong YH et al (2003) CBL1, a calcium sensor that differentially regulates salt, drought, and cold responses in Arabidopsis. Plant Cell 15:1833–1845
137. Tripathi V et al (2009) CIPK6, a CBL-interacting protein kinase is required for development and salt tolerance in plants. Plant J 58:778–790
138. Xiang Y, Huang YM, Xiong LZ (2007) Characterization of stress-responsive CIPK genes in rice for stress tolerance improvement. Plant Physiol 144:1416–1428
139. Zhao J et al (2009) Cloning and characterization of a novel CBL-interacting protein kinase from maize. Plant Mol Biol 69:661–674
140. Gisbert C et al (2000) The yeast *HAL1* gene improves salt tolerance of transgenic tomato. Plant Physiol 123:393–402
141. Garciadeblás B et al (2003) Sodium transport and HKT transporters: the rice model. Plant J 34:788–801
142. Golldack D et al (2002) Characterization of a HKT-type transporter in rice as a general alkali cation transporter. Plant J 31:529–542
143. Horie T et al (2001) Two types of HKT transporters with different properties of Na^+ and K^+ transport in *Oryza sativa*. Plant J 27:129–138
144. Wang T-B et al (1998) Rapid up-regulation of *HKT1*, a high-affinity potassium transporter gene, in roots of barley and wheat following withdrawal of potassium. Plant Physiol 118:651–659
145. Horie T et al (2006) Calcium regulation of sodium hypersensitivities of sos3 and athkt1 mutants. Plant Cell Physiol 47:622–633
146. Maser P et al (2002) Altered shoot/root Na^+ distribution and bifurcating salt sensitivity in Arabidopsis by genetic disruption of the Na^+ transporter *AtHKT1*. FEBS Lett 531:157–161
147. Rus A et al (2006) Natural variants of AtHKT1 enhance Na^+ accumulation in two wild populations of Arabidopsis. PLoS Genet 2:1964–1973
148. Møller I et al (2009) Salinity tolerance engineered by cell type-specific over-expression of a Na^+ transporter in the Arabidopsis root. Plant Cell 21:2163–2178
149. Garg AK et al (2002) Trehalose accumulation in rice plants confers high tolerance levels to different abiotic stresses. Proc Natl Acad Sci USA 99:15898–15903
150. Romero C et al (1997) Expression of the yeast *trehalose-6-phosphate synthase* gene in transgenic tobacco plants: pleiotropic phenotypes include drought tolerance. Planta 201: 293–297

Indoor Environmental Quality and Health Improvement, Evidence-Based Design for

Charlene W. Bayer
Georgia Tech Research Institute, Georgia Institute of Technology, Atlanta, GA, USA

Article Outline

Glossary

AIA (The American Institute of Architects) The AIA has been the leading professional membership association for licensed architects, emerging professionals, and allied partners since 1857.

Cfm (cubic feet per minute) A non-SI (non-International System) unit of measurement of the flow of a gas or liquid that indicates how much volume in cubic feet passes by a stationary point in one minute. The ASHRAE standards and guidelines give ventilation rates for the IEQ in a specified number of cfm/person. 1 cfm = 0.472 L/s.

EBD (evidence-based design) The process of basing decisions about the built environment on credible research to achieve the best possible outcomes.

Health A state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity.

HVAC (heating, ventilation, and air-conditioning system) The systems used to provide heating, cooling, and ventilation in buildings.

IAQ (indoor air quality) The air quality within buildings, related to conditions around buildings and structures, and its relationship to the health and comfort of building occupants.

IEQ (indoor environmental quality) Beyond IAQ to encompass all aspects of the indoor setting including air quality, thermal, visual, and acoustic quality. Focuses on the strategies and systems that result in a healthy indoor environment for building occupants.

WHO (World Health Organization) A United Nations agency that coordinates international health activities and aids governments in improving health services.

Definition of Evidence-Based Design

Evidence-based design (EBD), as defined by the Center for Health Design [1], is "the process of basing decisions about the built environment on credible research to achieve the best possible outcomes." EBD is an approach to facilities design that treats the building and its occupants as a system and gives importance to design features that impact health, well-being, mood and stress, safety, operational efficiency, and economics. To date, EBD has been applied primarily to healthcare facility design, where it has been shown to frequently reduce costs, improve staff productivity, and decrease the length of patient hospital stays. The evidence-based designer, in collaboration with the informed client, develops appropriate solutions to the individual design project based on the needs and expectations of the client, research on similar projects, and experience [2]. EBD provides data on successful strategies for the design process for healthy, high quality buildings.

Introduction

Concepts

Healthy, high-performance buildings should have positive outcomes in terms of energy, sustainability, health, and productivity. A healthy building should meet the World Health Organization (WHO) [3] definition of health, "a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity". The use of this definition of health is particularly applicable to green buildings, intent on not only reducing exposures to chemicals, but also promoting exercise, lowering stress, increasing social interactions, and otherwise fostering physical, social, and mental

health for the occupants. EBD not only meets the WHO health definition, but also encompasses productivity, operational efficiency, economic performance, and occupant/customer satisfaction. Effective EBD needs to be combined with sustainable design, incorporating all practices that reduce the negative impact of development on ecological health and indoor environmental quality [4].

Sustainable, creative design features for application of EBD fall into four major categories, which impact health, economic performance, and operational efficiency of the building system:

- Innovative building enclosures that incorporate load balancing, natural ventilation, and daylighting
- Advanced HVAC systems that incorporate natural conditioning, increased environmental contact, and local control
- Innovative data/voice/power "connectivity" and individual control
- New interior system designs in workstations and workgroup designs for improvements in spatial, thermal, acoustic, visual, and IAQ parameters [5]

Innovative enclosures and advanced HVAC systems particularly impact IAQ, health, productivity and learning, stress reduction, and operational economics. Innovative connectivity and new interior system designs chiefly impact health both as physical well-being and social well-being via connectivity to the organization as a whole, stress reduction, and health.

Indoor Environmental Quality

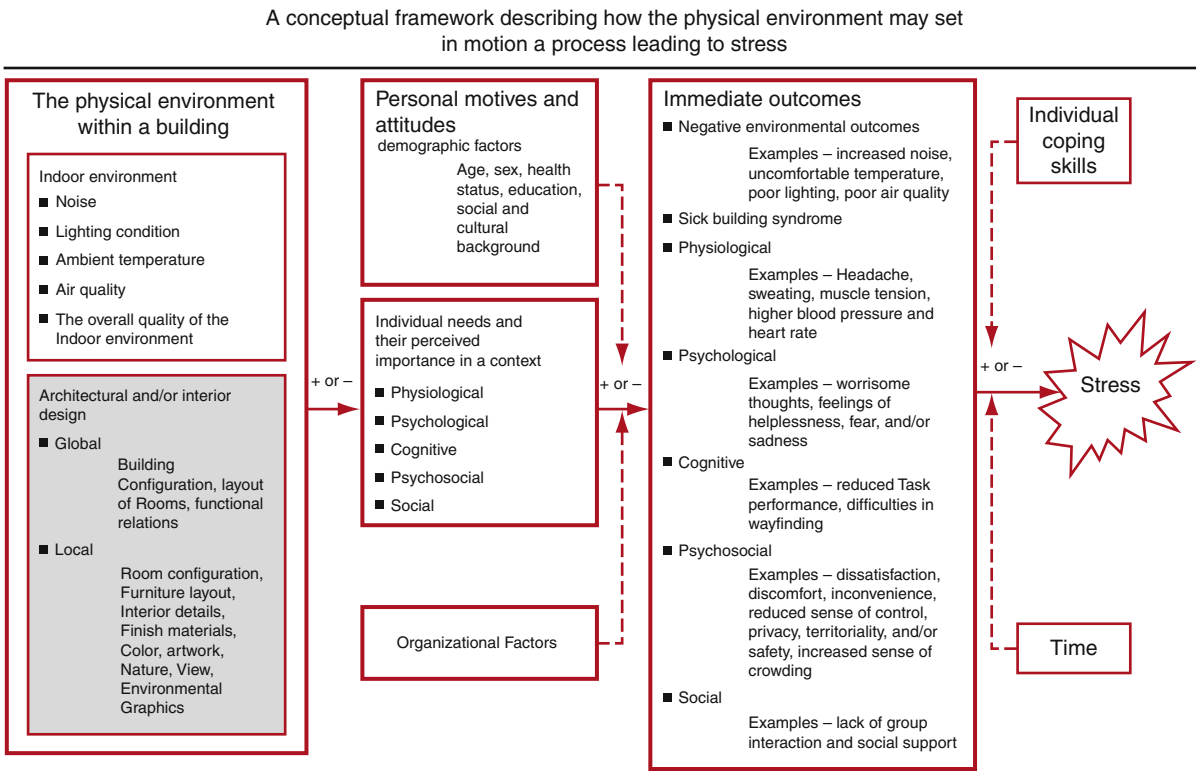
Healthy buildings encompass all aspects of indoor environmental quality (IEQ) including optimum thermal comfort, lighting with effective daylighting and access to views, IAQ, acoustical performance, ventilation effectiveness integrated with natural ventilation when applicable, and human comfort and health. Healthy buildings are designed for ease of operation and maintenance, because buildings with inadequate IEQ adversely impact occupants' overall health and productivity.

Rashid and Zimring [6] suggest that poor indoor environments may initiate a process leading to stress whenever the individual or workplace IEQ does not meet an occupant's needs, as is shown in Fig. 1. Their framework groups physical environmental variables into two primary groups: (1) IEQ variables including noise, lighting, ambient temperature, and IAQ, and (2) interior design variables including use of space, furniture, fixtures and equipment, finishing materials, color, artwork, natural views, and environmental graphics. These variables are interlinked in the design of the indoor environment and its conditioning systems. Factors leading to stress, similar to individual responses to odors, vary among individuals, further complicating the issues [7]. The collaboration between the designer and the user in the EBD design process is critical in reducing stressors in the indoor environment.

Examples of potential environmentally induced stressors that need to be assessed in the EBD process are:

1. Open office plans creating feelings of lack of privacy [8]
2. Open office plans, selection of hard-surfaced flooring and furnishing materials, office equipment location, HVAC system vibration, and/or or outdoor traffic that may increase noise levels resulting in difficulties in concentration, speech intelligibility, headaches, and other physical and emotional stress responses that impact learning and productivity [9–11]
3. Cafeteria, cleaning, furnishings, or systems odors permeating throughout the work areas of a building due to improper ventilation system design or poor materials selection [12, 13]
4. Daylight glare on work surfaces due to lack of effective window glazing or absence of blinds, and unshielded electric lighting that may result in headaches or eyestrain and poor productivity [14–16]

The Academy of Neuroscience for Architecture (www.anfarch.org) is using evidence-based design as a means to assess the linkage between neuroscience research and human responses to the built environment; thus seeking to relate behavioral changes to brain function changes based on the built environment. The Academy, in its studies, defines the dimensions of functional comfort as: (1) air quality, (2) thermal comfort, (3) spatial comfort, (4) collaborative or teamspace, (5) visual comfort, (6) workstation comfort, (7) lighting quality, (8) noise control, and (9) security. These nine parameters are used to direct the



Indoor Environmental Quality and Health Improvement, Evidence-Based Design for. Figure 1
Rashid and Zimring [6] conceptual framework describing how the physical environment may initiate a process leading to stress

evidence-based design practices to reduce human stress, poor behaviors and attitudes, and overall human health, as defined by WHO.

Indoor Air Quality The primary design strategies that are used to improve IAQ in green buildings are the use of low-emitting furnishings and building materials, designed to meet an iteratively tightening set of standards [17–20]. This strategy addresses one of the most important IAQ determinants that is clearly in the realm of the designer – source control. However, the construction process, including installation sequence and protection of materials prior to installation, is also an important factor to be addressed by the EBD team. Installation of carpet prior to painting of walls can result in long-term low level emissions of paint fumes due to adsorption by the carpet and slow reemission into the indoor environment. Key furnishing and material sources that must be specified as low-emitting and eco-friendly include office furniture, flooring, paints and coatings, adhesives and sealants, wall coverings, wood products, textiles, insulation, and cleaning products. The potential adverse health impacts of pollutants that may emit from these products has been determined though many emissions investigations [21].

Another strategy available for reducing exposures to airborne contaminants is source control of indoor equipment and activities. Office machines, stoves, and other appliances that are known to be active pollutant generators benefit from the use of local source control via the installation of dedicated exhaust fans. The use of local source control systems needs to be part of the design process and the location of the areas needing dedicated ventilation and exhausts need to be defined early in the design process. The use of well-maintained air cleaners is another strategy that may be appropriate to selected areas and types of facilities, such as in hospitals.

Ventilation systems are the primary method to dilute and transport airborne contaminants out of

the building. Natural and mixed-mode systems, if employed, must be designed to provide sufficient pollutant dilution and transport out of the building.

Ventilation System Design/Environmental Control The ventilation system is the primary means of transporting contaminants into, throughout, and out of the indoor environment. The placement and design of the system is critical to the quality of the indoor environment. Superior ventilation has been shown to improve learning, productivity, satisfaction, and health. At the same time the ventilation system can transport unwanted outdoor pollutants indoors, transfer indoor pollutants from one space to another, or transport infections [22].

In most buildings, the ventilation system is linked to the thermal conditioning (temperature control) system. Combined thermal comfort and ventilation systems may inadvertently compromise ventilation potentially adversely impacting IEQ, health, and occupant satisfaction. If a decision has to be made between thermal comfort and ventilation response, EBD reveals that the lack of temperature control is a primary stressor in the indoor environment, impacting productivity, learning, mood, and overall health [23].

On the other hand, lower temperatures, especially when combined with increased ventilation rates, tend to increase productivity and student performance. Wargocki and Wyon found that lowering the classroom temperature approximately 5°C improved elementary school students' performance on two numerical tasks and two language-based tasks [24, 25]. The children also reported lower incidence of headaches. When the classroom effective outdoor air supply rate was raised from 11 cfm/person (5 L/s) to 20 cfm/person (10 L/s), the students' performance was improved on four numerical tasks by improving the task performance speed. The children also reported feeling that the air felt fresher with the lower ambient temperatures. Similar results on the relationship of temperature and ventilation on productivity have been reported in adult work situations [26–29]. As a result, EBD reveals the importance in the design of the environmental control/ventilation system of separating the ventilation system from the thermal conditioning system and providing the ability for occupants to individually control the ambient temperature.

Numerous studies show health, productivity, and learning improvements with higher ventilation rates; however, this must also be balanced with sustainable design for greater energy efficiency through the use of innovative ventilation systems and maximizing ventilation efficiency. Haverinen-Shaughnessy et al. [30] found a linear association between classroom ventilation rates within the range of 0.9–7.1 L/s/person and students' academic achievement. In this study of fifth graders, it was determined that for every unit (1 L/s/person) increase in the ventilation rate, the proportion of students passing standardized tests increased by 2.9% for math and 2.7% for reading. Studies have shown that occupants in buildings or spaces with higher ventilation rates on average have fewer communicable respiratory illnesses, and lower asthma rates, and fewer absences from work or school [30–32]. The European Multidisciplinary Scientific Consensus Meeting (EUROVEN) [32] found that ventilation is strongly associated with perceived air quality and health (sick building syndrome symptoms, inflammation, infections, asthma, allergy, short-term sick leave) and that there is an association between ventilation and productivity in offices. The EUROVEN group also concluded that outdoor air supply rates below 25 L/s/person increased the risk of sick building syndrome symptoms, increases in short-term sick leave, and decreased productivity among occupants of an office building. Additionally improper maintenance, design, and functioning air-conditioning systems contribute to increased prevalence of sick building syndrome symptoms.

The research clearly demonstrates significant associations between ventilation system design that allows increased levels of ventilation, at least 10 L/s per person of outdoor air supply in buildings for optimized health, productivity/learning, and reduced stress. In order to meet sustainable design practices meeting the goal of energy efficiency and reduced operating costs, innovative ventilation strategies and systems must be used. Natural ventilation and hybrid systems are important innovative approaches, to be combined with next generation active systems.

Lighting/Daylighting/Access to Views Studies have shown that daylighting has a positive impact on humans, improving accuracy of work performance,

reducing stress and fatigue, and improving patient outcomes [32]. Loftness et al. [14] found that improved lighting quality design decisions are linked with 0.7–23% gains in individual productivity. The lighting quality design ranged from indirect–direct lighting systems, higher quality fixtures, and daylighting simulation. When daylight responsive dimming was employed energy savings of 27–87% were realized.

Access to the natural environment is associated with individual health and productivity. Design decisions for exposure to views include access to windows and view, daylighting through windows and skylights, natural and mixed-mode ventilation systems, and direct accessibility to landscaped indoor and outdoor spaces. Access to the natural environment has been shown to result in 3–18% increases in individual productivity [14] including access to operable windows.

Evidence from school lighting research indicates that improved school lighting can enhance both visual (healthy vision) and non-visual (achievement outcomes). Lighting conditions in classrooms have important non-visual effects on students including potentially raising test scores and faster responding on tests [33].

Acoustics/Noise Control Acoustics is an area of continued dissatisfaction in many green buildings [9]. In a number of projects, the open plan design, large areas of glass, hard-surface materials and furnishings, and natural ventilation strategies used in many green buildings have led to ongoing concerns with acoustic conditions. Building acoustical problems are generally classified in three categories: excessive noise, lack of speech privacy, and lack of speech clarity. Excessive noise is usually the result of high background noise emanating from outdoor noise sources that are transmitted through to the indoor environment, as well as noise from other rooms, building equipment, and/or noise from other occupants. Acoustical design strategies need to control noise levels at the source, reduce sound transmission pathways, and employ sound isolation techniques. Speech privacy is the extent to which speech is unintelligible to an *unintended* listener. The worst speech privacy situations are those where the background noise is very low. In open office plan environments, the lack of speech privacy may be a significant stressor. Design strategies to help improve speech privacy include possibly reconsideration of the open office plan, designing private areas adjacent to the open office area for use in private situations as needed, or increasing background noise. A lack of speech clarity occurs when the acoustics or a room design deteriorate the acoustical communication channel, rendering speech to the *intended* listener unintelligible, creating communication problems. This is particularly an issue in school classrooms and conference rooms. The problem may be caused by excessive background noise or excessive reverberation. EBD solutions to improve acoustics while maintaining sustainable design strategies include the use of acoustically absorbing materials, such as ceiling absorbers, acoustical ceiling tiles or wall-mounted panels.

Operation and Maintenance

A critical area that EBD needs to address for long-term building sustainability and occupant health is designing for maintainability. The life-cycle costing must include the maintenance and operating costs over the facilities lifetime, and EBD feedback on the long-term integrity and maintainability of the materials, components or systems. Metrics should be defined during the design process for the ability to maintain the facility in order to meet health and client economic performance needs. These metrics, at a minimum, should include:

- Labor hours per year that will be required to maintain each integral part of the facility, such as the HVAC system(s), the electrical system, lighting, windows, skylights, floors, and furnishings
- Frequency, extensiveness, and difficulty to perform required cleaning (including avoided toxicity)
- Cost of cleaning and replacement materials
- Equipment and furnishings life expectancies
- Training costs in labor hours and dollars for maintenance staff and occupants/building users

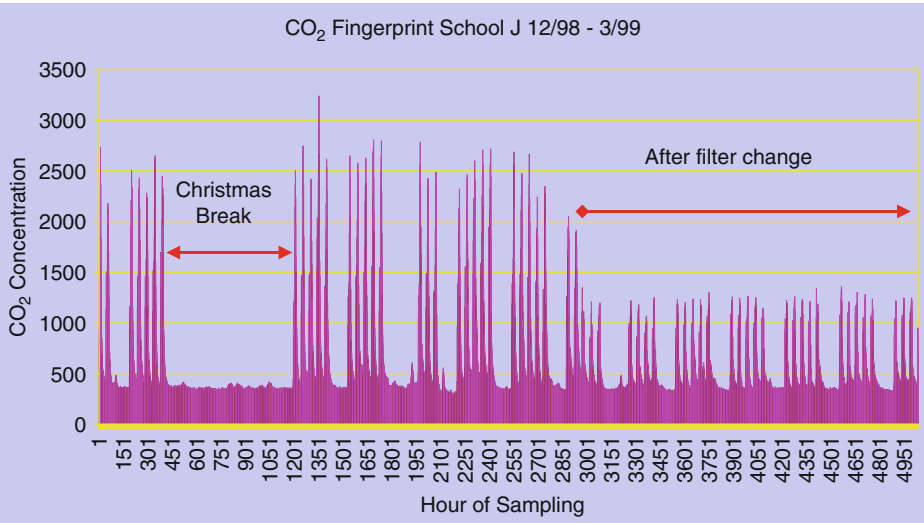
Magee [33] defined the specific maintenance objectives of the majority of facilities as follows:

- Perform necessary daily housekeeping and cleaning to maintain
- Promptly respond and repair minor discrepancies
- Develop and execute a system of regularly scheduled maintenance actions to prevent premature failure of the facility, its systems, and/or components

- Complete major repairs based on lowest life-cycle costs
- Identify and complete improvement projects to reduce and minimize total operating and maintenance costs without increasing indoor toxicity
- Operate the facility utilities in the most economical manner that achieves reliability and optimum functioning, while minimizing or eliminating indoor toxicity
- Provide for easy and complete reporting and identification of necessary repair and maintenance work
- Perform accurate cost estimating to ensure lowest cost and most effective solutions
- Maintain a proper level of materials and spare parts to minimize downtime
- Actively track all costs of maintenance work
- Schedule all planned work in advance allocating and anticipating staff requirements to meet planned and unplanned events
- Monitor progress of all maintenance work
- Maintain complete historical data concerning the facility in general and equipment and components in particular
- Continually seek workable engineering solutions to maintenance problems

Maintenance has a considerable impact on a building's performance and upon occupants' health and satisfaction. Maintenance-related problems over a building's lifetime can be minimized by making appropriate design decisions early in the process.

For example, maintainability is a critical measure for the performance for all ventilation systems including innovative high-performance ventilation systems and may have a significant impact on the health of the building occupants. In a study conducted by Bayer et al. [34] on the benefits of active humidity control and continuous ventilation at a minimum level of at least 15 cfm/person in schools using high-efficiency total energy heat recovery desiccant cooling ventilation system, the importance of system particulate filter maintenance was clearly demonstrated. As can be seen in Fig. 2, the carbon dioxide (CO_2) concentrations in the classroom exceeded 2,000 ppm during occupied times in the classrooms prior to replacement of the particulate filter in the system. Once the filter was changed, reducing the impedance to outside air delivery, the CO_2 levels dropped to approximately 800–1,000 ppm during occupied periods of the classroom. This result clearly demonstrates the necessity of system maintenance for effective ventilation even when a high-efficiency ventilation system is employed. In this school, filter replacement was inadequate due to difficulty in accessing the filter for replacement, a design and maintenance flaw.



Indoor Environmental Quality and Health Improvement, Evidence-Based Design for. Figure 2
CO_2 levels demonstrate the importance of particulate filter maintenance for effective ventilation in an occupied classroom

Maintaining the cleanliness of the ventilation filters has been found to impact productivity and learning in office buildings and schools. Wargocki et al. [35], in a study on the performance and subjective responses of call-center operators, found that replacing a used filter with a clean filter reduced operator talktime by about 10% at a outdoor air supply rate of approximately 34.4 L/s, but no effect was noted when the filter was replaced and the outdoor air supply rate was only 34.4 L/s. Additionally the operators reported a decrease in sick building syndrome symptoms with clean filters and the increased ventilation rates.

These investigations clearly demonstrate the importance of filter changeouts and ventilation system maintenance for IEQ, health, and productivity. The building systems need to be designed for easy performance of ventilation system maintenance tasks.

Arditi and Nawakorawit [36] surveyed 211 of the largest US building design firms to investigate the relationship between design practices and maintenance considerations. The study examined the extent to which maintenance issues are considered when designers specify building materials and service equipment; the level of designers' knowledge in maintenance-related issues; the degree to which design personnel are exposed to training in maintenance-related matters; the extent to which designers consult property managers and maintenance consultants; the relative importance of maintenance issues to other design factors; the level of difficulty in cleaning, inspecting, repairing, and replacing various building components; and the magnitude and frequency of maintenance-related complaints that designers receive from clients and tenants. Their findings indicate that maintenance consideration follow cost and aesthetics issues when designers specify building materials, but maintenance considerations constitute the number one issue when specifying service equipment. For most firms, the mechanical system was considered to be the most important consideration with regard to difficulty of cleaning, inspection, repair, and replacement with both the designers and the property managers. However ease of repair and replacement, access to cleaning area, and ease of cleaning were ranked by designers to be among the least important design factors for building systems and the facility. This in spite of the fact that the primary complaint that designers reported receiving from clients and tenants concerned issues of ease of repair, access to cleaning area, and ease of cleaning. Property managers also reported frequently receiving similar complaints. The design firms considered themselves to be knowledgeable in maintenance issues and design, and stated that they consulted property managers and maintenance consultants during the designing of selected projects, primarily in the schematic and preliminary design phases.

This is an area where EBD demands increased collaboration among all of the interested parties throughout the entire design process. EBD maintenance planning and design will enhance the life-long performance of the building.

Human Factor Impacts/Occupant/Customer Satisfaction on Sustainable Designs

Many sustainable design strategies reduce the use of walls and partitions – with more open space planning – to reduce material use, enhance views and daylight, and increase ventilation airflow, particularly when natural and hybrid ventilation strategies are used. Although this may increase satisfaction with daylight and access to views, it may also increase dissatisfaction with noise, privacy, and the ability to concentrate [37]. This situation was encountered in the LEED Platinum certified Philip Merrill Environmental Center in Annapolis, MD [38]. This facility placed the entire workforce into an open plan setting, regardless of status in the company, including the president and the key executives, without doors and low partitions for almost all employees (Fig. 3). This allows access to views and daylighting for all employees and the occupants' satisfaction ratings are very high. However, the primary complaints that remain are lack of privacy, noise, distractions, and interference with work concentration. At the same time, the occupants rated the views, daylighting, and interactive behaviors and communication highly.

Evidence-based design is an effective strategy for determining the potential effectiveness of open space planning in different types of buildings and task situations [39, 40]. For example, an elementary school in Atlanta, GA, organized in pods, uses four-foot high partitions among lower grade classrooms in each pod rather than floor-to-ceiling walls to increase interaction between grade classes. The partition heights

Indoor Environmental Quality and Health Improvement, Evidence-Based Design for. Figure 3
Open floor plan at Philip Merrill Environmental Center. Picture available at http://www.cbf.org/Page.aspx?pid=445

Indoor Environmental Quality and Health Improvement, Evidence-Based Design for. Figure 4
Open classroom style at Atlanta, GA, elementary school

increase as the grade level increases until in fifth grade (Fig. 4), the traditional classroom style is used. Staff interviews expressed mixed attitudes about this open design style. Noise between classrooms is a problem; however, as with the Philip Merrill Environmental Center, there was satisfaction with the feeling of community between the grade levels [41]. What has not been sufficiently studied at the school is the potential interference with student concentration in a school with an open floor plan such as is used in this school. The use of the lower partitions in the lower grade levels is actually the converse of what is needed for optimum acoustical performance for learning. Younger children in K-2 grades require a higher signal-to-noise ratio (clearer voices in a quieter environment) since they need to be able to carefully listen to develop the ability to discriminate among minor differences in words, which is extremely difficult in noisy environments [42].

Application to Healthcare Facilities

Hospitals are embracing evidence-based health care design for the promotion of therapeutic, supportive, and efficient environments. EBD is undertaken to develop appropriate solutions to design problems and unique situations in order to improve the organization's clinical outcomes, economic performance, efficiency, and customer satisfaction. EBD helps to provide solutions to the healthcare challenges of cost control, financial stability, avoidance of harm, quality improvements, sustainability, staff retention, and improved patient experience.

Ulrich et al. [43] reviewed the research literature on EBD healthcare design. Their overall findings indicated the importance of improving patient outcomes through a range of design characteristics including single-bed rooms, effective ventilation systems, good acoustical environments, increased views of nature, improved daylighting and interior lighting, better ergonomic design, acuity-adaptable rooms, and improved floor layouts and work settings. A number of significant results were found by optimization of environmental measures through the design process.

EBD can help eliminate hospital-acquired infections through better control of the three most significant vehicles for transmission: air, contact, and water. The most important design measures for infections controls are: (1) effective air quality control measures during construction and renovation using high-efficiency particulate air filters (HEPA) filtration and installation of barriers isolating construction areas (minimize airborne transmission); (2) installation and use of alcohol-based handrub dispensers at the bedside and other accessible locations (minimize contact transmission); (3) easy to clean floor, wall, and furniture coverings (minimize contact transmission);

(4) water system maintained at proper temperatures with adequate pressure to minimize stagnation and backflow (minimize waterborne transmission); and (5) single-bed rooms with private toilets for better patient isolation (minimize airborne and contact transmission).

Medical errors may be reduced through control of several environmental factors including noise, light, and acuity-adaptable single-patient rooms. Noise, both as unacceptable background and episodic interruptions, is responsible for loss of concentration, slower learning, and poor memorization. Additionally excessive noise adversely impacts patient recovery by increasing stress and interrupting sleep. Lighting levels impact task performance, which in a hospital may result in transcription errors [44]. Conversely, better lighting and daylighting design results in improved patient care and outcomes, staff satisfaction, safety, and decreased operational costs [45]. The acuity-adaptable rooms have adequate square footage in the room to accommodate several clinical activities without moving the patient, well-defined zones for patient care activities, strategic placement of handwashing sink and handrub dispensers, convenient access to medical supplies, headwalls designed with adequate critical care services, maximum patient visibility, and patient lifts to ease strain on staff. Another desirable feature is a family zone so that a visitor is able to stay with the patient comfortably [46] (Fig. 5).

Indoor Environmental Quality and Health Improvement, Evidence-Based Design for. Figure 5
Acuity-adaptable, well-lit hospital rooms improve patient care and staff satisfaction

Studies are showing that patient pain levels and length of hospital stays can be reduced by exposure to nature and exposure to higher levels of daylight [47]. Walch et al. [48] found that spinal surgery patients in bright daylight lit rooms required 22% less opioid-equivalent analgesic medications than those in rooms without the bright daylight. Beauchemin and Hays [49] found that myocardial infarction patients in bright daylight lit rooms had shorter hospital stays of at least a day shorter. Ulrich [50] showed that surgery patients with views of nature had reduced hospital stays and used lower levels of pain medicine. EBD reveals that providing patients with high levels of daylight and views of nature (even if only pictures of nature if access to actual outdoor views are not possible) offers an opportunity to reduce patient pain medicine use and length of hospital stays, improving overall patient outcomes.

Reduction in ambient noise levels has been shown through EBD studies to improve patient sleep and reduce patient stress [51, 52]. For example, studies have shown reduced wound healing with exposure to noise, primarily attributed to increased levels of stress [53, 54]. EBD strategies that are applicable to noise control in hospitals include single-patient rooms, use of high-performance sound absorbing materials (although these must be easily cleanable), reduced noise from carts in the hallways, and noiseless paging systems.

EBD has led to improvements in staff workspace design as well as in patient care. EBD reveals that staff workspace needs to be designed with closer alignment to work patterns to improve staff satisfaction, productivity, and reduce stress reduction, which in turn will improve patient outcomes [38]. Potential design features may include decentralized nursing stations, more efficient layouts that allow staff interaction with patients and family members, and decentralized supply locations. Early EBD studies also reveal that the location of family members near the patients may also improve patient outcomes and reduce hospital stay lengths [55].

Economic Performance

Salaries and worker benefits generally exceed energy costs by approximately a factor of 100 [56]. Healthy,

high-performance sustainable buildings that are based on EBD principles have a strong potential to have positive economic performance, as long as the EBD design principles meet the organizational and health needs of the users as well as sustainable design principles. Therefore, a significant potential exists for businesses and building owners to employ EBD principles that improve worker performance, improve health, reduce health insurance costs, and reduce absenteeism.

Heerwagen [57] examined the range of benefits of green building features and attributes in buildings. She found that

- Green buildings are relevant to business interests across the full spectrum of concerns, from portfolio issues to enhanced quality of individual workspaces.
- Outcomes of interest that research should address include workforce attraction and retention, quality of work life, work output, and customer relationships.
- Green buildings can provide both cost reduction benefits and value added benefits.
- The benefits are most likely to occur when the building and organization are treated as an integrated system from the initiation of the design process, as in Evidence-Based Design approaches.

The Carnegie Mellon Center for Building Performance and Diagnostics (CBPD) and the Advanced Building Systems Integration Consortium have developed a decision support tool (The Building Investment Decision Support Tool – BIDS) to enable building decision makers to calculate returns on investments in high-performance building systems and to advance the understanding of the relationship between land use and buildings and health [56]. BIDS is based on a collection of building case studies as well as laboratory and simulation study results to statistically link the quality of buildings. BIDS uses "soft" and hard life-cycle costs to calculate the return on investment. The diverse building-related costs in the USA, including salaries and health benefits, technological and spatial turnover, rent, energy, and maintenance costs, normalized in dollars per person per year, are shown in Fig. 6.

Using statistics from the Bureau of Labor Statistics, the CBPD [56] calculated that the average employer health insurance cost was approximately $5,000 per employee per year in 2003. The CBPD went on and linked the cost of several specific health conditions and illnesses to IEQ (colds, headaches, respiratory illnesses, musculoskeletal disorders, and back pain), which account for approximately $750 of the $5,000 annual costs per employee – 14% of all annual health insurance expenditures. These direct costs would be additionally multiplied by the indirect costs of lost productivity. The results from employing BIDS provide the impetus to demonstrate the financial benefits of using EBD to design better building environments.

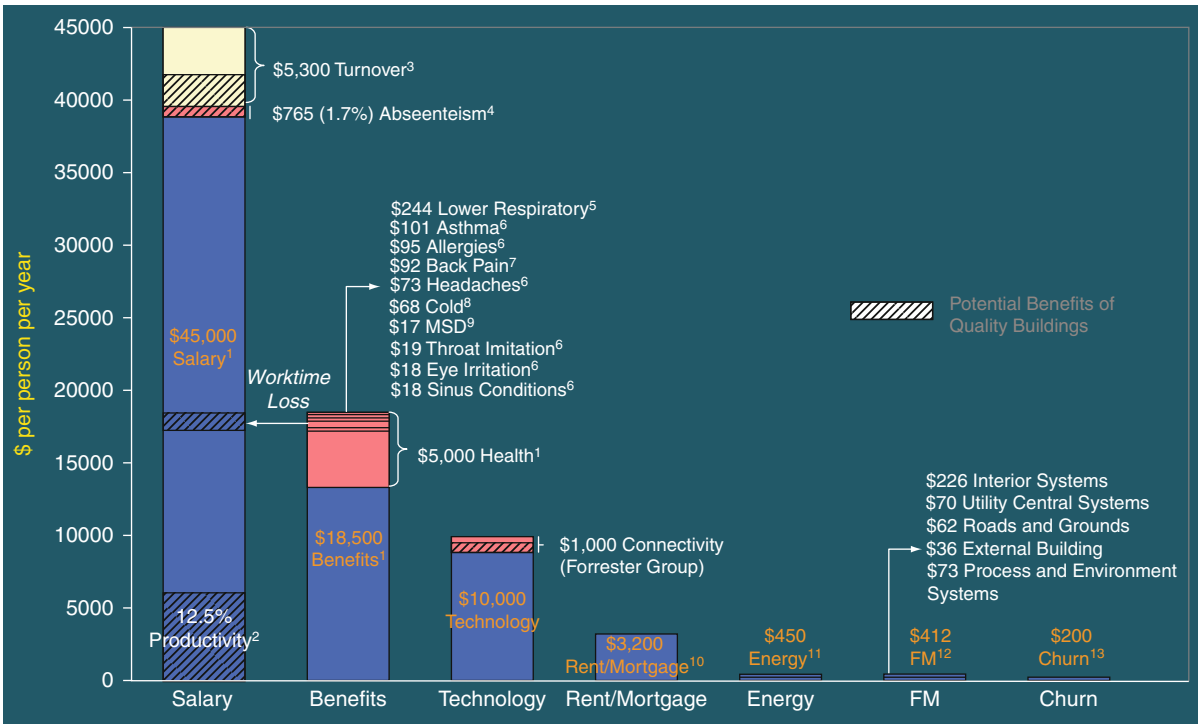
Fisk and Seppanen [58] demonstrated a benefit-cost ratio as high as 80 and an annual economic benefit as high as $700 per person when measures are made to improve indoor temperature control and increased ventilation rates based on a review of the existing literature of the health linkages between temperature control and increased ventilation rates. Table 1 shows the estimated productivity gains as a result of four categories of sources.

Application to Other Types of Facilities

The in-depth studies to support EBD in healthcare settings are readily adaptable to other types of facilities, particularly K-12 schools, including methods for infection control, better lighting, access to views and daylighting, improved acoustical performance, interior workspace layouts, and community design. The application of EBD in conjunction with sustainable design should result in optimal facilities for learning, healthcare, and work with maximum emphasis on human and ecological health as well as economic performance.

Schools

The impact of environmental design on the educational performance of students in the UK was investigated by Edwards [59]. In this study, Edwards investigated if "green" schools provide teaching and learning benefits beyond those in conventional schools, and what aspects of classroom design appear to be most critical in improving enhanced educational performance. Green schools were defined as being resource efficient particularly in terms of energy use; healthy both physically and psychologically; comfortable, responsive, and flexible; and based on ecological principles. In the study of



Indoor Environmental Quality and Health Improvement, Evidence-Based Design for. Figure 6
The true cost of least-cost buildings in the USA (US baselines from CMU BIDS) [50]

Indoor Environmental Quality and Health Improvement, Evidence-Based Design for. Table 1 Estimated potential productivity gains [58]

Source of productivity gain	Potential annual health benefits	Potential US annual savings or productivity gain (1996 US $$)
Reduced respiratory illness	16–37 million avoided cases of common cold or influenza	$6–14 billion
Reduced allergies and asthma	8–25% decrease in symptoms within 53 million allergy sufferers and 16 million asthmatics	$1–4 billion
Reduced sick building syndrome symptoms	Health symptoms experienced frequently at work by ~15 million workers	$10–30 billion
Improved worker performance from changes in thermal environment and lighting	Not applicable	$20–160 billion

54 schools built between 1975 and 1995, it was demonstrated that there is relationship between design, energy conservation, and educational performance. Overall the study demonstrated that green schools resulted in enhanced student performance and greater teacher satisfaction with the greatest impact on elementary schools. Benefits were greater in the newer schools with higher levels of ventilation. Absenteeism was reduced in the green schools. The student performance improvement appeared to be particularly related to the

level of daylight in the classroom, but also the level of ventilation, the temperature control, and noise level controls.

Elzeyadi [60] conducted a study to develop the Green Classroom Toolbox with green design guidelines for retrofitting existing educational spaces. The guidelines are based on carbon neutrality metrics and student achievement metrics, developed from a meta-analysis of reported studies and energy modeling simulations. The guidelines center on best practices that increase productivity, comfort, and health of students in retrofitted classrooms; facilitate integrated design and cooperation between designers; reduce environmental impacts and move toward carbon neutrality environments in schools; and are a model for future replication and dissemination. The strategic categories relevant to building professionals are based on the USGBC LEED criteria (1) energy and atmosphere (envelope, lighting, HVAC, and ventilation); (2) materials and resources (site construction, structural, and nonstructural); (3) environmental quality (IAQ, comfort, and acoustics); (4) sustainable sites (density, light pollution, and transportation); and (5) water and waste (building fixtures, landscaping, and recycling). Elzeyadi's method examined the facility as a whole system. He used a framework that treated the students and the school environment as interdependent elements of a system. The system is comprised of "people" and "buildings" on the macro-scale and "buildings" and "environment" on the megascale. This study resulted in three primary decision support tools of evidence-based guidelines to help architects, school designers, and school/school system staff to make informed decisions for implementing green retrofit measures in classrooms. The first tool is a check list of best practices compiled from focus groups and interviews of affected and interested parties. The second tool is a prioritization guide that provides a comparative analysis and ranking of the best practices list (in Tool 1) based on their impacts on building energy consumption and carbon emissions. The third tool is a meta-analysis guide that links the Tool 1 best practices to their impact on student and staff health and performance in schools. All of the tools were based on the specific climates and school typologies of the Pacific Northwest in the USA. The primary reason found for adoption of the best practices in schools was energy conservation followed by providing improved IEQ and connections to nature, reflected in energy and atmosphere, IEQ, and materials and resources gains. Better IAQ, based on the meta-analysis, was found to positively impact occupants' performance in a range of 5–20% improvement. This included reduced illnesses, both chronic and acute, and improved performance on testing. Improved temperature control was found to improve student performance in the range of 3–10%. Access to views and daylighting improved student performance in the range of 5–20%. This study emphasizes the need for evidence-based design guidelines for schools, especially to focus on improving IAQ, improved temperature control, and access to views and daylighting. The manner in which the study was conducted simulates the evidence-design process – interaction between the designers and the users, studying best practices and strategies in other successful facilities, and implementing the practices expected to have the most positive impact based on all of the stakeholders needs.

Office Buildings and Other Types of Facilities

The Academy of Neuroscience for Architecture has applied evidence-based design practices to office building design – focusing on the previously enumerated parameters (1) air quality, (2) thermal comfort, (3) spatial comfort, (4) collaborative or teamspace, (5) visual comfort, (6) workstation comfort, (7) lighting quality, (8) noise control, and (9) security. In their office building study [61], conducted via post-occupancy questionnaires, it was found that the office design features that support security, wayfinding, and feeling part of a cohesive organization created increased satisfaction and "workability" (considered to be neuro-environmental factors) among the employees over their previous office space. This was hypothesized to result in reducing stress, improving attention, focus, and mood. The office space design features included a centralized three-story open stairway connecting the three office floors, providing a naturally mapped sense of place, a "public square" housing centralized communications and meeting areas, a main entry area, centralized lunchroom, well-labeled directional signage, and use of porcelain tile paving across primary transit areas.

The Academy of Neuroscience for Architecture [62] also conducted a limited intervention study exploring the potential applications of neuroscience concepts and evidence-design based methods to correctional facilities. The specific focus topics were (1) daylight and views, (2) exposure to nature, (3) space size, (4) ambient noise levels, (5) color, and (6) environmental design features and their impact on inmate–staff relationships – reducing stress and aggressive behaviors. The overall goal of the study was to develop evidence-based design decisions for correctional settings and operations. The results of this study seemed to indicate that views of nature was the most effective measure of stress reduction, even if they were only projected nature views on a wall.

Future Directions

It is critical that EBD be applied much more widely across the spectrum of buildings. EBD has a tremendous potential to set a new paradigm for designing healthy, sustainable buildings, by including the building managers and occupants as a central player in the entire system's resolution of ecological and human health.

Even in the limited time that evidence-based design has been embraced, the data demonstrate important shifts for the building design and management community. For example, the need for increased ventilation rates significantly above those currently being used in the majority of buildings demands the development and implementation of innovative solutions that simultaneously meet reduced energy usage and cost. These include systems that separate ventilation and thermal conditioning, and new HVAC system types, such as underfloor air distribution and chilled beams. These also include improvements in system maintenance, such as the application of the ASHRAE Indoor Air Quality Procedure (IAQP) employing gaseous phase filtration to aid in air cleaning so that the ventilation level can be reduced. Ongoing research in more effective technologies and systems management is critical.

Future research must also include the development of protocols and metrics to accurately and realistically measure human impact improvements in health and productivity/learning, operational efficiencies, and sustainability. These metrics must consider the entire system in the occupied setting and not a just a single unit of the system. Metrics in specific will greatly aid in providing the necessary parameters for effective EBD studies in a wide range of buildings.

In the future, disparities between sustainable design practices and EBD will need to be resolved. Many practices are fully concurrent, but there are still areas where there is conflict, such as lack of acoustical satisfaction in open office planning and the potential energy costs of higher rates of ventilation for improved health and productivity/learning.

EBD takes the first step in rigorous research of "real" buildings by actively engaging in feedback through occupant questionnaires, and pursuing multi-configuration studies (in the form of layout or building system variations) or multi-building studies for comparative evaluation by end users. The lack of consistent feedback from building occupants and managers in the building design community has led for far too long to anecdotal design decision making, either in the form of untested shifts (such as open classrooms) or a dogged commitment to the status quo. EBD is an invaluable step forward, employing a range of post-occupancy tools – both qualitative and quantitative – to develop design innovations for human and environmental and economic benefit. EBD does not eliminate the need for controlled experimentation, both in the lab and in the field, to advance innovations in building materials, components, and systems design and operation.

Summary/Conclusions

The use of Evidence-Based Design to improve the IEQ in buildings has the potential to significantly impact the total health, productivity, learning, operational efficiency, and economic performance of a facility and its occupants. To begin with, a wide variety of studies have shown the importance of a connection to nature through access to views and daylighting to reduce stress, improve patient outcomes, improve health, and increase productivity. In the available literature, this connection to nature may be the most important design feature for overall impact studied to date. Secondly, an improved, innovative ventilation system has been shown to be critical to improving health and

productivity in buildings, of at least 25 L/s per person. Thirdly, the separation of temperature control from the ventilation system is another important component for improving thermal comfort without compromising ventilation air delivery. Finally, acoustical control is one of the most challenging parameters for EBD innovation, yet critically needed to achieve occupant satisfaction, stress reduction, and optimum learning in schools. EBD combined with sustainable design principles is an important tool for retrofitting and designing healthy, high-performance buildings.

Bibliography

1. The Center for Health Care Design (2008) Defined on their website http://clinicdesign.healthdesign.org/about#cfhd. Accessed 26 Jan 2011
2. Vischer JC (2009) Applying knowledge on building performance: from evidence to intelligence. Intell Build Int 2009: 239–248
3. WHO (1948) Preamble to the Constitution of the World Health Organization as adopted by the International Health Conference, New York, 19–22 June, 1946; signed on 22 July 1946 by the representatives of 61 States (Official Records of the World Health Organization, no. 2, p. 100) and entered into force on 7 April 1948
4. Lawson B (2010) Healing architecture. Arts Health Int J Res Policy Pract 2(2):95–108
5. Hartkofp V, Loftness V, Mahdavi A, Lee S, Shankavaram J (1997) An integrated approach to design and engineering of intelligent buildings – The intelligent workplace at Carnegie Mellon University. Autom Constr 6:401–415
6. Rashid M, Zimring C (2008) A review of the empirical literature on the relationships between indoor environment and stress in health care and office settings. Environ Behav 40(2):151–190
7. McEwen BS, Stellar E (1993) Stress and the individual. Mechanisms leading to disease. Arch Intern Med 157:2093–2101
8. Kaarlela-Tuomaala AA, Helenius RR, Keskinen EE, Hongisto VV (2009) Effects of acoustic environment on work in private office rooms and open-plan offices - longitudinal study during relocation. Ergonomics 52(11):1423–1444
9. Muehleisen RT (2010) Acoustics of green buildings, implications. A newsleter by InformeDesign vol 8(1). www.informedesign.umn.edu. Accessed 3 Oct 2010
10. Nabelek AK, Robinson PK (1982) Monaural and binaural speech-perception in reverberation for listeners of various ages. J Acoust Soc Am 71(5):1242–1248
11. Ryherd EE, Wang LM (2008) Implications of human performance and perception under noise conditions on indoor noise criteria. J Acoust Soc America 124(1):218–226
12. Ryherd EE, Wang LM (2007) Effects of exposure duration and type of task on subjective performance and perception in noise. Noise Control Eng J 55(1):334–347
13. Wilkins KK, Wolkoff PP, Knudsen HN, Clausen PA (2007) The impact of information on perceived air quality –"organic" vs. "synthetic" building materials. Indoor Air 17(2):130–134
14. Wolkoff PP, Wilkins CK, Clausen PA, Nielsen GD (2006) Organic compounds in office environments – sensory irritation, odor, measurements and the role of reactive chemistry. Indoor Air 16(1):7–19
15. Loftness V, Hartkopf V, Gurtekin B, Hansen D, Hitchcock R (2003) Linking energy to health and productivity in the built environment. Presented at the 2003 Greenbuild Conference, Chicago, IL, Nov 2003
16. Okcu S, Ryherd E, Bayer C (2011) The role of the physical environment on student health and education in green schools. Rev Environ Health (submitted)
17. Azuma K, Uchiyama IIK (2008) The regulations for indoor air pollution in Japan: a public health perspective. J Risk Res 11:301–314
18. Koistinen K, Kotzias D, Kephalopoulos S, Schlitt C, Carrer P, Jantunen M, Kirchner S, McLaughlin J, Mølhave L, Fernandes EO, Seifert B (2008) The INDEX project: executive summary of a European Union project on indoor air pollutants. Allergy 63:810–819
19. Olesen BW (2004) International standards for the indoor environment. Indoor Air 14:18–26
20. Kristen K (2005) California looks at tackling indoor air quality. Environ Sci Technol 39:256A
21. Spengler JD, Chen Q (2000) Indoor air quality factors in designing a healthy building. Ann Rev Energy Environ 25:567–601
22. Liddament MW (2000) A review of ventilation and the quality of ventilation air. Indoor Air 10:193–199
23. Wargocki P, Wyon DP (2007) The effects of moderately raised classroom temperatures and classroom ventilation rate on the performance of schoolwork by children (RP-1257). HVAC&R Res 13(2):193–220
24. Wargocki P, Wyon DP (2006) Effects of HAVC on student performance. ASHRAE J 48:22–28
25. Wyon DP, Wargocki P (2005) Room temperature effects on office work. In: Clements-Croome D (ed) Creating the productive workplace, 2nd edn. Taylor and Francis, London, pp 181–192
26. Tanabe Shin-ichi, Kobayashi K, Kiyota O, Nishihara N, Haneda M (2009) The effect of indoor thermal environment on productivity by a year-long survey of a call centre. Intell Buildings Int 1:184–194
27. Milton DK, Glencross PM, Walters MD (2000) Risk of sick leave associated with outdoor air supply rate, humidification, and occupant complaints. Indoor Air 10(4):212–221
28. Shendell DG, Prill R, Fisk WJ, Apte MG, Blake D, Faulkner D (2004) Associations between classroom CO_2 concentrations and student attendance in Washington and Idaho. Indoor Air 14(5):333–341
29. Sundell J, Levin H, Nazaroff WW, Cain WS, Fisk WJ, Grimsrud DT, Gyntelberg F, Li Y, Persily AK, Pickering AC, Samet JM, Spengler JD, Taylor ST, Weschler CJ (2010) Ventilation rates and health: multidisciplinary review of the

I

scientific literature. *Indoor Air*, an accepted article, Available online at http://onlinelibrary.wiley.com/doi/10.1111/j.1600-0668.2010.00703.x/pdf. Accessed 27 Jan 2011
30. Haverinen-Shaughnessy U, Moschandreas DJ, Shaughnessy RJ (2010) Association between substandard classroom ventilation rates and students' academic achievement. Indoor Air, Article first published online: 28 Oct 2010. http://onlinelibrary.wiley.com/doi/10.1111/j.1600-0668.2010.00686.x/pdf
31. Wargocki P, Sundell J, Bischof W, Brundrett G, Fanger PO, Gyntelberg F, Hanssen SO, Harrison P, Pickering A, Seppanen O, Wouters P (2002) Ventilation and health in non-industrial indoor environments: report from a European Multidisciplinary Scientific Consensus Meeting (EUROVEN). Indoor Air 12:113–128
32. Baron RA, Rea MS, Daniels SG (1992) Effects of indoor lighting (illuminance and spectral distribution) on the performance of cognitive tasks and interpersonal behaviors: the potential mediating role of positive affect. Motiv Emotion 16:1–33
33. Magee GH (1988) Facilities maintenance management. R.S. Means, Kingston
34. Bayer C, Hendry RJ, Cook A, Downing C, Crow SC, Hagen S, Fischer JC (2002) Active humidity control and continuous ventilation for improved air quality in schools. Presented at the DOE integrated energy systems peer review meeting, Nashville, TN, April 30–May 2 2002
35. Wargocki P, Wyon DP, Fanger PO (2004) The performance and subjective responses of call-center operators with new and used supply air filters at two outdoor air supply rates. Indoor Air 14(Suppl 8):7–16
36. Arditi D, Nawakorawit M (1999) Designing building for maintenance: Designer's perspective. J Architect Eng 5(4):107–116
37. Molhave L, Bach B, Pederson OF (1986) Human reactions to low concentrations of volatile organic compounds. Environ Int 12:167–175
38. Smith SW, Rea MS (1982) Performance of a reading test under different levels of illumination. J Illum Eng Soc 1:29–33
39. Veitch JA (1990) Office noise and illumination effects on reading comprehension. J Environ Psychol 10:209–217
40. Heerwagen J, Zagreus L (2005) The human factors of sustainable building design: post occupancy evaluation of the Philip Merrill Environmental Center, Annapolis, MD. Report for Drury Crawley, US Department of Energy, Building Technology Program. http://www.cbe.berkeley.edu/research/publications.htm
41. Bayer CW (2009) Unpublished work on IAQ investigations in Atlanta elementary schools
42. Anderson K (2004) The problem of classroom acoustics: the typical classroom soundscape is a barrier to learning. Semi Hear 25:117–129
43. Ulrich RS, Zimring C, Zhu X, DuBose J, Seo HB, Choi YS, Quan X, Joseph A (2008) A review of the research literature on evidence-based healthcare design. Health Environ Res Des J 1(3):101–532
44. Buchanan TL, Barker KN, Gibson JT, Jiang BC, Pearson RE (1991) Illumination and errors in dispensing. Am J Hosp Pharm 48(10):2137–2145
45. Zigmond J (2006) Built-in benefits. Mod Healthc 36(11):30–38
46. Lipschutz LN (2008) Acuity-adaptable rooms: design considerations can improve patient care. Healthcare Construction and Operations, Jan/Feb 10–11
47. Malenbaum S, Keefe FJ, Williams AC, Ulrich R, Somers TJ (2008) Pain in it environmental context: implications for designing environments to enhance pain control. Pain 134:241–244
48. Walch JM, Rabin BS, Day R, Williams JN, Choi K, Kang JD (2005) The effect of sunlight on postoperative analgesic medication use: a prospective study of patients undergoing spinal surgery. Psychosom Med 67:156–163
49. Beauchemin KM, Hays P (1998) Dying in the dark: sunshine, gender, and outcomes in myocardial infarction. J R Soc Med 91:352–354
50. Ulrich RS (1984) View through a window may influence recovery from surgery. Science 224:420–421
51. Topf M, Bookman M, Arand D (1996) Effects of critical care unit noise on the subjective quality of sleep. J Adv Nurs 24:545–551
52. Freedman N, Gazendam J, Levan L, Pack A, Schwab R (2001) Abnormal sleep/wake cycles and the effect of environmental noise on sleep disruption in the intensive care unit. Am J Respir Crit Care Med 163(2):451–457
53. Wysocki A (1996) The effect of intermittent noise exposure on wound healing. Adv Wound Care 9(1):35–39
54. Toivanen S, Hulkko S, Naatanen E (1960) Effect of psychic stress and certain hormone factors on the healing of wounds in rats. Ann Med Exp Bio Fenn 38:343–349
55. France D, Throop P, Joers B, Allen L, Parekh A, Rickard D, Deshpande JK (2009) Adapting to family-centered hospital design: changes in providers' attitudes over a two-year period. Health Environ Res Des J 3:79–96
56. Loftness V, Hartkopf V, Poh LK, Snyder M, Hua Y, Gu Y, Choi J, Yang X (2006) Sustainability and health are integral goals for the built environment. Presented at the 2006 Healthy Buildings Conference, Lisbon, Portugal, 4–8 June 2006
57. Heerwagen J (2000) Green buildings, organizational success and occupant productivity. Build Res Inf 28(5/6): 353–367
58. Fisk W, Seppanen O (2007) Providing better indoor environmental quality brings economic benefits. In: Proceedings of Clima 2007: well-being indoors. FINVAC, Helsinki, Finland, 10–14 June 2007
59. Edwards BW (2006) Environmental design and educational performance. Res Educ 76:14–32
60. Elzeyadi, Ihab MK (2008) Green classroom retrofit toolbox (GCRT): Evidence-based design guidelines to adapt K-12 school facilities for climate change. http://www.aia.org/aiaucmp/groups/aia/documents/pdf/aiab079900.pdf. Accessed 26 Jan 2011

61. Zeise J, Vischer J (2006) Executive summary: environment/behavior/neuroscience pre and post-occupancy evaluation of new offices for society of neuroscience. http://www.anfarch.org/pdf/SfNExecutiveSummary.pdf. Accessed 26 Jan 2011
62. Farbstein J, Farling M, Wener R (2009) The evidence for evidence-based design: nature views reduce stress in a jail intake area. Presented at The American Institute of Architects Annual Meeting, Orlando, FL, June 2009. http://www.aia.org/akr/Resources/Documents/AIAB086133?dvid=&recspec=AIAB086133. Accessed 26 Jan 2011

Infectious Disease Modeling

Angela R. McLean
Zoology Department, Institute of Emerging Infections, University of Oxford, Oxford, UK

Article Outline

Glossary

Basic reproductive number A summary parameter that encapsulates the infectiousness of an infectious agent circulating in a population of hosts.
Host An organism that acts as the environment within which an infectious agent replicates.
Infectious agent A microorganism that replicates inside another organism.
Pathogen An infectious agent that damages its host.
Variant One of several types of an infectious agent, often closely related to and sometimes evolved from other variants under consideration.

Definition of the Subject

Infectious disease models are mathematical descriptions of the spread of infection. The majority of infectious disease models consider the spread of infection from one host to another and are sometimes grouped together as "mathematical epidemiology." A growing body of work considers the spread of infection within an individual, often with a particular focus on interactions between the infectious agent and the host's immune responses. Such models are sometimes grouped together as "within-host models." Most recently, new models have been developed that consider host–pathogen interactions at two levels simultaneously: both within-host dynamics and between-host transmissions. Infectious disease models vary widely in their complexity, in their attempts to refer to data from real-life infections and in their focus on problems of an applied or more fundamental nature. This entry will focus on simpler models tightly tied to data and aimed at addressing well-defined practical problems.

Introduction

Why is it that smallpox was eradicated in 1979 [1] but measles, once scheduled for eradication by the year 2000, still kills over a hundred thousand children each year [2]? Both diseases can be prevented with cheap, safe, and effective vaccines which probably induce lifelong immunity, and neither virus has an environmental or animal reservoir.

One way to address this question is to consider the comparative ease of spread of the two infections. A useful parameter that summarizes this ease of spread is the "basic reproductive number" always denoted as R_0. The definition of the basic reproductive number is the number of secondary infections caused during the entire duration of one infection if all contacts are susceptible (i.e., can be infected). The concept has widespread currency in the literature on infectious disease models with varying degrees of affection [3]. There is no question that it has been a useful, simple rule of thumb for characterizing how easily an infection can spread [4]. Furthermore, the simplest of calculations relate R_0 to the degree of intervention needed to bring an infection under control and, eventually, eradicate it. The relationship between the basic reproductive number and disease control arises from the simple fact that if each infectious person causes less than one secondary case, then the number of infections must fall. If it is always true, even when there is no infection circulating,

that each new case causes less than one secondary case, then the infection will die out. This simple observation leads to a straightforward calculation for the proportion of a population that must be vaccinated in order to achieve eradication, p_c:

$$p_c = 1 - \frac{1}{R_0} \quad (1)$$

This relationship arises from the fact that if $(R_0 - 1)$ out of the R_0 people a case might have infected have been vaccinated, then each case, in a population vaccinated to that degree, will cause less than one secondary case. For example, say $R_0 = 10$, if 9/10 of the population are successfully immune following vaccination, then infection cannot spread. Thus, in general, $p_c = (R_0 - 1)/R_0$, as stated in Eq. 1. If the fraction of the population that are successfully vaccinated is greater than p_c, then each case will cause, on average, less than one case, and infection cannot spread.

Comparing estimated values for R_0 for measles and smallpox and inferred values for the proportion that need to be vaccinated to ensure eradication (Table 1) leads to a simple answer to the question why has smallpox been eradicated, but not measles? Smallpox, with a basic reproductive number around three, was eradicated with vaccination coverage of around 67%. The higher R_0 for measles, nearer to 15, requires vaccination coverage close to 95% to ensure eradication. Many parts of the world remain unable to achieve such high coverage; measles remains suppressed by tremendous efforts at vaccination but is not yet eradicated.

These calculations are so straightforward that they can be made without recourse to any formal modeling. However, embedding these ideas inside a formal modeling framework has proven very useful. The next section describes the simplest applicable model.

Infectious Disease Modeling. Table 1 The basic reproductive number, R_0, and the critical vaccination proportion for eradication, p_c, for measles and smallpox

Infection	Place	Time	R_0	p_c (%)
Smallpox	India	1970s	3	67
Measles	India	1970s	15	93
Measles	UK	1960s	15	93

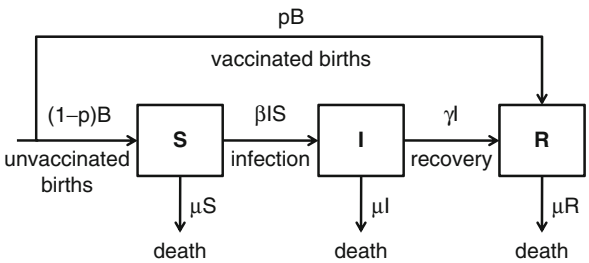
The SIR Model

The "plain vanilla" model of mathematical epidemiology is called the SIR model because it splits the host population into three groups:

- The susceptible (S) can be infected if exposed
- The infectious (I) are both infected and infectious to others
- The recovered (R) are no longer infectious and are immune to further infection

The SIR model's structure then consists of a set of assumptions about how people flow into, out of, and between these three groups. Those assumptions can be represented graphically as in Fig. 1.

The assumptions of the SIR model with vaccination are the following: People are born at a constant rate B, and a proportion p of them are vaccinated at birth. Vaccinated newborns are immune for life and so they join the recovered class. Unvaccinated newborns enter the susceptible class. Susceptibles are infected at a per capita rate proportional to I, the number of infectious people in the population. This gives rise to a transfer from the susceptible to the infectious class at rate βIS. Susceptibles are also subject to a per capita background death rate μ. Infectious people recover into the recovered class I at per capita rate γ or die at the per capita background death rate μ. Recovered individuals are immune for the rest of their lives, so the only exit from the recovered class is at the per capita background death rate μ.



Infectious Disease Modeling. Figure 1
The SIR model in graphical form. The host population is divided into three groups, and transitions of people between those groups are described. Those transitions represent the five processes: birth, vaccination, death, infection, and recovery

These assumptions can be written in several different forms of equations, for example, difference equations, ordinary differential equations, or stochastic differential equations. The difference equation form is as follows:

$$S(t+1) = S(t) + (1-p)B - \beta I(t)S(t) - \mu S(t) \quad (2)$$

$$I(t+1) = I(t) + \beta I(t)S(t) - \gamma I(t) - \mu I(t) \quad (3)$$

$$R(t+1) = R(t) + pB + \gamma I(t) - \mu R(t) \quad (4)$$

This difference equation form is particularly easy to handle numerically and can be straightforwardly solved in a spreadsheet. Figure 2a shows the solutions to Eqs. 2–4 over 50 years with a 1-week timestep. Parameters are set so that an infection with a basic reproductive number of 5 and a 1-week duration of infection is spreading in a population of 100,000 individuals. The figure illustrates how this model shows damped oscillations towards a stable state. The same is true for the ODE version of this model.

This model is useful for understanding the impact of vaccination. In Fig. 2b, the solutions to Eqs. 2–4 are shown when vaccination at birth is introduced 10 years into the model run. With a basic reproductive number of 5, Eq. 1 tells us that vaccination of over 80% of newborns will lead to eradication. This is exemplified in the pink line where 90% vaccination leads to no further cases. Vaccination coverage below this threshold value reduces the numbers of cases and increases the inter-epidemic period but does not lead to eradication. Notice the very long inter-epidemic period at 70% vaccination. This phenomen occurs when vaccine coverage levels are close to but do not achieve the critical coverage level. Under these circumstances, it takes a very long time to accumulate enough susceptibles to trigger the first epidemic after vaccination is introduced. It may therefore appear as though eradication has been achieved even though vaccination coverage is below the critical level. This phenomenon, named "the honeymoon period," [5] was first described in modeling studies and later identified in field data [6].

The ability to identify target vaccination levels predicted to lead to disease eradication has been widely influential in policy circles [7]. Models with the same fundamental structure as the SIR model are used to set targets for vaccination coverage in many settings [8]. Similar models are also used to understand the likely impact of different interventions of other sorts, for example, drug treatment [9] or measures for social distancing [10]. However, models for informing policy need to explore more of the wrinkles and complexities of the real world than are acknowledged in the simple equations of the SIR model. The next section describes some of the types of host heterogeneity that have been explored in making versions of the SIR model that aim to be better representations of the real world.

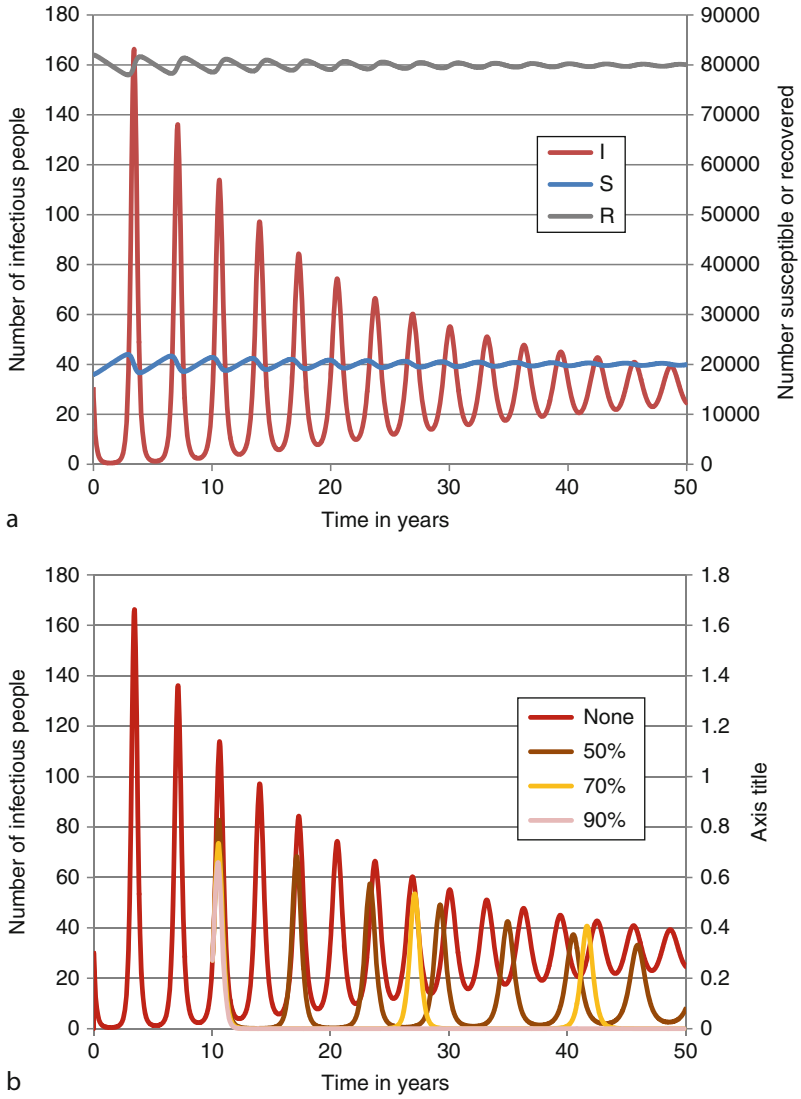
Host Heterogeneity

There are many aspects of host heterogeneity that have bearing on the transmission and impact of infections. Two of the most important are host age and spatial distribution. In this section, the modeling of these two types of host heterogeneity is introduced with reference to two specific infections: rubella and foot-and-mouth disease.

Rubella is a directly transmitted viral infection that usually causes mild disease when contracted during childhood. However, infection of a woman during early pregnancy can lead to serious birth defects for her unborn child. The set of consequent conditions is labeled "congenital rubella syndrome" or CRS. Because vaccination acts to extend the time between epidemics (Fig. 1b), it also acts to increase the average age at infection. This sets up a complex trade-off when introducing rubella vaccination to a community. On the one hand, vaccinated girls are protected from catching rubella at any age, but on the other hand, the girls who remain unvaccinated are likely to catch rubella when they are older, more likely to be in their childbearing years and so at greater risk of CRS. This means that vaccination with low coverage can actually lead to more CRS, and only when coverage levels get above a certain level do the benefits of vaccinating the community outweigh the costs. Calculating where that level lies then becomes an important public health question.

Because age is such an important component of the risks associated with rubella infection, models of this system need to take account of host age. The relevant versions of Eqs. 2–4 are difference equations with two independent variables, age (a) and time (t):

$$S(a+1, t+1) = S(a,t) - S(a,t)[\Sigma_{a'}\beta(a,a')I(a,t)] - \mu(a)S(a,t) \quad (5)$$



Infectious Disease Modeling. Figure 2

Numerical solutions to the SIR difference equation model. Infection circulates in a population of 100,000 individuals, with an expectation of life at birth of 50 years. The infectious period is 1 week, and the basic reproductive number is 5. This gives the following model parameters: B = 38 per week, $\mu = 0.00385$ per person per week, $\gamma = 1$ per person per week, and $\beta = 0.00005$ per infected per susceptible per week. (**a**) shows damped oscillations in all three classes after an initial perturbation of 20% of the susceptible class into the recovered class. In (**b**), vaccination of 50%, 70%, or 90% of newborns is introduced at time 10 years. With $R_0 = 5$, the critical vaccination proportion $p_c = 0.8$. Vaccination coverage above this level (at 90%) leads to eradication

$$I(a+1,t+1) = I(a,t) + S(a,t)[\Sigma_{a'}\beta(a,a')I(a,t)] - \gamma(a)I(a,t) - \mu(a)I(a,t) \quad (6)$$

$$R(a+1,t+1) = R(a,t) + \gamma(a)I(a,t) - \mu(a)R(a,t) \quad (7)$$

Notice how these equations, by taking account of age as well as time, allow consideration of several different kinds of age dependence. Firstly, Eq. 6 calculates the number of cases of infection *of a given age* over time. Since the main consideration in balancing up the

pros and cons of rubella vaccination is the number of cases in women of childbearing age, this is an essential model output. Secondly, the per capita rate at which susceptibles become infected depends on their age and on the age of all the infected people. This model is thus able to take account of the complexities of family, school, and working life which drive people of different ages to age-dependent patterns of mixing. Thirdly, the recovery rate $\gamma(a)$ and, more importantly, the background death rate $\mu(a)$ can both be made to depend on age. Since a fixed per capita death rate is a particularly bad approximation of human survival, this is another important advance on models without age structure.

Models with age structure akin to that presented in Eqs. 5–7 have been essential components of the planning of rubella vaccination strategies around the world [11, 12]. A model as simple as these equations would never be used for formulating policy; furthermore, most age-structured models use the continuous time and age versions and so have the structure of partial differential equations. Nevertheless, Eqs. 5–7 illustrate the fundamentals of how to include age in an epidemiological model.

The spatial distribution of hosts is another important aspect of their heterogeneity. If the units of infection are sessile (e.g., plants), the assumption that all hosts are equally likely to contact each other becomes particularly egregious and models that acknowledge the spatial location of hosts more important. One example of units of infection that do not move is farms. If trade between farms has been halted because of a disease outbreak, then disease transmission between farms is likely to be strongly dependent upon their location. This was the case during the 2001 foot-and-mouth disease epidemic in the UK, and spatial models of that epidemic are nice examples of how to explicitly include the distance between hosts in a model epidemic.

On February 19, 2001, a vet in Essex reported suspected cases of foot-and-mouth disease (FMD) in pigs he had inspected at an abattoir. FMD is a highly infectious viral disease of cloven-hoofed animals. Because of its economic and welfare implications for livestock, FMD had been eradicated from Western Europe. The FMD outbreak that unfolded in the UK over the ensuing months had a huge impact with millions of farm animals killed and major economic impact in the countryside as tourism was virtually shut down.

There was heated debate about the best way to control the spread of infection from farm to farm. FMD virus is so very infectious that no attempt was made to control its spread within a farm. Once infection of livestock on a farm was detected, all susceptible animals were slaughtered. Mathematical models of the spread of this epidemic thus treat each farm as a unit of infection, and, as before, farms can be categorized as susceptible, infectious, etc. The best of these models [13] keeps track of every single farm in the United Kingdom, characterizing farms by their location and the number of sheep and cattle they hold. The model classifies farms into four groups: susceptible, incubating, infectious, or slaughtered. As in all epidemic models, the heart of the model is the per capita rate at which susceptible farms become infected – the so-called force of infection. Because this FMD model is an individual-based, stochastic simulation, it is not possible to write out its equations in a simple form as before, but the probability of infection for a single farm can easily be written.

Suppose all farms in the UK are listed and indexed with i. Then p_i, the probability that an individual farm i becomes infected during one unit of time, is:

$$p_i = \beta_i \Sigma_{\text{all infectious farms}j} \tau_j K(d_{ij}) \quad (8)$$

where β_i is the susceptibility of farm i, determined by the number of sheep and cows it holds; τ_j is the infectiousness of farm j, also determined by the number of sheep and cows it holds; and $K(d_{ij})$ is a function of the distance between the pair of farms i and j which determines how quickly infectiousness falls off with increasing distance. K is known as the "infection kernel." In the FMD example, the infection kernel was estimated from contact tracing data on farms that were sources of infection and their secondary cases. This observed relationship shows a very sharp falling off of infectiousness, with a farm just 2 km distant being less than tenfold as infectious to a susceptible farm than one that is adjacent.

This section describes just two of the possible heterogeneities that are often included when making models of the spread of epidemics. There is almost no end to how complex an epidemiological model can become. However, it is very easy for complex models

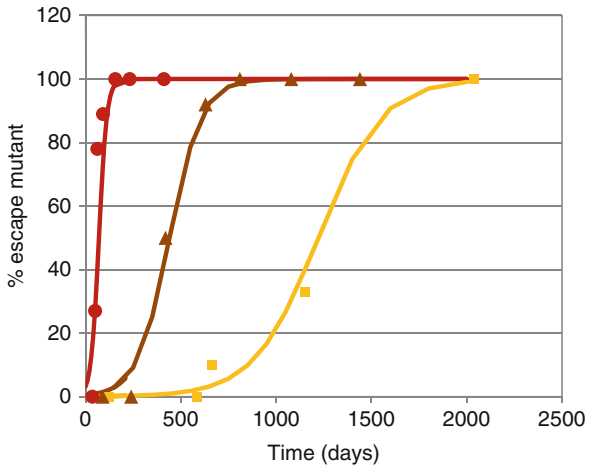
to outstrip the data available to calculate their parameters. In some cases, this can mean that models become black boxes concealing ill-informed guesswork, rather than prisms unveiling the implications of well-sourced and well-understood data.

Within-Host Dynamics

Mathematical models can also be used to investigate the dynamics of events that unfold within infected hosts. In these models, the units of study are often infected cells and immune cells responding to infection. As with epidemiological models, there is a wide range of modeling styles: Some models detail many different interacting components; others make a virtue of parsimony in their description of within-host interactions. In this section, a simple model of the within-host evolution of HIV is used to illustrate how pared-down, within-host models of infection can address important practical questions.

Several trials of prophylactic HIV vaccines have shown little or no effect [14–16], and understanding why these vaccines failed is a major research priority [17]. A quantitative description of the interaction between HIV and host immune cells would be an asset to such understanding. For one component of host immunity – the cytotoxic T cell (CTL) response – such a description can be derived. The question is how effective are host CTL responses at killing HIV-infected cells? Not how many CTLs are present, nor which cytokines they secrete, but how fast do they kill HIV-infected cells?

During the course of a single infection, HIV evolves to escape from the selection pressure imposed by host CTLs [18]. In this process, new HIV variants emerge that are not recognized by the host CTLs. These variants are called "CTL escape mutants." These CTL escape mutants can be seen to grow out in hosts who mount relevant CTL responses (Fig. 3) and to revert in hosts who do not. The rate of reversion in hosts without relevant CTL responses reflects the underlying fitness cost of the mutation. The rate of outgrowth in hosts who do mount relevant CTL responses is a balance between the efficacy of those responses and the fitness cost of the mutations. These costs and benefits need to be examined in the context of the underlying rate of turnover of HIV-infected cells. All this can be represented in a two-line mathematical model [19].



Infectious Disease Modeling. Figure 3
The outgrowth of HIV CTL escape mutants through time. Data sets from three different patients (reviewed in [19]) are shown as *red*, *brown*, and *yellow* symbols. Equation 12 is fitted to these data, yielding rates of outgrowth, $c-(a-\hat{a})$, of 0.048 (*red*), 0.012 (*brown*), and 0.006 (*yellow*)

Let x be the number of host cells infected with "wild-type" virus – that is, virus that can be recognized by the relevant host CTL responses. Let y be the number of host cells infected with escape mutant virus. The model then consists of a pair of ordinary differential equations describing the growth rate of each population of infected cells. The wild-type population grows at rate a, is killed by the CTL response in question at rate c, and is killed by all other processes at rate b. The escape mutant population grows at rate â ($\hat{a} < a$, reflecting the underlying fitness cost of the mutation) and is killed by all other processes at rate b. Escape-mutant-infected cells are not killed by the CTL response in question because of the presence of the escape mutation in the viral genome. These assumptions give rise to the following pair of linear ordinary differential equations:

$$x' = ax - bx - cx \tag{9}$$

$$y' = \hat{a}y - by \tag{10}$$

The observed quantity, call it p, is the fraction of virus that is of the escape mutant type; $p = y/(x + y)$. Simple application of the quotient rule for differentiation yields the single differential equation

$$p' = (c - (a - \hat{a}))p(1 - p) \tag{11}$$

with solution:

$$p = (k \exp(-(c - (a - \hat{a}))t) + 1)^{-1} \tag{12}$$

where for p_0, the fraction escaped at time 0:

$$k = \frac{(1 - p_0)}{p_0} \tag{13}$$

It is straightforward to fit the analytic expression (12) to data on the outgrowth of escape mutants to obtain estimates of the quantity $c - (a - \hat{a})$. Figure 3 shows fitted curves with estimates of $c - (a - \hat{a})$ of 0.048, 0.012, and 0.006. The quantity of interest is the parameter c – the rate at which CTL kills cells infected with wild-type virus. Fortunately, independent estimates of the fitness cost of the escape mutation $(a - \hat{a})$ are available. The median of several such observations yields $(a - \hat{a}) = 0.005$ [19]. Taken together and combined with further data, the inference is that on average, a single CTL response kills infected cells at rate 0.02 per day.

The half-life of an HIV-infected cell is about 1 day. This figure was itself derived from the application of elegantly simple models to data on the post-treatment dynamics of HIV [20, 21]. If a single CTL response kills infected cells at rate 0.02 per day and their overall death rate is one, then just 2% of the death of infected cells can be attributed to killing by one CTL response. Patients will typically mount many responses – but probably not more than a dozen. This analysis shows that even though CTL responses are effective enough to drive viral evolution, they are, in quantitative terms, very weak. A vaccine to protect against HIV infection would have to elicit immune responses that are many-fold stronger than the natural responses detected in ongoing infection. This simple, model-based observation greatly helps understand why the vaccines trialed so far have failed.

Multilevel Models

The models discussed so far deal either with events inside individuals or with transmission amongst individuals (people or farms) in a population. Some questions require simultaneous consideration of events at both levels of organization. This is particularly true for questions about the evolution of infectious agents as their evolution proceeds within individual hosts, but they are also transmitted between hosts. Models that capture events at both the within-host and between-host levels are fairly recent additions to the literature on infectious disease modeling. Here, they are illustrated with two examples, a set of models that consider the emergence of a zoonotic infection in humans and a model of the within-host evolution and between-host transmission of HIV.

Emerging infections are a continuing threat to human well-being. The pandemics of SARS in 2003 and H1N1 swine flu in 2009 illustrated how quickly a new infectious agent spreads around the world. Neither of these was as devastating as some predicted, but the continuing pandemic of HIV is ample proof that emerging infectious diseases can have devastating consequences for human communities. Many novel emerging infections arise as zoonoses – that is, infections that cross from animals into humans [22]. To become a successful emerging infection of humans – that is, one that spreads widely amongst people – is a multi-step process [23]. First, the pathogen must cross the species barrier into people, then it must transmit between people, and finally, it must transmit *efficiently enough* that epidemics arise. This latter step amounts to having a basic reproductive number, R_0, that is greater than 1. The emerging infections mentioned already, SARS, swine flu, and HIV, have transited all these steps. But there are other zoonoses that transmit to humans without emerging as epidemics or pandemics. For example, simian foamy virus, a retrovirus that is endemic in most old-world primates [24], can be detected in people who work with primates [25] or hunt them [26]. There is no record of any human-to-human transmission, implying that this zoonosis only completes the first step in becoming an emerging infection. Other infections, whilst spreading from person to person, still do not cause epidemics because that spread is insufficiently efficient. An example of such an infection is the newly discovered arenavirus from Southern Africa called "Lujo virus" [27]. This virus caused a small outbreak in the autumn of 2008. Very dramatically, four out of the five known cases died, but with five cases and just four transmission events, the basic reproductive number stayed below one, and there was no epidemic.

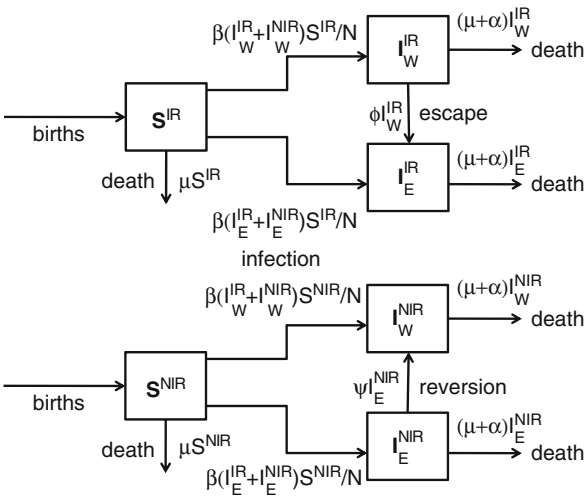
Acquiring $R_0 > 1$ is thus an important threshold that zoonoses must breach before they can become emerging infections. Antia and colleagues [28] developed an elegant model of the within-host evolution and between-host transmission of a zoonotic infection that initially has $R_0 < 1$, but through within-host adaptation in humans can evolve to become efficient enough at transmitting from one human to another that R_0 increases above 1 and epidemics become possible. They developed a multi-type branching process model of the transmission and evolution of a zoonosis. They found that the probability of emergence depends very strongly on the basic reproductive number of the pathogen as it crosses into humans. This is because, even when $R_0 < 1$, short chains of transmission are still possible (as exemplified with Lujo virus described above). During ongoing infections in humans, the zoonosis has opportunities to evolve towards higher transmissibility. The higher its initial R_0, the more opportunities there are for such ongoing evolution and hence for emergence.

This model of the emergence of a novel infection has been extended by other authors to address questions about the interpretation of surveillance data [29] and the role of host heterogeneity in the process of emergence [30]. These extensions confirm the original finding that the transmission efficiency (R_0) of the introduced variant (and any intermediate variants) is a very important driver of the probability of emergence. Kubiak and colleagues explored the emergence of a novel infection in populations split into several communities, with commuters acting to join those communities together. They found that most communities are sufficiently interconnected to show no effect of spatial distribution on the emergence process, even a small number of commuters being sufficient to successfully transmit any novel pathogen between settlements. Thus, although many zoonotic events happen in isolated parts of the world, unless they are really cut off from urban centers, that isolation offers little barrier to the transmission of newly emerged infections.

HIV emerged as a human infection sometime during the end of the 1800s and the early 1900s [31]. It was only recognized as a new human infection in the 1980s when cases of immunodeficiency in young Americans were unusual enough to warrant investigation [32]. As discussed above, during the course of a single infection, HIV is able to adapt to escape from the selection pressures imposed by its host's immune response. HIV variants that cannot be recognized by current host CTLs are termed "CTL escape mutants." These mutants yield important information about the strength of the immune responses that they evade. However, since they were shown to transmit from one host to another, their status has been raised to potential drivers of evolutionary change across the global HIV pandemic [33, 34].

Different hosts respond to different parts of HIV's proteins (known as epitopes). For CTL responses, it is the host class 1 human leukocyte antigen (HLA) type that determines which epitopes are recognized. When CTL escape mutants are transmitted into a host who does not make immune response to that epitope, the mutations are no longer advantageous, and the virus can revert to the wild type [35]. Global change in the prevalence of CTL escape mutants is therefore driven by three parallel processes: the selection of escape mutants in some hosts, transmission between hosts, and reversion of escape mutants in other hosts. Once again, this is a process that takes place across multiple levels of organization, evolution and reversion of escape mutations within infected hosts, and transmission between hosts.

Fryer and colleagues [36] developed a multilevel model of the three processes of within-host evolution, within-host reversion, and between-host transmission. The model is a version of the so-called SI model which is a simplified version of the SIR model presented above which does not allow recovery. The model allows heterogeneity in hosts and in the infecting virus so that there are hosts who do and do not mount immune responses to a given epitope and there are viruses that do and do not have escape mutations in that epitope. This model is represented in Fig. 4. As in the SIR model described in section "The SIR Model," the rate at which susceptibles become infected is determined by the number of infectious people present. However, in this model, because it represents the spread of a sexually transmitted disease, it is the proportion of hosts who are infectious that drives new infections. Furthermore, there are now two virus types circulating – wild type and escape mutant. Within-host adaptation allows hosts who do mount immune responses to the epitope to drive the evolution of escape mutants, and conversely, hosts who do not mount such responses can



Infectious Disease Modeling. Figure 4
A model of the within-host evolution and between-host transmission of HIV escape mutants [36]. Hosts are divided into two types: immune responders (superscript IR) and nonimmune responders (superscript NIR). There are also two variants of virus, wild type (subscript WT) and escape mutant (subscript E). Hosts are either susceptible, S, or infectious, I, and the type of virus with which they are infected is denoted by the subscript. Rates of infection are determined by the number of people infectious with each virus type. Immune responding hosts infected with the wild-type virus drive immune escape at per capita rate ϕ, whilst nonimmune responding hosts infected with escape mutant virus drive reversion at per capita rate ψ. All hosts are prone to per capita death rate μ, and infected hosts have an additional death rate α

drive the reversion of escape mutant viruses back to the wild type.

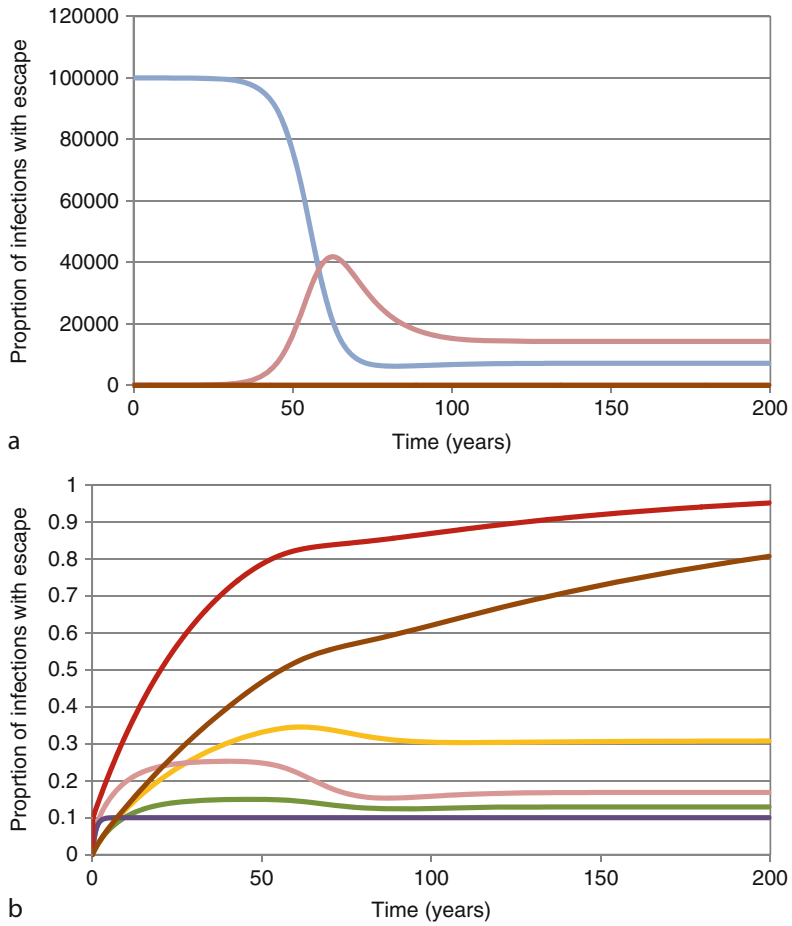
This model's behavior is easy to understand. The total numbers of susceptible and infectious people simply follow the well-characterized SI model. Figure 5a shows total cases through time. The total epidemic goes through three phases: an initial exponential growth, a saturation phase, and then settling to a long-term equilibrium. Figure 5b shows the proportion of all cases that are escape mutants through time. Not surprisingly, faster escape rates and slower reversion rates lead to higher prevalence of escape mutants. Less intuitive are the following characteristics of Fig. 5b. Whilst the epidemic is in its exponential growth phase, so long as reversion rates are reasonably fast (say once in 10 years or faster), the prevalence of escape is expected to stabilize quite quickly. However, this is not a long-term equilibrium, and as the total epidemic turns over, the escape prevalence shifts again. For an epitope that escapes fast but reverts at an intermediate rate, this leads to a substantial drop in the prevalence of escape. Secondly, fixation of escape variants only occurs if they never revert, and even then fixation takes a very long time – much longer than it takes for the underlying epidemic to equilibrate. Thirdly, the predicted dynamics and equilibrium are very sensitive to the reversion rate when that is slow. Notice the big difference, in the long term, between no reversion (brown line) and average time to reversion of 50 years (yellow line).

As well as predicting the future spread of escape mutations for different rates of escape and reversion, this model can be used to infer escape and reversion rates from data on their current prevalence. This exercise reveals a surprisingly slow average rate of escape. Across 26 different epitopes, the median time to escape was over 8 years. There is close agreement between rates of escape inferred using this model and those estimated from a longitudinal cohort study. These slow rates are in marked contrast to the general impression given by a large number of case reports in which escape is described as occurring during the first year of infection. However, a collection of case reports is a poor basis upon which to estimate an average rate of escape.

These are just two examples from the new family of infectious disease models that encapsulate processes at multiple levels of organization. As data on pathogen evolution continues to accrue, this approach will doubtlessly continue to yield new insights.

Future Directions

It seems likely that infectious diseases will continue to trouble both individuals and communities. Whilst technological advances in new drugs, new vaccines, and better methods for surveillance will undoubtedly assist with the control of infection, several trends in society pull in the opposite direction. Chief amongst these is a growing population, and second is increasing population density as more and more people live in towns and cities. What can infectious disease modeling do to help?



Infectious Disease Modeling. Figure 5
Predictions of a model of within-host evolution and between-host transmission of HIV. (**a**) shows total numbers of susceptible (*blue*) and infectious (*red*) people through time. (**b**) shows the proportion of infections that are with escape mutant virus for a range of escape and reversion rates. The mean times to escape and reversion for each curve are as follows: *red* – escape 1 month, reversion never; *brown* – escape 5 years, reversion never; *yellow* – escape 5 years, reversion 50 years; *pink* – escape 1 year, reversion 10 years; *green* – escape 5 years, reversion 10 years; *mauve* – escape 1 year, reversion 1 year

Models can help in two different ways. The first is to assist the understanding of systems that are intrinsically complicated. Many different interacting populations, events that occur on multiple timescales, and systems with multiple levels of organization can all be better understood when appropriate models are used as an organizing principle and a tool for formal analysis. Sometimes, the problem is that there is not enough data. A systematic description can be very revealing in searching for which new data are most needed. There are also situations where the problem is a deluge of data. In these circumstances, well-constructed models provide a useful organizing scheme with which to interrogate those data.

The second use of models is as representations of well-understood systems used as tools for comparing different intervention strategies. The model of the farm-to-farm spread of FMD described at section "Host Heterogeneity" is a fine example of this use of modeling. It includes enough detail to be a useful tool

for comparing different interventions, but is still firmly rooted in available data so does not rest on large numbers of untested assumptions.

Bibliography

Primary Literature

1. Fenner F (1982) A successful eradication campaign. Global eradication of smallpox. Rev Infect Dis 4(5):916–930
2. Moss WJ, Griffin DE (2006) Global measles elimination. Nat Rev Microbiol 4:900–908
3. Farrington CP, Kanaan MN, Gay NJ (2001) Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. J R Stat Soc Ser C Appl Stat 50:251–292
4. Anderson RM, May RM (1991) Infectious diseases of humans. Oxford University Press, Oxford
5. McLean AR, Anderson RM (1988) Measles in developing countries. Part II: the predicted impact of mass vaccination. Epidemiol Infect 100:419–442
6. McLean AR (1995) After the honeymoon in measles control. Lancet 345(8945):272
7. Babad HE et al (1995) Predicting the impact of measles vaccination in England and Wales: model validation and analysis of policy options. Epidemiol Infect 114:319–344
8. McLean AR (1992) Mathematical modelling of the immunisation of populations. Rev Med Virol 2:141–152
9. Arinaminpathy N, McLean AR (2009) Logistics for control for an influenza pandemic. Epidemics 1(2):83–88
10. Longini IM et al (2005) Containing pandemic influenza at the source. Science 309(5737):1083–1087
11. Anderson RM, Grenfell BT (1986) Quantitative investigations of different vaccination policies for the control of congenital rubella syndrome (CRS) in the United Kingdom. J Hyg 96:305–333, Cambridge University Press
12. Metcalf CJE et al (2010) Rubella metapopulation dynamics and importance of spatial coupling to the risk of congenital rubella syndrome in Peru. J R Soc Interface 8(56):369–376
13. Keeling MJ et al (2001) Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. Science 294(5543):813–817
14. Flynn NM et al (2005) Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. J Infect Dis 191(5):654–665
15. Buchbinder SP et al (2008) Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step Study): a double-blind, randomised, placebo-controlled, test-of-concept trial. Lancet 372(9653):1881–1893
16. Rerks-Ngarm S et al (2009) Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. N Engl J Med 361(23):2209–2220, Epub 2009 Oct 20
17. Council of the Global HIV Vaccine Enterprise (2010) The 2010 scientific strategic plan of the global HIV vaccine enterprise. Nat Med 16(9):981–989
18. Phillips RE et al (1991) Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. Nature 354:453–459
19. Asquith B et al (2006) Inefficient cytotoxic T lymphocyte-mediated killing of HIV-1-infected cells in vivo. PLoS Biol 4:e90
20. Wei X et al (1995) Viral dynamics in human immunodeficiency virus type 1 infection. Nature 373(6510):117–122
21. Ho DD et al (1995) Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. Nature 373(6510):123–126
22. Woolhouse MEJ (2002) Population biology of emerging and re-emerging pathogens. Trends Microbiol 10(10):s3–s7
23. Wolfe ND et al (2007) Origins of major human infectious diseases. Nature 447:279–283
24. Meiering CD, Linial ML (2001) Historical perspective of foamy virus epidemiology and infection. Clin Microbiol Rev 14:165–176
25. Switzer WM et al (2004) Frequent simian foamy virus infection in persons occupationally exposed to nonhuman primates. J Virol 78(6):2780–2789
26. Wolfe ND et al (2004) Naturally acquired simian retrovirus infections in central African hunters. Lancet 363:932–937
27. Briese T et al (2009) Genetic detection and characterization of Lujo virus, a new hemorrhagic fever–associated arenavirus from Southern Africa. PLoS Pathog 5(5):e1000455. doi:10.1371/journal.ppat.1000455
28. Antia R et al (2004) The role of evolution in the emergence of infectious diseases. Nature 426:658–661
29. Arinaminpathy N, McLean AR (2009) Evolution and emergence of novel human infections. Proc R Soc B 273:3075–3083
30. Kubiak R et al (2010) Insights into the evolution and emergence of a novel infectious disease. PLoS Comput Biol 6(9): e10000947
31. Worobey M et al (2008) Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature 455(7213):661–664
32. CDC (1981) Kaposi's sarcoma and *Pneumocystis pneumonia* among homosexual men – New York City and California. MMWR 30:305–308
33. Kawashima Y et al (2009) Adaptation of HIV-1 to human leukocyte antigen class I. Nature 458:641–645
34. Goulder PJ et al (2001) Evolution and transmission of stable CTL escape mutations in HIV infection. Nature 412:334–338
35. Leslie AJ et al (2004) HIV evolution: CTL escape mutation and reversion after transmission. Nat Med 10:282–289
36. Fryer HR et al (2010) Modelling the evolution and spread of HIV immune escape mutants. PLoS Pathog 6(11):e1001196

Books and Reviews

Anderson RM, May RM (1991) Infectious diseases of humans. Oxford University Press, Oxford

Keeling MJ, Rohani R (2008) Modeling infectious diseases in humans and animals. Princeton University Press, Princeton

Nowak MA, May RM (2000) Virus dynamics. Oxford University Press, Oxford

I

Infectious Diseases, Climate Change Effects on

Matthew Baylis, Claire Risley
LUCINDA Group, Institute of Infection and Global Health, University of Liverpool, Neston, Cheshire, UK

Article Outline

Glossary

Climate The weather averaged over a long time or, succinctly, climate is what you expect, weather is what you get!

El Niño Southern Oscillation (ENSO) A climate phenomenon whereby, following reversal of trade winds approximately every 4–7 years, a vast body of warm water moves slowly west to east across the Pacific, resulting in "an El Niño" event in the Americas and leading to a detectable change to climate (mostly disruption of normal rainfall patterns) across 70% of the earth's surface.

Emerging disease An infection or disease that has recently increased in incidence (the number of cases), severity (how bad the disease is), or distribution (where it occurs).

Endemic stability The counter-intuitive situation where the amount of disease rises as the amount of infection falls, such that controlling infection can exacerbate the problem.

Infection The body of a host having been invaded by microorganisms (mostly viruses, bacteria, fungi, protozoa, and parasites).

Infectious disease A pathology or disease that results from infection. Note that many diseases are not infectious and not all infections result in disease.

Intermediate host A host in which a parasite undergoes an essential part of its lifecycle before passing to a second host, and where this passing is passive, that is, not by direct introduction into the next host (see *vector*).

Vector Usually, an arthropod that spreads an infectious pathogen by directly introducing it into a host. For diseases of humans and animals, the most important vectors are flies (like mosquitoes, midges, sandflies, tsetse flies), fleas, lice, and ticks. Aphids are important vectors of diseases in plants. In some instances, other means of carriage of pathogens, such as human hands, car wheels, etc., are referred to as vectors.

Vector competence The proportion of an arthropod vector population that can be infected with a pathogen.

Zoonosis An infection of animals that can spread to, and cause disease in, humans (plural, zoonoses).

Definition of the Subject and Its Importance

Infectious diseases of humans continue to present a significant burden to our health, disproportionately so in the developing world. Infectious diseases of livestock affect their health and welfare, are themselves important causes of human disease and, exceptionally, can threaten our food security. Wildlife infections again present a zoonotic risk to humans, but additionally, such diseases may threaten vulnerable populations and be a cause of extinction and biodiversity loss. Wild populations are inherently more susceptible to environmental change, largely lacking any human protective influence that domesticated species and human populations may benefit from.

Many infectious diseases of humans and farmed or wild animals are influenced by weather or climate, affecting where or when disease occurs, or how severe outbreaks are, and it is therefore likely that future climate change, whether human caused or natural, will have an impact on future disease burdens. Understanding the processes involved may enable prediction of how disease burdens will change in the future and, therefore, allow mitigative or adaptive measures to be put in place.

While climate change will likely be an important cause of change in some infectious diseases in the

future, there are other disease drivers which will also change over similar time scales and which may exacerbate or counteract any effects of climate change. Assessment of the future importance of climate change as an influence over future disease burdens must therefore be considered alongside other causes of change.

Introduction

The impact of infectious diseases of humans and animals seems as great now as it was a century ago. While many disease threats have disappeared or dwindled, at least in the developed world, others have arisen to take their place. Important infectious diseases of humans that have emerged in the last 30 years, for a range of reasons, include Acquired immune deficiency syndrome (AIDS), variant Creutzfeldt-Jakob disease (vCJD), multidrug resistant tuberculosis, severe acute respiratory syndrome (SARS), *E. coli* O157, avian influenza, swine flu, West Nile fever, and Chikungunya [1, 2]. The same applies to diseases of animals: Indeed, all but one of the aforementioned human diseases have animal origins – they are zoonoses – and hence the two subjects of human and animal disease, usually studied separately by medical or veterinary scientists, are intimately entwined.

What will be the global impact of infectious diseases at the end of the twenty-first century? Any single disease is likely to be affected by many factors that cannot be predicted with confidence, including changes to human demography and behavior, new scientific or technological advances including cures and vaccines, pathogen evolution, livestock management practices and developments in animal genetics, and changes to the physical environment. A further, arguably more predictable, influence is climate change.

Owing to anthropogenic activities, there is widespread scientific agreement that the world's climate is warming at a faster rate than ever before [3], with concomitant changes in precipitation, flooding, winds, and the frequency of extreme events such as El Niño. Innumerable studies have demonstrated links between infectious diseases and climate, and it is unthinkable that a significant change in climate during this century will not impact on at least some of them.

How should one react to predicted changes in diseases ascribed to climate change? The answer depends on the animal populations and human communities affected, whether the disease changes in severity, incidence, or spatiotemporal distribution and, of course, on the direction of change: Some diseases may spread but others may retreat in distribution. It also depends on the relative importance of the disease. If climate change is predicted to affect mostly diseases of relatively minor impact on human society or global biodiversity/ecosystem function, while the more important diseases are refractory to climate change's influence, then our concerns should be tempered.

To understand climate change's effects on infectious diseases in the future it is necessary to first understand how climate affects diseases today. This entry begins by first presenting examples of climate's effects on diseases of humans and livestock today and, from the understanding gained, then describes the processes by which climate change might affect such diseases in the future. Diseases of wildlife are important, to some extent, for different reasons to those of humans and livestock, and are therefore considered separately. The relative importance of climate change as a disease driver, compared to other forces, is considered, with examples provided of where climate change both is, and is not, the major force. Finally, the future prospects and the uncertainties surrounding them are considered.

Weather, Climate, and Disease

Many diseases are affected directly or indirectly by weather and climate. Remarkably, no systematic surveys of links between diseases and weather/climate seem to exist and, therefore, it is not possible to indicate whether these diseases represent a minority or majority.

The associations between diseases and weather/climate fall broadly into three categories. The associations may be *spatial*, with climate affecting the distribution of a disease; *temporal* with weather or climate affecting the timing of outbreaks; or they may relate to the *intensity* of an outbreak. Temporal associations can be further broken into at least two subcategories: *seasonal*, with weather or climate affecting the seasonal occurrence of a disease, and *interannual*, with weather or climate affecting the timing, or frequency of years in which outbreaks occur. Here a selection of these associations is presented, which is by no means exhaustive

but is, rather, intended to demonstrate the diversity of effects. Furthermore, the assignment of diseases into the different categories should not be considered hard-and-fast as many diseases could come under more than one heading.

Spatial

- Schistosomosis is an important cause of human mortality and morbidity in Africa and, to a lesser extent, in Asia. The disease is caused by species of *Schistosoma* trematode parasite, for which water-living snails are intermediate hosts. The distribution of suitable water bodies is therefore important for its distribution. However, there must also be suitable temperature: In China, *Oncomelania hupensis* snail intermediate hosts cannot live north of the January 0°C isotherm (the "freezing line") while *Schistosoma japonicum* only develops within the snail at temperatures above 15.4°C. Shistosomosis risk in China is therefore restricted to the warmer southeastern part of the country [4].
- Diseases transmitted by tsetse flies (sleeping sickness, animal trypanosomosis) and ticks (such as anaplasmosis, babesiosis, East Coast fever, heartwater) impose a tremendous burden on African people and their livestock. Many aspects of the vectors' life cycles are sensitive to climate, to the extent that their spatial distributions can be predicted accurately using satellite-derived proxies for climate variables [5].
- Mosquitoes (principally *Culex* and *Aedes*) transmit several viruses of birds that can also cause mortality in humans and horses. Examples are West Nile fever (WNF) and the viral encephalitides such as Venezuelan, western, and eastern equine encephalitis (VEE, WEE, and EEE, respectively) [6]. The spatial distributions of the mosquito vectors are highly sensitive to climate variables.

Temporal-Seasonal

All of the previous examples of spatial associations between diseases and climate can also be classified as temporal-seasonal, as the effects of climate on the seasonal cycle of the intermediate hosts (snails, tsetse flies, and mosquitoes, respectively) also determines in part the seasonal cycle of disease. There are other diseases where the associations can be described as seasonal-temporal.

- Salmonellosis is a serious food-borne disease caused by *Salmonella* bacteria, most often obtained from eggs, poultry, and pork. Salmonellosis notification rates in several European countries have been shown to increase by about 5–10% for each 1°C increase in ambient temperature [7]. Salmonellosis notification is particularly associated with high temperatures during the week prior to consumption of infected produce, implicating a mechanistic effect via poor food handling.
- Foot-and-mouth disease (FMD) is a highly contagious, viral infection of cloven-footed animals, including cattle, sheep, and pigs. Most transmission is by contact between infected and susceptible animals, or by contact with contaminated animal products. However, FMD can also spread on the wind. The survival of the virus is low at relative humidity (RH) below 60% [8], and wind-borne spread is favored by the humid, cold weather common to temperate regions. In warmer drier regions, such as Africa, wind-borne spread of FMD is considered unimportant [9].
- Peste des petits ruminants (PPR) is an acute, contagious, viral disease of small ruminants, especially goats, which is of great economic importance in parts of Africa and the Near East. It is transmitted mostly by aerosol droplets between animals in close contact. However, the appearance of clinical PPR is often associated with the onset of the rainy season or dry cold periods [10], a pattern that may be related to viral survival. The closely related rinderpest virus survives best at low or high relative humidity, and least at 50–60% [11].
- Several directly transmitted human respiratory infections, including those caused by rhinoviruses (common colds) and seasonal influenza viruses (flu) have, in temperate countries, seasonal patterns linked to the annual temperature cycle. There may be direct influences of climate, such as the effect of humidity on survival of the virus in aerosol [12], or indirect influences via, for example, seasonal changes in the strength of the human immune system or more indoor crowding during cold weather [13].

Temporal-Interannual

The previous examples of spatial associations between diseases and climate, which were further categorized as temporal-seasonal, can also be classified as temporal-interannual, as the effects of climate on the intermediate hosts (snails, tsetse flies, and mosquitoes, respectively) will determine in part the risk or scale of a disease outbreak in a given year. There are other diseases where the associations can be described as seasonal-interannual.

- Anthrax is an acute infectious disease of most warm-blooded animals, including humans, with worldwide distribution. The causative bacterium, *Bacillus anthracis* forms spores able to remain infective for 10–20 years in pasture. Temperature, relative humidity, and soil moisture all affect the successful germination of anthrax spores, while heavy rainfall may stir up dormant spores. Outbreaks are often associated with alternating heavy rainfall and drought, and high temperatures [14].
- Cholera, a diarrheal disease which has killed tens of millions of people worldwide, is caused by the bacterium *Vibrio cholerae*, which lives amongst sea plankton [15]. High temperatures causing an increase in algal populations often precede cholera outbreaks. Disruption to normal rainfall helps cholera to spread further, either by flooding, leading to the contamination of water sources, such as wells, or drought which can make the use of such water sources unavoidable. Contaminated water sources then become an important source of infection in people.
- Plague is a flea-borne disease caused by the bacterium *Yersinia pestis*; the fleas' rodent hosts bring them into proximity with humans. In Central Asia, large scale fluctuations in climate synchronize the rodent population dynamics over large areas [16], allowing population density to rise over the critical threshold required for plague outbreaks to commence [17].
- African horse sickness (AHS), a lethal infectious disease of horses, is caused by a virus transmitted by *Culicoides* biting midges. Large outbreaks of AHS in the Republic of South Africa over the last 200 years are associated with the combination of drought and heavy rainfall brought by the warm phase of the El Niño Southern Oscillation (ENSO) [18].
- Rift Valley Fever (RVF), an important zoonotic viral disease of sheep and cattle, is transmitted by *Aedes* and *Culex* mosquitoes. Epizootics of RVF are associated with periods of heavy rainfall and flooding [19–21] or, in East Africa, with the combination of heavy rainfall following drought associated with ENSO [20, 22]. ENSO-related floods in 1998, following drought in 1997, led to an epidemic of RVF (and some other diseases) in the Kenya/Somalia border area and the deaths of more than 2000 people and two-thirds of all small ruminant livestock [23]. Outbreaks of several other human infections, including malaria and dengue fever have, in some parts of the world, been linked to ENSO events.

Intensity

In addition to associations between climate and the spatial and temporal distributions of disease outbreaks, there are some examples of associations between climate and the intensity or severity of the disease that results from infection. It is theoretically possible, for example, that climate-induced immunosuppression of hosts may favor the multiplication of some microparasites (viruses, bacteria, rickettsia, fungi, protozoa), thereby increasing disease severity or, alternatively, reduce the disease-associated immune response to infection, thereby reducing disease severity.

However, the clearest examples pertain to macroparasites (helminth worms) which, with the notable exception of *Strongyloides* spp., do not multiply within the host. Disease severity is therefore directly correlated with the number of parasites acquired at the point of infection or subsequently, and in turn this is frequently associated with climate, which affects both parasite survival and seasonal occurrence.

- Fasciolosis, caused by the *Fasciola* trematode fluke, is of economic importance to livestock producers in many parts of the world and also causes disease in humans. In sheep, severe pathology, including sudden death, results from acute fasciolosis which occurs after ingestion of more than 2,000 metacercariae (larval flukes) of *Fasciola hepatica* at pasture, while milder pathology associated with

subacute and chronic fasciolosis occurs after ingestion of 200–1,000 metacercariae [24]. Acute fasciolosis is therefore most common in places or in years when rainfall and temperature favor the survival of large numbers of metacercariae.

- *Nematodirus battus* is a nematode parasite of the intestine of lambs. Eggs deposited in the feces of one season's lambs do not hatch straightaway; instead, they build up on the pasture during summer and remain as eggs over winter, not hatching until temperatures the following spring exceed 10°C [25]. Once the mean daily temperature exceeds this threshold the eggs hatch rapidly, leading to a sharp peak in the number of infective larvae on the pasture. If this coincides with the new season's lambs grazing on the pasture, there is likely to be a large uptake of larvae and severe disease, called nematodiriasis. If, however, there is a warm spell early in the year, the peak in larvae on pasture may occur while lambs are still suckling rather than grazing, such that fewer larvae are ingested and the severity of nematodiriasis is reduced.

Climate Change and Disease

There is a substantial scientific literature on the effects of climate change on health and disease, but with strong focus on human health and vector-borne disease [5, 26–42]. By contrast, the effects of climate change on diseases spread by other means, or animal diseases in general, have received comparatively little attention [43–48]. Given the global burden of diseases that are not vector-borne, and the contribution made by animal diseases to poverty in the developing world [49], attention to these areas is overdue.

The previous section demonstrates the range of climate influences upon infectious disease. Such influences are not the sole preserve of vector-borne diseases: Food-borne, water-borne, and aerosol-transmitted diseases are also affected. A common feature of non-vector-borne diseases affected by climate is that the pathogen or parasite spends a period of time outside of the host, subject to environmental influence. Examples include the infective spores of anthrax; FMD viruses in temperate regions; the *Salmonella* bacteria that contaminate food products; the cholera-causing *vibrio* bacteria in water; and the moisture- and temperature-dependent survival of the parasites causing schistosomosis and fasciolosis.

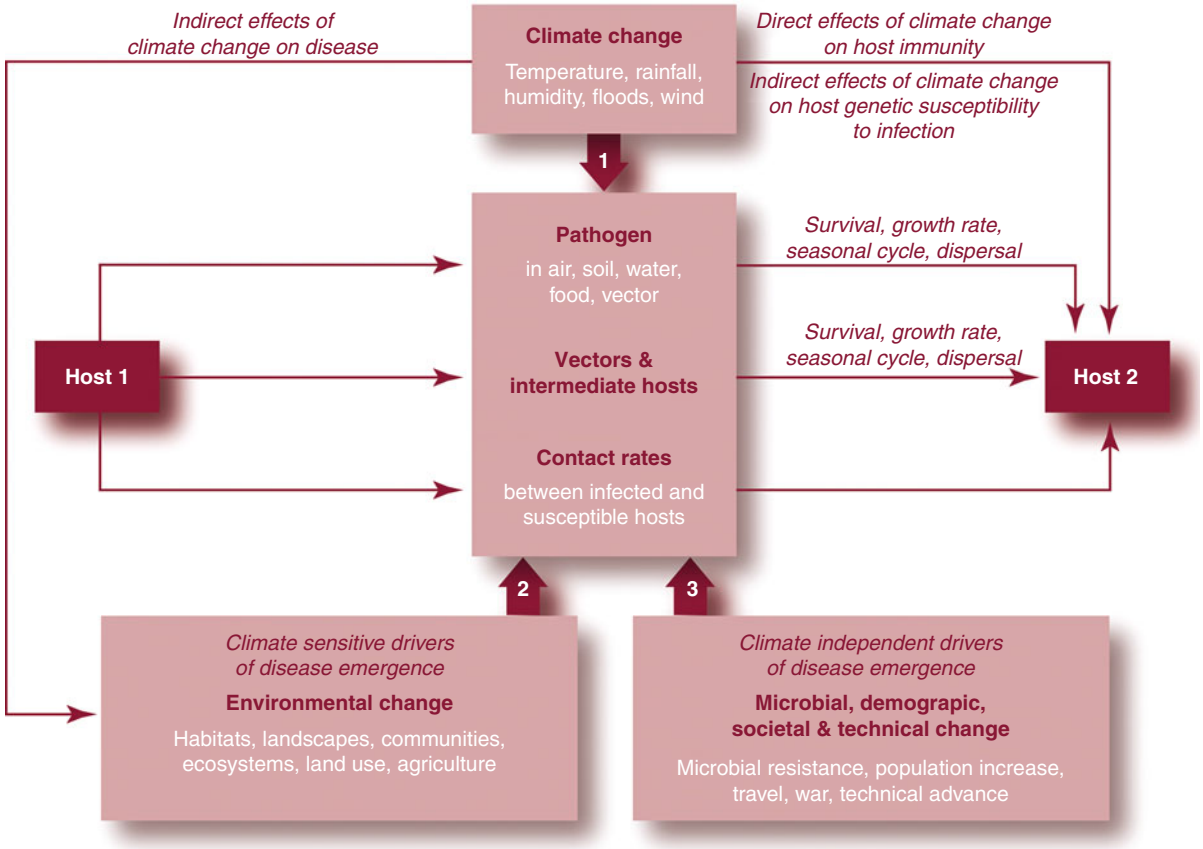
By contrast, most diseases transmitted directly between humans (for example, human childhood viruses, sexually transmitted diseases (STDs), tuberculosis) have few or no reported associations with climate. This is also the case for animal infections such as avian influenza, bovine tuberculosis, brucellosis, Newcastle's disease of poultry, and rabies. Clear exceptions are the viruses that cause colds and seasonal flu in humans, and PPR in small ruminants; these viruses are spread by aerosol between individuals in close contact but are nevertheless sensitive to the effects of ambient humidity and possibly temperature.

The influence of climate on diseases that are not vector-borne appears to be most frequently associated with the timing (intra- or interannual) of their occurrence rather than their spatial distribution. There are examples of such diseases that occur only in certain parts of the world (for example, PPR) but most occur worldwide. By contrast, the associations of vector-borne diseases with climate are equally apparent in time and space, with very few vector-borne diseases being considered a risk worldwide. This is a reflection of the strong influence of climate on both the spatial and temporal distributions of intermediate vectors. If there are exceptions to this rule, the vectors are likely to be lice or fleas, with lives so intimately associated with humans or animals that they are relatively protected from climate's influences.

In the scientific literature, many processes have been proposed by which climate change might affect infectious diseases. These processes range from the clear and quantifiable to the imprecise and hypothetical. They may affect pathogens directly or indirectly, the hosts, the vectors (if there is an intermediate host), epidemiological dynamics, or the natural environment. A framework for how climate change can affect the transmission of pathogens between hosts is shown in Fig. 1. Only some of the processes can be expected to apply to any single infectious disease.

Effects on Pathogens

Higher temperatures resulting from climate change may increase the rate of development of certain pathogens or parasites that have one or more lifecycle stages



Infectious Diseases, Climate Change Effects on. Figure 1

A schematic framework of the effects of climate change on the transmission of diseases of humans and animals. Climate change can act directly on pathogens in a range of external substrates, or their vectors and intermediate hosts, thereby affecting the processes of survival, growth, seasonality, and dispersal. It can also directly affect hosts themselves or the contact rates between infected and susceptible individuals. Climate change can have indirect effects on disease transmission via its effects on the natural or anthropogenic environment, and via the genetics of exposed populations. Environmental, demographic, social, and technical change will also happen independently of climate change and have as great, or much greater, influence on disease transmission than climate change itself. The significance of climate change as a driver of disease will depend on the scale of *arrow 1*, and on the relative scales of *arrows 2* and *3*

outside their human or animal host. This may shorten generation times and, possibly, increase the total number of generations per year, leading to higher pathogen population sizes [48]. Conversely, some pathogens are sensitive to high temperature and their survival may decrease with climate warming.

Phenological evidence indicates that spring is arriving earlier in temperate regions [50]. Lengthening of the warm season may increase or decrease the number of cycles of infection possible within 1 year for warm or cold-associated diseases respectively. Arthropod vectors tend to require warm weather so the infection season of arthropod-borne diseases may extend. Some pathogens and many vectors experience significant mortality during cold winter conditions; warmer winters may increase the likelihood of successful overwintering [41, 48].

Pathogens that are sensitive to moist or dry conditions may be affected by changes to precipitation, soil moisture, and the frequency of floods. Changes to winds could affect the spread of certain pathogens and vectors.

Effects on Hosts

A proposed explanation for the tendency for human influenza to occur in winter is that the human immune system is less competent during that time; attributable to the effects of reduced exposure to light on melatonin [51] or vitamin D production [52]. The seasonal light/dark cycle will not change with climate change, but one might hypothesize that changing levels of cloud cover could affect exposure in future. A second explanation, the tendency for people to congregate indoors during wintertime, leads to a more credible role for a future influence of climate change.

Mammalian cellular immunity can be suppressed following heightened exposure to ultraviolet B (UV-B) radiation – an expected outcome of stratospheric ozone depletion [53, 54]. In particular, there is depression of the number of T helper 1 lymphocytes, cells which stimulate macrophages to attack pathogen-infected cells and, therefore, the immune response to intracellular pathogens may be particularly affected. Examples of such intracellular pathogens include many viruses, rickettsia (such as *Cowdria* and *Anaplasma*, the causative agents of heartwater and anaplasmosis), *Brucella, Listeria monocytogenes* and *Mycobacterium tuberculosis*, the bacterial agents of brucellosis, listeriosis, and tuberculosis, respectively, and the protozoan parasites *Toxoplasma gondii* and *Leishmania* which cause toxoplasmosis and visceral leishmanosis (kala-azar), respectively, in humans [55].

A third host-related effect worthy of consideration is genetic resistance to disease. Some human populations and many animal species have evolved a level of genetic resistance to some of the diseases to which they are commonly exposed. Malaria presents a classic example for humans, with a degree of resistance to infection in African populations obtained from heterozygosity for the sickle-cell genetic trait. Considering animals, wild mammals in Africa may be infected with trypanosomes, but rarely show signs of disease; local Zebu cattle breeds, which have been in the continent for millennia, show some degree of trypanotolerance (resistance to disease caused by trypanosome infection); by contrast, recently introduced European cattle breeds are highly susceptible to trypanosomosis. In stark contrast, African mammals proved highly susceptible to rinderpest which swept through the continent in the late nineteenth century, and which they had not previously encountered. It seems unlikely that climate change will directly affect genetic or immunologic resistance to disease in humans or animals. However, significant shifts in disease distributions driven by climate change pose a greater threat than simply the exposure of new populations. Naïve populations may, in some cases, be particularly susceptible to the new diseases facing them.

Certain diseases show a phenomenon called *endemic stability*. This occurs when the severity of disease is less in younger than older individuals, when the infection is common or endemic and when there is life-long immunity after infection. Under these conditions most infected individuals are young, and experience relatively mild disease. Counter-intuitively, as endemically stable infections become rarer, a higher proportion of cases are in older individuals (it takes longer, on average, to acquire infection) and the number of cases of severe disease rises. Certain tick-borne diseases of livestock in Africa, such as anaplasmosis, babesiosis, and cowdriosis, show a degree of endemic stability [56], and it has been proposed to occur for some human diseases like malaria [57]. If climate change drives such diseases to new areas, nonimmune individuals of all ages in these regions will be newly exposed, and outbreaks of severe disease could follow.

Effects on Vectors

Much has already been written about the effects of climate change on invertebrate disease vectors. Indeed, this issue, especially the effects on mosquito vectors, has dominated the debate so far. It is interesting to bear in mind, however, that mosquitoes are less significant as vectors of animal disease than they are of human disease (Table 1). Mosquitoes primarily, and secondarily lice, fleas, and ticks, transmit between them a significant proportion of important human infections. By contrast, biting midges, brachyceran flies (e.g., tabanids, muscids, myiasis flies, hippoboscids), ticks, and mosquitoes (and, in Africa, tsetse) all dominate as vectors of livestock disease. Therefore, a balanced debate on the effects of climate change on human and animal disease must consider a broad range of vectors.

There are several processes by which climate change might affect disease vectors. First, temperature and moisture frequently impose limits on their distribution. Often, low temperatures are limiting because of high winter mortality, or high temperatures because they involve excessive moisture loss. Therefore, cooler regions which were previously too cold for certain vectors may begin to allow them to flourish with

Infectious Diseases, Climate Change Effects on. Table 1 The major diseases transmitted by arthropod vectors to humans and livestock (Adapted from [58])

Vector	Diseases of humans	Diseases of livestock
Phthiraptera (Lice)	Epidemic typhus Trench fever Louse-borne relapsing fever	
Reduvidae (Assassin bugs)	Chagas' disease	
Siphonaptera (Fleas)	Plague Murine typhus Q fever Tularaemia	Myxomatosis
Psychodidae (Sand flies)	Leishmanosis Sand fly fever	Canine leishmanosis Vesicular stomatitis
Culicidae (Mosquitoes)	Malaria Dengue Yellow fever West Nile Filiariasis Encephalitides (WEE, EEE, VEE, Japanese encephalitis, Saint Louis encephalitis) Rift Valley fever	West Nile fever Encephalitides Rift Valley fever Equine infectious anemia
Simulidae (Black flies)	Onchocercosis	Leucocytozoon (birds) Vesicular stomatitis
Ceratopogonidae (Biting midges)		Bluetongue African horse sickness Akabane Bovine ephemeral fever
Glossinidae (Tsetse flies)	Trypanosomosis	Trypanosomosis
Tabanidae (Horse flies)	Loiasis	Surra Equine infectious anemia *Trypanosoma vivax*
Muscidae (Muscid flies)	Shigella *E. coli*	Mastitis Summer mastitis Pink-eye (IBK)
Muscoidae, Oestroidae (Myiasis-causing flies)	Bot flies	Screwworm Blow fly strike Fleece rot
Hippoboscoidae (Louse flies, keds)		Numerous protozoa
Acari (Mites)	Chiggers Scrub typhus (tsutsugamushi) Scabies	Mange Scab Scrapie?

Infectious Diseases, Climate Change Effects on. Table 1 (Continued)

Vector	Diseases of humans	Diseases of livestock
Ixodidae (Hard ticks) Argasidae (Soft ticks)	Human babesiosis Tick-borne encephalitis Tick fevers Ehrlichiosis Q fever Lyme disease Tick-borne relapsing fever Tularaemia	Babesiosis East coast fever (Theileriosis) Louping ill African Swine Fever Ehrlichiosis Q fever Heartwater Anaplasmosis Borreliosis Tularaemia

climate change. Warmer regions could become even warmer and yet remain permissive for vectors if there is also increased precipitation or humidity. Conversely, these regions may become less conducive to vectors if moisture levels remain unchanged or decrease, with concomitant increase in moisture stress.

For any specific vector, however, the true outcome of climate change will be significantly more complex than that outlined above. Even with a decrease in future moisture levels, some vectors, such as certain species of mosquito, could become more abundant, at least in the vicinity of people and livestock, if the response to warming is more water storage and, thereby, the creation of new breeding sites. One of the most important vectors of the emerging Chikungunya virus (and to a lesser extent dengue virus) is the Asian tiger mosquito (*Aedes albopictus*) which is a container breeder and therefore thrives where humans store water. Equally, some vectors may be relatively insensitive to direct effects of climate change, such as muscids which breed in organic matter or debris, and myiasis flies which breed in hosts' skin.

Changes to temperature and moisture will also lead to increases or decreases in the abundance of many disease vectors. This may also result from a change in the frequency of extreme weather events such as ENSO. Outbreaks of several biting midge and mosquito-borne diseases, for example, have been linked to the occurrence of ENSO [18, 22, 59–62] and mediated, at least in part, by increase in the vector population size in response to heavy rainfall, or rainfall succeeding drought, that ENSO sometimes brings [18, 22]. Greater intra- or interannual variation in rainfall, linked or unlinked to ENSO, may lead to an increase in the frequency or scale of outbreaks of such diseases.

The ability of some insect vectors to become or remain infected with pathogens (their vector competence) varies with temperature [63, 64]. In addition to this effect on vector competence, an increase in temperature may alter the balance between the lifespan of an infected vector, its frequency of feeding, and the time necessary for the maturation of the pathogen within it. This balance is critical, as a key component of the risk of transmission of a vector-borne disease is the number of blood meals taken by a vector between the time it becomes infectious and its death [65]. Accordingly, rising ambient temperature can increase the risk of pathogen transmission by shortening the time until infectiousness in the vector and increasing its feeding frequency at a faster rate than it shortens the vector's lifespan, such that the number of feeds taken by an infectious vector increases. However, at even higher temperatures this can reverse [66] such that the number of infectious feeds then decreases relative to that possible at lower temperatures. This point is extremely important, as it means that the risk of transmission of vector-borne pathogens does not uniformly increase with rising temperature, but that it can become too hot and transmission rates decrease. This effect will be most important for short-lived vectors such as biting midges and mosquitoes [30].

Lastly, there may be important effects of climate change on vector dispersal, particularly if there is a change in wind patterns. Wind movements have been associated with the spread of epidemics of many *Culicoides*- and mosquito-borne diseases [67–72].

Effects on Epidemiological Dynamics

Climate change may alter transmission rates between hosts by affecting the survival of the pathogen or the intermediate vector, but also by other, indirect, forces that may be hard to predict with accuracy. Climate change may influence human demography, housing, or movement or be one of the forces that leads to changes in future patterns of international trade, local animal transportation, and farm size. All of these can alter the chances of an infected human or animal coming into contact with a susceptible one. For example, a series of droughts in East Africa between 1993 and 1997 resulted in pastoral communities moving their cattle to graze in areas normally reserved for wildlife. This resulted in cattle infected with a mild lineage of rinderpest transmitting disease both to other cattle and to susceptible wildlife, causing severe disease, for example, in buffalo, lesser kudu, and impala, and devastating certain populations [73].

Indirect Effects

No disease or vector distribution can be fully understood in terms of climate only. The supply of suitable hosts, the effects of co-infection or immunological cross-protection, the presence of other insects competing for the same food sources or breeding sites as vectors [74], and parasites and predators of vectors themselves, could have important effects [48]. Climate change may affect the abundance or distribution of hosts or the competitors/predators/parasites of vectors and influence patterns of disease in ways that cannot be predicted from the direct effects of climate change alone.

Equally, it has been argued that climate change-related disturbances of ecological relationships, driven perhaps by agricultural changes, deforestation, the construction of dams, and losses of biodiversity, could give rise to new mixtures of different species, thereby exposing hosts to novel pathogens and vectors and causing the emergence of new diseases [40]. A possible "example in progress" is the reemergence in the UK of bovine tuberculosis, for which the badger (*Meles meles*) is believed to be a carrier of the causative agent, *Mycobacterium bovis*. Farm landscape, such as the density of linear features like hedgerows, is a risk factor for the disease, affecting the rate of contact between cattle and badger [75]. Climate change will be a force for modifying future landscapes and habitats, with indirect and largely unpredictable effects on diseases.

Other Drivers of Disease

The future disease burden of humans and animals will depend not only on climate change and its direct and indirect effects on disease, but also on how other drivers of disease change over time. Even for diseases with established climate links, it may be the case that in most instances these other drivers will prove to be more important than climate. A survey of 335 events of human disease emergence between 1940 and 2004 classified the underlying causes into 12 categories [2]. One of these, "climate and weather," was only listed as the cause of ten emergence events. Six of these were non-cholera *Vibrio* bacteria which cause poisoning via shellfish or exposure to contaminated seawater; the remaining four were a fungal infection and three mosquito-borne viruses. The other 11 categories included, however, "land use changes" and "agricultural industry changes," with 36 and 31 disease emergence events, respectively, and both may be affected by climate change. The causes of the remaining 77% of disease emergence events – "antimicrobial agent use," "international travel and commerce," "human demography and behavior," "human susceptibility to infection," "medical industry change," "war and famine," "food industry changes," "breakdown of public health," and "bushmeat" – would be expected to be either less or not influenced by climate change. Hence, climate change's indirect effects on human or animal disease may exceed its direct effects, while drivers unsusceptible to climate change may be more important still at determining our disease future.

Climate Change and Disease in Wildlife

Wildlife disease is important for different reasons to those which make disease in humans and domestic

animals important. It has the potential for endangerment of wildlife and can be a source of zoonoses and livestock disease.

Wildlife Disease as a Cause of Endangerment

Disease in wild populations may limit or cause extreme fluctuations in population size [76] and reduce the chances of survival of endangered or threatened species [77]. Indeed, disease can be the primary cause of extinction in animals or be a significant contributory factor toward it. For example:

- The Christmas Island rat, *Rattus macleari*, is believed to have been extinct by 1904. There is molecular evidence that this was caused by introduction of murine trypanosomes apparently brought to the island by black rats introduced shortly before 1904 [78].
- Similarly, the last known Po'o-uli bird (*Melamprosops phaeosoma*) in Hawaii died from avian malaria brought by introduced mosquitoes [79].
- Canine Distemper in the Ethiopian Wolf (*Canis simensis*) has brought about its decline [80].
- Devil facial tumor disease, an aggressive nonviral transmissible parasitic cancer, continues to threaten Tasmanian Devil (*Sarcophilus harrisii*) populations [81].
- The white nose fungus G*eomyces destructans* is decimating bat populations in Northeastern US states and is currently spreading in Europe [82]. This is perhaps the most recent emergence of a disease of concern to wildlife endangerment.

Although disease can cause endangerment and extinction, its importance relative to other causes is uncertain. A review of the causes of endangerment and extinction in the International Union for Conservation of Nature (IUCN) red list of plant and animal species found that disease was implicated in a total of 254 cases, some 7% of the total examined [83]. Although the other factors may be more important overall, disease remains an important cause of endangerment and extinction for certain animal groups. A contender for the single issue of greatest current conservation concern is the epidemic of the chytrid fungus *Batrachochytrium dendrobatidis* in amphibians. With a broad host range and high mortality, this pathogen is likely to be wholly or partly responsible for all recent amphibian extinction events, which, remarkably, comprise 29% of all extinctions attributable to disease since the year 1500 [77].

Wildlife Disease as a Source of Infections for Humans and Livestock

Many diseases of wildlife frequently cross the species barrier to infect humans or domestic animals [84, 85]. Closely related organisms often share diseases. A particular risk to humans is presented by diseases of primates: The human immunodeficiency viruses (HIV) that cause AIDS originated as simian immunodeficiency viruses in African monkeys and apes. Humans acquire or have acquired many other pathogens from mammals other than primates, especially those that humans choose to live close to (livestock, dogs, cats), or that choose to live close to us (rodents, bats) [86]. In addition, there are examples of human infections shared with birds (e.g., avian influenza), and reptiles and fishes (e.g., *Salmonella* spp.). Insects are frequent carriers of pathogens between vertebrate hosts, and there may even be a pathogen transmissible from plants to humans [87].

Wildlife populations are the primary source of emerging infectious diseases in humans. A search of the scientific literature published between 1940 and 2004 attempted to quantify the causes of disease introductions into human populations and found that about 72% were introduced from wildlife [2].

How Climate Change Can Influence Wildlife Disease

The effects of climate change on wildlife disease are important when the changes produced lead to increased risk of endangerment or extinction of the wildlife, or increased transmission risk to humans or domestic animals.

Climate change can increase the threat of endangerment or extinction, via reduction in population size of the wildlife host (by altering habitats, for example), or increase in pathogen range or virulence, such that the persistence of a host population is at risk, and climate change can increase the risk of disease emergence and spread to humans or livestock via change to

the distribution of wildlife hosts, such that encroachment on human or livestock populations is favored.

Changes in species' distribution may arise directly under climate change as a result of an organism's requirement for particular climatic conditions or indirectly via ecological interactions with other species which are themselves affected. Climate change can cause the appearance of new assemblages of species and the disappearance of old communities [88], which can create new disease transmission opportunities or end existing ones.

Climatic factors potentially affecting wildlife disease transmission more directly include the growth rate of the pathogen in the environment or in a cold-blooded (ectothermic) wildlife host (e.g., fish, amphibian, reptile). Therefore, effects which are more marked in wildlife disease in comparison to human or livestock disease include the occurrence of ectothermic hosts, and also the vast range of potential vectors that may transmit disease. It is therefore more difficult to generalize about the effects of climate change on wildlife.

In colder climates, the parasite that causes the most severe form of human malaria, *Plasmodium falciparum* does not develop rapidly enough in its mosquito vectors for there to be sufficient transmission to maintain the parasite. Avian malaria (*Plasmodium relictum*) exhibits an elevational gradient due mainly to temperature and is subject to similar constraints [89].

An example of a wildlife pathogen constrained due to its dependence on environmental transmission is anthrax. Infective spores of anthrax bacilli can lie dormant in soils for decades, becoming dangerous when climatic conditions, particularly precipitation, favor it. It is well established that when a host population size is reduced, the pool of susceptible individuals may be too small for pathogen survival. This effect is particularly acute for host-specific diseases, such as the transmissible cancer of Tasmanian devils, *Sarcophilus harrisii*, DFTD or Devil facial tumor disease (discussed in [81]); at low host population sizes, DFTD may become extinct. By contrast, diseases with a broad host range may threaten individual species down to the last individual. As anthrax has both a broad host range and can lie dormant in the environment, it is a particular threat for species with very low numbers, and is currently a conservation consideration for many species. For example the Javan rhinoceros population is down to fewer than 60 individuals and identified as a priority for conservation (an "EDGE" species) because of its uniqueness and scarcity [90].

Climate change may have particular impact on marine animals, because of the preponderance of ectothermic animals in the sea, the multiple ways in which climate change is predicted to affect the marine environment, and the multiple stresses that marine organisms and ecosystems are already experiencing due to anthropogenic influence. Disease is an important part of this impact. For example, warming of the Pacific in the range of the oyster *Crassostrea virginica* caused range expansion of the protozoan parasite *Perkinsus marinus* probably due to a combination of increased overwinter survival, greater summer proliferation, and increased host density [91]. Coral reefs are also sensitive to at least 12 potential factors associated with climate change: [92] CO_2 concentration, sea surface temperature, sea level, storm intensity, storm frequency, storm track, wave climate, run off, seasonality, overland rainfall, and ocean and air circulation [92]. Although these factors might not all increase levels of disease, the synergism between disease, climate, and other stressors might lead to accelerating degradation of the coral reef habitat.

From a geographic perspective, there is evidence that the greatest change in ecosystems attributable to climate change is likely to be in the tropics; the second being the arctic [88]. The impacts of this change on wildlife disease and its consequences may be particularly great in these two regions, and there is evidence that it is already occurring. The tropics have the most species in imminent danger of extinction [93] while tropical coral reefs comprise much of the biodiversity of the oceans. In addition to extinction risks, tropical forests may also pose a zoonosis risk. An increase in animal-human interaction is likely in tropical forests, which have a diverse fauna subject to increasing human encroachment.

With regard to the Arctic, a model of the effect of global warming on a protostrongylid lung-dwelling nematode *Umingmakstrongylus pallikuukensis*, in Canadian Arctic muskoxen *Ovibos moschatus*, found that warming was likely to significantly influence infection, making the muskoxen more susceptible to predation [94]. Muskoxen were also subject to climate-influenced outbreaks of fatal pneumonia [95]. Indeed,

the combination of climate change's effects on pathogen survival and transmission, and the stress to host species from changing environmental conditions, may have serious impact on arctic populations of fish, muskoxen, sheep, and others [96].

Dependency of Disease on Climate: The Example of Chytridiomycosis in Amphibians

In wildlife epidemiology, the host may be of equal importance to the pathogen and vector when considering the impact of climate, as wildlife may be impacted by climate in more diverse ways than humans or domestic animals, and are subject to much reduced human mitigation of those impacts. The importance of climate in *Batrachochytrium dendrobatidis* epidemiology, the cause of chytridiomycosis, and numerous amphibian extinctions is fiercely debated. Although the pathogen *B. dendrobatidis* is neither vector-borne nor has a prolonged environmental phase, it is affected by temperature because the environment of its ectothermic hosts is not kept constant when external temperature varies. It belongs to a basal group within the fungi and has a brief motile zoospore stage for dispersal, followed by the penetration of the outer layers of amphibian skin and asexual intracellular multiplication [97]. Its growth is limited by warmer temperatures, perhaps because amphibians shed their outer layers of skin more frequently in warmer temperatures [97]. Pounds et al. [98] were the first to make a connection between climate and *B. dendrobatidis*-mediated amphibian extinctions, reporting spatiotemporal associations between warming and last year of detection of frog species. The development rate of the *B. dendrobatidis* fungus depends on summer temperature [99], and its survival is dependent on winter freezing [100]. The fungus appears to cause more mortality in mountainous regions [101], yet may be limited at the upper extremes of altitude. Climate may also affect the impact of the disease due to host factors. The habitat of the golden toad *Bufo periglenes* in 1987, the last year of its existence, was much reduced due to an especially dry summer. This may have caused crowded conditions in the remaining ponds, facilitating the spread of chytridiomycosis [102].

In addition, climate may affect mortality associated with the disease. The mortality of frogs exposed to *B. dendrobatidis* spores as adults has been shown to be dependent on the condition of the frogs [103]. In a changing climate where amphibians are shifting their ranges into suboptimal areas, hosts are likely to be more susceptible to the damaging effects of *B. dendrobatidis* infection.

On the other hand, it has been argued that climate is not important in *B. dendrobatidis* outbreaks [104]. The authors contrast the climate-linked epidemic hypothesis with the hypothesis that disease outbreaks occur largely due to introduction into unexposed, naive populations, and describe spatiotemporal "waves" of declines across Central America as evidence that it is disease introduction (and not climatic variables) causing declines.

Evidence of Climate Change's Impact on Disease

Climate warming has already occurred in recent decades. If diseases are sensitive to such warming, then one might expect a number of diseases to have responded by changing their distribution, frequency, or intensity. A major difficulty, however, is the attribution of any observed changes in disease occurrence to climate change because, as shown above, other disease drivers also change over time. It has been argued that the minimum standard for attribution to climate change is that there must be known biological sensitivity of a disease or vector to climate, and that the change in disease or vector (change in seasonal cycle, latitudinal or altitudinal shifts) should be statistically associated with observed change in climate [28]. This has been rephrased as the need for there to be change in both disease/vector and climate at the same time, in the same place, and in the "right" direction [105]. Given these criteria, few diseases make the standard: Indeed, only a decade ago one group concluded that the literature lacks strong evidence for an impact of climate change on even a single vector-borne disease, let alone other diseases.

This situation has changed. One disease in particular, bluetongue, has emerged dramatically in Europe over the last decade and this emergence can be attributed to recent climate change in the region. It satisfies the right time, right place, right direction criterion [64], but in fact reaches a far higher standard: A model for the disease, with variability in time driven

only by variation in climate, produces quantitative estimates of risk which fit closely with the disease's recent emergence in both space and time.

Bluetongue

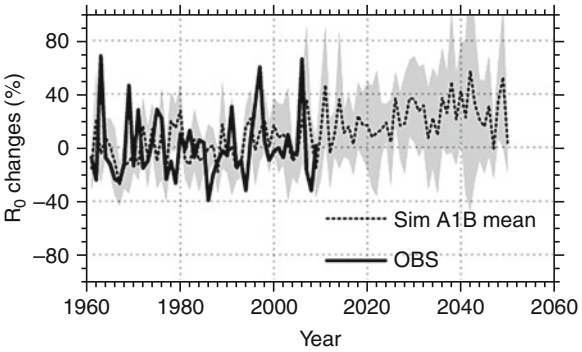
Bluetongue is a viral disease of sheep and cattle. It originated in Africa, where wild ruminants act as natural hosts for the virus, and where a species of biting midge, *Culicoides imicola*, is the major vector [106]. During the twentieth century bluetongue spread out of Africa into other, warm parts of the world, becoming endemic in the Americas, southern Asia, and Australasia; in most of these places, indigenous *Culicoides* became the vectors. Bluetongue also occurred very infrequently in the far extremes of southern Europe: once in the southwest (southern Spain and Portugal, 1955–1960), and every 20–30 years in the southeast (Cyprus, 1924, 1943–46); Cyprus and Greek islands close to Turkey (1977–1978); the presence of *C. imicola* was confirmed in both areas and this species was believed to be the responsible vector. Twenty years after this last 1977–1978 outbreak, in 1998 bluetongue once again reappeared in southeastern Europe [107]. Subsequent events, however, are unprecedented.

Between 1998 and 2008 bluetongue accounted for the deaths of more than one million sheep in Europe – by far the longest and largest outbreak on record. Bluetongue has occurred in many countries or regions that have never previously reported this disease or its close relatives. There have been at least two key developments. First, *C. imicola* has spread dramatically, now occurring over much of southern Europe and even mainland France. Second, indigenous European *Culicoides* species have transmitted the virus. This was first detected in the Balkans where bluetongue occurred but no *C. imicola* could be found [108]. In 2006, however, bluetongue was detected in northern Europe (The Netherlands) from where it spread to neighboring countries, the UK and even Scandinavia. The scale of this outbreak has been huge, yet the affected countries are far to the north of any known *C. imicola* territory [109].

Recently, the outputs of new, observation based, high spatial resolution (25 km) European climate data, from 1960 to 2006 have been integrated within a model for the risk of bluetongue transmission, defined by the basic reproduction ratio R_0 [110]. In this model, temporal variation in transmission risk is derived from the influence of climate (mainly temperature and rainfall) on the abundance of the vector species, and from the influence of temperature alone on the ability of the vectors to transmit the causative virus. As described earlier, this arises from the balance between vector longevity, vector feeding frequency, and the time required for the vector to become infectious. Spatial variation in transmission risk is derived from these same climate-driven influences and, additionally, differing densities of sheep and cattle. This integrated model successfully reproduces many aspects of bluetongue's distribution and occurrence, both past and present, in Western Europe, including its emergence in northwest Europe in 2006. The model gives this specific year the highest positive anomaly (relative to the long-term average) for the risk of bluetongue transmission since at least 1960, but suggests that other years were also at much higher-than-average risk. The model suggests that the risk of bluetongue transmission increased rapidly in southern Europe in the 1980s and in northern Europe in the 1990s and 2000s.

These results indicate that climate variability in space and time are sufficient to explain many aspects of bluetongue's recent past in Europe and provide the strongest evidence to date that this disease's emergence is, indeed, attributable to changes in climate. What then of the future? The same model was driven forward to 2050 using simulated climate data from regional climate models. The risk of bluetongue transmission in northwestern Europe is projected to continue increasing up to at least 2050 (Fig. 2). Given the continued presence of susceptible ruminant host populations, the models suggest that by 2050, transmission risk will have increased by 30% in northwest Europe relative to the 1961–1999 mean. The risk is also projected to increase in southwest Europe, but in this case only by 10% relative to the 1961–1999 mean.

The matching of observed change in bluetongue with quantitative predictions of a climate-driven disease model provides evidence for the influence of climate change far stronger than the "same place, same time, right direction" criterion described earlier. Indeed, it probably makes bluetongue the most convincing example of a disease that is responding to climate change. In this respect, bluetongue differs remarkably from another vector-borne disease, malaria.



Infectious Diseases, Climate Change Effects on. Figure 2 Projections of the effect of climate change on the future risk of transmission of bluetongue in northern Europe. The *y-axis* shows relative anomalies (%) with respect to the 1961–1999 time period for the risk of bluetongue transmission, during August-October in northwest Europe, as defined by the basic reproductive ratio, R_0. R_0 was estimated from climate observations (*OBS – thick black line*), and an ensemble of 11 future climate projections (*SimA1B*), for which the *dashed line* presents the mean and the *grey* envelope the spread (Adapted from [110])

Malaria

Some 3.2 billion people live with the risk of malaria transmission, between 350 and 500 million clinical episodes of malaria occur each year and the disease kills at least one million people annually [111]. Of these, each year about 12 million cases and 155,000–310,000 deaths are in epidemic areas [112]. Interannual climate variability primarily drives the timing of these epidemics.

Malaria is caused by *Plasmodium* spp. parasites. Part of the parasite's life cycle takes place within anopheline mosquitoes while the remainder of the life cycle occurs within the human host. The parasite and mosquito life cycles are affected by weather and climate (mainly rain, temperature, and humidity), allowing models of the risk of malaria transmission to be driven by seasonal forecasts from ensemble prediction systems [113], thereby permitting forecasts of potential malaria outbreaks with lead times of up to 4–6 months [114, 115].

Among scientists there are contrasting views about the overall importance of climate on the transmission of malaria, and therefore on the importance of future climate change. Some argue that climate variability or change is the primary actor in any changing transmission pattern of malaria, while others suggest that any changing patterns today or in the foreseeable future are due to non-climate factors [35, 116, 117].

A key insight is that while global temperatures have risen, there has been a net reduction in malaria in the tropics over the last 100 years and temperature or rainfall change observed so far cannot explain this reduction [118]. Malaria has moved from being climate sensitive (an increasing relationship between ambient temperature and the extent of malaria transmission) in the days before disease interventions were widely available to a situation today where regions with malaria transmission are warmer than those without, but within the malaria-affected region, warmer temperatures no longer mean more disease transmission. Instead, other variables affecting malaria, such as good housing, the running of malaria control schemes, or ready access to affordable prophylaxis, now play a greater role than temperature in determining whether there are higher or lower amounts of transmission. This would suggest that the importance of climate change in discussions of future patterns of malaria transmission is likely to have been significantly overplayed.

What is clearly recognized, by all sides in the malaria and climate debate, is that mosquitoes need water to lay their eggs in, and for larval development, and that adult mosquitoes need to live long enough in an environment with high humidity and with sufficiently high temperature for transmission to be possible to the human host. Hence, while the spatial distribution of higher versus lower degrees of malaria transmission appears to have become, in a sense, divorced from ambient temperature, it seems likely that the weather plays as important a role as ever in determining when seasonal transmission will start and end. Climate change may therefore still have a role to play in malaria: not so much affecting where it occurs but, via changing rainfall patterns and mosquito numbers, when or for how long people are most at risk.

Malaria has only recently become confined to the developing world and tropics. It is less than 40 years since malaria was eradicated in Europe and the United States; and the 15°C July isotherm was the northern limit until the mid nineteenth century [119]. Changes in land use and increased living standards, in particular,

acted to reduce exposure to the mosquito vector in these temperate zones, leading ultimately to the final removal of the disease. In the UK, a proportion of the reduction has been attributed to increasing cattle numbers and the removal of marshland [120]. In Finland, changes in family size, improvements in housing, changes in farming practices, and the movement of farmsteads out of villages lead to the disappearance of malaria [121], where it had formerly been transmitted indoors in winter. While future increases in temperature may, theoretically, lead to an increased risk of malaria transmission in colder climes than at present [120, 122], the much-altered physical and natural environment may preclude this risk increasing to a level that merits concern. Once again, a more important future driver of malaria risk, in the UK at least, may be the pressure to return some of our landscape to its former state, such as the reflooding of previously drained marshland.

Future Directions

Climate change is widely considered to be a major threat to human and animal health, and the viability of certain endangered species, via its effects on infectious diseases. How realistic is this threat? Will most diseases respond to climate change, or just a few? Will there be a net increase in disease burden or might as many diseases decline in impact as increase?

The answers to these questions are important, as they could provide opportunities to mitigate against new disease threats, or may provide the knowledge-base required for policy makers to take necessary action to combat climate change itself. However, both the methodology to accurately predict climate change's effects on diseases and, in most cases, the data to apply the methodology to a sufficiently wide-range of pathogens is currently lacking.

The majority of pathogens, particularly those not reliant on intermediate hosts or arthropod vectors for transmission, either do occur, or have the potential to occur, in most parts of the world already. Climate change has the capacity to affect the frequency or scale of outbreaks of these diseases: Good examples would be the frequency of food poisoning events from the consumption of meat (such as salmonellosis) or shellfish (caused by *Vibrio* bacteria).

Vector-borne diseases are usually constrained in space by the climatic needs of their vectors, and such diseases are therefore the prime examples of where climate change might be expected to cause distributional shifts. Warmer temperatures usually favor the spread of vectors to previously colder environments, thereby exposing possibly naïve populations to new diseases.

However, altered rainfall distributions have an important role to play. Many pathogens or parasites, such as those of anthrax, haemonchosis, and numerous vector-borne diseases, may in some regions be subject to opposing forces of higher temperatures promoting pathogen or vector development, and increased summer dryness leading to more pathogen or vector mortality. Theoretically, increased dryness could lead to a declining risk of certain diseases. A good example is fasciolosis, where the lymnaeid snail hosts of the *Fasciola* trematode are particularly dependent on moisture. Less summer rainfall and reduced soil moisture may reduce the permissiveness of some parts of the UK for this disease. The snail and the free-living fluke stages are, nevertheless, also favored by warmer temperatures and, in practice, current evidence is that fasciolosis is spreading in the UK [123].

One way to predict the future for disease in a specific country is to learn from countries that, today, are projected to have that country's future climate [37, 39]. At least some of the complexity behind the multivariate nature of disease distributions should have precipitated out into the panel of diseases that these countries currently face.

For example, in broad terms, the UK's climate is predicted to get warmer, with drier summers and wetter winters, becoming therefore increasingly "mediterranean." It would seem reasonable, therefore, to predict that the UK of the future might experience diseases currently present in, or that occur periodically in, southern Europe. For humans, the best example would be leishmanosis (cutaneous and visceral) [124], while for animals, examples include West Nile fever [125], *Culicoides imicola*–transmitted bluetongue and African horse sickness [41], and canine leishmanosis [126]. The phlebotomid sandfly vectors of the latter do not currently occur in the UK, but they are found widely in southern continental Europe, including France, with recent reports of their detection in

Belgium [127]. The spread of the Asian tiger mosquito into Europe and the recent transmission in Europe of both dengue fever [128] and Chikungunya [129] by this vector are further cause for alarm.

However, the contrasting examples of bluetongue and malaria – one spreading because of climate change and one retreating despite it – show that considerations which focus entirely on climate may well turn out to be false. Why are these two diseases, both vector-borne and subject to the similar epidemiological processes and temperature dependencies, so different with respect to climate change? The answer lies in the relative importance of other disease drivers. For bluetongue, it is difficult to envisage epidemiologically relevant drivers of disease transmission, other than climate, that have changed significantly over the time period of the disease's emergence [64]. Life on the farm for the midges that spread bluetongue is probably not dramatically different today from the life they enjoyed 30 years ago. Admittedly, changes in the trade of animals or other goods may have been important drivers of the increased risk of introduction of the causative viruses into Europe, but after introduction, climate change may be the most important driver of increased risk of spread.

For malaria, change in drivers other than climate, such as land use and housing, the availability of prophylaxis, insecticides and, nowadays, insecticide-treated bed nets, have played far more dominant roles in reducing malaria occurrence than climate change may have played in increasing it. Two key reasons, then, for the difference between the two diseases are, first, that life for the human hosts of malaria has changed more rapidly than that of the ruminant hosts of bluetongue, and second, the human cost of malaria was so great that interventions were developed; while the (previously small) economic burden of bluetongue did not warrant such effort and our ability to combat the disease 5 years ago was not very different from that of 50 years before. The very recent advent of novel inactivated vaccines for bluetongue may now be changing this situation.

This entry began by asking whether climate change will affect most diseases or just a few. The examples of malaria and bluetongue demonstrate that a better question may be as follows: Of those diseases that are sensitive to climate change, how many are relatively free from the effects of other disease drivers such that the pressures brought by a changing climate can be turned into outcomes?

Bibliography

Primary Literature

1. Morens DM, Folkers GK, Fauci AS (2004) The challenge of emerging and re-emerging infectious diseases. Nature 430:242–249
2. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, Daszak P (2008) Global trends in emerging infectious diseases. Nature 451:990–993
3. IPCC (2001) Climate change 2001: the scientific basis. Intergovernmental Panel on Climate Change, Cambridge
4. Zhou XN, Yang GJ, Yang K, Wang XH, Hong QB, Sun LP, Malone JB, Kristensen TK, Bergquist NR, Utzinger J (2008) Potential impact of climate change on schistosomiasis transmission in China. Am J Trop Med 78:188–194
5. Rogers DJ, Packer MJ (1993) Vector-borne diseases, models, and global change. Lancet 342:1282–1284
6. Weaver SC, Barrett ADT (2004) Transmission cycles, host range, evolution and emergence of arboviral disease. Nat Rev Microbiol 2:789–801
7. Kovats RS, Edwards SJ, Hajat S, Armstrong BG, Ebi KL, Menne B (2004) The effect of temperature on food poisoning: a time-series analysis of salmonellosis in ten European countries. Epidemiol Infect 132:443–453
8. Donaldson AI (1972) The influence of relative humidity on the aerosol stability of different strains of foot-and-mouth disease virus suspended in saliva. J Gen Virol 15:25–33
9. Sutmoller P, Barteling SS, Olascoaga RC, Sumption KJ (2003) Control and eradication of foot-and-mouth disease. Virus Res 91:101–144
10. Wosu LO, Okiri JE, Enwezor PA (1992) Optimal time for vaccination against peste des petits ruminants (PPR) disease in goats in the humid tropical zone in southern Nigeria. In: Rey B, Lebbie SHB, Reynolds L (eds) Small ruminant research and development in Africa: proceedings of the first biennial conference of the African small ruminant research network. International Laboratory for Research in Animal Diseases (ILRAD), Nairobi
11. Anderson J, Barrett T, Scott GR (1996) Manual of the diagnosis of Rinderpest. Food and Agriculture Organization of the United Nations, Rome
12. Soebiyanto RP, Adimi F, Kiang RK (2010) Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters. PLoS ONE 5:e9450
13. Lowen AC, Mubareka S, Steel J, Palese P (2007) Influenza virus transmission is dependent on relative humidity and temperature. PLoS Pathog 3:e151
14. Parker R, Mathis C, Looper M, Sawyer J (2002) Guide B-120: anthrax and livestock. Cooperative Extension Service, College of Agriculture and Home Economics, University of New Mexico, Las Cruces

15. Eiler A, Johansson M, Bertilsson S (2006) Environmental influences on *Vibrio* populations in northern temperate and boreal coastal waters (Baltic and Skagerrak Seas). Appl Environ Microbiol 72:6004–6011
16. Kausrud KL, Viljugrein H, Frigessi A, Begon M, Davis S, Leirs H, Dubyanskiy V, Stenseth NC (2007) Climatically driven synchrony of gerbil populations allows large-scale plague outbreaks. Proc R Soc B Biol Sci 274:1963–1969
17. Davis S, Begon M, De Bruyn L, Ageyev VS, Klassovskiy NL, Pole SB, Viljugrein H, Stenseth NC, Leirs H (2004) Predictive thresholds for plague in Kazakhstan. Science 304: 736–738
18. Baylis M, Mellor PS, Meiswinkel R (1999) Horse sickness and ENSO in South Africa. Nature 397:574
19. Davies F, Linthicum K, James A (1985) Rainfall and epizootic Rift Valley fever. Bull World Health Org 63:941–943
20. Linthicum KJ, Anyamba A, Tucker CJ, Kelley PW, Myers MF, Peters CJ (1999) Climate and satellite indicators to forecast Rift Valley fever epidemics in Kenya. Science 285:397–400
21. Linthicum KJ, Bailey CL, Davies FG, Tucker CJ (1987) Detection of Rift Valley fever viral activity in Kenya by satellite remote sensing imagery. Science 235:1656–1659
22. Anyamba A, Linthicum KJ, Mahoney R, Tucker CJ, Kelley PW (2002) Mapping potential risk of Rift Valley fever outbreaks in African savannas using vegetation index time series data. Photogramm Eng Remote Sens 68:137–145
23. Little PD, Mahmoud H, Coppock DL (2001) When deserts flood: risk management and climatic processes among East African pastoralists. Clim Res 19:149–159
24. Behm CA, Sangster NC (1999) Pathology, pathyophysiology and clinical aspects. In: Dalton JP (ed) Fasciolosis. CAB International, Wallingford, pp 185–224
25. Christie MG (1962) On hatching of *Nematodirus battus*, with some remarks on *N. filicollis*. Parasitology 52:297
26. Githeko AK, Lindsay SW, Confalonieri UE, Patz JA (2000) Climate change and vector-borne diseases: a regional analysis. Bull World Health Org 78:1136–1147
27. Hay SI, Cox J, Rogers DJ, Randolph SE, Stern DI, Shanks GD, Myers MF, Snow RW (2002) Climate change and the resurgence of malaria in the East African highlands. Nature 415:905–909
28. Kovats RS, Campbell-Lendrum DH, McMichael AJ, Woodward A, Cox JS (2001) Early effects of climate change: do they include changes in vector-borne disease? Philos Trans R Soc Lond B 356:1057–1068
29. Kovats RS, Haines A, Stanwell-Smith R, Martens P, Menne B, Bertollini R (1999) Climate change and human health in Europe. Br Med J 318:1682–1685
30. Lines J (1995) The effects of climatic and land-use changes on insect vectors of human disease. In: Harrington R, Stork NE (eds) Insects in a changing environment. Academic Press, London, pp 157–175
31. McMichael AJ, Githeko AK (2001) Human health (Chapter 9). In: McCarthy OFCJJ, Leary NA, Dokken DJ, White KS (eds) Climate change 2001: impacts, adaptation, and vulnerability: contribution of working group II to the third assessment report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, pp 453–485
32. Patz JA, Kovats RS (2002) Hotspots in climate change and human health. Br Med J 325:1094–1098
33. Randolph SE (2004) Evidence that climate change has caused 'emergence' of tick- borne diseases in Europe? Int J Med Microbiol 293:5–15
34. Reeves WC, Hardy JL, Reisen WK, Milby MM (1994) Potential effect of global warming on mosquito-borne arboviruses. J Med Entomol 31:323–332
35. Reiter P, Thomas CJ, Atkinson PM, Hay SI, Randolph SE, Rogers DJ, Shanks GD, Snow RW, Spielman A (2004) Global warming and malaria: a call for accuracy. Lancet Infect Dis 4:323–324
36. Rogers DJ, Randolph SE (2000) The global spread of malaria in a future, warmer world. Science 289:1763–1766
37. Rogers DJ, Randolph SE, Lindsay SW, Thomas CJ (2001) Vector-borne diseases and climate change. Department of Health, London
38. Semenza JC, Menne B (2009) Climate change and infectious diseases in Europe. Lancet Infect Dis 9:365–375
39. Sutherst RW (1998) Implications of global change and climate variability for vector-borne diseases: generic approaches to impact assessments. Int J Parasitol 28:935–945
40. WHO (1996) Climate change and human health. World Health Organisation, Geneva
41. Wittmann EJ, Baylis M (2000) Climate change: effects on *Culicoides*-transmitted viruses and implications for the UK. Vet J 160:107–117
42. Zell R (2004) Global climate change and the emergence/ re-emergence of infectious diseases. Int J Med Microbiol 293:16–26
43. Baylis M, Githeko AK (2006) State of science review: the effects of climate change on infectious diseases of animals. Office of Science and Innovation, London
44. Cook G (1992) Effect of global warming on the distribution of parasitic and other infectious diseases: a review. J R Soc Med 85:688–691
45. Gale P, Adkin A, Drew T, Wooldridge M (2008) Predicting the impact of climate change on livestock disease in Great Britain. Vet Rec 162:214–215
46. Gale P, Drew T, Phipps LP, David G, Wooldridge M (2009) The effect of climate change on the occurrence and prevalence of livestock diseases in Great Britain: a review. J Appl Microbiol 106:1409–1423
47. Harvell CD, Kim K, Burkholder JM, Colwell RR, Epstein PR, Grimes DJ, Hofmann EE, Lipp EK, Osterhaus A, Overstreet RM, Porter JW, Smith GW, Vasta GR (1999) Review: marine ecology – emerging marine diseases – climate links and anthropogenic factors. Science 285:1505–1510
48. Harvell CD, Mitchell CE, Ward JR, Altizer S, Dobson AP, Ostfeld RS, Samuel MD (2002) Ecology – climate warming and disease risks for terrestrial and marine biota. Science 296:2158–2162

49. Perry BD, Randolph TF, McDermott JJ, Sones KR, Thornton PK (2002) Investing in animal health research to alleviate poverty. International Livestock Research Institute, Nairobi
50. Walther GR, Post E, Convey P, Menzel A, Parmesan C, Beebee TJC, Fromentin JM, Hoegh-Guldberg O, Bairlein F (2002) Ecological responses to recent climate change. Nature 416:389–395
51. Dowell SF (2001) Seasonal variation in host susceptibility and cycles of certain infectious diseases. Emerg Infect Dis 7:369–374
52. Cannell JJ, Vieth R, Umhau JC, Holick MF, Grant WB, Madronich S, Garland CF, Giovannucci E (2006) Epidemic influenza and vitamin D. Epidemiol Infect 134:1129–1140
53. Aucamp PJ (2003) Eighteen questions and answers about the effects of the depletion of the ozone layer on humans and the environment. Photochem Photobiol Sci 2:9–24
54. de Gruijl FR, Longstreth J, Norval M, Cullen AP, Slaper H, Kripke ML, Takizawa Y, van der Leun JC (2003) Health effects from stratospheric ozone depletion and interactions with climate change. Photochem Photobiol Sci 2:16–28
55. Jankovic D, Liu ZG, Gause WC (2001) Th1- and Th2-cell commitment during infectious disease: asymmetry in divergent pathways. Trends Immunol 22:450–457
56. Eisler MC, Torr SJ, Coleman PG, Machila N, Morton JF (2003) Integrated control of vector-borne diseases of livestock – pyrethroids: panacea or poison? Trends Parasitol 19:341–345
57. Coleman PG, Perry BD, Woolhouse MEJ (2001) Endemic stability – a veterinary idea applied to human public health. Lancet 357:1284–1286
58. Mullen G, Durden L (2002) Medical and veterinary entomology. Academic, Orlando
59. Gagnon AS, Bush ABG, Smoyer-Tomic KE (2001) Dengue epidemics and the El Nino Southern Oscillation. Clim Res 19:35–43
60. Gagnon AS, Smoyer-Tomic KE, Bush ABG (2002) The El Nino Southern Oscillation and malaria epidemics in South America. Int J Biometeorol 46:81–89
61. Hales S, Weinstein P, Souares Y, Woodward A (1999) El Niño and the dynamics of vector borne disease transmission. Environ Health Perspect 107:99–102
62. Kovats RS (2000) El Nino and human health. Bull World Health Org 78:1127–1135
63. Kramer LD, Hardy JL, Presser SB (1983) Effect of temperatures of extrinsic incubation on the vector competence of *Culex tarsalis* for western equine encephalomyelitis virus. Am J Trop Med 32:1130–1139
64. Purse BV, Mellor PS, Rogers DJ, Samuel AR, Mertens PPC, Baylis M (2005) Climate change and the recent emergence of bluetongue in Europe. Nat Rev Microbiol 3:171–181
65. Macdonald G (1955) The measurement of Malaria transmission. Proc R Soc Med Lond 48:295–302
66. De Koeijer AA, Elbers ARW (2007) Modelling of vector-borne disease and transmission of bluetongue virus in North-West Europe. In: Takken W, Knols BGJ (eds) Emerging pests and vector-borne diseases in Europe. Wageningen Academic, Wageningen, pp 99–112
67. Sellers RF (1992) Weather, *Culicoides*, and the distribution and spread of bluetongue and African horse sickness viruses. In: Walton TE, Osburn BI (eds) Bluetongue, African horse sickness and related Orbiviruses. CRC Press, Boca Raton, pp 284–290
68. Sellers RF, Maarouf AR (1991) Possible introduction of epizootic hemorrhagic disease of deer virus (serotype 20) and bluetongue virus (serotype 11) into British Columbia in 1987 and 1988 by infected *Culicoides* carried on the wind. Can J Vet Res 55:367–370
69. Sellers RF, Pedgley DE (1985) Possible windborne spread to Western Turkey of bluetongue virus in 1977 and of Akabane virus in 1979. J Hyg Camb 95:149–158
70. Sellers RF, Pedgley DE, Tucker MR (1977) Possible spread of African horse sickness on the wind. J Hyg Camb 79:279–298
71. Sellers RF, Pedgley DE, Tucker MR (1978) Possible windborne spread of bluetongue to Portugal, June-July 1956. J Hyg Camb 81:189–196
72. Sellers RF, Pedgley DE, Tucker MR (1982) Rift Valley fever, Egypt 1977: disease spread by windborne insect vectors? Vet Rec 110:73–77
73. Kock RA, Wambua JM, Mwanzia J, Wamwayi H, Ndungu EK, Barrett T, Kock ND, Rossiter PB (1999) Rinderpest epidemic in wild ruminants in Kenya 1993–97. Vet Rec 145:275–283
74. Davis AJ, Jenkinson LS, Lawton JH, Shorrocks B, Wood S (1998) Making mistakes when predicting shifts in species range in response to global warming. Nature 391:783–786
75. White PCL, Brown JA, Harris S (1993) Badgers (*Meles meles*), cattle and bovine tuberculosis (*Mycobacterium bovis*) – a hypothesis to explain the influence of habitat on the risk of disease transmission in southwest England. Proc R Soc Lond B 253:277–284
76. Tompkins DM, Dobson AP, Arneberg P, Begon ME, Cattadori IM, Greenman JV, Heesterbeek JAP, Hudson PJ, Newborn D, Pugliese A, Rizzoli AP, Rosa R, Rosso F, Wilson K (2001) Parasites and host population dynamics. In: Hudson PJ, Dobson AP (eds) Ecology of wildlife diseases. Oxford University Press, Oxford, pp 45–62
77. Smith KF, Sax DF, Lafferty KD (2006) Evidence for the role of infectious disease in species extinction and endangerment. Conserv Biol 20:1349–1357
78. Wyatt KB, Campos PF, Gilbert MTP, Kolokotronis SO, Hynes WH, DeSalle R, Daszak P, MacPhee RDE, Greenwood AD (2008) Historical mammal extinction on christmas island (Indian ocean) correlates with introduced infectious disease. PLoS ONE 3(11):e3602
79. Freed LA, Cann RL, Goff ML, Kuntz WA, Bodner GR (2005) Increase in avian malaria at upper elevation in Hawai'i. Condor 107:753–764
80. Haydon DT, Laurenson MK, Sillero-Zubiri C (2002) Integrating epidemiology into population viability analysis: managing the risk posed by rabies and canine distemper to the Ethiopian wolf. Conserv Biol 16:1372–1385

81. McCallum H, Jones M (2006) To lose both would look like carelessness: Tasmanian devil facial tumour disease. PLoS Biol 4:1671–1674
82. Frick WF, Pollock JF, Hicks AC, Langwig KE, Reynolds DS, Turner GG, Butchkoski CM, Kunz TH (2010) An emerging disease causes regional population collapse of a common North American bat species. Science 329:679–682
83. Smith KF, Acevedo-Whitehouse K, Pedersen AB (2009) The role of infectious diseases in biological conservation. Anim Conserv 12:1–12
84. Daszak P, Cunningham AA, Hyatt AD (2000) Wildlife ecology – emerging infectious diseases of wildlife – threats to biodiversity and human health. Science 287:443–449
85. Gortazar C, Ferroglio E, Hofle U, Frolich K, Vicente J (2007) Diseases shared between wildlife and livestock: a European perspective. Eur J Wildl Res 53:241–256
86. Wolfe ND, Dunavan CP, Diamond J (2007) Origins of major human infectious diseases. Nature 447:279–283
87. van der Riet FD (1997) Diseases of plants transmissible between plants and man (phytonoses) exist. Med Hypotheses 49:359–361
88. Williams JW, Jackson ST, Kutzbacht JE (2007) Projected distributions of novel and disappearing climates by 2100 AD. Proc Natl Acad Sci USA 104:5738–5742
89. Benning TL, LaPointe D, Atkinson CT, Vitousek PM (2002) Interactions of climate change with biological invasions and land use in the Hawaiian islands: modeling the fate of endemic birds using a geographic information system. Proc Natl Acad Sci USA 99:14246–14249
90. Isaac NJB, Turvey ST, Collen B, Waterman C, Baillie JEM (2007) Mammals on the EDGE: conservation priorities based on threat and phylogeny. PLoS ONE 2(3):e296
91. Ford SE, Smolowitz R (2007) Infection dynamics of an oyster parasite in its newly expanded range. Mar Biol 151:119–133
92. Sokolow S (2009) Effects of a changing climate on the dynamics of coral infectious disease: a review of the evidence. Dis Aquat Organ 87:5–18
93. Ricketts TH, Dinerstein E, Boucher T, Brooks TM, Butchart SHM, Hoffmann M, Lamoreux JF, Morrison J, Parr M, Pilgrim JD, Rodrigues ASL, Sechrest W, Wallace GE, Berlin K, Bielby J, Burgess ND, Church DR, Cox N, Knox D, Loucks C, Luck GW, Master LL, Moore R, Naidoo R, Ridgely R, Schatz GE, Shire G, Strand H, Wettengel W, Wikramanayake E (2005) Pinpointing and preventing imminent extinctions. Proc Natl Acad Sci USA 102:18497–18501
94. Kutz SJ, Hoberg EP, Polley L, Jenkins EJ (2005) Global warming is changing the dynamics of Arctic host-parasite systems. Proc R Soc Lond B 272:2571–2576
95. Ytrehus B, Bretten T, Bergsjo B, Isaksen K (2008) Fatal pneumonia epizootic in musk ox (*Ovibos moschatus*) in a period of extraordinary weather conditions. EcoHealth 5:213–223
96. Bradley M, Kutz SJ, Jenkins E, O'Hara TM (2005) The potential impact of climate change on infectious diseases of Arctic fauna. Int J Circumpolar Health 64:468–477
97. Berger L, Hyatt AD, Speare R, Longcore JE (2005) Life cycle stages of the amphibian chytrid *Batrachochytrium dendrobatidis*. Dis Aquat Organ 68:51–63
98. Pounds JA, Bustamante MR, Coloma LA, Consuegra JA, Fogden MPL, Foster PN, La Marca E, Masters KL, Merino-Viteri A, Puschendorf R, Ron SR, Sanchez-Azofeifa GA, Still CJ, Young BE (2006) Widespread amphibian extinctions from epidemic disease driven by global warming. Nature 439:161–167
99. Ribas L, Li MS, Doddington BJ, Robert J, Seidel JA, Kroll JS, Zimmerman LB, Grassly NC, Garner TWJ, Fisher MC (2009) Expression profiling the temperature-dependent amphibian response to infection by *Batrachochytrium dendrobatidis*. PLoS ONE 4(12):e8408
100. Gleason FH, Letcher PM, McGee PA (2008) Freeze tolerance of soil chytrids from temperate climates in Australia. Mycol Res 112:976–982
101. Fisher MC, Garner TWJ, Walker SF (2009) Global emergence of *Batrachochytrium dendrobatidis* and amphibian chytridiomycosis in space, time, and host. Annu Rev Microbiol 63:291–310
102. Pounds JA, Crump ML (1994) Amphibian declines and climate disturbance – the case of the golden toad and the harlequin frog. Conserv Biol 8:72–85
103. Garner TWJ, Walker S, Bosch J, Leech S, Rowcliffe JM, Cunningham AA, Fisher MC (2009) Life history tradeoffs influence mortality associated with the amphibian pathogen *Batrachochytrium dendrobatidis*. Oikos 118: 783–791
104. Lips KR, Diffendorfer J, Mendelson JR, Sears MW (2008) Riding the wave: reconciling the roles of disease and climate change in amphibian declines. PLoS Biol 6:441–454
105. Rogers DJ, Randolph SE (2003) Studying the global distribution of infectious diseases using GIS and RS. Nat Rev Microbiol 1:231–237
106. Mellor PS, Boorman J, Baylis M (2000) Culicoides biting midges: their role as arbovirus vectors. Annu Rev Entomol 45:307–340
107. Mellor PS, Wittmann EJ (2002) Bluetongue virus in the Mediterranean basin 1998–2001. Vet J 164:20–37
108. Purse BV, Nedelchev N, Georgiev G, Veleva E, Boorman J, Denison E, Veronesi E, Carpenter S, Baylis M, Mellor PS (2006) Spatial and temporal distribution of bluetongue and its Culicoides vectors in Bulgaria. Med Vet Entomol 20:335–344
109. Mellor PS, Carpenter S, Harrup L, Baylis M, Mertens PPC (2008) Bluetongue in Europe and the Mediterranean basin: history of occurrence prior to 2006. Prev Vet Med 87:4–20
110. Guis H, Caminade C, Calvete C, Morse AP, Tran A, Baylis M (2011) Modelling the effects of past and future climate on the risk of bluetongue emergence in Europe. J Roy Soc Interface (in press)
111. WHO (2005) World Malaria report, rollback Malaria programme. World Health Organisation, Geneva

I

112. Worrall E, Rietveld A, Delacollette C (2004) The burden of malaria epidemics and cost-effectiveness of interventions in epidemic situations in Africa. Am J Trop Med 71:136–140
113. Palmer TN, Alessandri A, Andersen U, Cantelaube P, Davey M, Delecluse P, Deque M, Diez E, Doblas-Reyes FJ, Feddersen H, Graham R, Gualdi S, Gueremy JF, Hagedorn R, Hoshen M, Keenlyside N, Latif M, Lazar A, Maisonnave E, Marletto V, Morse AP, Orfila B, Rogel P, Terres JM, Thomson MC (2004) Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). Bull Am Meteorol Soc 85:853–872
114. Morse AP, Doblas-Reyes FJ, Hoshen MB, Hagedorn R, Palmer TN (2005) A forecast quality assessment of an end-to-end probabilistic multi-model seasonal forecast system using a malaria model. Tellus Ser A 57:464–475
115. Jones AE, Morse AP (2010) Application and validation of a seasonal ensemble prediction system using a dynamic malaria model. J Clim 23:4202–4215
116. Lafferty KD (2009) The ecology of climate change and infectious diseases. Ecology 90:888–900
117. Epstein P (2010) The ecology of climate change and infectious diseases: comment. Ecology 91:925–928
118. Gething PW, Smith DL, Patil AP, Tatem AJ, Snow RW, Hay SI (2010) Climate change and the global malaria recession. Nature 465:342–344
119. Reiter P (2008) Global warming and malaria: knowing the horse before hitching the cart. Malar J 7:S3
120. Kuhn KG, Campbell-Lendrum DH, Armstrong B, Davies CR (2003) Malaria in Britain: past, present, and future. Proc Natl Acad Sci USA 100:9997–10001
121. Hulden L (2009) The decline of malaria in Finland – the impact of the vector and social variables. Malar J 8:94
122. Lindsay SW, Hole DG, Hutchinson RA, Richards SA, Willis SG (2010) Assessing the future threat from vivax malaria in the United Kingdom using two markedly different modelling approaches. Malar J 9:70
123. Pritchard GC, Forbes AB, Williams DJL, Salimi-Bejestani MR, Daniel RG (2005) Emergence of fasciolosis in cattle in East Anglia. Vet Rec 157:578–582
124. Dujardin JC, Campino L, Canavate C, Dedet JP, Gradoni L, Soteriadou K, Mazeris A, Ozbel Y, Boelaert M (2008) Spread of vector-borne diseases and neglect of leishmaniasis, Europe. Emerg Infect Dis 14:1013–1018
125. Gould EA, Higgs S, Buckley A, Gritsun TS (2006) Potential arbovirus emergence and implications for the United Kingdom. Emerg Infect Dis 12:549–555
126. Shaw SE, Langton DA, Hillman TJ (2008) Canine leishmaniosis in the UK. Vet Rec 163:253–254
127. Depaquit J, Naucke TJ, Schmitt C, Ferté H, Léger N (2005) A molecular analysis of the subgenus *Transphlebotomus* Artemiev, 1984 (Phlebotomus, Diptera, Psychodidae) inferred from ND4 mtDNA with new northern records of *Phlebotomus mascittii* Grassi, 1908. Parasitol Res 95:113–116
128. La Ruche G, Souares Y, Armengaud A, Peloux-Petiot F, Delaunay P, Despres P, Lenglet A, Jourdain F, Leparc-Goffart I, Charlet F, Ollier L, Mantey K, Mollet T, Fournier JP, Torrents R, Leitmeyer K, Hilairet P, Zeller H, Van Bortel W, Dejour-Salamanca D, Grandadam M, Gastellu-Etchegorry M (2010) First two autochthonous dengue virus infections in metropolitan France, September 2010. Euro Surveill 15:2–6
129. Eitrem R, Vene S (2008) Chikungunya fever-a threat for Europeans. A review of the recent literature. Parasitol Res 103: S147–S148

Books and Reviews

International Panel on Climate Change (2007) Climate change 2007: impacts, adaptation and vulnerability. Cambridge University Press, Cambridge

Infectious Diseases, Introduction

Phyllis J. Kanki
Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, MA, USA

Infectious diseases of humans and animals are illnesses resulting from an infection caused by the presence or growth of a biological organism, often termed a pathogen, for its disease-causing behavior. The term derives from the transmissibility of the pathogen to others and when this results in large numbers of infections in a region can be responsible for epidemics. Pathogens responsible for infectious diseases can be viruses, bacteria, protozoa, fungi, multicellular parasites, and prions. While antibiotics and vaccines have made major progress in the treatment and prevention of major infectious diseases, largely in the developed world, the developing world still bears a significant burden of disease due to infectious disease pathogens such as malaria, tuberculosis, and the Human Immunodeficiency Virus (HIV). Changes in the environment, zoonotic pathogens and their interaction with human populations, and medical practice including treatment and vaccines are just some examples of determinants that can modulate the impact of infectious diseases, in terms of spread, ability to cause disease, or even response to prevention or treatment measures. The ever-changing dynamic nature of infectious diseases is not only due to some pathogen's intrinsic propensity for diversity and fitness but also complex

lifecycles involving intermediate nonhuman hosts. Therefore, our ability to control or eradicate various infectious diseases must entail new technologies and analytic methods.

There is significant disparity in the burden of infectious diseases globally. According to the 2008 Global Health Observatory report, infectious diseases only account for one of the top ten causes of death in high-income countries of the world, whereas in low- and middle-income countries, four of the ten leading causes of death are infectious diseases [1]. However, the mobility of populations globally has resulted in infectious disease outbreaks such as the H1N1 influenza outbreak in Mexico in March 2009 that led to 28,000 confirmed cases in the United States just 3 months later. The WHO raised the pandemic alert level to phase 6, the highest level indicating a global pandemic, because of widespread infection beyond North America to Australia, the United Kingdom, Argentina, Chile, Spain, and Japan. Six months after the initial outbreak in Mexico, H1N1 infection had been confirmed in over 200,000 people from more than 100 countries and several thousand deaths [2]. While influenza virus infections are found in both high- and low- and middle-income countries, the virus responsible for this pandemic appeared to be a novel virus with characteristics of North American and Asian swine influenza viruses, as well as human and avian influenza viruses. Thus, the viruses' propensity for variation through genetic reassortment, various animal reservoirs and their contact with human populations, and the mobility of populations led to an epidemic of global proportions in a short time period.

In the past decade, international donor agencies have supported large-scale programs to address the gap in prevention and treatment of HIV/AIDS, malaria and tuberculosis. The burden of these three infectious diseases is disproportionately high in Africa, where health systems are weak and heavily dependent on foreign aid. The President's Emergency Plan For AIDS Relief initiated in 2005 is the single largest funded program for a disease in the history of US government support, active in 30 countries primarily in Africa and responsible for the initiation of antiretroviral therapy to 3.2 million adults and children with AIDS. A summary of the "► HIV/AIDS Global Epidemic" describes the many challenges posed by the HIV virus, first described in the early 1980s. The HIV/AIDS epidemic most severe in Africa has also led to a concurrent increase in tuberculosis, where the presence of either infection increases the risk of coinfection, and as a result in the past decade, TB incidence has tripled. HIV and TB coinfected patients are more difficult to treat and are responsible for the highest mortality rates in both untreated and treated populations. The complex "► Tuberculosis, Epidemiology of" described by Mario Raviglione and colleagues illustrates both the severity of the public health problem and the efforts by the WHO Stop TB alliance in its control. Development of improved, cost-effective, and point-of-care diagnostics is an emphasis for all three of these pathogens.

The development of drug resistance is another feature common to many infectious disease pathogens. The widespread use of chloraquine to treat malaria in the 1940s and 1950s, led to the detection of chloraquine resistant malaria first in South America and Asia and later in Africa by the late 1970s. It became widespread across Africa within a decade. Continued surveillance for drug resistance is critical to adjust treatment policies and the need for more effective drugs is ever present. In 2006 in Tugela Ferry, South Africa, the interaction between tuberculosis and HIV resulted in the recognition of an "extensively drug resistant" tuberculosis strain (XDR), where the bacteria was not only resistance to the common first line drugs, rifampicin and isoniazid but also to drugs in the quinolone family and at least one of the second line drugs [3]. The XDR tuberculosis epidemic in Tugela Ferry was unusually severe with rapid (~2 weeks) mortality, demonstrating the grave consequences of pathogens that readily evolve under drug pressure. As a result of these biologic propensities, the need for new drugs that target resistant strains is an ongoing process. The cost of second and third line drug therapy is prohibitive in most low-income countries and the need for more efficacious and cost-effective drugs is an important priority. Unfortunately, despite the importance of these pathogens like malaria and TB primarily in low-income countries, major biotechnology firms do not prioritize these diseases agents for diagnostic, vaccine, or drug development. The example of malaria and the structural barriers to solutions for low-income (tropical) settings is well described by J. Kevin Baird in "► Tropical Health and Sustainability."

It is widely believed that prevention measures including vaccines are the most effective means of combating infectious diseases whenever possible and this becomes of paramount importance in infectious diseases with high burden and mortality. In the case of malaria, the ubiquity of the mosquito vector, difficulty in its control, and prevalent drug resistance all lend support for the search for an effective malaria vaccine. As described by Christopher Plowe in "► Malaria Vaccines," a study conducted in a single African village, documented more than 200 variants of blood stage malaria antigens. Thus evidence of the difficulty in developing vaccines that must elicit cross-protective immunity to an ever-expanding set of antigens, such as the multiple parasitic stages of malaria. Despite these many challenges, Christopher Plowe describes progress toward a malaria vaccine that would reduce parasite burden, rather than sterilizing protection, such a vaccine would be an important milestone to be reached in the short-term future of malaria control.

While effective vaccines against poliomyelitis have been available since the 1950s, the global eradication campaign is still in effect, with >99% reduction in the number of cases since 1988 and the inception of the Global Poliomyelitis Eradication Initiative by the World Health Assembly. Indigenous poliovirus remains in only four countries of the world, including Afghanistan, Pakistan, Nigeria, and India. "► Polio and its Epidemiology" by Lester Shulman describes the complexity of a disease system with both natural and vaccine strains of the poliovirus, and the many challenges to its future eradication. The use of the live oral polio vaccine has generated vaccine-derived poliovirus, which contributes to the complex molecular epidemiology of polioviruses in countries with residual infection. The cost and implementation considerations for polio's ultimate eradication are therefore far from simple. It is possible that alternative inactivated vaccines may need to be developed if the ultimate phase out of the current oral polio vaccine is to be considered.

Worldwide, one billion people are infected with pathogens termed neglected tropical diseases, largely in low-income countries. Many infectious diseases in this category are considered waterborne. A comprehensive review of major waterborne diseases is covered in "► Waterborne Infectious Diseases, Approaches to Control" by Fenwick and colleagues. Where the water serves as the habitat for the intermediate animal host or vector and the proximity of human populations facilitates the lifecycle. These include diseases such as schistosomiasis, a protozoa transmitted by snails and guinea worm, transmitted by contaminated water, onchocerciasis or river blindness transmitted by flies, as examples. Protozoal and parasitic infections often have complex lifecycles involving multiple hosts, creating challenges to prevention and treatment. Since 1986, the Carter Foundation has devoted its efforts to neglected tropical diseases such as guinea worm in Africa. More than 3.5 million people were affected by this parasitic roundworm untreatable infection caused by *Dracunculus medenisis* in the 1980s and today, the eradication of this disease through prevention is imminent, despite its neglect in the global health agenda.

Zoonotic diseases are infectious diseases transmitted from animals to humans, and constitute more than half of infectious diseases to humans [4]. There are examples of viruses, bacteria, protozoa, parasites, and prions (transmissible proteins) that are considered zoonotic diseases, where their biology and epidemiology are influenced by the animal host, its behavior, and ecology. Examples such as anthrax (*Bacillus anthrasis*), bovine tuberculosis (*Mycobacterium bovis*), brucellosis (*Brucella* sp), cysticercosis (*Taenia solium*, the pork tapeworm), echinococcosis (*Echinococcus* sp), and rabies virus are endemic in many developing countries of Africa, Asia, and South and Central America. Many of which have poor human and veterinary infrastructure to control these important pathogens. Interdisciplinary research is needed to develop novel and more effective control measures. The divided responsibilities between veterinary and medical governing bodies and resources needs to be further integrated as envisioned by the "One Health" initiatives that study the risks of biological pollution on wildlife and humans.

Climate change has long been considered an important determinant of many infectious diseases but the field has been recently expanding in its scope. Pathogens requiring an intermediate host or insect vector may be particularly sensitive to climate change. Warmer temperatures will be predicted to provide an expanded environment for vectors such as mosquitoes, potentially changing the distribution of vector borne human disease. Climate change has also been associated with the frequency or magnitude of outbreaks of food poisoning

due to salmonellosis in meat or Vibrio infection in shellfish. This field is expanding to consider infectious diseases that are nonvector borne with consideration of climate's impact on seasonality, pathogen replication, dispersal, and survival. However, as described in "► Infectious Diseases, Climate Change Effects on" by Matthew Baylis and Claire Risley, the methodology for predicting climate change's impact on disease is yet to be fully developed and more research is needed to collect data on pathogens that might be influenced.

Disease control in humanitarian emergencies should rely on joint situation analysis and technical support involving experts from related specialties and include the development of standards, guidelines, and tools adapted for field use. Communicable disease epidemiological profiles and risk assessments specific to countries or crisis situations prioritize interventions and provide policy guidance to national authorities and humanitarian partner agencies for the control of communicable diseases in specific settings [5]. As an example, in an 8-week period in 2011, a cholera outbreak was reported in the Democratic Republic of Congo (DRC) and Republic of Congo, a poliomyelitis outbreak in Pakistan, and cases of avian influenza in humans in Egypt. Thus highlighting the ever-changing threat of infectious disease infections on a global and temporal scale.

The dynamic nature of various infectious disease agents is thus evident from a variety of examples, and the harnessing of new technologies for the rapid diagnosis and response to infectious disease agents is described in "► Infectious Diseases, Vibrational Spectroscopic Approaches to Rapid Diagnostics" by Jeremy Driskell and Ralph Tripp. These new high-resolution approaches are being developed and evaluated for both bacterial and viral pathogens. Their further instrumentation and commercialization envisions point-of-care, mobile, and cost-effective spectroscopic based diagnostic methods, which has great potential for the sustainability of infectious disease agent control in our ever-changing environment.

The development of new treatments for current, emerging, and drug resistant infectious disease pathogens is also a priority. In "► Antibiotics for Emerging Pathogens," Vinayak Agarwal and Satish Nair describe improvements and innovations to the approach of identification of antibiotics through metabolic connections between the host and microbe, as well as synthesis and mining of new potential antibiotic candidates. Added to these more conventional approaches is the use of genomics and bioinformatics to identify antibiotic gene clusters and microbial ecological evaluations to better understand the interactions of natural antibiotic with their microbial targets. Future emphasis on narrow spectrum antibiotics coupled with more discriminating diagnostic methods may reduce the emergence of drug resistance already associated with use of broad-spectrum agents.

"► Infectious Disease Modeling," as described by Angela McLean, has become an important methodology to characterize disease spread, both in populations and within a single host. While within-host modeling, often considers the spread of infection within an individual and its interface with the host's immune responses, newer models employ multiple levels simultaneously; such as within-host dynamics and between host transmissions. Modeling has become an even more important tool in characterizing infectious diseases particularly with the challenges of growing population and densities. These methods can organize available data and identify critical missing data. Perhaps most important is the use of modeling techniques to compare or project the impact of various intervention strategies.

Globally, infectious diseases account for more than 17 million deaths each year. While modern medicine and technology have diminished the threat of many of these pathogens in high-income countries, the ever present threats of reemerging infections, population mobility, and pathogen genetic variability are but some of the reasons for the dynamic threat of this broad category of risks to human health. The majority of infectious disease burden remains in the tropics, in low- and middle-income countries with scare resources, infrastructure, and health systems to mount or sustain control efforts in the absence of outside support. It is therefore critical that efforts from the scientific research community and international donor agencies continue to increase their efforts with integrated goals of vigilant surveillance, improved and cost-effective diagnostics, and treatment with a goal of sustainable control.

Cross-References

► Airborne Toxic Chemicals
► Biomarkers and Metabolomics, Evidence of Stress

► Bioremediation and Mitigation
► Biosensors and Bioassays for Ecological Risk Monitoring and Assessment
► CERCLA, Sustainability and Public and Environmental Health
► Ecological and Health Risks at Low Doses
► Ecological Risk Assessment and Animal Models
► Environmental Toxicology: Carcinogenesis
► Environmental Toxicology: Children at Risk
► Microbial Risk Assessment of Pathogens in Water
► Recreational Water Risk: Pathogens and Fecal Indicators
► Science, Policy, and Risk Management: Case of Seafood Safety
► Xenobiotic Protection/Resistance Mechanisms in Organisms

Bibliography

1. WHO. http://www.who.int/gho/mortality_burden_disease/causes_death_2008/en/index.html
2. WHO. Influenza A (H1N1): special insights. http://www.who.int/en/
3. Gandhi NR, Moll A, Sturm AW, Pawinski R, Govender T, Lalloo U, Zeller K, Andrews J et al (2006) Extensively drug-resistant tuberculosis as a cause of death in patients co-infected with tuberculosis and HIV in a rural area of South Africa. Lancet 368: 1575–1580
4. Taylor LH, Latham SM, Woolhouse MEJ (2001) Risk factors for human disease emergence. Philos Trans R Soc Lond B 356:983–989
5. WHO. Global alert and response. http://www.who.int/csr/en/

Infectious Diseases, Vibrational Spectroscopic Approaches to Rapid Diagnostics

Jeremy D. Driskell, Ralph A. Tripp
Department of Infectious Diseases, College of Veterinary Medicine, Animal Health Research Center, University of Georgia, Athens, GA, USA

Article Outline

Glossary

Chemometrics A term to describe the use of multivariate statistics used to extract chemical information.

Fourier-transform infrared spectroscopy (FTIR) A specific technique for acquiring IR absorption spectra in which all wavelengths are simultaneously measured.

Infrared spectroscopy An absorption-based vibrational spectroscopic technique which primarily probes non-polar bonds.

Polymerase chain reaction (PCR) An enzymatic method for amplifying a specific nucleic acid sequence.

Raman spectroscopy A scattering vibrational spectroscopic technique which primarily probes polar bonds.

Surface-enhanced Raman spectroscopy (SERS) A technique used to amplify Raman scattered signal via adsorption to a nanometer-scale metallic surface.

Vibrational (molecular) spectroscopy A general term for the use of light to probe vibrations in a sample as a means of determining chemical composition and structure.

Definition of the Subject and Its Importance

Importance of Rapid Diagnosis of Infectious Diseases

Infectious diseases are a major burden on human health with the World Health Organization (WHO) reporting that infectious diseases are responsible for one in ten deaths in the world's richest nations. The impact of infectious diseases is even greater in poorer regions of the world where six of every ten deaths are caused by a spectrum of infectious diseases that include bacteria, viruses, parasites and fungi. These infectious agents can further be described as classical pathogens, e.g., tuberculosis and malaria, seasonal epidemics, e.g., influenza and rhinoviruses, emerging infectious

disease, e.g., highly pathogenic avian influenza and hemorrhagic fevers, or global pandemics such as the most recent outbreak of novel H1N1 influenza virus. Central to the management of each of these diseases are diagnostics. Early and rapid detection of an infectious agent is not only imperative to prevent the spread of disease, but it is also an essential first step to identify appropriate therapeutics that target the disease, as well as to overcome inappropriate administration of ineffective drugs that may drastically lead to drug-resistant pathogens such as methicillin-resistant *Staphylococcus aureus* (MRSA). This is just a succinct example which highlights the importance of diagnostic testing; however, the sections that follow discuss the current status of diagnostics and introduce an emerging approach to diagnostics based on vibrational spectroscopy which has tremendous potential to significantly advance the field.

Introduction

Classical Culture-Based Diagnostics

Despite the importance of diagnostic tests for infectious diseases, relatively few technological advances have supplanted classical microbiological approaches, e.g., in vitro culture, as a diagnostic standard. Clinical laboratories routinely rely on selective and chromogenic growth media to identify bacterial agents. For example, an αβ−chromogenic medium, which includes two substrates, has been developed to selectively isolate *Salmonella* spp. with 100% sensitivity and 90.5% specificity [1]. More recently chromogenic media have been developed to identify *Staphylococcus aureus* and distinguish methicillin-resistant strains (MRSA) [2, 3]. Culture-based diagnostics provide a method for definitive identification of many bacteria, and the tests are relatively inexpensive; however, the identification process has generally low throughput and substantial time is required for diagnostics. Typically, culture requires 24–72 h to allow the bacteria to grow while slow-growing organisms such as mycobacteria require substantially longer (6–12 week) incubation times. Obviously, this is not ideal as the time frame can delay patient treatment. It should also be noted that not all pathogens can be cultured in a laboratory environment, thus the technique cannot be universally applied. There are several additional drawbacks of culture-based diagnostic methods including the requirement for species specific reagents, appropriate culture and storage environments, and labor intensive procedures.

Antibody-Based Diagnostics

Antigen detection and serology are common approaches used in clinical laboratories as alternative or complementary tools to culture-based detection. Common to both of these methods is the use of antibodies either to directly label and detect the antigen or to capture the host response, e.g., antibody responses to infection. Typically, an enzyme-linked immunosorbent assay (ELISA) is employed for antigen detection in diagnostic laboratories. As a first step, ELISA requires an unknown amount of antigen in a sample to be specifically, via a capture antibody in a sandwich assay format, or nonspecifically, via adsorption, immobilized to a solid phase such as a microtiter plate. After removing excess antigen, a known amount of detection antibody specific to the pathogen is then introduced to bind any immobilized antigen. The detection antibody is either directly labeled with an enzyme, or in an additional step, detected with an enzyme-labeled secondary antibody. After removal of excess reagent, a substrate is introduced to react with the enzyme producing a quantifiable color change. A slight modification of this approach replaces the enzyme with a fluorophor for fluorescence-based readout, eliminating the final substrate incubation step. A similar approach is taken for ELISA-based serological assays in which a known amount of purified antigen is immobilized onto the solid phase and incubated with serum to detect the presence of antibodies. While the procedure requires multitudinous steps, reagents, and substantial labor, ELISA is considered rapid relative to culture-based diagnostics as in many cases the assay can be completed within several hours. ELISA-based assays continue to be an integral part of laboratory diagnostics, but in their original form they are limited to the laboratory.

Lateral flow immunoassays, also called dipstick assays, immunochromatography, sol particle immunoassays, or rapid diagnostic tests, have been developed to overcome many limitations of ELISAs by eliminating the complex multi-step procedure, reducing labor, and allowing field or point-of-care testing. Lateral

I

flow assays, like ELISAs, utilize pathogen-specific antibodies for the direct detection of antigen or detection of antibody response. However, for the case of lateral-flow assays the labeled detection antibody, capture antibody, and control reagents are dried on a prefabricated carrier strip. By design, these assays overcome diffusion-limited kinetics to exploit the rapid kinetics of antibody-antigen recognition [4, 5] to yield results in 10–20 min. Thus, because of the "reagentless" nature and rapid results, these assays are well suited for field use and resource-poor regions where reagent storage and test sites are severely limited. It should be noted, however, that these benefits are at the loss of quantitative information and often a lower threshold of detection.

Numerous lateral flow immunoassays have been developed commercially for clinical diagnostics. Several competing manufacturers offer rapid diagnostic tests for influenza virus in which a conserved antigen is detected in a lateral flow assay format. Some detect influenza A and influenza B without distinction of the subtypes, e.g., QuickVue Influenza Test (Quidel), others detect and differentiate A and B strains, e.g., QuickVue Influenza A + B Test (Quidel) and 3 M™ Rapid Detection Flu A + B Test, and some only identify A or B strains, e.g., SAS Influenza A Test, (SA Scientific). Similarly, commercial rapid diagnostic tests are available for detection of a conserved protein for rotavirus A, e.g., IVD Rotavirus A Testing Kit. Not all lateral flow assays are designed for antigen detection; a rapid diagnostic test developed for leptospirosis diagnosis target anti-*Leptospira* IgM antibodies [6].

Despite continued advancements in antibody-based diagnostics these platforms will always be limited by the need for species-specific reagents, i.e., antibodies where the assays can only perform with the sensitivity and specificity inherent to the antibody. For example, commercial lateral flow assays for influenza only provide 50–70% sensitivity and 90–95% specificity with respect to culture-based diagnosis [7]. While the lateral flow assays may be performed rapidly, their low sensitivity may preclude early diagnosis due to low levels or unsustained levels of antigen through disease available for detection. Moreover, serological-based assays developed to detect agent-specific antibodies require that the infection elicit a detectable sustained immune response before the assay can be performed, a feature which substantially delays diagnosis.

Molecular Diagnostics

Nucleic acid and sequence-based methods for diagnostics offer significant advantages over conventional culture- and antibody-based diagnostics with regard to sensitivity, specificity, speed, cost, and portability. Central to molecular diagnostics is the use of a complementary nucleic acid probe that hybridizes to a unique species-specific region of the infectious agents RNA or DNA. While several molecular platforms have been developed for infectious diseases diagnostics, e.g., fluorescence in situ hybridization (FISH), polymerase chain reaction (PCR) is the most commonly employed molecular method for diagnostics. PCR is a method of amplifying targeted segments of nucleic acid by several orders of magnitude to facilitate detection. In principal, complementary primer sequences are used to hybridize to a target nucleic acid sequence. A thermostable DNA polymerase, e.g., Taq polymerase, is then employed to extend the primer sequence. Thermal control facilitates extension, melting, and annealing, and via temperature-controlled cycling, the number of target sequences increases exponentially with each cycle. Amplification of the target sequence can be read out in an ethidium bromide-stained agarose gel or in real-time via cleavage of a fluorescent tag from the primer during the extension step. Appropriate selection of the primers provides extremely specific detection, while the amplification of the target nucleic acid provides excellent sensitivity.

PCR has been demonstrated to be sensitive to single-copy numbers of DNA/RNA targets. In these most sensitive PCR assays, primers are chosen to fully complement a region of the target sequence. Perfect complement probes are also ideal for maximizing the assay specificity to a particular infectious agent. However, in practice, genetic mutations, particularly prevalent in RNA viruses such as influenza, can render a primer/probe set ineffective for diagnosis. Thus, degenerate probes are sometimes chosen to encompass some genetic diversity at the expense of sensitivity.

Multiplexed PCR utilizes multiple primer/probe sets that target different pathogens. Multiplexed assays are implemented when the sample size is limited, preventing multiple individual singleplex PCR analyses, and the clinician is unable to determine the most likely causative agent based on early clinical

presentation. Multiplexed PCR assays are not quantitative due to target competition for reagents, are typically less sensitive than singleplex assays, and because of increased reagents, are more expensive to perform than singleplex assays. Moreover, multiple PCR products cannot be simultaneously read out by fluorescence, thus microarray analysis or electrophoresis to identify PCR products of different lengths is required to detect multiple PCR products. However, breakthroughs in multiplexed detection and quantitation are forthcoming [8–10].

Thus, for detection and diagnosis of many diseases such as viruses, PCR offers many advantages over classical methods of diagnostics, and its role will continue to expand in clinical diagnostic laboratories. However, there are challenges associated with PCR. For pathogens in which culture and microscopy can be used, molecular diagnostics are not the most cost effective. For example, the cost of a commercial PCR assay for tuberculosis ranges from $40 to $80, whereas microscopy and culture can be implemented for $1 and $5, respectively. Another consideration is how to interpret PCR results. Due to the sensitivity afforded by PCR, extremely low levels of infectious agent can be detected which may be below clinically relevant thresholds for disease presentation. Thus, quantitative PCR rather than qualitative PCR is typically more informative when correlating to clinical diagnosis. PCR assays developed in the laboratory are not always translational to analysis of clinical samples. In general, PCR cannot be performed directly on biological fluids such as blood because compounds such as hemoglobin, lactoferrin, heme, and IgG inhibit amplification [11–13]. Therefore, DNA and RNA are extracted as a first step, prior to PCR, but inefficiency of extraction kits often lead to a decrease in analytical performance compared to laboratory cultures [14, 15]. Moreover, isolation of nucleic acids is time consuming and technically challenging unless automated, which requires expensive equipment and reagents. A final consideration is the need for temperature-sensitive reagents, thermocyclers, skilled workers, and a clean laboratory environment to prevent contamination leading to false-negative results. While tremendous efforts are focused on PCR automation, incorporation of microfluidics [16, 17], and isothermal amplification [18–21], e.g., loop-mediated isothermal amplification (LAMP), to overcome these challenges, the current status of PCR precludes widespread use of PCR diagnostics in point-of-care settings and developing nations.

Limitations of Classical and Conventional Diagnostics

As discussed above, diagnostics are essential to healthcare and disease management. While there is merit in further advancing current diagnostic methods, there are shortcomings for each approach. As noted above, culture is sensitive, but the length of time prevents rapid and early diagnosis. Antibody-based techniques are often limited in sensitivity, and like molecular diagnostics, are expensive, and require species-specific detection reagents. It is likely that any improvements in these methods to address these challenges will be incremental; however there are several next generation diagnostics that offer potential paradigm shifting approaches to detection and diagnoses that are currently being investigated. One important area is the use of molecular or vibrational spectroscopy for "whole-organism" fingerprinting. This innovative approach to diagnostics promises to be rapid, specific, and truly reagentless.

Spectroscopy-Based Diagnostics

Vibrational spectroscopy includes a number of nondestructive analytical techniques which provide molecular information about the chemical makeup, e.g., functional groups, of a sample. Subtle changes in the frequency of a particular functional group vibration, e.g., group frequency, provides additional details of chemical structure, local environment surrounding the bond, bond angle, length, geometry, and conformation. These attributes of vibrational spectroscopy have led to the development of vibrational spectroscopic approaches to generate whole-organisms fingerprints to serve as unique biochemical signatures for pathogen identification. Unique to this approach of infectious agent identification is rapid readout, and perhaps most importantly, there is no need for species-specific reagents or other reagents of any kind. Three of the most developed vibrational spectroscopic techniques include infrared absorption spectroscopy, Raman scattering spectroscopy, and surface-enhanced Raman spectroscopy (SERS). These three methods, as

well as their development for diagnosing infectious diseases are described in detail below.

Infrared Spectroscopy

Infrared spectroscopy is an absorption technique in which the sample is irradiated most commonly with mid-infrared light with wavelengths in the range of 2.5–50 μm. Photons of appropriate energy to bring about a transition from one vibrational state to an excited vibrational state are absorbed by the analyte. Selection rules govern which vibrations are allowed to absorb IR photons, providing chemical and structural information. These rules require a net change in the dipole moment of the molecule as a result of the vibration. Thus, IR absorption spectra are dominated by asymmetrical vibrations. With consideration to these selection rules, proteins and nucleic acids (building blocks of bacteria, viruses, etc.) are excellent absorbers of IR radiation making IR spectroscopy an ideal tool for characterization of infectious agents.

The chemical complexity and sheer size of bacteria and viruses tend to produce complex spectra with broad and overlapping bands. Yet subtle changes in band shape, slight shifts in band peak positions, and variation in relative band intensities provide significant insight into chemical structure. Thus, careful evaluation of the full spectral fingerprint of whole-organisms, rather than analysis of single peaks common to small molecule studies, can lead to identification and classification of microorganisms.

IR spectroscopy as a technique for whole-organism fingerprinting dates back to 1952 [22]. In this early study, Stevenson and Boulduan showed that the IR spectra for *Escherichia coli*, *Pseudomonas aeruginosa*, *Bacillus subtilis*, and six other species of cultured bacteria are unique to each species. In addition to species identification, six strains of *Bacterium tularense* were spectroscopically differentiated. This initial work did not immediately translate to the diagnostic applications of IR spectroscopy. Two major breakthroughs that did not occur for several decades after the original findings were essential to further increase the utility of IR for whole-organism fingerprinting. First, prior to the 1980s when Fourier transform infrared spectroscopy (FTIR) was introduced, dispersive instruments were typically used. Dispersive instruments do not provide the speed or analytical performance required for IR-based diagnostics. FTIR instruments provide much better signal-to-noise spectra with improved spectral resolution, and are acquired in less time. These advantages provided by FTIR were essential to accurately analyze these complex biological spectra by distinguishing subtle changes in spectral band shape and to rapidly collect data. Additional developments in the methods of data analysis were also essential to move IR whole-organism fingerprinting forward. The large number of variables, i.e., wavelengths, each containing relevant information, inherent to IR spectra and rather subtle changes in intensity prevented traditional single-peak analysis for bacterial identification. Moreover, visual inspection of the spectra by the analyst is too labor intensive, prone to operator error, and unrealistic for large numbers of spectra and/or organisms from which to identify. The introduction of chemometrics, i.e., multivariate statistics applied to spectroscopy, led to the continued interest and advancement of IR-based diagnostics. These methods, including principal component analysis [23, 24], discriminant analysis, multiple regression analysis [25], and artificial neural networks [26, 27], function to simplify the high dimensional dataset by identifying the most significant variables with the ultimate goal of sample identification or quantification.

FTIR has received the greatest attention with respect to vibrational spectroscopic techniques over the last 2 decades and has been developed to the point that it can be considered an established method for the identification of both bacterial species and strains [28–33]. Researchers continue to investigate laboratory cultures of bacteria in an effort to standardize sampling protocols so that spectral databases can be shared among laboratories and to standardize the methods of analysis including spectral preprocessing, feature selection, and classification algorithms. As the method has matured, and while it continues to be tweaked and standardized, FTIR is now being applied to the analysis of clinical specimens. For example, FTIR has been successfully used to analyze clinical sputum samples collected from cystic fibrosis patients [26]. This FTIR capability is important as historically, lung infection caused by *Pseudomonas aeruginosa*, *Staphylococcus aureus*, and *Haemophilus influenzae* were the major causes of morbidity and mortality in cystic

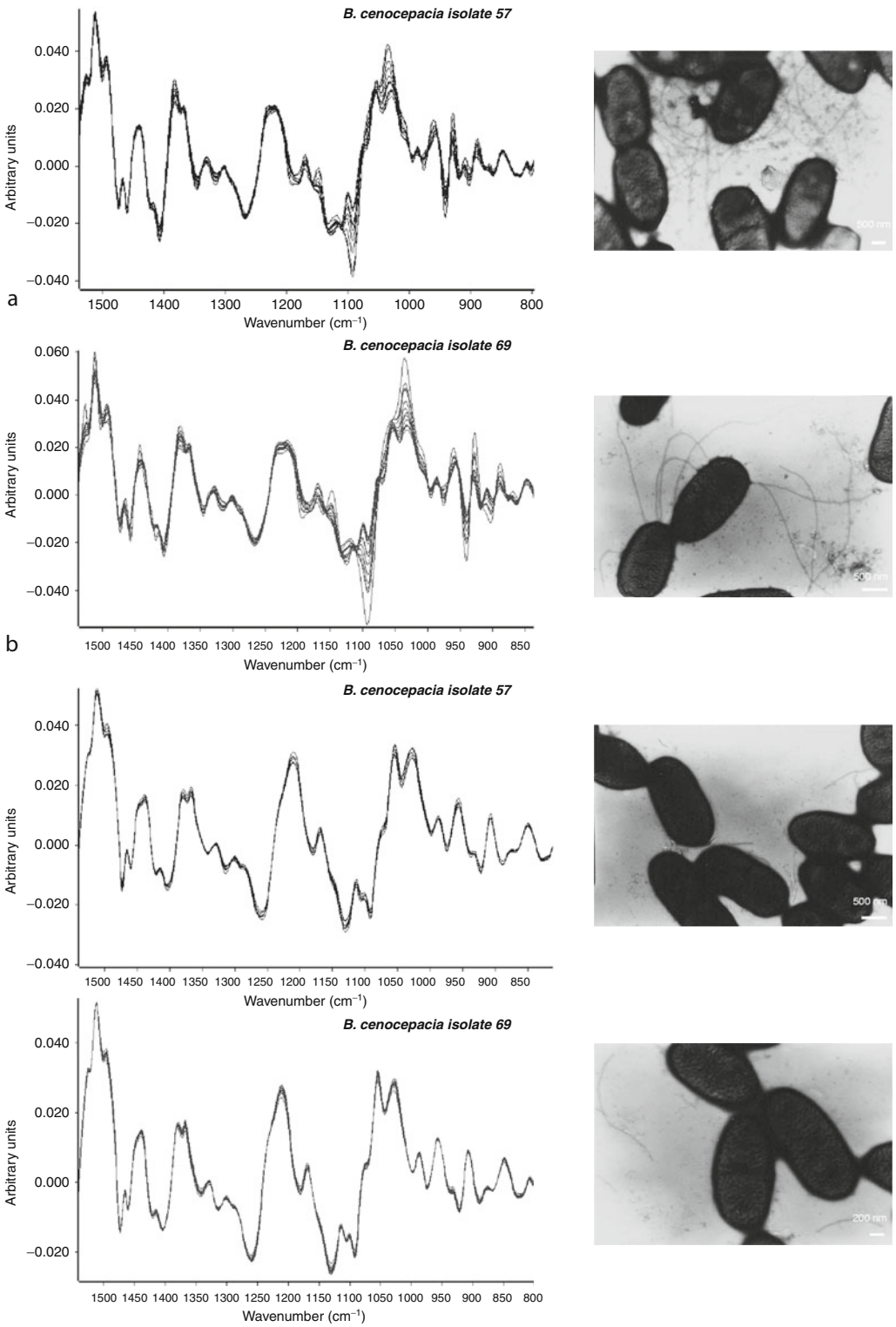
fibrosis patients; however, the number of emerging nonfermenting species is on the rise [34], and many of these species are closely related and not appropriately identified using typical clinical diagnostics and microbiological approaches [35]. Using a FTIR spectral library and an artificial neural network built for pathogen identification, the results from the FTIR method were compared to conventional microbiology detection methods. A two-tiered ANN classification scheme was built in which the top-level network identified *P. aeruginosa*, *S. maltophilia*, *Achromobacter xylosoxidans*, *Acinetobacter* spp., *R. pickettii*, and *Burkholderia cepacia* complex (BCC) bacteria. The second-level network differentiated among four species of BCC, *B. cepacia*, *B. multivorans*, *B. cenocepacia*, and *B. stabilis*. Ultimately, this method resulted in identification success rates of 98.1% and 93.8% for the two ANN levels, respectively. However, before this optimized method was established, the research highlighted three important considerations. First, not all bacterial isolates produce poly-β-hydroxybutyric acid (PHB) which contributes to the IR spectra and confounds classification. To overcome this, each isolate was cultured on TSA medium and harvested after 5 h of growth, prior to the expression of PHB. This step enriches the bacteria for analysis and eliminates interference from PHB. Second, flagella or pilus fibers were determined to contribute to spectral heterogeneity. Vigorous vortexing and subsequent centrifugation removes the fibers to significantly improve spectral reproducibility and classification results (Fig. 1). Third, the classification algorithm significantly affects the classification results. The authors show that hierarchical clustering algorithms (HCA) discriminate between reference and clinical strains rather than based on bacterial identity. Advanced methods, such as ANN, that determine spectral variables that vary only as a result of the bacteria was necessary to correctly classify according to strain. This example work demonstrates the power of IR-based diagnostics, but suggests that these methods may require problem-specific standardization of experimental protocols and data analysis.

These groundbreaking efforts to develop IR for bacterial analysis have led to the realization that spectroscopic methods have advantages for exploring detection of other pathogens. For example, FTIR has been employed for the distinction of yeast and fungi with success [36, 37]. More recently IR has been investigated as a method to detect viral infections, although current experiments are limited to viral infection of cells in culture [38–41]. While mock-infected and herpes simplex virus type 1-infected Vero cells are readily distinguished via IR, infection-induced spectral changes are inconsistent [39, 40]. Thus, substantially more research effort is necessary to standardize protocols and correlate the spectral response to the biochemical response upon infection.

Reports continue to support the utility of FTIR-based diagnostics in the clinical laboratory, but there are certain limitations to consider. First, water is a particularly strong absorber of IR light. Thus, care must be taken to completely dehydrate the sample prior to data acquisition. This obviously does not prevent IR-based diagnostics, it is merely an inconvenience. Second, IR absorption spectroscopy is not an inherently sensitive method and trace levels of a pathogen are not readily apparent. Hence, clinical samples will likely require a culture step to generate sufficient biomass for IR analysis. As noted above, this sample enrichment can be as short as 5 h, and with IR data acquisition on the order of minutes, the total analysis time is still more rapid, less labor intensive, and more informative in many cases than conventional diagnostic methods and does not require species-specific reagents.

Raman Spectroscopy

Raman spectroscopy is a scattering technique, in which the sample is irradiated with a monochromatic light source, almost always a laser. The majority of the scattered photons are elastically scattered and maintain the same frequency as the excitation source; however, a small fraction of the photons are shifted in frequency relative to the excitation source. The difference in the energy between the excitation and inelastically, i.e., Raman, scattered photons correspond to the energy necessary to bring about a transition from one vibrational state to an excited vibrational state. Thus, much like IR spectroscopy, Raman spectra provide insight into the chemical structure, local environment, geometry, and conformation of the sample and can serve as a whole-organism fingerprinting method. Selection rules also govern which vibrations are Raman active.



Infectious Diseases, Vibrational Spectroscopic Approaches to Rapid Diagnostics. Figure 1 (Continued)

These rules require a change in the polarizability during the vibration to be Raman active. Thus, Raman spectra are dominated by symmetrical vibrations and the technique is often seen as a complementary rather than competing technique with IR spectroscopy. However, for application to the analysis of biological materials and whole-organism fingerprinting methods, Raman offers many inherent advantages over IR spectroscopy.

Because of the selection rules, the main chain and aromatic side chains of peptides rather than aliphatic side chains are probed via Raman scattering in contrast to IR. Raman bands of nucleic acids are limited to heterocyclic bases or phosphodiester groups making up the backbone. Raman bands are narrower and less likely to overlap, thus the spectra are much less complicated compared to IR spectra because of the many more nonsymmetric vibrations that are possible. Another major advantage of Raman is that water does not interfere since its vibrations do not fit the selection rules criteria. This is an extremely important consideration when analyzing biological samples which are endemic to aqueous environments. Other advantages of Raman include the flexibility to analyze samples in any state, e.g., gas, liquid, or solid, and the ability to analyze small sample volumes and masses because of the tight focus of incident laser light (square microns) compared to the incident IR beams (square centimeters).

Viruses were the first infectious agent analyzed by Raman spectroscopy, although not in a diagnostic capacity [42]. In this first work, Raman spectroscopy was used to probe the RNA and protein structure upon viral packaging. In the 1970s, Raman spectroscopy suffered from poor sensitivity due to instrument limitations. The first evaluation of Raman spectroscopy for pathogen detection was not until 1987 when spectra were collected for five species of bacteria including *E. coli*, *P. fluorescens*, *S. epidermidis*, *B. subtilis*, and *E. cloacae* [43]. To overcome the limited sensitivity of the instruments at the time, an ultraviolet laser was used for excitation to enhance spectral features of RNA, DNA, tyrosine, and tryptophan via resonance Raman. Unique spectra were observed for each bacterium, although analysis relied on visual interpretation since chemometrics had not been implemented for spectral analysis yet. UV Raman instruments, while producing the requisite sensitivity for pathogen analysis, is quite expensive and non-resonant vibrations are not observed which results in a significant loss in information that is valuable for differentiation.

Despite the recognized benefits of Raman-based diagnostics, particularly when compared to conventional and IR-based diagnostics, instrumentation has limited the maturation of Raman-based diagnostics. After development of UV Raman for pathogen detection [43–46], Fourier transform Raman (FT-Raman) instruments were introduced for microbiological studies which increased instrument sensitivity [47, 48]. Raman instruments have now evolved to include NIR lasers to reduce fluorescence from biological and NIR-sensitive CCD detectors. These modern instruments have only been developed in this decade to fully explore the potential of Raman as a diagnostic technique [49–55]. Thus Raman-based whole-organism fingerprinting is less developed than IR-based methods and examples are generally limited to the analysis of laboratory cultures.

In an early study, Maquelin et al. [54] utilized Raman spectroscopy to directly analyze five bacterial strains, including three strains of *Staphylococcus* spp., *E. coli*, and *E. faecium*, on solid culture medium. The flexibility in sample type afforded by Raman spectroscopy allowed direct measurement on the culture plate that would not be possible using IR spectroscopy.

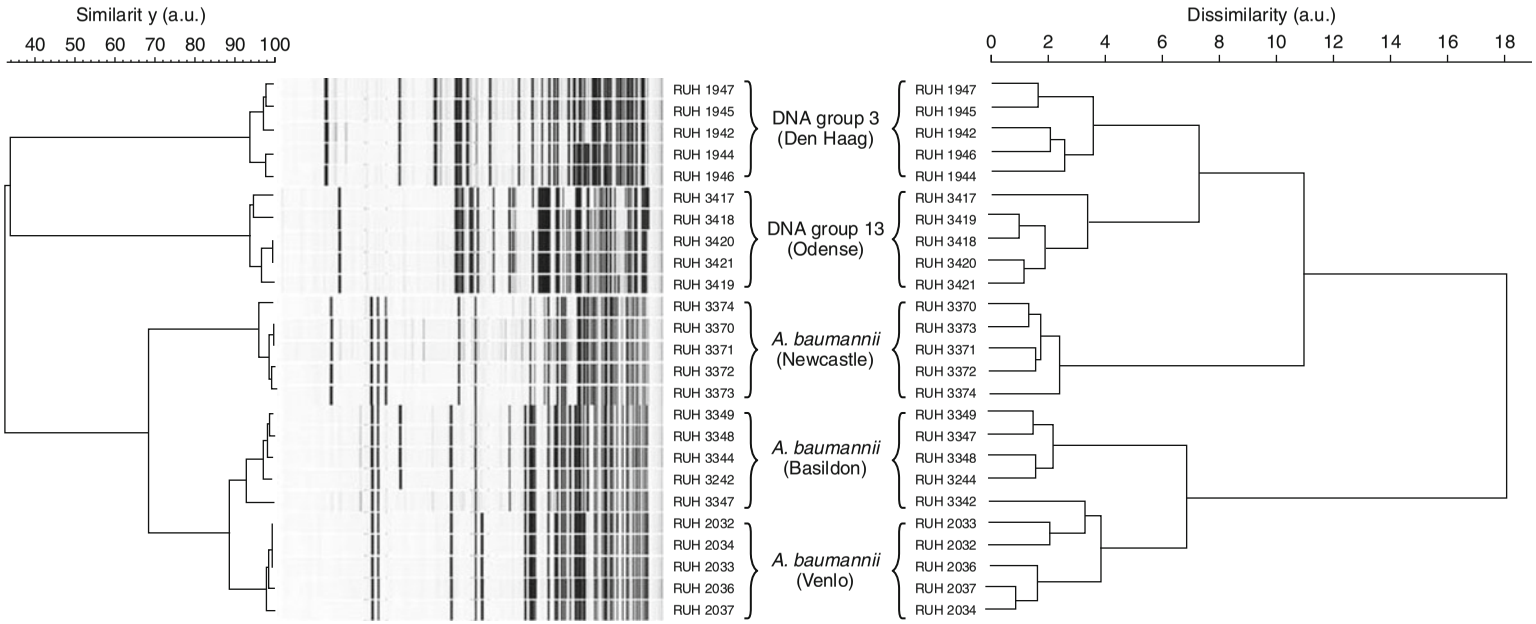
Infectious Diseases, Vibrational Spectroscopic Approaches to Rapid Diagnostics. Figure 1
Vector-normalized first-derivative spectra of two *B. cenocepacia* clinical isolates (isolates 57 and 69) in the 1,500–800 cm^{-1} range. (**a**) The heterogeneity of 15 replicate measurements for each strain in the spectral ranges of 1,200–900 cm^{-1} and 1,500–1,300 cm^{-1} and the corresponding micrographs obtained by TEM are shown. (**b**) Vector-normalized first-derivative spectra measured after vortexing of similar cells at the maximum intensity for 15 min and subsequent centrifugation at 8,000 × g for 5 min to separate the cells from free pilus appendages in the supernatants. Micrographs of the cells obtained by TEM after they were vortexed without centrifugation show the small fragments of pili or fibers suspended in the supernatants (From [26])

The background Raman spectrum resulting from the culture medium was subtracted from those spectra collected from the bacterial microcolonies. Hierarchical cluster analysis yielded two major groupings, one consisting of the three *Staphylococcus* strains and one consisting of the *E. coli* and *E. faecium*. The *E. coli* and *E. faecium* spectra clearly grouped according to species within the latter subcluster while spectra in the *Staphylococcus* subcluster grouped according to strain. While chemometric analysis of these spectra collected from same-day cultures yielded a successful classification rate of 100% for external validation samples, combined data collected from 3 days dropped the accuracy to 83% for classification of two *S. aureus* strains (ATCC 29213 and UHR 28624). However, these two strains are extremely similar and in general the results demonstrate the utility of Raman-based diagnostics.

The most rigorous evaluation of Raman spectroscopy for reagentless detection and identification of pathogens was performed in collaboration with a US government laboratory. In this work, a comprehensive library of Raman spectra has been established for over 1,000 species, including 281 CDC category A and B biothreats, 146 chemical threats, 310 environmental interferents, and numerous others [52]. Spectral signatures were collected using Raman chemical imaging spectroscopy (RCIS) [56]. RCIS technology combines digital imaging and Raman spectroscopy. Digital imaging automatically discriminates against background particulates and identifies regions of interest on a sample platform that are then targeted for Raman analysis. Sample analysis is faster and completely automated using this approach. Two commercially available instruments were tested, one in the laboratory (ChemImage Corp., Falcon) and the other in the field (ChemImage Corp., Eagle). To test the robustness of the Raman spectral library and classification scheme, blinded samples containing one of four *Bacillus* strains were analyzed and identified. The predictive performance ranged from 89.4% to 93.1% for these closely related bacteria. It was concluded that key to the success of this diagnostic approach is the extensiveness of the spectral library. There are many more bacterial phenotypes than genotypes, and it has been found that Raman fingerprints correlate with cell phenotype, thus an all-inclusive library must contain spectra for each bacterial strain grown under different conditions and at different stages of development. In a subsequent study untrained personnel at the Armed Forces Institute of Pathology evaluated 14 bacteria to generate a spectral library and sent 20 blinded samples to ChemImage for external validation in which all 20 samples were correctly identified. This comprehensive study is the first to establish the true utility of automated Raman-based diagnostics carried out off-site by untrained personnel. However, these samples were prepared in water, cell culture media, or spiked nasal swabs, none of which are truly clinical samples.

An early study to evaluate clinical samples for *Acinetobacter* by Raman spectroscopy and compare the results with an established diagnostic method were among the first showing the power and speed of Raman-based detection [55]. In this study, 25 *Acinetobacter* isolates from five hospitals in three countries were analyzed using selective amplification of restriction fragments (AFLP), an established molecular technique for typing bacteria strains. Dendograms resulting from the hierarchical cluster analysis of Raman and AFLP fingerprints for the isolates were generated and compared (Fig. 2). Both dendograms resulted in five clusters that separate the strains according to the five outbreaks, with the exception of one Basildon isolate RUH 3242 which clustered with isolates from Venlo in the Raman-based dendogram. Overall results from Raman fingerprinting of these clinical isolates were very similar to those obtained for established methods, but with the advantage of faster analysis and less complicated procedures.

Despite the advancement of Raman spectroscopy instrumentation and methods for pathogen fingerprinting, Raman is still often limited by poor sensitivity. Only $\sim$1 in 10^6–10^8 photons are inelastically scattered as the vast majority are elastically scattered. This means that high quality spectra with the requisite signal-to-noise can take minutes to acquire. While this may not be a limitation in laboratory experiments, or developmental stages in research, it prohibits its usefulness in clinical diagnostic laboratories which analyze hundreds to thousands of samples per day. Thus, there is great interest in enhancing the Raman signal. One such method is to excite the sample with a frequency that resonates with an electronic transition, so called resonance Raman spectroscopy. For biological samples, this requires UV lasers for excitation, and as noted



Infectious Diseases, Vibrational Spectroscopic Approaches to Rapid Diagnostics. Figure 2

Dendrograms resulting from the hierarchical cluster analysis of (*left*) AFLP analysis and (*right*) Raman analysis of the isolates. The asterisk marks the strain RUH 3242 misclassified via Raman fingerprinting (From [55])

above, is cost prohibitive to widespread adoption of this method. Moreover, chemical information is lost when performing resonance Raman which would likely reduce classification accuracy of closely related pathogens. An alternative method to amplify Raman scattering is surface-enhanced Raman spectroscopy (SERS). SERS has received a great deal of attention, particularly with respect to whole-organism fingerprinting and is the subject of the next section.

SERS

Surface-enhanced Raman spectroscopy is a technique in which the Raman signal of a sample is significantly amplified via adsorption onto a metallic nanostructured surface. A laser excitation frequency is selected such that it is in resonance with the collective oscillation of the conduction electrons in the nanostructures, i.e., surface plasmon resonance. When resonance conditions are met, the local electromagnetic field experienced by molecules in close proximity to the surface is significantly increased to yield rather large enhancements in the Raman scattering. While the signal enhancement is substrate and sample dependent, typical enhancements are on the order of 10^4–10^{14} with respect to normal Raman intensities, with several studies reporting the detection of single molecules using this technique [57, 58]. SERS offers the benefits of normal Raman compared to IR spectroscopy while providing a markedly improved sensitivity. Recent advances in nanofabrication methods and SERS theory has led to significant improvements in SERS substrates in the last several years and has driven increased efforts to develop SERS for whole-organism fingerprinting [59–78].

The major focus of whole-organism fingerprinting via SERS has been on bacteria identification [51, 64–74, 77, 78]. Most of these studies report differentiation among bacteria species, with many demonstrating discrimination of different strains of the same species. However, there are several inconsistencies that have been noted by researchers, particularly in the earlier studies. For example, Grow et al. found SERS spectra for strains that belong to the same species were sometimes less similar than spectra collected from different species [65], and Jarvis and Goodacre observed similar spectra for the same bacteria using different preparations of silver nanoparticles, but noted subtle changes in signal intensities among nanoparticle batches [68]. These discrepancies evident in these early studies highlight the primary challenge of SERS-based diagnostics, i.e., the enhancing substrate. The SERS signal is highly dependent on the enhancing substrate, thus a reliable means of fabricating nanostructured materials is vital to the success of SERS-based diagnostics.

Several research laboratories have analyzed and published SERS spectra for both *Bacillus subtilis* and *E. coli*; however, each reported incongruent spectral fingerprints [67, 68, 71, 72]. The experimental protocols, however, varied among each study. For example, in two different reports Jarvis et al. used two different chemical synthesis preparations to generate colloidal silver, citrate reduction [67] and borohydride reduction [51], to serve as the SERS substrate. The SERS spectra were drastically different in each study. It is well known that spectra are dependent on the enhancing nanostructure, e.g., material, size, shape, interparticle spacing, etc., but given the same final nanostructure similar spectra were expected. The authors attributed the differences to the effect of diverse chemistries used to prepare each silver colloid [79]. However, it should be noted that different excitation sources, 532 nm and 785 nm, were employed in the two studies. For normal Raman, the Raman shifts should be independent of the excitation source, thus spectral fingerprinting should not be affected by the choice of the laser. SERS spectra, however, can be influenced by the excitation source because of the requisite pairing of the excitation frequency and plasmon resonance of the substrate. Therefore, it is perhaps more probable that spectral differences observed by Jarvis et al. are due to greater signal enhancement for the 532 nm excitation source rather than due to differences in chemical preparation of the colloidal silver. This interpretation is supported by a study in which a third variation in experimental parameters was implemented utilizing citrate-reduced silver colloid but acquired spectra with a 647 nm laser [71]. Results from this study closely resembled the results for *B. subtilis* obtained by Jarvis et al. employing borohydride reduced silver nanoparticles and 532 nm excitation. Collectively, these studies also demonstrate the need for procedural consistency.

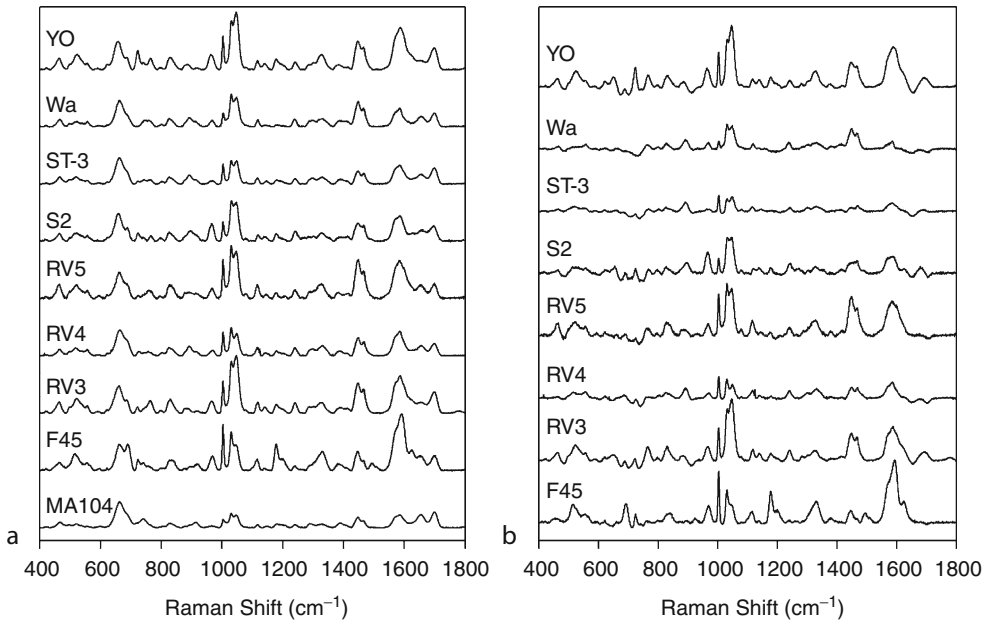
In a pivotal study, scientists at a US Army research laboratory evaluated the SERS signatures for many bacteria using a standardized sampling protocol and instrumentation. To date, three SERS substrates were directly compared using the standardized protocol: silver nanoparticles, silver film over nanospheres (FONS), and commercially available Klarite. Interaction between substrate and bacteria vary significantly as visualized with electron microscopy which likely results in different spectral fingerprints. Moreover the signal intensities varied significantly among the substrates reflecting differences in enhancing quality. Details of these experiments are approved for public release as a technical report (ARL-TR-4957).

In another key study, SERS and Raman fingerprints were directly compared to assess the advantages of SERS analysis [72]. Raman and SERS spectra were collected for several bacteria, including four strains of *Bacillus*, *S. typhimurium*, and *E. coli*. As noted above, the substrate is a critical factor in SERS analysis, and in this study aggregated gold nanoparticle films were grown in-house and established as a reliable means of substrate preparation for acquisition of repeatable spectra. As anticipated, SERS yielded much greater signal-to-noise spectra compared to normal Raman. The study also identified two unexpected benefits of SERS. Normal Raman signal for *Bacillus* species was overwhelmed by native fluorescence of the sample; however, in the SERS analysis, the metal substrate functioned to quench the fluorescence component in addition to enhancing the Raman signal. It was also observed that normal Raman spectra are more complex than SERS spectra. This is explained by the fact that bulk Raman interrogates all components throughout the entire bacterium equally, while the distance-dependence of SERS enhancement preferentially probes the region of the bacterium closest to the metal substrate and bands for the internal components are not detected. Fortunately, most chemical variation among bacterial strains and species are expressed on the cell surface, thus greater spectral differences are observed among SERS spectra of different samples than compared to bulk Raman spectra. This added advantage is exemplified by greater discrimination of bacteria when utilizing SERS spectra as compared to Raman spectra [72].

A number of novel nanofabrication methods have recently emerged for producing SERS substrates with the potential for addressing the issues noted above due to substrate heterogeneity. These include electron beam lithography [80, 81], nanosphere lithography [82–84], a template method [85–88], oblique angle vapor deposition (OAD) [89–91], and a proprietary wet-etching technology used to produce commercially available Klarite (D3 technologies). It should be noted, however, that with the exception of OAD and Klarite, these fabrication methods are not adaptable to large-scale production due to the complexity of the fabrication procedure. Not only is it likely that these substrates will lead to significant advances in SERS-diagnostics of bacteria, the use of OAD and Klarite substrates has already lead to successful application to virus identification [59, 60, 62, 63, 75, 76].

In the most recent investigation of SERS-based viral fingerprinting, eight strains of rotavirus were analyzed [63]. These isolates were recovered from clinical fecal samples and propagated in MA104 cells and represent the 5 G and 3 P genotypes responsible for the most severe infections. Unique SERS fingerprints were acquired for each strain when adsorbed onto OAD-fabricated silver nanorods. Representative spectra for each strain and negative control, as well as the difference spectra which subtract out the background cell lysate signal are displayed in Fig. 3. Classification algorithms based on partial least squares discriminant analysis were constructed to identify the samples according to (1) rotavirus positive or negative, (2) P4, P6, or P8 genotype, (3) G1, G2, G3, G4, or G9 genotype, or (4) strain. Respectively, these four classification models resulted in 100%, 98–100%, 96–100%, and 100% sensitivity and 100%, 100%, 99–100%, and 99–100% specificity.

Compilation and critical analysis of reports to date demonstrate the potential of Raman-based diagnostics and its advantages over IR, normal Raman spectroscopy, and convention diagnostic methods, but also highlight the need for standardization. The challenge in the future is standardization of substrates and sampling protocols since background can “quench” signal from the analyte. For example, blood analysis requires sample processing to remove some competing elements [92], yet SERS spectra highly dependent on the sample pretreatment procedure as remaining chemical species will also contribute signal and degrade the performance of matching in spectral library databases. The outlook



Infectious Diseases, Vibrational Spectroscopic Approaches to Rapid Diagnostics. Figure 3
(**a**) Average SERS spectra for eight strains of rotavirus and the negative control (MA104 cell lysate). Spectra were baseline corrected, normalized to the band at 633 cm^{-1}, and offset for visualization. (**b**) Difference SERS spectra for eight strains after subtraction of MA104 spectrum (From [63])

of SERS is not a question of spectral quality and reproducibility in a controlled environment, the question is how to control the environment across laboratories.

Future Directions

The future of spectroscopic-based diagnostics is bright as demonstrated by the many studies cited and discussed above. In addition to the success found in these studies, areas of improvement have also been identified. An important area of potential development is the methods used for statistical analysis. Well-established algorithms such as PCA, HCA, and discriminant analysis continue to provide high predictive accuracy, but recent examples have shown that more creative and novel approaches such as artificial neural networks, "bar-coding" [70], or innovative uses of PLS [59] can further improve the predictive value. A revolution in instrumentation is also occurring. Vibrational spectroscopy has recently filled niches in quality control of pharmaceuticals and raw materials as well as identification of chemical threats. The nature of these applications and explosion in interest have driven the instrumentation industry to invest in the development and production of high-quality yet affordable handheld instruments for mobile, on-site analysis. This market-driven commercialization, in effect, is paving the way for point-of-care, mobile, and cost-effective spectroscopy-based diagnostics. The most important factor for widespread realization of spectroscopic diagnosis will be the emergence of a universal protocol for sampling, and for the case of SERS, a standard substrate. The accepted protocol must then be used to build a spectral database covering a variety of phenotypes and developmental stages as illustrated above. Implementing a standard practice is crucial for the success of the technique, but once developed this technology has the potential to become the first and immediate response to clinical cases in which infection is suspected.

Bibliography

Primary Literature

1. Perry JD, Ford M, Taylor J, Jones AL, Freeman R et al (1999) ABC medium, a new chromogenic agar for selective isolation of *Salmonella* spp. J Clin Microbiol 37:766–768

2. Hedin G, Fang H (2005) Evaluation of two new chromogenic media, CHROMagar MRSA and S. aureus ID, for identifying Staphylococcus aureus and screening methicillin-resistant S-aureus. J Clin Microbiol 43:4242–4244
3. Perry JD, Davies A, Butterworth LA, Hopley ALJ, Nicholson A et al (2004) Development and evaluation of a chromogenic agar medium for methicillin-resistant *Staphylococcus aureus*. J Clin Microbiol 42:4519–4523
4. Nygren H, Stenberg M (1985) Kinetics of antibody-binding to surface-immobilized antigen – influence of mass-transport on the enzyme-linked immunosorbent-assay (ELISA). J Colloid Interface Sci 107:560–566
5. Nygren H, Werthen M, Stenberg M (1987) Kinetics of antibody-binding to solid-phase-immobilized antigen – effect of diffusion rate limitation and steric interaction. J Immunol Methods 101:63–71
6. Gussenhoven GC, vanderHoorn M, Goris MGA, Terpstra WJ, Hartskeerl RA et al (1997) LEPTO dipstick, a dipstick assay for detection of Leptospira-specific immunoglobulin M antibodies in human sera. J Clin Microbiol 35:92–97
7. Gordon A, Videa E, Saborio S, Lopez R, Kuan G et al (2010) Diagnostic accuracy of a rapid influenza test for pandemic influenza A H1N1. Plos One 5:e10364
8. Erdman DD, Weinberg GA, Edwards KM, Walker FJ, Anderson BC et al (2003) GeneScan reverse transcription-PCR assay for detection of six common respiratory viruses in young children hospitalized with acute respiratory illness. J Clin Microbiol 41:4298–4303
9. Lassauniere R, Kresfelder T, Venter M (2010) A novel multiplex real-time RT-PCR assay with FRET hybridization probes for the detection and quantitation of 13 respiratory viruses. J Virol Methods 165:254–260
10. Li H, McCormac MA, Estes RW, Sefers SE, Dare RK et al (2007) Simultaneous detection and high-throughput identification of a panel of RNA viruses causing respiratory tract infections. J Clin Microbiol 45:2105–2109
11. Abu al-Soud W, Radstrom P (2001) Purification and characterization of PCR-inhibitory components in blood cells. J Clin Microbiol 39:485–493
12. Abu Al-Soud W, Radstrom P (2001) Effects of amplification facilitators on diagnostic PCR in the presence of blood, feces, and meat. J Clin Microbiol 38:4463–4470
13. Widjojoatmodjo MN, Fluit AC, Torensma R, Verdonk G, Verhoef J (1992) The magnetic immunopolymerase chain-reaction assay for direct detection of salmonellae in fecal samples. J Clin Microbiol 30:3195–3199
14. Chui LW, King R, Lu P, Manninen K, Sim J (2004) Evaluation of four DNA extraction methods for the detection of *Mycobacterium avium* subsp *paratuberculosis* by polymerase chain reaction. Diagn Microbiol Infect Dis 48:39–45
15. McOrist AL, Jackson M, Bird AR (2002) A comparison of five methods for extraction of bacterial DNA from human faecal samples. J Microbiol Methods 50:131–139
16. Kaigala GV, Hoang VN, Stickel A, Lauzon J, Manage D et al (2008) An inexpensive and portable microchip-based platform for integrated RT-PCR and capillary electrophoresis. Analyst 133:331–338
17. Zhang NY, Tan HD, Yeung ES (1999) Automated and integrated system for high-throughput DNA genotyping directly from blood. Anal Chem 71:1138–1145
18. Aryan E, Makvandi M, Farajzadeh A, Huygen K, Bifani P et al (2010) A novel and more sensitive loop-mediated isothermal amplification assay targeting IS6110 for detection of *Mycobacterium tuberculosis* complex. Microbiol Res 165:211–220
19. Fang XE, Liu YY, Kong JL, Jiang XY (2010) Loop-mediated isothermal amplification integrated on microfluidic chips for point-of-care quantitative detection of pathogens. Anal Chem 82:3002–3006
20. Notomi T, Okayama H, Masubuchi H, Yonekawa T, Watanabe K et al (2000) Loop-mediated isothermal amplification of DNA. Nucl Acids Res 28:e63
21. Shivakoti S, Ito H, Murase T, Ono E, Takakuwa H et al (2010) Development of reverse transcription-loop-mediated isothermal amplification (RT-LAMP) assay for detection of avian influenza viruses in field specimens. J Vet Med Sci 72:519–523
22. Stevenson HJR, Bolduan OEA (1952) Infrared spectrophotometry as a means for identification of bacteria. Science 116:111–113
23. Lin MS, Al-Holy M, Al-Qadiri H, Kang DH, Cavinato AG et al (2004) Discrimination of intact and injured *Listeria monocytogenes* by Fourier transform infrared spectroscopy and principal component analysis. J Agric Food Chem 52:5769–5772
24. Ngo-Thi NA, Kirschner C, Naumann D (2003) Characterization and identification of microorganisms by FIF-IR microspectrometry. J Mole Struct 661:371–380
25. Janbu AO, Moretro T, Bertrand D, Kohler A (2008) FT-IR microspectroscopy: a promising method for the rapid identification of *Listeria* species. FEMS Microbiol Lett 278:164–170
26. Bosch A, Minan A, Vescina C, Degrossi J, Gatti B et al (2008) Fourier transform infrared spectroscopy for rapid identification of nonfermenting gram-negative bacteria isolated from sputum samples from cystic fibrosis patients. J Clin Microbiol 46:2535–2546
27. Rebuffo-Scheer CA, Schmitt J, Scherer S (2007) Differentiation of Listeria monocytogenes serovars by using artificial neural network analysis of Fourier-transformed infrared spectra. Appl Environ Microbiol 73:1036–1040
28. Bouhedja W, Sockalingum GD, Pina P, Allouch P, Bloy C et al (1997) ATR-FTIR spectroscopic investigation of *E coli* transconjugants beta-lactams-resistance phenotype. FEBS Lett 412:39–42
29. Goodacre R, Timmins EM, Rooney PJ, Rowland JJ, Kell DB (1996) Rapid identification of *Streptococcus* and *Enterococcus* species using diffuse reflectance-absorbance Fourier transform infrared spectroscopy and artificial neural networks. FEMS Microbiol Lett 140:233–239
30. Helm D, Labischinski H, Naumann D (1991) Elaboration of a procedure for identification of bacteria using

Fourier-transform IR spectral libraries – a stepwise correlation approach. J Microbiol Methods 14:127–142
31. Helm D, Labischinski H, Schallehn G, Naumann D (1991) Classification of bacteria by Fourier-transform infrared-spectroscopy. J Gen Microbiol 137:69–79
32. Naumann D, Fijala V, Labischinski H, Giesbrecht P (1988) The rapid differentiation and identification of pathogenic bacteria using Fourier-transform infrared spectroscopic and multivariate statistical-analysis. J Mole Struct 174:165–170
33. Naumann D, Helm D, Labischinski H (1991) Microbiological characterizations by FT-IR spectroscopy. Nature 351:81–82
34. Ferroni A, Sermet-Gaudelus I, Abachin E, Quesne G, Lenoir G et al (2002) Use of 16 S rRNA gene sequencing for identification of nonfermenting gram-negative bacilli recovered from patients attending a single cystic fibrosis center. J Clin Microbiol 40:3793–3797
35. Miller MB, Gilligan PH (2003) Laboratory aspects of management of chronic pulmonary infections in patients with cystic fibrosis. J Clin Microbiol 41:4009–4015
36. Fischer G, Braun S, Thissen R, Dott W (2006) FT-IR spectroscopy as a tool for rapid identification and intra-species characterization of airborne filamentous fungi. J Microbiol Methods 64:63–77
37. Sandt C, Sockalingum GD, Aubert D, Lepan H, Lepouse C et al (2003) Use of Fourier-transform infrared spectroscopy for typing of *Candida albicans* strains isolated in intensive care units. J Clin Microbiol 41:954–959
38. Erukhimovitch V, Karpasasa M, Huleihel M (2009) Spectroscopic detection and identification of infected cells with Herpes viruses. Biopolymers 91:61–67
39. Erukhimovitch V, Mukmanov I, Talyshinsky M, Souprun Y, Huleihel M (2004) The use of FTIR microscopy for evaluation of herpes viruses infection development kinetics. Spectrochimica Acta Part A-Mole Biomol Spectrosc 60:2355–2361
40. Hastings G, Krug P, Wang RL, Guo J, Lamichhane HP et al (2009) Viral infection of cells in culture detected using infrared microscopy. Analyst 134:1462–1471
41. Salman A, Erukhimovitch V, Talyshinsky M, Huleihil M, Huleihel M (2002) FTIR spectroscopic method for detection of cells infected with herpes viruses. Biopolymers 67:406–412
42. Hartman KA, Clayton N, Thomas GJ (1973) Studies of virus structure by Raman spectroscopy 1. R17 virus and R17 RNA. Biochem Biophys Res Commun 50:942–949
43. Dalterio RA, Baek M, Nelson WH, Britt D, Sperry JF et al (1987) The resonance Raman microprobe detection of single bacterial-cells from a chromobacterial mixture. Appl Spectrosc 41:241–244
44. Chadha S, Manoharan R, Moenneloccoz P, Nelson WH, Peticolas WL et al (1993) Comparison of the UV resonance Raman-spectra of bacteria, bacteria-cell walls, and ribosomes excited in the deep UV. Appl Spectrosc 47:38–43
45. Ghiamati E, Manoharan R, Nelson WH, Sperry JF (1992) UV resonance Raman-spectra of bacillus spores. Appl Spectrosc 46:357–364
46. Manoharan R, Ghiamati E, Dalterio RA, Britton KA, Nelson WH et al (1990) UV resonance Raman-spectra of bacteria, bacterial-spores, protoplasts and calcium dipicolinate. J Microbiol Methods 11:1–15
47. Edwards HGM, Russell NC, Weinstein R, Wynnwilliams DD (1995) Fourier-transform Raman-spectroscopic study of fungi. J Raman Spectrosc 26:911–916
48. Williams AC, Edwards HGM (1994) Fourier-transform Raman-spectroscopy of bacterial-cell walls. J Raman Spectrosc 25:673–677
49. Choo-Smith LP, Maquelin K, van Vreeswijk T, Bruining HA, Puppels GJ et al (2001) Investigating microbial (micro) colony heterogeneity by vibrational spectroscopy. Appl Environ Microbiol 67:1461–1469
50. Huang WE, Griffiths RI, Thompson IP, Bailey MJ, Whiteley AS (2004) Raman microscopic analysis of single microbial cells. Anal Chem 76:4452–4458
51. Jarvis RM, Brooker A, Goodacre R (2004) Surface-enhanced Raman spectroscopy for bacterial discrimination utilizing a scanning electron microscope with a Raman spectroscopy interface. Anal Chem 76:5198–5202
52. Kalasinsky KS, Hadfield T, Shea AA, Kalasinsky VF, Nelson MP et al (2007) Raman chemical imaging spectroscopy reagentless detection and identification of pathogens: signature development and evaluation. Anal Chem 79:2658–2673
53. Maquelin K, Choo-Smith LP, Endtz HP, Bruining HA, Puppels GJ (2002) Rapid identification of *Candida* species by confocal Raman micro spectroscopy. J Clin Microbiol 40:594–600
54. Maquelin K, Choo-Smith LP, van Vreeswijk T, Endtz HP, Smith B et al (2000) Raman spectroscopic method for identification of clinically relevant microorganisms growing on solid culture medium. Anal Chem 72:12–19
55. Maquelin K, Dijkshoorn L, van der Reijden TJK, Puppels GJ (2006) Rapid epidemiological analysis of *Acinetobacter* strains by Raman spectroscopy. J Microbiol Methods 64:126–131
56. Schaeberle MD, Morris HR, Turner JF, Treado PJ (1999) Raman chemical imaging spectroscopy. Anal Chem 71:175A–181A
57. Kneipp K, Wang Y, Kneipp H, Perelman LT, Itzkan I et al (1997) Single molecule detection using surface-enhanced Raman scattering (SERS). Phys Rev Lett 78:1667–1670
58. Nie SM, Emory SR (1997) Probing single molecules and single nanoparticles by surface-enhanced Raman scattering. Science 275:1102–1106
59. Alexander TA (2008) Development of methodology based on commercialized SERS-active substrates for rapid discrimination of Poxviridae virions. Anal Chem 80:2817–2825
60. Alexander TA (2008) Surface-enhanced Raman spectroscopy: a new approach to rapid identification of intact viruses. Spectroscopy 23:36–42
61. Bao PD, Huang TQ, Liu XM, Wu TQ (2001) Surface-enhanced Raman spectroscopy of insect nuclear polyhedrosis virus. J Raman Spectrosc 32:227–230
62. Driskell JD, Shanmukh S, Liu YJ, Hennigan S, Jones L et al (2008) Infectious agent detection with SERS-active silver

nanorod arrays prepared by oblique angle deposition. IEEE Sens J 8:863–870
63. Driskell JD, Zhu Y, Kirkwood CD, Zhao YP, Dluhy RA et al (2010) Rapid and sensitive detection of rotavirus molecular signatures using surface enhanced Raman spectroscopy. Plos One 5(4):e10222
64. Goeller LJ, Riley MR (2007) Discrimination of bacteria and bacteriophages by Raman spectroscopy and surface-enhanced Raman spectroscopy. Appl Spectrosc 61:679–685
65. Grow AE, Wood LL, Claycomb JL, Thompson PA (2003) New biochip technology for label-free detection of pathogens and their toxins. J Microbiol Methods 53:221–233
66. Guicheteau J, Christesen SD (2006) Principal component analysis of bacteria using surface-enhanced Raman spectroscopy. Proc SPIE 6218:62180G
67. Jarvis RM, Brooker A, Goodacre R (2006) Surface-enhanced Raman scattering for the rapid discrimination of bacteria. Faraday Discuss 132:281–292
68. Jarvis RM, Goodacre R (2004) Discrimination of bacteria using surface-enhanced Raman spectroscopy. Anal Chem 76:40–47
69. Laucks ML, Sengupta A, Junge K, Davis EJ, Swanson BD (2005) Comparison of Psychro-active arctic marine Bacteria and common Mesophillic bacteria using surface-enhanced Raman spectroscopy. Appl Spectrosc 59:1222–1228
70. Patel IS, Premasiri WR, Moir DT, Ziegler LD (2008) Barcoding bacterial cells: a SERS-based methodology for pathogen identification. J Raman Spectrosc 39:1660–1672
71. Pearman WF, Fountain AW (2006) Classification of chemical and biological warfare agent simulants by surface-enhanced Raman spectroscopy and multivariate statistical techniques. Appl Spectrosc 60:356–365
72. Premasiri WR, Moir DT, Klempner MS, Krieger N, Jones G et al (2005) Characterization of the surface enhanced Raman scattering (SERS) of bacteria. J Phys Chem B 109:312–320
73. Premasiri WR, Moir DT, Lawrence DZ (2005) Vibrational fingerprinting of bacterial pathogens by surface enhanced Raman scattering (SERS). Proc SPIE 5795:19–29
74. Sengupta A, Laucks ML, Davis EJ (2005) Surface-enhanced Raman spectroscopy of bacteria and pollen. Appl Spectrosc 59:1016–1023
75. Shanmukh S, Jones L, Driskell J, Zhao Y, Dluhy R et al (2006) Rapid and sensitive detection of respiratory virus molecular signatures using a silver nanorod array SERS substrate. Nano Lett 6:2630–2636
76. Shanmukh S, Jones L, Zhao Y-P, Driskell JD, Tripp RA et al (2008) Identification and classification of respiratory syncytial virus (RSV) strains by surface-enhanced Raman spectroscopy and multivariate statistical techniques. Anal Bioanal Chem 390:1551–1555
77. Yan F, Vo-Dinh T (2007) Surface-enhanced Raman scattering detection of chemical and biological agents using a portable Raman integrated tunable sensor. Sens Actuat B-Chem 121:61–66
78. Zeiri L, Bronk BV, Shabtai Y, Czege J, Efrima S (2002) Silver metal induced surface enhanced Raman of bacteria. Colloids Surf A-Physicochem Eng Aspects 208:357–362
79. Jarvis RM, Goodacre R (2008) Characterisation and identification of bacteria using SERS. Chem Soc Rev 37:931–936
80. DeJesus MA, Giesfeldt KS, Oran JM, Abu-Hatab NA, Lavrik NV et al (2005) Nanofabrication of densely packed metal-polymer arrays for surface-enhanced Raman spectrometry. Appl Spectrosc 59:1501–1508
81. Kahl M, Voges E, Kostrewa S, Viets C, Hill W (1998) Periodically structured metallic substrates for SERS. Sens Actuat B-Chem 51:285–291
82. Haynes CL, Van Duyne RP (2001) Nanosphere lithography: a versatile nanofabrication tool for studies of size-dependent nanoparticle optics. J Phys Chem B 105:5599–5611
83. Hulteen JC, Treichel DA, Smith MT, Duval ML, Jensen TR et al (1999) Nanosphere lithography: size-tunable silver nanoparticle and surface cluster arrays. J Phys Chem B 103: 3854–3863
84. Jensen TR, Malinsky MD, Haynes CL, Van Duyne RP (2000) Nanosphere lithography: tunable localized surface plasmon resonance spectra of silver nanoparticles. J Phys Chem B 104:10549–10556
85. Broglin BL, Andreu A, Dhussa N, Heath JA, Gerst J et al (2007) Investigation of the effects of the local environment on the surface-enhanced Raman spectra of striped gold/silver nanorod arrays. Langmuir 23:4563–4568
86. Lombardi I, Cavallotti PL, Carraro C, Maboudian R (2007) Template assisted deposition of Ag nanoparticle arrays for surface-enhanced Raman scattering applications. Sensor Actuat B-Chem 125:353–356
87. Ruan CM, Eres G, Wang W, Zhang ZY, Gu BH (2007) Controlled fabrication of nanopillar arrays as active substrates for surface-enhanced Raman spectroscopy. Langmuir 23:5757–5760
88. Yao JL, Pan GP, Xue KH, Wu DY, Ren B et al (2000) A complementary study of surface-enhanced Raman scattering and metal nanorod arrays. Pure Appl Chem 72:221–228
89. Chaney SB, Shanmukh S, Zhao Y-P, Dluhy RA (2005) Randomly aligned silver nanorod arrays produce high sensitivity SERS substrates. Appl Phys Lett 87:31908–31910
90. Driskell JD, Shanmukh S, Liu Y, Chaney SB, Tang XJ et al (2008) The use of aligned silver nanorod arrays prepared by oblique angle deposition as surface enhanced Raman scattering substrates. J Phys Chem C 112:895–901
91. Liu YJ, Fan JG, Zhao YP, Shanmukh S, Dluhy RA (2006) Angle dependent surface enhanced Raman scattering obtained from a Ag nanorod array substrate. Appl Phys Lett 89:173134
92. Premasiri WR, Moir DT, Klempner MS, Ziegler LD (2007) Surface-enhanced Raman scattering of microorganisms. New Approaches Biomed Spectrosc 963:164–185

Books and Reviews

Carter EA, Marshall CP, Ali MHM, Ganendren R, Sorrell TC, Wright L, Lee Y-C, Chen C-I, Lay PA (2007) Infrared spectroscopy of microorganisms: characterization, identification, and differentiation. In: Kneipp K, Aroca R, Kneipp H, Wentrup-Byrne E (eds)

New approaches in biomedical spectroscopy. American Chemical Society, Washington, DC, pp 64–84

Huang WE, Li M, Jarvis RM, Goodacre R, Banwart SA (2010) Shining light on the microbial world: the application of Raman microspectroscopy. In: Laskin A, Sariaslani S, Gadd GM (eds) Advances in applied microbiology, vol 70. Elsevier, San Diego, pp 153–186

Ince J, McNally A (2009) Development of rapid, automated diagnostics for infectious disease: advances and challenges. Expert Rev Med Devices 6(6):641–651

Posthuma-Trumpie G, Korf J, van Amerongen A (2009) Lateral flow (immune)assay: its strengths, weakness, opportunities and threats. A literature survey. Anal Bioanaly Chem 393: 569–582

Premasiri WR, Moir DT, Klempner MS, Ziegler LD (2007) Surface-enhanced Raman scattering of microorganisms. In: Kneipp K, Aroca R, Kneipp H, Wentrup-Byrne E (eds) New approaches in biomedical spectroscopy. American Chemical Society, Washington, DC, pp 164–199

Quan P-L, Briese T, Palacios G, Lipkin WI (2008) Rapid sequence-based diagnosis of viral infection. Antiviral Res 79:1–5

Sharaf MA, Illman DL, Kowalski BR (1986) Chemometrics. Wiley, New York

Tuma R, Thomas GJ Jr (2002) Raman spectroscopy of viruses. In: Chalmers JM, Griffiths PR (eds) Handbook of vibrational spectroscopy applications in life, pharmaceutical and natural sciences, vol 5. Wiley, West Sussex, pp 3519–3535

Integrated Assessment Modeling

James A. Edmonds, Katherine V. Calvin, Leon E. Clarke, Anthony C. Janetos, Son H. Kim, Marshall A. Wise, Haewon C. McJeon

Joint Global Change Research Institute (JGCRI), Pacific Northwest National Laboratory (PNNL), College Park, MD, USA

Article Outline

Glossary

Climate policy (greenhouse gas mitigation policy) A climate policy refers to a policy scheme designed to deliberately limit the magnitude of climate change, often involving mitigation of greenhouse gases. Integrated assessment models (IAMs) represent climate policies in abstract forms. The most commonly modeled climate policy is attaching a universal price on emissions of carbon dioxide (or carbon dioxide equivalent of other greenhouse gases). Such policy represents a universal carbon tax or an economy-wide cap-and-trade policy. Other forms of climate policies, such as differential carbon price by sector or renewable portfolio standards, have also been used in IAMs.

Cost of greenhouse gas mitigation (economic cost) Integrated assessment models (IAMs) employ varies metrics for estimating the economic cost of mitigation policy. One common approach estimates reduction in GDP, a proxy for slowdown in economic activity due to increased price of energy and agricultural products. Another approach estimates the (gross) loss in social welfare due to a policy by measuring the area under the marginal abatement cost curve. Other metrics include foregone consumption, compensated variation, and equivalent variation.

Integrated assessment model (IAM) Integrated assessment model (IAM) in climate change research is a model which simulates the interactions of human decision-making about energy systems and land use with biogeochemistry and the natural Earth system. IAMs can be divided into two categories.

Higher resolution IAMs focus on explicitly representing processes and process interactions among human and natural Earth systems.

Highly aggregated IAMs use highly reduced-form representations of the link between human activities, impacts from climate change, and the cost of emissions mitigation.

Integrated earth system model (iESM) Integrated Earth System Models (iESMs) are a class of models under development by collaboration between integrated assessment modeling community and

climate modeling community. By fully integrating the human dimension from an IAM and the natural dimension from a climate model, iESM allows simultaneously estimating human system impacts on climate change and climate change impacts on human systems, as well as examining the effects of feedbacks between the components.

Land use (land-use emissions) Land use is one of the largest anthropogenic sources of emissions of greenhouse gases, aerosols, and short-lived species. Emissions, as well as sequestration of emissions, may occur from land-use practices, changes in land cover, or changes in forested area or the density. On the other hand, land-use patterns are affected by the changes in the climate. As such, modeling land use has been an important component of the integrated assessment modeling of climate change.

Representative concentration pathways (RCPs) The Representative Concentration Pathways (RCPs) are the most recent set of emission scenarios generated by integrated assessment models. Four scenarios explicitly considering emission mitigation efforts that were sufficiently differentiated in terms of radiative forcing at the end of the century were selected from published literature. RCPs are designed to facilitate the interactions with climate models by including geospatially resolved emissions and land-use data.

Definition of the Subject

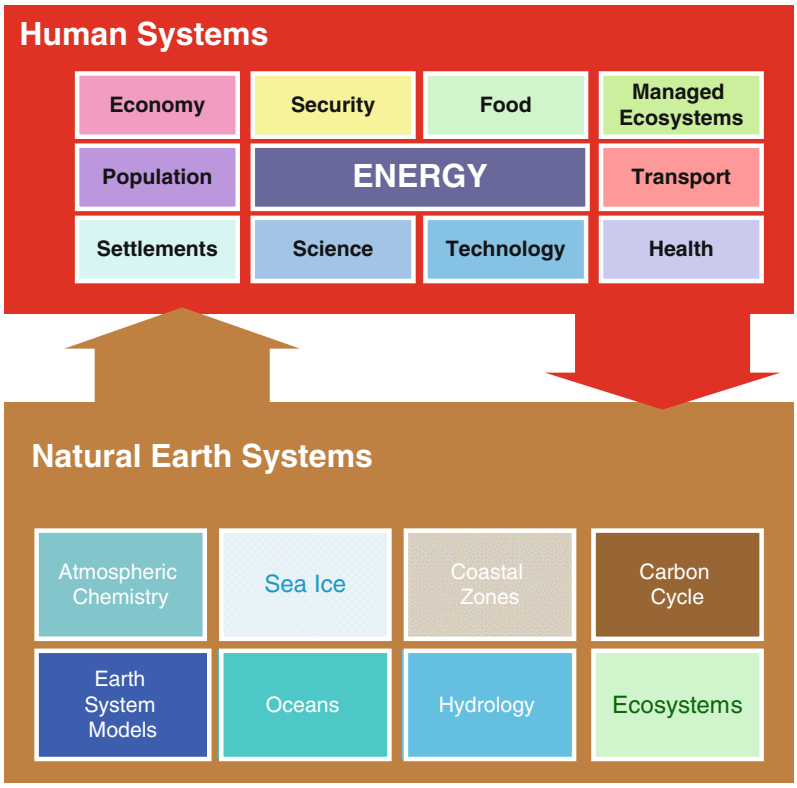
This entry discusses the role of integrated assessment models (IAMs) in climate change research. IAMs are an interdisciplinary research platform, which constitutes a consistent scientific framework in which the large-scale interactions between human and natural Earth systems can be examined. In so doing, IAMs provide insights that would otherwise be unavailable from traditional single-discipline research. By providing a broader view of the issue, IAMs constitute an important tool for decision support. IAMs are also a home of human Earth system research and provide natural Earth system scientists information about the nature of human intervention in global biogeophysical and geochemical processes.

Introduction

Integrated assessment models (IAMs) are a class of models which simulate the interactions of human decision-making about energy systems and land use with biogeochemistry and the natural Earth system (see Fig. 1). In so doing, IAMs provide insights that would otherwise be unavailable from investigating either natural systems or human systems, or their various components, alone. By their nature, IAMs capture interactions between complex and highly nonlinear systems. IAMs serve multiple purposes. One purpose is to provide natural science researchers with information about human systems such as greenhouse gas emissions, land use, and land cover. Another purpose of IAMs is to help human system researchers – such as social scientists – better understand the nature of the human impacts on the natural Earth systems.

Traditionally, researchers have relied on models that are each built on the foundations of a single discipline – such as economics, geography, meteorology, etc. By integrating research methods from various disciplines that characterize both the human and natural Earth systems, IAMs produce insights that would not otherwise be available from disciplinary research. The work of Wigley, Richels, and Edmonds [1] provides a classic example of the nature of insights that are available from the explicit linking of human and Earth systems. Wigley et al. showed that the consideration of economic efficiency in the context of the physical carbon cycle carried important implications for the timing of emissions and emissions mitigation in a world seeking to stabilize the concentration of atmospheric CO_2. In other words, the imposition of human system considerations – in this case economic efficiency considerations – led to a different and smaller set of emissions pathways for consideration than were indicated by Earth system considerations alone.

This entry discusses a range of selected topics associated with the development and use of IAMs. This is not an extensive survey of the literature and the available models. Instead, it focuses on a selected set of topics required to understand the various types and uses of IAMs as well as those required to understand the direction of cutting-edge IAM research. In addition, the entry focuses more heavily on the strain



Integrated Assessment Modeling. Figure 1
Integrated assessment models integrate human and physical Earth system climate science (Source: Janetos et al. [72])

of IAMs and integrated assessment modeling (IA modeling) research focused on more effectively modeling human and Earth system processes (higher resolution IAMs) than on the strain of IAMs and IA modeling research that focuses on more aggregate representations of these systems to allow for cost-benefit analysis. The remainder of this entry proceeds as follows. Section "The Variety of Integrated Assessment Models" focuses on the emerging distinction between highly aggregated and higher resolution IAMs. Section "GCAM as an Example of a Higher Resolution IAM" then follows with a discussion of the Global Change Assessment Model (GCAM) as an example of a higher resolution IAM. Section "Using Higher Resolution IAMs to Analyze the Impact of Policies to Mitigate Greenhouse Gas Emissions" discusses the long history of using IAMs to explore the costs of greenhouse gas policies as well as several of the most important conceptual issues that the IAMs have had to wrestle with in this regard. Section "Future Directions: Integrating Climate Impacts with IAM" then explores an important cutting-edge research direction for higher resolution IAMs: the inclusion of structural or process models of climate impacts.

The Variety of Integrated Assessment Models

There are many approaches that have been used to develop and use IAMs. Indeed, every IAM is different. One of the most important ways that IAMs are distinguished from one another is the level of resolution at which they model the underlying human and natural Earth system process. At one end of the spectrum are *highly aggregated* IAMs. Highly aggregated IAMs use highly reduced-form representations of the link between human activities, impacts from climate change, and the cost of emissions mitigation. At the other end of the spectrum are *higher resolution* IAMs. Higher resolution IAMs focus on explicitly representing processes and process interactions among human and natural Earth systems. The following two subsections provide background on each of these two classes of IAMs.

Highly Aggregated IAMs

The highly aggregated class of IAMs was developed to be able to explore the general shape of optimal climate policy, taking into account both the economic costs of mitigation and the economic damages from a changing climate. Highly aggregated IAMs typically frame the climate change mitigation problem in a cost-benefit framework, choosing emission pathways by explicitly weighing the economic costs of mitigation with the economic benefits of reduced impacts. For this reason, highly aggregated IAMs often focus on issues such as the social cost of carbon or optimal tradeoffs over time between mitigation and impacts. Simplicity and parsimony are main virtues of highly aggregated IAMs.

The oldest of the highly aggregated IAMs is the DICE (Dynamic Integrated model of Climate and the Economy) model, whose antecedents have roots in the work of Nordhaus and Yohe [2]. The original DICE model [3] was utilized to explore the integration of human and natural Earth systems as part of a cost-benefit calculation. Originally developed as a one-region global model, DICE was soon followed by a multiregional version, RICE (Regional dynamic Integrated model of Climate and the Economy) [4]. Other such models also emerged building on the Nordhaus-Yohe and DICE paradigm of combining economic costs and benefits in a single framework. These models include, among others, ICAM (the Integrated Climate Assessment Model) [5], PAGE (Policy Analysis of the Greenhouse Effect) [6], and FUND (Climate Framework for Uncertainty, Negotiation and Distribution) [7] (Weyant et al. [8] and Parson and Fisher-Vanden [9] provide good sources of information on pioneering IAMs).

Highly aggregated IAMs are generally composed of three parts: emissions and mitigation, atmosphere and climate, and climate impacts. Mitigation cost and climate change damages are typically monetized (i.e., expressed in dollars or another currency) to allow comparison between mitigation and impacts on a common basis. Highly aggregated IAMs do not attempt to describe in detail either the energy system or the land-use systems that generate emissions. Similarly, detailed descriptions of the physical process links between climate change and emissions are generally beyond their scope. Instead these models use emissions mitigation supply schedules and climate damage functions. The former maps the relationship between the degree of emissions mitigation and associated cost, while the latter represents the relationship between a measure of climate change and the economic value of damages including both damages from market and nonmarket activities. The strength of these reduced-form representations is that they allow highly aggregated IAMs to weigh costs and benefits explicitly. The drawback is that they cannot provide insight into the actual processes that lead to these costs and benefits.

The technical structure of highly aggregated models is simple, but the equations and associated parameterizations are carefully estimated to capture the behavior of more complex systems. These functions are parameterized by either approximating the behavior of more complex process models, or by fitting simple equations to highly aggregated variables. Analyses using FUND, for example, often produce simple equations that capture the behavior of systems that are represented in more complex models and data. Some models use a simpler approach, in which the economic damages from a prescribed level of climate change are first estimated – for example, a 2°C global mean surface temperature change (GMST) relative to preindustrial level – and a simple function that passes through the estimate – for example, a power function – is assigned to represent the relationship between GMST and total economic damages.

A principle role of highly aggregated IAMs is to integrate and to compare in a common metric, both mitigation effort and climate change impact – each estimated from different disciplines – in order to determine the optimal pathway of emissions reductions or the social cost of carbon. Valuation of damages provides substantial conceptual challenges for highly aggregated IAMs. For example, they must put a value on the loss of human lives as well as nonmarket damages. Another difficult challenge faced by highly aggregated IAMs concerns the relative valuation of impacts that occur at different points in time. See Box 1 for details.

Other issues that arise within the highly aggregated IAM paradigm include the problem of interactive effects, that is, the state of one system directly affects the state of another. For example, emissions mitigation may have large-scale effects on land use, which in turn

affect the climate, or the climate system may change as a consequence of land-use policy. A challenge for highly aggregated IAMs is to represent such complex interactions in a simple model structure.

Box 1. Valuation over Time and Across Generations

Climate change is an issue that is inherently long term as well as global. The nature of carbon cycle processes and their associated time scales create a cumulative relationship between CO_2 emissions and concentrations in the atmosphere (and ocean). Thus, unlike traditional atmospheric pollution problems, control of emissions to a level is insufficient to control the concentration of greenhouses in the atmosphere. In other words, CO_2 and other greenhouse gases are stock pollutants.

One of the most important determinants of the social cost of carbon is the rate at which future events are discounted back to the present. Nordhaus [10] argues that the order of magnitude difference between his estimate of the social value of carbon, derived using DICE, and the value estimated in the Stern Report [11], derived using PAGE, is predominantly a result of the differences in valuing the present relative to the future.

The problem is that there is no consensus on precisely how to approach discounting over periods of time long enough to connect multiple generations. The issues are laid out in Portney and Weyant [12], where the editors note in their overview chapter that "those looking for guidance on the choice of a discount rate could find justification for a rate at or near zero, as high as 20% and any and all values in between" ([12], p. 4). The range of estimates for the appropriate discount rate is generally nonnegative, though even that generalization has its exceptions, for example [13].

Methods for determining the appropriate method for discounting the future can be grouped into two general categories – those which are *prescriptive* and those which are *descriptive*. The prescriptive approach appeals to ethical and moral grounds for choosing a discount rate, while the descriptive approach appeals to observed rates of return on assets in economic markets. It is frequently observed that prescriptive approaches tend to generate lower discount rates than descriptive approaches.

Another challenge for highly aggregated IAMs is to determine how to treat impacts occurring outside of the country undertaking the valuation. Early work with highly aggregated IAMs looked at the problem of climate change from the perspective of a single, global, infinitely lived decision maker. But, more recent work has shifted from the perspective of the globe (e.g., [3, 11]) to the perspective of a single country, for example, the United States [14].

The Higher Resolution IAMs

The higher resolution IAMs have roots in the same era as the highly aggregated IAMs. However, they were developed along different lines to serve different purposes. The higher resolution IAMs were developed to provide detailed information about human and natural Earth system processes and the interactions between these processes. The initial focus of these models was the determinants of anthropogenic carbon emissions. To address this problem, IAMs developed detailed representations of the key features determining long-term energy production, transformation, and end use. The higher resolution models distinguished different forms of energy, their supplies, demands, and their transformation from primary energy to fuels and electricity for use in end-use sectors such as buildings, transportation, and industry. Examples of higher resolution IAMs are provided in Table 1.

Over time these models have grown in complexity. The models have added increasing detail to their representations of both the energy system and the economy. They also broadened their scope, adding natural Earth system processes such as carbon cycle. The current generations of higher resolution IAMs also typically contain representations of agriculture, land use, land cover, and terrestrial carbon cycle processes in addition to representations of atmosphere and climate processes. While all of the higher resolution IAMs model both human and natural Earth system processes, each model was developed independently and each IAM development path emphasized different features of the climate change problem. Some emphasized the development of detailed atmosphere and climate system models. Some focused on detailed representations of technology. Others focused on regional differences in emission patterns and energy systems data.

Integrated Assessment Modeling. Table 1 Some higher resolution integrated assessment models

Some higher resolution integrated assessment models with interdisciplinary research teams		
Model	**Home institution**	**Web link**
AIM Asia-Pacific integrated model	National Institutes for Environmental Studies, Tsukuba, Japan	http://www-iam.nies.go.jp/aim/
GCAM Global change assessment model	Joint Global Change Research Institute, PNNL, College Park, MD	http://www.globalchange.umd.edu/models/gcam/
IGSM Integrated global system model	Joint Program on the Science and Policy of Global Change, MIT, Cambridge, MA	http://globalchange.mit.edu/igsm/
IMAGE The integrated model to assess the global environment	PBL Netherlands Environmental Assessment Agency, Bildhoven, The Netherlands	http://themasites.pbl.nl/en/themasites/image/
MERGE Model for evaluating the regional and global effects of GHG reduction policies	Electric Power Research Institute, Palo Alto, CA	http://www.stanford.edu/group/MERGE/
MESSAGE Model for energy supply strategy alternatives and their general environmental impact	International Institute for Applied Systems Analysis; Laxenburg, Austria	http://www.iiasa.ac.at/Research/ENE/model/message.html
ReMIND Refined model of investments and technological development	Potsdam Institute for Climate Impact Research; Potsdam, Germany	http://www.pik-potsdam.de/research/sustainable-solutions/models/remind/

The complex nature of the models requires interdisciplinary research and modeling teams, some of which are listed in Table 1.

Because the higher resolution IAMs have grown in their complexity over time, describing the structure of each model in detail is beyond the scope of this entry. For a reference, comparison of three IAMs – IGCM, MERGE, and MiniCAM (the direct ancestor of GCAM) – can be found in [14]. Here, we present the summary comparison table from the report in Table 2. All three of these modeling systems have evolved considerably in the subsequent years.

GCAM as an Example of a Higher Resolution IAM

Introduction to GCAM

Rather than try to describe and compare the set of higher resolution IAMs, we have chosen to describe here the Global Change Assessment Model (GCAM) as an example of the higher resolution IAM genre. GCAM is the oldest of the higher resolution IAMs. It traces its roots to work initiated in the late 1970s. The model's first applications were completed in the early 1980s by Edmonds and Reilly [15–18]. Over time the model has developed and evolved through a series of advances documented in a variety of papers including [19–22]. Documentation for GCAM under its previous name, MiniCAM, can be found at http://www.globalchange.umd.edu/models/MiniCAM.pdf/. Other higher resolution IAMs, such as IMAGE and MESSAGE, also use MAGICC to represent atmosphere and climate processes.

At the top level the GCAM model is broken into two interacting system, human Earth system and natural Earth systems. Each of these systems in turn is made up of subsystems. This is the basic structure of all IAMs. GCAM and the other higher resolution IAMs are distinguished from the highly aggregated IAMs in the degree of detail that is incorporated in describing human and natural Earth systems.

All higher resolution IAMs emphasize the representation of human activities and their connection to the sources of greenhouse gas emissions. However, each modeling team has taken a different approach.

Integrated Assessment Modeling. Table 2 Characteristics of the three integrated assessment models

Feature	IGSM (with EPPA economics component)	MERGE	MiniCAM
Regions	16	9	14
Time horizon, time steps	2100, 5-year steps	2200, 10-year steps	2095, 15-year steps
Model structure	General equilibrium	General equilibrium	Partial equilibrium
Solution	Recursive dynamic	Inter-temporal optimization	Recursive dynamic
Final energy demand sectors in each region	Households, private transportation, commercial transportation, service sector, agriculture, energy-intensive industries, and other industry	A single, nonenergy production sector	Buildings, transportation, and industry (including agriculture)
Capital turnover	Five vintages of capital with a depreciation rate	A putty clay approach wherein the input-output coefficients for each cohort are optimally adjusted to the future trajectory of prices at the time of investment	Vintages with constant depreciation rate for all electricity-sector capital; capital structure not explicitly modeled in other sectors
Goods in international trade	All energy and nonenergy goods as well as emissions permits	Energy, energy-intensive industry goods, emissions permits, and representative tradable goods	Oil, coal, natural gas, biomass, agricultural goods, and emissions permits
Emissions	CO_2, CH_4, N_2O, HFCs, PFCs, SF_6, CO, NOx, SOx, NMVOCs, BC, OC, NH_3	CO_2, CH_4, N_2O, long-lived F-gases, short-lived F-gases, and SOx	CO_2, CH_4, N_2O, CO, NOx, SO_2, NMVOCs, BC, OC, HFC245fa, HFC134a, HFC125, HFC143a, SF_6, C_2F_6, and CF_4
Land use	Agriculture (crops, livestock, and forests), biomass land use, and land use for wind and/or solar energy	Reduced-form emissions from land use; no explicit land-use sector; assume no net terrestrial emissions of CO_2	Agriculture (crops, pasture, and forests) as well as biomass land use and unmanaged land; the agriculture-land-use module directly determines land-use change emissions and terrestrial carbon stocks
Population	Exogenous	Exogenous	Exogenous
GDP growth	Exogenous productivity growth assumptions for labor, energy, and land; exogenous labor force growth determined from population growth; endogenous capital growth through savings and investment	Exogenous productivity growth assumptions for labor and energy; exogenous labor force growth determined from population growth; endogenous capital growth through savings and investment	Exogenous productivity growth assumptions for labor; exogenous labor force growth based on population demographics
Energy efficiency change	Exogenous	Proportional to the rate of GDP growth in each region	Exogenous

Integrated Assessment Modeling. Table 2 (Continued)

Feature	IGSM (with EPPA economics component)	MERGE	MiniCAM
Energy resources	Oil (including tar sands), shale oil, gas, coal, wind and/or solar, land (biomass), hydro, and nuclear fuel	Conventional oil, unconventional oil (coal-based synthetics, tar sands, and shale oil), gas, coal, wind, solar, biomass, hydro, and nuclear fuel	Conventional oil, unconventional oil (including tar sands and shale oil), gas, coal, wind, solar, biomass (waste and/or residues and crops), hydro, and nuclear fuel (uranium and thorium); includes a full representation of the nuclear fuel cycle
Electricity technologies	Conventional fossil (coal, gas, and oil), nuclear, hydro, natural gas combined cycle (NGCC) with and without capture, integrated coal gasification with capture, and wind and/or solar, biomass	Conventional fossil (coal, gas, and oil), nuclear, hydro, new coal and gas with and without CCS, other renewables	Conventional fossil (coal, gas, and oil) with and without capture; IGCCs with and without capture; NGCC with and without capture; Gen II, III, and IV reactors and associated fuel cycles; hydro, wind, solar, and biomass (traditional and modern commercial)
Conversion technologies	Oil refining, coal gasification, and bio-liquids	Oil refining, coal gasification and liquefaction, bio-liquids, and electrolysis	Oil refining, natural gas processing, natural gas to liquids conversion, coal, and biomass conversion to synthetic liquids and gases; hydrogen production using liquids, natural gas, coal, biomass; and electrolysis, including direct production from wind and solar, and nuclear thermal conversion
Atmosphere – ocean	2-dimensional atmosphere with a 3-dimensional ocean general circulation model, resolved at 20 minute time steps, 4° latitude, 4 surface types, and 12 vertical layers in the atmosphere	Parameterized ocean thermal lag	Global multi-box energy balance model with upwelling-diffusion ocean heat transport
Carbon cycle	Biogeochemical models of terrestrial and ocean processes; depends on climate and/or atmospheric conditions with 35 terrestrial ecosystem types	Convolution ocean carbon cycle model assuming a neutral biosphere	Globally balanced carbon cycle with separate ocean and terrestrial components, with terrestrial response to land-use changes
Natural emissions	CH_4, N_2O, and weather and/or climate dependent as part of biogeochemical process models	Fixed natural emissions over time	Fixed natural emissions over time
Atmospheric fate of GHGs, pollutants	Process models of atmospheric chemistry resolved for urban and background conditions	Single box models with fixed decay rates. No consideration of reactive gases	Reduced-form models for reactive gases and their interactions
Radiation code	Radiation code accounting for all significant GHGs and aerosols	Reduced form, top-of-the-atmosphere forcing	Reduced form and top-of-the-atmosphere forcing; including indirect forcing effects

Source: Clarke et al. [14]

For example, the IGSM employs a computable general equilibrium (CGE) model of the economy [23]. CGE models emphasize the structure of the economy and the interaction of economic sectors with each other and with labor and capital markets. The MERGE model also employs a highly aggregated CGE model in combination with more highly disaggregated energy sector models all embedded in an intertemporal optimization framework [24, 25]. The Asia-Pacific Integrated Model (AIM) employs a set of models that are used in combination [26]. The GCAM model uses a partial equilibrium framework, rather than a CGE framework. Partial equilibrium models delve into more detail in sectors that are directly related to the analysis in question (e.g., energy supply and demand, agricultural production, land use, and land-use change), and treat other sectors of the economy in aggregate.

The GCAM model drives the scale of human activities for each of its 14 geopolitical regions utilizing assumptions about future labor force – determined by working-age population, labor participation, and unemployment rate assumptions – along with the assumptions about labor productivity growth. The highly disaggregated energy, agriculture, and land-use components of GCAM are driven by the scale of human activity. The GCAM geopolitical regions are explicitly linked through international trade in energy commodities, agricultural and forest products, and other goods such as emissions permits.

The human dimension of the Earth system as shown in Fig. 2 integrates the energy system and the agriculture and land-use system, as well as the economic system that drives the activity in both systems. An important feature of the GCAM architecture is that the GCAM terrestrial carbon cycle model is embedded within the agriculture-land-use system model; that is, the agriculture-land-use system model explicitly calculates net land-use-change emissions from changes in land-use patterns over time. The energy system model produces and transforms energy for use in three end-use sectors: buildings, industry, and transport. The global human Earth systems are modeled for 14 geopolitical regions.

GCAM is a dynamic-recursive market equilibrium model. In each period of time the model's solution algorithm reconciles the supplies and demands for goods and services in all markets by finding a set of market-clearing prices. That market solution establishes the foundation from which the model steps forward to the next time period. Other IAMs, such as MERGE and MESSAGE, are built on an intertemporal optimization framework. These models solve all periods simultaneously so that expectations about the future are consistent with the model's future realizations in each time period. In contrast, GCAM, and other dynamic-recursive models, do not assume such intertemporal optimization takes place. Decisions taken in one period contain only expectations about future market conditions. These expectations will not necessarily be realized in the future. In other words, the economic agents in GCAM make decisions based on a less-than-perfect foresight, and the agents' only recourse in the subsequent period is to make another set of decisions, which can also be suboptimal.

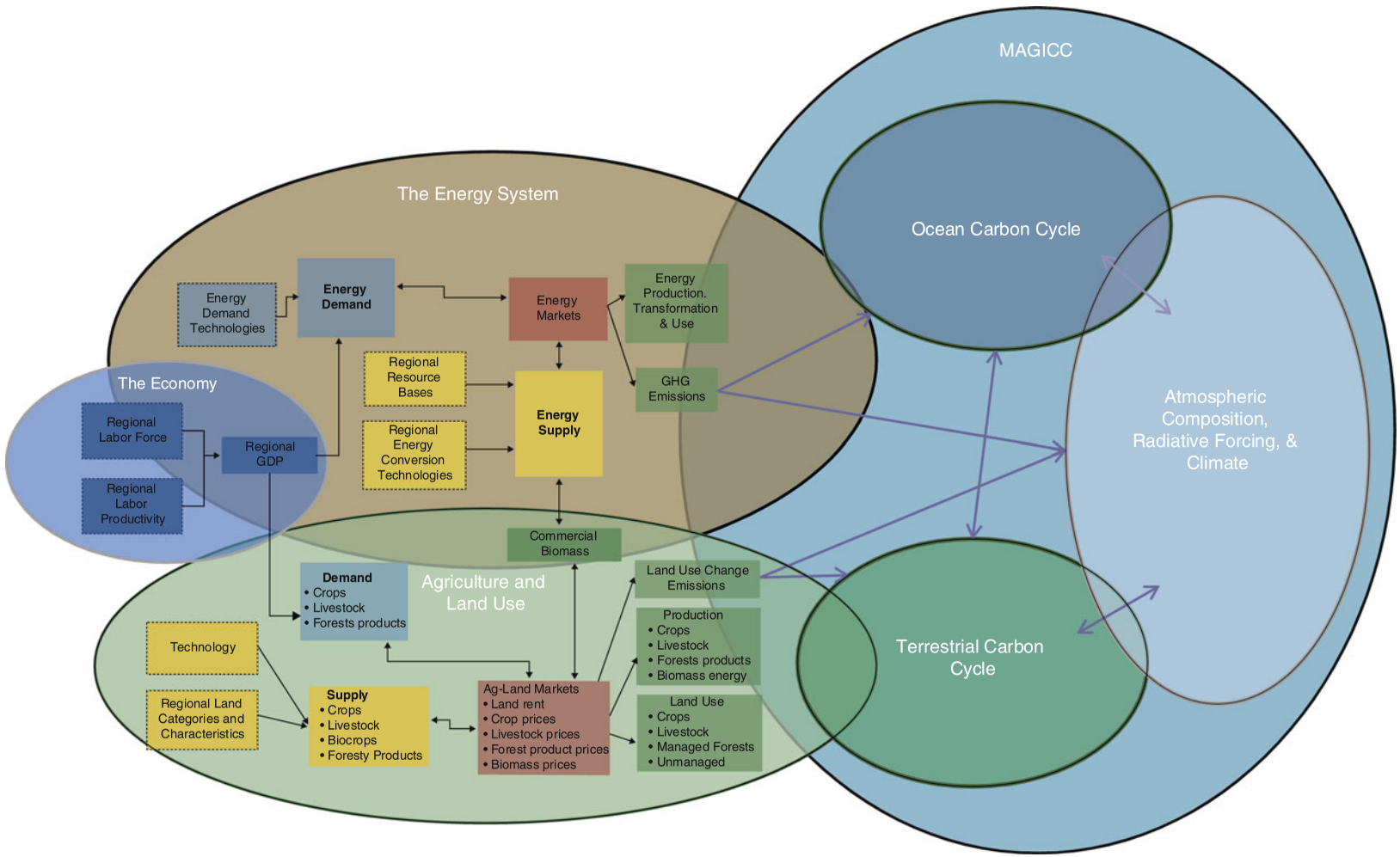
The GCAM's time step is variable, but in general is set to 5 years, which is relatively common among integrated assessment models. GCAM tracks 16 different greenhouse gases, aerosols, and short-lived species. The GCAM physical atmosphere and climate are represented by the Model for the Assessment of Greenhouse-Gas Induced Climate Change (MAGICC) [27–29].

In the remainder of this section, we discuss in more detail two of the most important model components in GCAM: the representation of the energy system and the representation of agriculture and land use more generally.

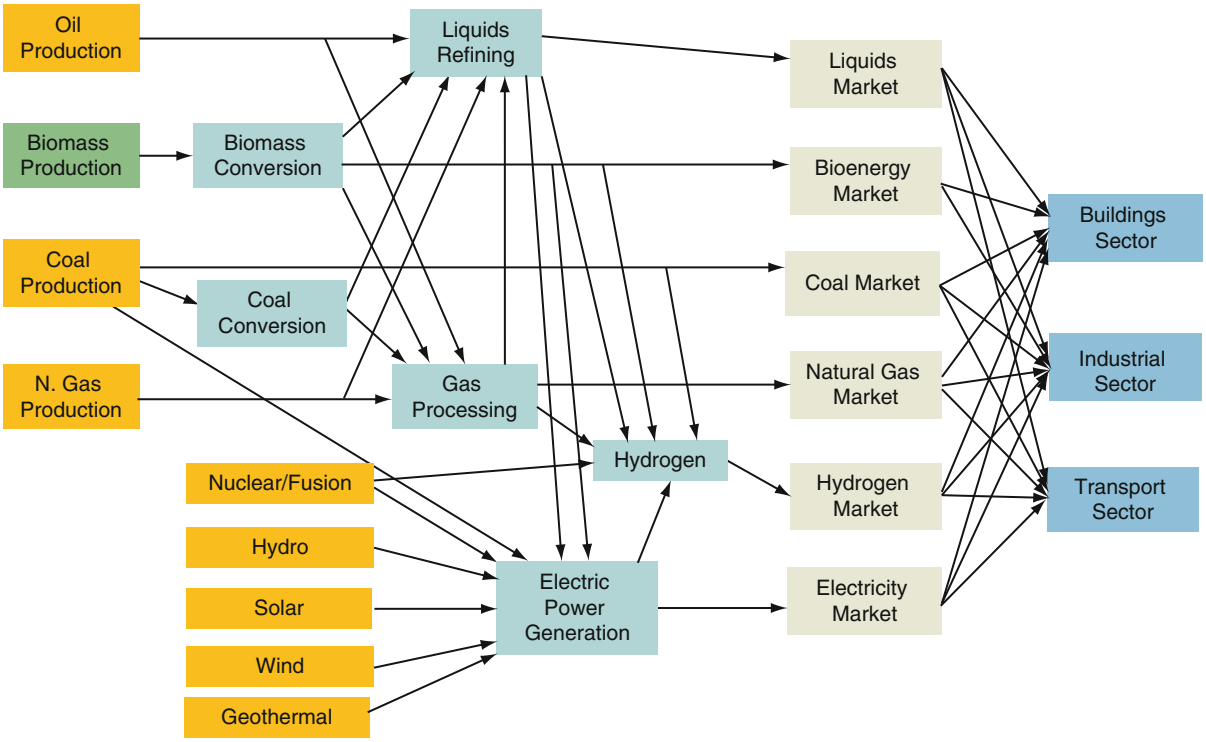
The Energy System in GCAM

In GCAM, the energy system represents processes of energy resource extraction, transformation, and delivery, ultimately producing services demanded by end users (Fig. 3). In each time period, the market prices of all goods and services, including primary energy resources, land, agricultural goods, and other products, are determined by the market equilibrium.

Primary energy production is limited by regional resource availability. Fossil fuel and uranium resources are finite, graded, and depletable. Wind, solar, hydro, and geothermal resources are also finite and graded, but renewable. Bioenergy is also renewable, but is treated as an explicit product of the agriculture-land-use portion of the model. Extraction costs for



Integrated Assessment Modeling. Figure 2
Human and natural Earth systems of the global change assessment model



Integrated Assessment Modeling. Figure 3
The energy system in GCAM

graded resources rise as the resource consumption increases, but can fall with improvement in extraction technologies, and can rise or fall depending on other environmental costs.

Primary energy forms can be transformed into six final energy products:

- Refined liquid energy products (oil and oil substitutes)
- Processed gas products (natural gas and other artificially gasified fuels)
- Coal
- Bioenergy solids (various forms of biomass)
- Electricity
- Hydrogen

Energy transformation sectors convert resources initially into fuels, which may be consumed by either other energy transformation sectors or ultimately into goods and services consumed by end users. In each energy sector, multiple technologies compete for market share; shares are allocated among competing technologies using a logit choice formulation [30–32]. The cost of a technology in any period is determined by two key exogenous input parameters – the nonenergy cost and the efficiency of energy transformation – as well as the prices of the fuels it consumes. The nonenergy cost represents all fixed and variable costs incurred over the lifetime of the equipment (except for fuel costs), amortized into a unit cost of output. For example, a coal-fired electricity plant incurs a range of costs associated with construction (a capital cost) and annual operations and maintenance. The efficiency of a technology determines the amount of fuel required to produce each unit of output (e.g., the fuel efficiency of a vehicle in passenger-km per GJ, or the electricity generation efficiency of a coal-fired power plant). The prices of different fuels are calculated endogenously in each time period based on supplies, demands, and resource depletion.

The representation of energy technologies in GCAM is highly disaggregated. Table 3 shows, for example, the set of technologies with accompanying assumptions of

Integrated Assessment Modeling. Table 3 Residential sector efficiencies by service and technology (Source: Kyle et al. [79])

Residential				Reference		Advanced	
Service	Technology	unit	2005	2050	2095	2050	2095
	Building shell	W/m^2	0.232	0.182	0.150	0.163	0.125
Heating	Gas furnace	Out/in	0.82	0.90	0.97	Same as Ref	
	Gas heat pump	Out/in	n/a	n/a	n/a	1.75	2.45
	Electric furnace	Out/in	0.98	0.99	0.99	Same as Ref	
	Electric heat pump	Out/in	2.14	2.49	2.79	2.94	4.12
	Oil furnace	Out/in	0.82	0.86	0.93	Same as Ref	
	Wood furnace	Out/in	0.40	0.42	0.44	Same as Ref	
Cooling	Air conditioning	Out/in	2.81	3.90	4.88	4.59	7.19
Water heating	Gas water heater	Out/in	0.56	0.61	0.64	0.79	0.88
	Gas HP water heater	Out/in	0.89	1.09	1.22	1.75	2.45
	Electric water heater	Out/in	0.88	0.93	0.97	Same as Ref	
	Electric HP water heater	Out/in	n/a	2.46	2.75	2.75	3.45
	Oil water heater	Out/in	0.55	0.56	0.59	Same as Ref	
Lighting	Incandescent lighting	Lumens/W	14	15	16	Same as Ref	
	Fluorescent lighting	Lumens/W	60	75	94	Same as Ref	
	Solid-state lighting	Lumens/W	100	112	125	156	245
Appliances	Gas appliances	Indexed	1.00	1.12	1.25	Same as Ref	
	Electric appliances	Indexed	1.00	1.23	1.38	1.44	2.01
Other	Other gas	Indexed	1.00	1.12	1.25	Same as Ref	
	Other electric	Indexed	1.00	1.08	1.21	1.40	1.96
	Other oil	Indexed	1.00	1.12	1.25	Same as Ref	

technology change over time, for the detailed US representation of residential buildings in GCAM.

Other energy sectors in GCAM have similar, high degrees of technology disaggregation. There are, for example, multiple technology options for generating electric power which include a variety of technologies utilizing solar energy as well as technology options to capture, transport, and store CO_2 in geologic repositories (CCS). The deployment of CCS technology in conjunction with bioenergy is of special interest in the consideration of very low long-term limits on CO_2 concentrations in that this combination potentially allows the production of energy with negative net CO_2 emissions. We discuss this particular technology combination in greater detail in a subsequent section of this entry.

Agriculture and Land Use in GCAM

Overview of the Agriculture and Land-Use Model in GCAM Land use is one of the largest anthropogenic sources of emissions of greenhouse gases, aerosols, and short-lived species. The conversion of grasslands and forests to agricultural land results in a net emission of CO_2 to the atmosphere. In the nineteenth century, the conversion of forests to agricultural land was the largest source of anthropogenic carbon emissions. In the future, biomass energy crops could compete for

agricultural land with traditional agricultural crops, providing a crucial linkage between land use and the energy system. Efforts to sequester carbon in terrestrial reservoirs, such as forests, may limit deforestation activities, and potentially lead to afforestation or reforestation activities. Interactions with crop prices may also prove important. Since land is limited, increasing the demand for land either to protect forests or to plant bioenergy crops could put upward pressure on crop prices that would not otherwise occur [33].

Many higher resolution IAMs include representations of agriculture, land use, and land cover. For some models, such as IGSM or IMAGE, a separate ecosystem model is used to represent terrestrial systems, which is then loosely coupled to the other elements of the IAM. These models represent land use, land cover, and the terrestrial carbon cycle. The IGSM model employs the Terrestrial Ecosystems Model [23], while IMAGE employs their terrestrial environment system submodel [34]. Since these models represent terrestrial processes at fine geographic scales – ½ degree by ½ degree gridded maps, for example – land use is determined by coupling an aggregated model of agriculture with a downscaling algorithm.

GCAM uses a model of land use and land cover, which allocates land area within each of its 14 global geopolitical regions among different land uses and tracks production from these uses and corresponding carbon flows into and out of terrestrial reservoirs. The GCAM agriculture, land use, land cover, terrestrial carbon cycle module determines the demands for and production of agricultural products, the prices of these products, the allocation of land to competing ends, and the carbon stocks and flows associated with land use.

Land is allocated between alternative uses based on expected profitability, which in turn depends on the productivity of the land-based product (e.g., mass of harvestable product per hectare), product price, and non-land costs of production (labor, fertilizer, etc.). The allocation of land types takes place in the model through global and regional markets for agricultural products. These markets include those for raw agricultural products as well as those for intermediate products such as poultry and beef. Demands for most agricultural products, with the exception of biomass products, are driven primarily by income and population. Land allocations evolve over time through the operation of these markets, in response to changes in income, population, technology, and prices.

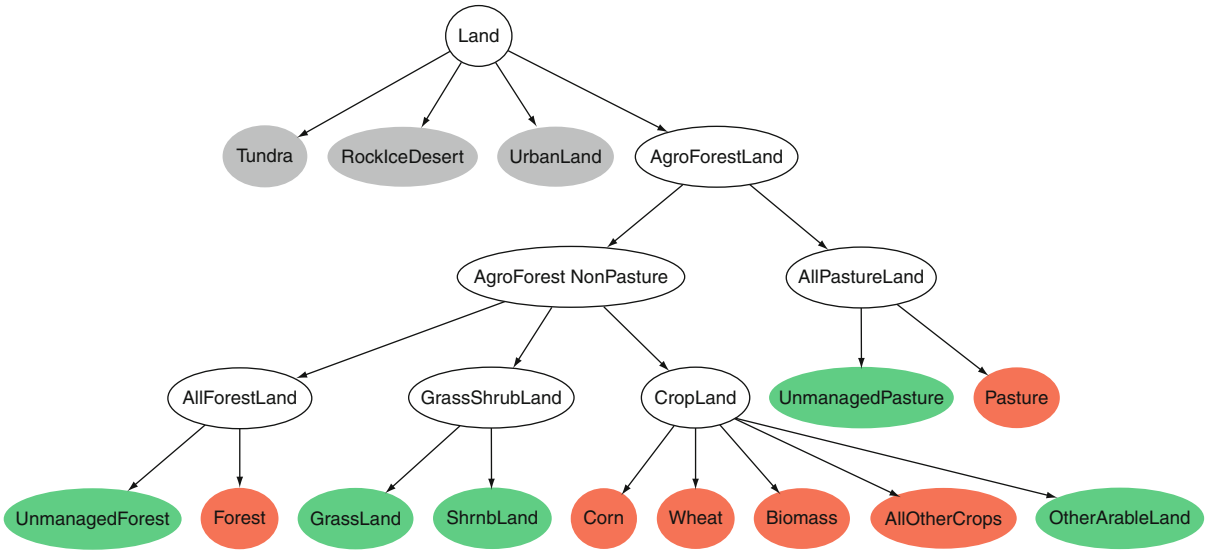
The boundary between managed and unmanaged ecosystems is assumed to be elastic in GCAM. The area of land under cultivation expands and contracts as crops become more or less profitable. Thus, increased demands for land result in higher cropland profitability and expansion into unmanaged ecosystems and vice versa. Competition between alternative land uses in the GCAM is modeled using a nested logit architecture [30–32] as depicted in Fig. 4.

The costs of supplying agricultural products are based on regional characteristics, such as the productivity of land and the variable costs of producing the crop. The productivity of land-based products is subject to change over time based on future estimates of crop productivity change. It has been shown that the rate of crop yield improvement is a critical determinant of land-use change emissions [33, 35–37].

Bioenergy in GCAM's Agriculture and Land-Use Model Bioenergy supply is determined by the agriculture-land-use component (AgLU) of GCAM, while bioenergy demand is determined in the energy component of the model. For example, the larger the value of carbon, the more valuable biomass is as an energy source and hence the greater the price the energy markets will be willing to pay for biomass. Conversely, as populations grow and incomes increase, competing demands for land may drive down the amount of land that would be available for biomass production at a given price.

There are three types of bioenergy produced in GCAM: traditional bioenergy production and use, bioenergy from waste products, and purpose-grown bioenergy. Traditional bioenergy consists of straw, dung, fuel wood, and other energy forms that are utilized in an unrefined state in the traditional sector of an economy. Traditional bioenergy use, although significant in developing nations, is a relatively small component of global energy. Traditional bioenergy is modeled as a function of regional income levels with its use diminishing as per capita incomes rise.

Other two types of bioenergy products are fuels that are consumed in the modernized sectors of the economy. Bioenergy from waste products are



Integrated Assessment Modeling. Figure 4

Competition for land in GCAM. Gray exogenous in future periods, *Green* unmanaged land use, *Red* managed land use. AgLU tracks carbon content in different land uses. Changes in land use result in carbon flux to the atmosphere. Land owners compare economic returns across crops, biomass, pasture, and (future) forest, based on underlying probability distribution of yields per hectare

by-products of another activity. Examples in the model include forestry and milling by-products, crop residues in agriculture, and municipal solid waste. The availability of byproduct energy feedstocks is determined by the underlying production of primary products and the cost of collection. The total potential agricultural waste available is calculated as the total mass of the crop less the portion that is harvested for food, grains, and fibers, and the amount of bioenergy needed to prevent soil erosion and nutrient loss and sustain the land productivity. The amount of potential waste that is converted to bioenergy is based on the price of bioenergy.

The third category of bioenergy is purpose-grown energy crops. Purpose-grown bioenergy refers to crops whose primary purpose is the provision of energy. These would include, for example, switchgrass and woody poplar. The profitability of purpose-grown bioenergy depends on the expected profitability of growing and selling that crop relative to other land-use options in GCAM. This in turn depends on numerous other model factors: in the agricultural sector, bioenergy crop productivity (which in turn depends on the character of available land as well as crop type and technology) and nonenergy costs of crop production, and in the fuel processing sector, cost and efficiency of transformation of purpose-grown bioenergy crops to final energy forms (including liquids, gases, solids, electricity, and hydrogen), cost of transportation to the refinery, and the price of final energy forms. Furthermore, the price of final energy forms is determined endogenously as a consequence of competition between alternative energy resources, transformation technologies, and end-use energy service delivery technologies. In other words, prices are determined so as to simultaneously match demand and supplies in all energy markets as well as all land-use markets.

A variety of crops could potentially be grown as bioenergy feedstocks. The productivity of those crops will depend on where they are grown – which soils they are grown in, climate characteristics and their variability, whether or not they are fertilized or irrigated, the availability of nitrogen and other minerals, ambient CO_2 concentrations, and their latitude. GCAM typically include a generic bioenergy crop, with its characteristics similar to switchgrass that is assumed to be grown in all regions. Productivity is based on

region-specific climate and soil characterizes and varies by a factor of three across the GCAM regions. GCAM allows for the possibility that bioenergy could be used in the production of electric power and in combination with technologies to provide CO_2 emissions captured and stored in geological reservoirs (CCS). This particular technology combination is of interest because bioenergy obtains its carbon from the atmosphere and if that carbon were to be captured and isolated permanently from the atmosphere the net effect of the two technologies would be to produce energy with negative CO_2 emissions. See, for example, [33, 38].

Pricing Carbon in Terrestrial Systems Efficient climate policies are those that apply an identical price to greenhouse gas emissions wherever they occur. Hence, an efficient policy is one that applies identical prices to land-use change emissions and fossil and industrial emissions. This efficient approach is used as the default for emissions mitigation scenarios, though other policy options have also been modeled (A change in atmospheric CO_2 concentration has the same impact on climate change no matter what the source. Thus, to a first approximation land-use emissions have the same impact as fossil emissions. But, there are important differences. Land-use emissions do not have the same impact on atmospheric concentrations as fossil emissions because land-use emissions also imply changes in the future behavior of the carbon cycle. A tonne of carbon emitted due to deforestation, for example, is associated with a decrease in forest that would otherwise act as a carbon sink in the future. This effect, however, is not currently captured in GCAM).

Carbon in terrestrial systems can be priced using either a flow approach or a stock approach. The flow approach is analogous to the pricing generally discussed for emissions in the energy sector: landowners would receive either a tax or a subsidy based on the net flow of carbon in or out of their land. If they cut down a forest to grow bioenergy crops, then they would pay a tax on the CO_2 emissions from the deforestation. In contrast, the stock approach applies a tax or a subsidy to landowners based on the carbon content of their land. If the carbon content of the land changes, for example, by cutting forests to grow bioenergy crops, then the tax or subsidy that the landowner receives is adjusted to represent the new carbon stock in the land. The stock approach can be viewed as applying a "carbon rental rate" on the carbon in land. Both approaches have strengths and weaknesses. Real-world approaches may not be explicitly one or the other. By default, GCAM uses the stock approach.

Using Higher Resolution IAMs to Analyze the Impact of Policies to Mitigate Greenhouse Gas Emissions

A Brief Overview of IAMs in Mitigation Policy Analysis

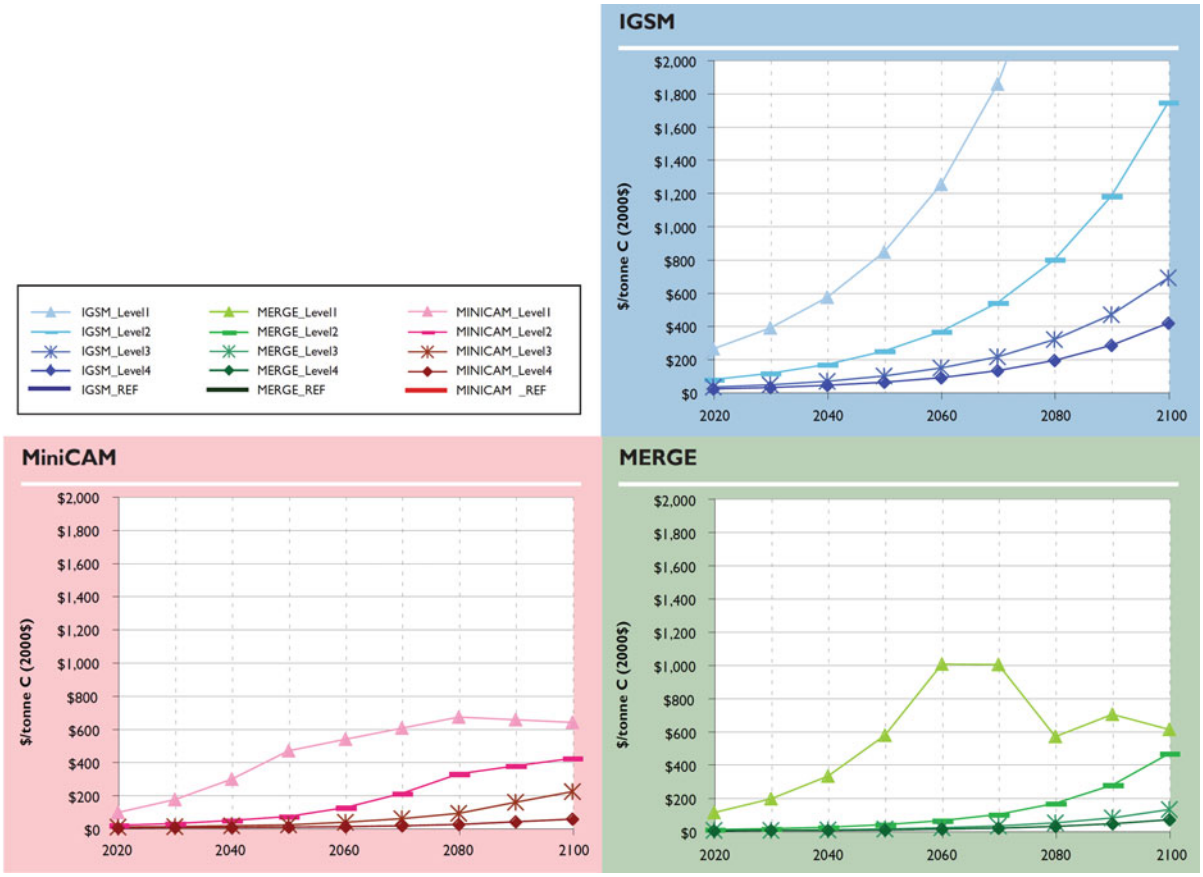
Higher resolution IAMs have been used extensively to estimate the effects of measures to reduce greenhouse gas emissions. Until recently, the great bulk of the literature focused on the analysis of idealized policy instruments, particularly carbon taxes and cap-and-trade policies. For example, an important vein of early analysis focused on the question of emissions trading. In general, this literature showed that emissions mitigation undertaken with tradable permits resulted in lower costs to all parties without any reduction in overall emissions mitigation (see, for example, [39, 78]. The basic architecture of the Kyoto Protocol [40] reflected this line of thought. The application of these idealized pollution pricing mechanisms was inherently straightforward in higher resolution IAMs because these IAMs' representations of the energy and terrestrial systems are all built on economic principles. Furthermore, these mechanisms were of interest because they were theoretically attractive for the efficiency with which they reduced emissions.

In all the stabilization scenarios, the carbon price rises, by design, over time until stabilization is achieved (or the end-year 2100 is reached), and the prices are higher the more stringent is the stabilization level. There are substantial differences in carbon prices between MERGE and MiniCAM stabilization scenarios, on the one hand, and the IGSM stabilization scenarios on the other. Differences between the models reflect differences in the emissions reductions necessary for stabilization and differences in the technologies that might facilitate carbon emissions reductions, particularly in the second half of the century.

Whether for CO_2 or for multiple gases, a major focus of analysis has been to compute minimum-cost emissions trajectories for meeting long-term

stabilization goals. The minimum cost is generally calculated on the assumption that all regions of the world undertake emissions mitigation in a coordinated, intertemporal program that reduces emissions in an economically efficient manner. One key characteristic of this pathway is that the marginal cost of emissions mitigation is equal in all sectors and in all regions at any point in time. It also means that the price of CO_2 rises at the rate of interest plus the rate of removal of CO_2 from the atmosphere until stabilization is reached [41]. After stabilization is reached, the CO_2 price no longer rises at this roughly constant rate, but instead is determined so as to ensure that at any point in time emissions match uptake so concentrations remain constant. Examples of classic stabilization CO_2 price pathways are shown in Fig. 5.

While mitigation cost may be one of the core questions addressed by the higher resolution IAMs, it is not the only question. A second and complementary set of questions focuses on implications for energy and agricultural systems, the next level of detail upon which higher resolution IAMs focus. How fast must the energy system change? Which technologies need to be deployed and when (see, e.g., [42, 43])? Stabilization of the concentration of CO_2 at any level requires that net anthropogenic carbon emissions must peak and decline indefinitely toward zero [1], but an almost infinite set of combinations of technology could in



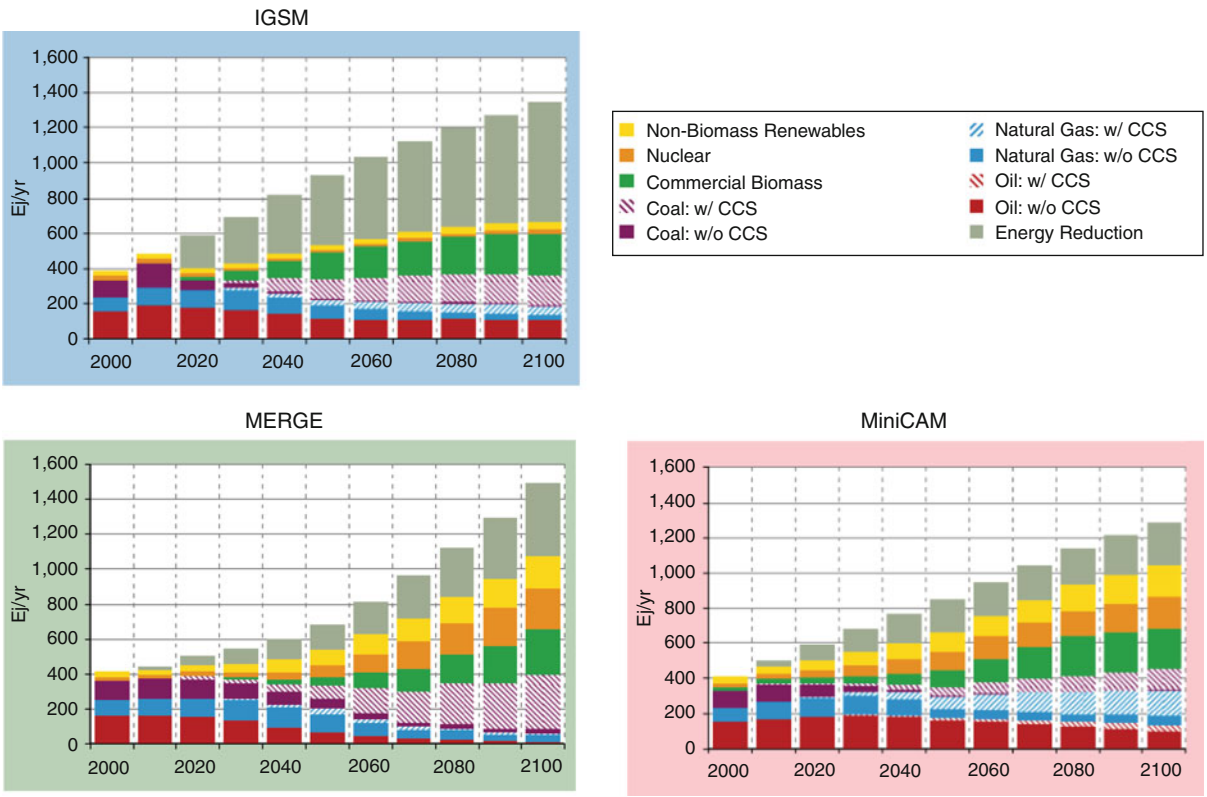
Integrated Assessment Modeling. Figure 5
Carbon prices across stabilization scenarios ($/ton C, 2000$) from three higher resolution IAMs leading to stabilization at approximately 750 ppmv CO_2 (Level 4), 650 ppmv CO_2 (Level 3), 550 ppmv CO_2 (Level 2), and 450 ppmv CO_2 (Level 1) (Source: Clarke et al. [14])

principle deliver that outcome. For example, fossil fuel use could be replaced with renewable energy forms in combination with energy efficiency improvements. Alternatively, fossil fuels could continue to be deployed in the global energy system in combination with CO_2 capture and storage (CCS), nuclear power, renewable energy, and energy efficiency. The combinations that emerge from different models depend on assumptions about technology performance and availability, scale of the economic system, and climate policy. A wide range of studies has made evolution of the energy system to meet long-term goals a focus of analysis (see, e.g., Fig. 6 from [14]).

Stabilization in IAMs with Multiple Greenhouse Gases

The United Nations Framework Convention on Climate Change (UNFCCC) has as its goal the stabilization of the concentration of greenhouse gases in the atmosphere. As discussed above, examination of the cost of stabilization of CO_2 and other gases has been the focus of a great number of papers utilizing higher resolution IAMs. Early studies focused exclusively on stabilization. However, more recent efforts have explored stabilization considering multiple greenhouse gases [14, 44].

When multiple greenhouse gases are considered simultaneously the problem emerges as to how to compare the greenhouse effects across the various constituents. In terms of climate change, the natural aggregate measure is radiative forcing (see Box 2). It is relatively straightforward to compute the radiative forcing for a group of gases, aerosols, and short-lived species and then to estimate what concentration of CO_2 would yield that radiative forcing level if all other species were set at their preindustrial levels. The answer to that question is the CO_2-equivalent concentration for that bundle of gases.



Integrated Assessment Modeling. Figure 6
Global primary energy production across scenarios from three higher resolution IAMs leading to approximately 450 ppmv CO_2 (Source: Clarke et al. [14])

Box 2. Radiative Forcing

Most of the Sun's energy that reaches the Earth is absorbed by the oceans and land masses and radiated back into the atmosphere in the form of heat or infrared radiation. Some of this infrared energy is absorbed and reradiated back to the Earth by atmospheric gases, including water vapor, CO_2, and other substances. As concentrations of GHGs increase, there are direct and indirect effects on the Earth's energy balance. The direct effect is often referred to as a radiative forcing, a subset of a more general set of phenomena referred to as climate forcings. The National Research Council [45] offers the following set of definitions:

- Factors that affect climate change are usefully separated into forcings and feedbacks. . . . A climate forcing is an energy imbalance imposed on the climate system either externally or by human activities. Examples include changes in solar energy output, volcanic emissions, deliberate land modification, or anthropogenic emissions of greenhouse gases, aerosols, and their precursors. A climate feedback is an internal climate process that amplifies or dampens the climate response to an initial forcing. An example is the increase in atmospheric water vapor that is triggered by an initial warming due to rising carbon dioxide (CO_2) concentrations, which then acts to amplify the warming through the greenhouse properties of water vapor.. . .

  Climate forcing: An energy imbalance imposed on the climate system either externally or by human activities.

- *Direct radiative forcing*: A climate forcing that directly affects the radiative budget of the Earth's climate system; for example, added carbon dioxide (CO_2) absorbs and emits infrared radiation. Direct radiative forcing may be due to a change in concentration of radiatively active gases, a change in solar radiation reaching the Earth, or changes in surface albedo. Radiative forcing is reported in the climate change scientific literature as a change in energy flux at the tropopause, calculated in units of watts per square meter (W/m^2); model calculations typically report values in which the stratosphere was allowed to adjust thermally to the forcing under an assumption of fixed stratospheric dynamics.
- *Indirect radiative forcing*: A climate forcing that creates a radiative imbalance by first altering climate system components (e.g., precipitation efficiency of clouds), which then almost immediately lead to changes in radiative fluxes. Examples include the effect of solar variability on stratospheric ozone and the modification of cloud properties by aerosols.
- *Nonradiative forcing*: A climate forcing that creates an energy imbalance that does not immediately involve radiation. An example is the increasing evapotranspiration flux resulting from agricultural irrigation.

Source: Clarke et al. [14], Box 1.1; NRC [45]

Two approaches have been used to determine the optimal mix of abatement across gases in stabilization. One approach is to minimize the total costs of meeting a long-term radiative forcing target, based on the combined mitigation costs for all greenhouse gases using intertemporal optimization. This is the approach employed by intertemporal optimization models such as MERGE. In this structure, all of the prices of the different greenhouse gases rise at relatively constant rates until stabilization is reached, consistent with the general result for minimum-cost CO_2 pathways discussed in the previous section [41], but the rates vary among gases. This leads to different timing of mitigation across gases. Indeed, one of the outcomes of this sort of approach to multi-gas stabilization is that the rate of increase in greenhouse gas prices is higher for gases with shorter lifetimes, with the implication that mitigation for these gases is delayed relative to CO_2. For example, this approach leads to scenarios in which mitigation of CH_4 is relatively modest in the early term and then increases dramatically as the total radiative forcing target gets close.

An alternative, though less rigorous methodology that is used to compare greenhouse gases in multi-gas emissions mitigation programs is the application of Global Warming Potential (GWP) coefficients. This is the approach generally used by dynamic-recursive models such as GCAM. The GWP was developed as an analogue to the Ozone Depletion Potential (ODP) coefficients employed to compare the various

stratospheric ozone depleting substances [46]. GWPs are defined as the effect on radiative forcing of the release of an additional kilogram of a gas, relative to the simultaneous release of a kilogram of CO_2, integrated over one of three time horizons: 20 years, 100 years, and 500 years. Values for the GWPs calculated by IPCC Working Group I in the Fourth Assessment Report [47] are given in Table 4. GWPs are something of a mixture between the relative contribution of a gas to radiative forcing, which would be better calculated directly if possible, and an incomplete estimate of climate damage associated with the release of an additional kilogram of a greenhouse gas.

The primary virtue in the GWP is its application as an estimate of the relative importance of various greenhouse gases by national, local, and regional parties. Multi-gas policy instruments often employ GWPs as a means of comparing emissions of different greenhouse gases. The ratio of any pair of GWPs serves as the inverse of the relative price of any pair of greenhouse gases.

In application to stabilization studies in IAMs, GWPs yield constant estimates of the relative contributions of various greenhouse gases to climate change. In other words, since the GWPs are assumed to be constant over time, the relative prices of CO_2 and other gases are also constant over time. Hence, in studies that use GWPs to achieve multi-gas stabilization, mitigation for gases with shorter lifetimes generally takes place more quickly that would be the case in models that employ an intertemporal optimization approach. In this sense, although GWPs are a reality in policy design, they are an imperfect tool for comparing greenhouse gases over time. Manne and Richels [52] showed that if the total cost is the only criteria by which emissions pathways are judged then GWPs were not constant, but would rather change systematically with time. Peck and Wan [41] showed that if minimizing the total cost of limiting radiative forcing were the sole criterion by which greenhouse gas concentrations were controlled then the shadow price of each greenhouse gas rises at the interest rate plus the rate of removal from the atmosphere. Hence the corresponding GWP ratio of any two gases changes over time at a rate equal to the removal rate difference between the two gases. This notion is profoundly different than the concept of the GWP as a constant.

Manne and Richels [52] did show that the inclusion of secondary criteria, in addition to limiting radiative forcing, such as limiting the rate of change of radiative forcing, could produce very different GWPs and rates of change in GWPs over time. Some combinations of objective criteria could generate relatively stable GWPs.

Integrated Assessment Modeling. Table 4 Direct global warming potential coefficients

Industrial designation or common name (years)	Chemical formula	Lifetime (years)	Radiative efficiency (Wm^{-2} ppb^{-1})	IPCC [48] (100^{-year})	20^{-year}	100^{-year}	500^{-year}
Carbon dioxide	CO_2	See notes[a]	[b]1.4×10^{-5}	1	1	1	1
Methane[c]	CH_4	12[c]	3.7×10^{-4}	21	72	25	7.6
Nitrous oxide	N_2O	114	3.03×10^{-3}	310	289	298	153

[a]The CO_2 response function used in this report is based on the revised version of the Bern carbon cycle model (Bern2.5CC) [49] used in IPCC [47] Chap. 10 Global Climate Projections using a background CO_2 concentration value of 378 ppm. The decay of a pulse of CO_2 with time t is given by $a_0 + \sum_{i=1}^{3} a_i \cdot e^{-t/\tau_i}$

where $a_0 = 0.217$, $a_1 = 0.259$, $a_2 = 0.338$, $a_3 = 0.186$, $\tau_1 = 172.9$ years, $\tau_2 = 18.51$ years, and $\tau_3 = 1.186$ years

[b]The radiative efficiency of CO_2 is calculated using the IPCC [50] simplified expression as revised in the TAR, with an updated background concentration value of 378 ppm and a perturbation of +1 ppm (see IPCC [47], Sect. 2.10.2)

[c]The perturbation lifetime for methane is 12 years as in the IPCC [48] (see also [47], Sect. 7.4). The GWP for methane includes indirect effects from enhancements of ozone and stratospheric water vapor (see [47], Sect. 2.10.3.1)

Source: IPCC [43], Table 2.14, pp. 212–213

The Economic Costs of Implementing the Framework Convention on Climate Change

As mentioned above, estimating the costs of meeting long-term targets is a primary function of IAMs. Typical estimates for global costs of limiting CO_2 equivalent concentrations to alternative levels from the IPCC [43] are shown below for two representative years, 2030 (Table 5) and 2050 (Table 6).

While the question of the measurement of the economic cost of emissions mitigation has not generated as much debate as questions about discounting, there are important differences in methodology that different modeling teams employ. Perhaps the most commonly used metric comparable across models is the price of carbon. This metric is useful for comparing across models when simple policy instruments to mitigate emissions are employed – specifically either an economy-wide carbon tax or the carbon price emerging from an economy-wide cap-and-trade. As policy assumptions become more complex the usefulness of this metric fades. In fact, in mixed emissions mitigation systems, where only part of the economy is controlled

Integrated Assessment Modeling. Table 5 Estimated global macroeconomic costs in 2030[a] for least-cost trajectories toward different long-term stabilization levels[b, c]

Stabilization levels (ppm CO_2-eq)	Median GDP reduction[d] (%)	Range of GDP reduction[d, e] (%)	Reduction of average annual GDP growth rates[d, f] (percentage points)
590–710	0.2	−0.6 to 1.2	<0.06
535–590	0.6	0.2 to 2.5	<0.1
445–535[g]	Not available	<3	<0.12

[a]For a given stabilization level, GDP reduction would increase over time in most models after 2030. Long-term costs also become more uncertain
[b]Results based on studies using various baselines
[c]Studies vary in terms of the point in time stabilization is achieved; generally this is in 2100 or later
[d]These are global GDP-based market exchange rates
[e]The median and the 10th and 90th percentile range of the analyzed data are given
[f]The calculation of the reduction of the annual growth rate is based on the average reduction during the period till 2030 that would result in the indicated GDP decrease in 2030
[g]The number of studies that report GDP results is relatively small and they generally use low baselines
Source: IPCC [43], SPM, p. 12

Integrated Assessment Modeling. Table 6 Estimated global macroeconomic costs in 2050 for least-cost trajectories toward different long-term stabilization levels[a]

Stabilization levels (ppm CO_2-eq)	Median GDP reduction[b] (%)	Range of GDP reduction[b, c] (%)	Reduction of average annual GDP growth rates[b, d] (percentage points)
590–710	0.5	−1 to 2	<0.05
535–590	1.3	Slightly negative-4	<0.1
445–535[e]	Not available	<5.5	<0.12

[a]This corresponds to the full literature across all baselines and mitigation scenarios that provide GDP numbers
[b]These are global GDP-based market exchange rates
[c]The median and the 10th and 90th percentile range of the analyzed data are given
[d]The calculation of the reduction of the annual growth rate is based on the average reduction during the period until 2050 that would result in the indicated GDP decrease in 2050
[e]The number of studies is relatively small and they generally use low baselines. High emissions baselines generally lead to higher costs
Source: IPCC [43], SPM, p. 18

by a tax or cap-and-trade program, the carbon price and real economic cost can move in opposite directions. That is, as more of the high-cost sectors of the economy are controlled with less-efficient nonmarket-based policies, the price of carbon may fall while the total economic cost rises.

A variety of approaches have been applied to obtain the total economic cost. These include integration under the marginal abatement cost schedule, measurement of foregone consumption, and compensated/equivalent variation. Each of these approaches traces its method back to welfare economics. While measures that directly link to welfare functions are in principle best, welfare cannot be directly observed and unless highly unlikely circumstances prevail, Arrow [53] has shown that a welfare function with the properties needed to get a measure on real economic cost cannot exist – a distinct disadvantage for numerical simulations.

While the choice of methodological approach to measuring real economic cost will doubtless affect valuation, two larger sources of variation in cost estimates are the policy instruments applied and the assumed rate of technological improvement. It is well known that different policy instruments can attain the same mitigation level with different costs [54]. Differences in technology assumptions can also produce substantial differences in cost (see, e.g., [42, 55]). Exploring the implication of different policy instruments and technology availability are two important directions of future work by the higher resolution IAM research community.

The principal research question which the higher resolution IAMs addressed has been different from that of the highly aggregated IAMs. Whereas the highly aggregated IAMs focused on the problem of determining the optimal balance between emissions mitigation and adaptation to climate change, the higher resolution IAMs focused more on the cost of implementing a policy to limit emissions, concentrations, or combined radiative forcing of greenhouse gases. The higher resolution IAM community has generally taken an agnostic position on the question of whether the policy instrument or the policy goal in question was desirable or not and simply went about the task of calculating the cost of achieving the given goal of implementing the prescribed policy.

As time has passed, the political conversation has moved away from the question of the use of cap-and-trade to control emissions to consider hybrid policy architectures in which emissions mitigation is pursued through a combination of policy measures some of which differ substantially from the conventional market mechanisms, such as carbon taxes or cap-and-trade. For example, many current emissions mitigation proposals contain renewable portfolio standards (RPS). These policy instruments require a minimum fraction of total power generation to be provided by renewable energy forms such as wind and solar.

There are many reasons for the shift. The prospects for a comprehensive international agreement based on the principles of cap-and-trade have diminished. Many parties in the international negotiations were less concerned with economic efficiency and cost-minimization than they were with a sense of moral obligation to achieve domestic emissions mitigation targets without resorting to emission trading. Within the United States similar forces are at work. Efforts to develop a comprehensive countrywide emissions cap-and-trade system show little prospect for entering into effect. Also, the European Union and Japan have either chosen alternatives to cap-and-trade or employ cap-and-trade within limited sectors of the economy. Such policies have pushed IAMs to develop more sophisticated representations of policies in order to estimate the policy effects [56]. In the same context, the IAMs have begun to explore the implications of international regimes in which nations begin emissions mitigation at different times [57, 58].

Future Directions: Integrating Human Earth Systems with Natural Earth Systems

Integrated Assessment modeling research is a continuously evolving field. As the models have matured and diversified, researchers have pushed the development frontiers in multiple directions simultaneously in order to answer a wide range of research questions. For example, researchers have broadened the scope of the models to include more sectors of the human Earth system such as land use and agriculture. They have expanded coverage of various types of the greenhouse gases by including an increasingly diverse set of their sources and activities. They have also lengthened the

time horizon of analysis, pressing past the year 2100 and multiple centuries beyond. At the same time, the researchers have elaborated the key model components by slicing each of them in smaller pieces, for example, by adding finer spatial and temporal resolution and disaggregated representation of technologies.

An increasingly prominent research frontier has been the formal integration with other fields of climate change research, namely climate modeling (CM) and impacts, adaptation, and vulnerability (IAV) research. Although many research questions do not require the use of IPCC-class models of human and natural Earth systems, others cannot be addressed adequately without the development of integrated Earth systems models. The development of integrated Earth systems models opens the door to formally modeling the simultaneous interactions between human activities, climate change, and climate change impacts on human systems.

The Representative Concentration Pathways: An Example of Interactions with Climate Models

The assessment of climate change has traditionally been a linear research process. IA researchers produce emissions scenarios which in turn are transferred to the climate modeling community for use as inputs. The climate modeling community employs these scenarios to force future climate calculations. These climate calculations are then used by IAV researchers to produce estimates of the consequences of climate change. In the past, there has been little communication or feedback between research communities. Each community conducted its research independently and left it for others to figure out how or whether to use it. Beginning in 1990, the integrated assessment modeling community began to interact with the climate modeling community, though interactions with the carbon cycle and other natural Earth system researchers go back even further (see, e.g., [59], and more generally, [60]). Moss et al. [61] provide a succinct history of scenario development, which is summarized in Fig. 7.

There have been numerous long-term scenarios of global greenhouse gas emissions. Three important benchmarks were the publication of scenarios referred to as SA90 [51], IS92 [62], and SRES [63]. These scenarios are notable in that the climate modeling community used them to simulate potential effect of future emissions paths on the climate system. The earliest scenarios considered only fossil fuel CO_2 emissions. Over time scenarios became richer, including land-use change emissions, non-CO_2 greenhouse gases, and short-lived species. While these scenarios span a wide range of potential future emissions, none considered limitations on emissions, that is, until Moss et al. [61] and the publication of the Representative Concentration Pathways (RCPs).

The RCPs are the most recent set of scenarios developed for use in the climate models. They were chosen to initiate an assessment cycle by providing the climate modeling community with a set of scenarios that were sufficiently differentiated by the end of the century to be scientifically relevant and to provide detailed information on the sources of emissions of greenhouse gases and short-lived species from all anthropogenic sources. RCPs differ from earlier scenario development activities in that they were selected from existing scenarios that were available in the peer-reviewed literature rather than being developed de novo. Selected scenarios from the open literature were named corresponding to their century's end radiative forcing levels: 8.5, 6.0, 4.5, and 2. 6 Wm^{-2} (see Table 7).

Subsequent to selection, the four scenarios were updated and harmonized to include the most recent observational data and downscaled to produce harmonized gridded outputs for emissions, land use, and land cover. The resulting time-paths for radiative forcing are given in Fig. 8 (The detailed scenario data are available at www.iiasa.ac.at/web-apps/tnt/RcpDb/).

The RCPs differ from previous scenarios employed by the climate modeling community in that they

1. Include scenarios with explicit emissions mitigation
2. Provide geospatially resolved emissions at ½ degree by ½ degree
3. Provide geospatially resolved land use and land cover at ½ degree by ½ degree

The most recent set of scenarios, while highly useful to the climate modeling community, are less useful from the perspective of the impacts, adaptation, and vulnerability community. While the scenarios contain detailed information that would be of interest to climate modelers, they do not carry associated socioeconomic information, or energy or commodity prices.

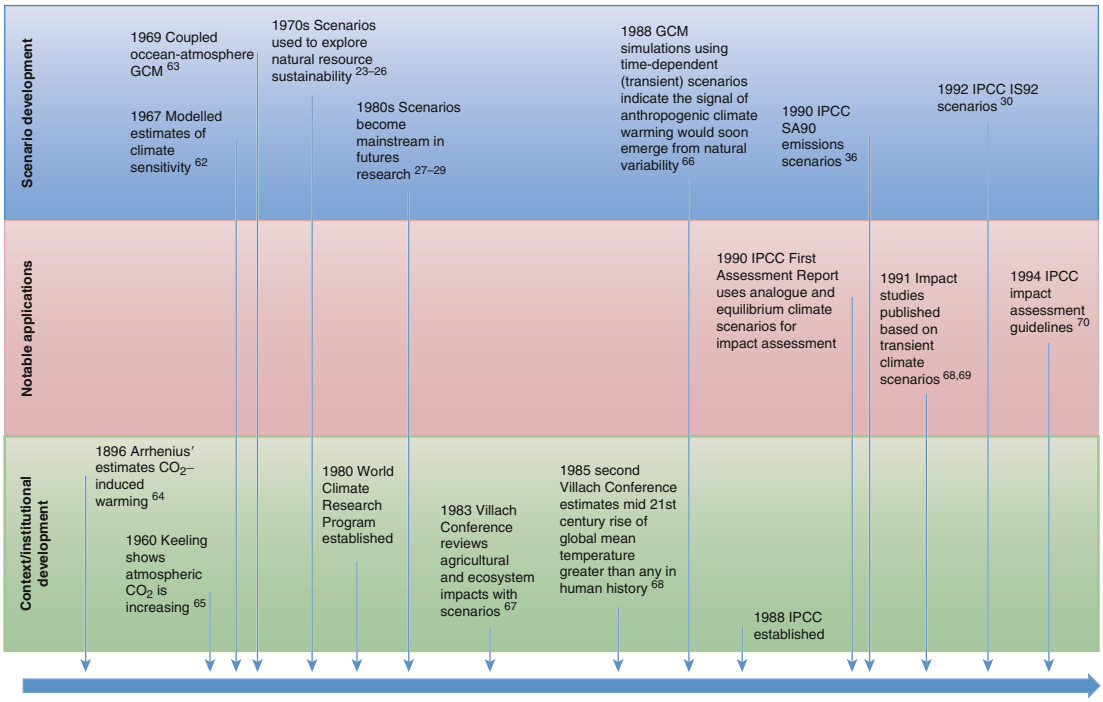
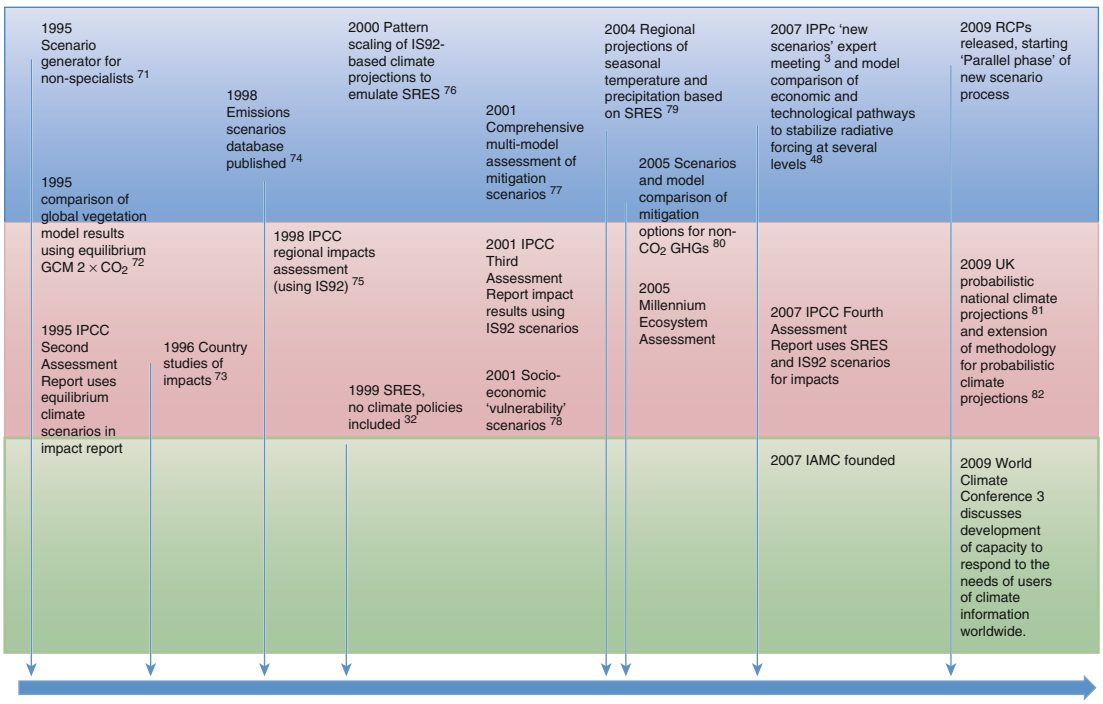


Figure 1 | Timeline highlighting some notable developments in the creation and use of emissions and climate scenarios. The entries are illustrative of the Overall course of model-based scenario development (blue) and application (beige) described in this Perspective, and also give some context (green); they do



not provide a comprehensive account of all major scenarios and significant studies or assessments that have used them. See Supplementary Information for details. GCM, general circulation model; GHG, greenhouse gas; IAMC, Integrated Assessment Modelling Consortium.

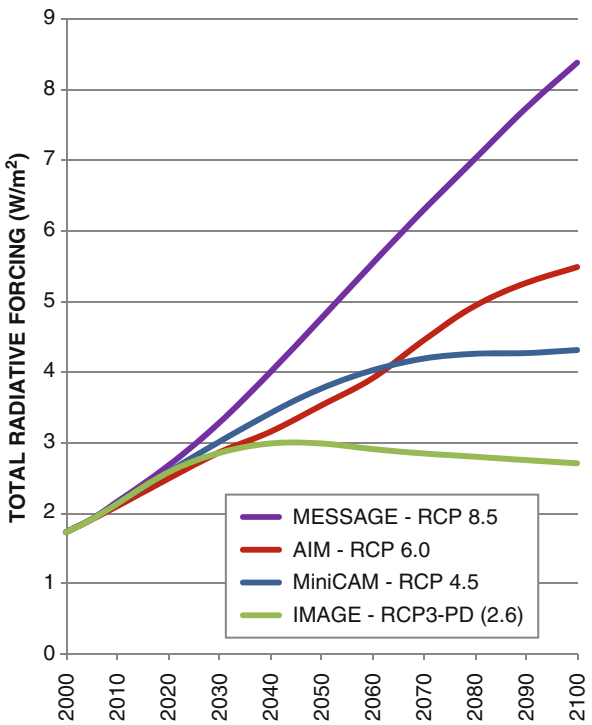
Integrated Assessment Modeling. Figure 7

Timeline highlighting some notable developments in the creation and use of emissions and climate scenarios (Source: Moss et al. [61], pp 748–749)

Integrated Assessment Modeling. Table 7 The four representative concentration pathways

Name	Radiative forcing	Concentration	Pathway	Model providing RCP	References
RCP8.5	>8.5 W/m^2 in 2100	>1,370 CO_2-eq in 2100	Rising	MESSAGE	Riahi et al. [64]
RCP6.0	~6 W/m^2 at stabilization after 2100	~850 CO_2-eq (at stabilization after 2100)	Stabilization without overshoot	AIM	Fujino et al. [65], Hijioka et al. [66]
RCP4.5	~4.5 W/m^2 at stabilization after 2100	~650 CO_2-eq (at stabilization after 2100)	Stabilization without overshoot	MiniCAM (GCAM)	Clarke et al. [14], Smith and Wigley [67]
RCP2.6	Peak at ~3 W/m^2 before 2100 and then decline	Peak at ~490 CO_2-eq before 2100 and then decline	Peak and decline	IMAGE	Van Vuuren et al. [68, 69]

Source: Moss et al. [61], p. 753



Integrated Assessment Modeling. Figure 8
The radiative forcing trajectories of the four RCP scenarios (Source: Moss et al. [61], P. 748–749)

Furthermore, even if the socioeconomic data were included for these scenarios, each of the scenarios was crafted by a different modeling team, using different assumptions about key socioeconomic and other variables. For instance, it would be difficult, if not impossible to determine if the difference in estimated impacts of climate change associated with RCP4.5 and RCP2.6 was the result of differences in the magnitude

of climate change or that of differences in the underlying human Earth systems that characterize the GCAM and IMAGE scenarios, respectively.

In order to establish a framework, in which the human system impact of climate change could be coupled with emissions scenario and climate model, a new scenario matrix architecture is under development. This architecture would create a suite of scenarios that are defined in terms of two bundles of descriptors: shared socio-ecosystem pathways (SSPs) and shared climate policy assumptions (SPAs).

SSPs have three components: a set of quantitative assumptions that are used by IAMs, such as population and economic growth; a set of quantified assumptions about variables that are not part of IAMs, for example, governance index; and a narrative which describes the general state of the world and its evolution over the course of the twenty-first century.

SPAs define the state of climate policy and its evolution around the world. They are defined with quantitative descriptors, where appropriate, and a qualitative narrative. The quantitative descriptors could be, for example, a limit on radiative forcing, such as was used to define the RCPs. In addition, information regarding the nature of policies that are to be employed to affect the prescribed outcome could be included.

The virtue of harmonizing SPAs with RCPs is that the new scenarios could be coupled smoothly with climate model output from ensemble calculations. This in turn would facilitate analysis that could potentially be fully integrated across three broad research communities: climate modeling, integrated assessment modeling and impacts, and adaptation and vulnerability. Two examples of such scenario matrix architectures can be found in [70, 71].

Climate Impacts in Higher Resolution IAMs

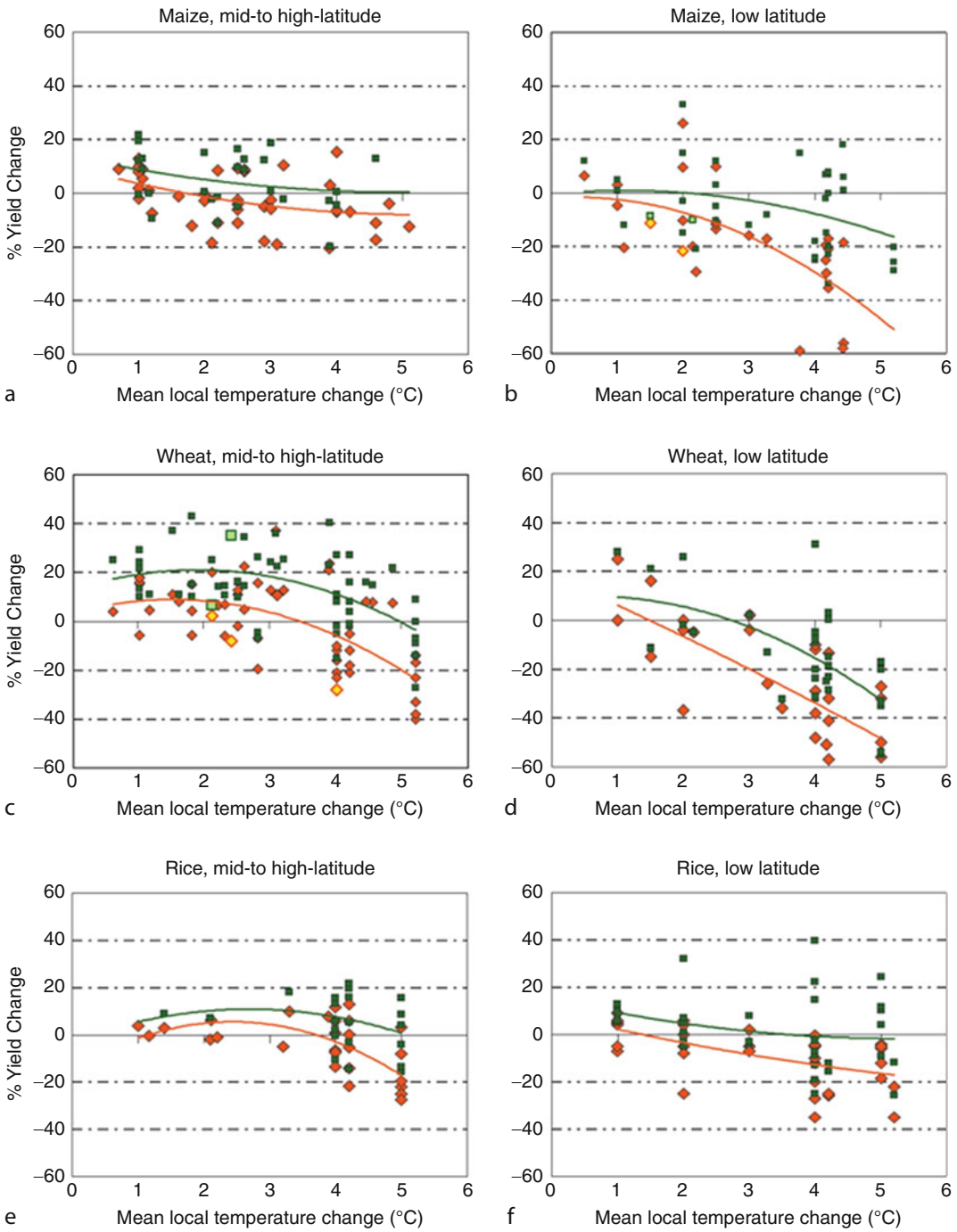
Higher resolution IAMs are increasingly focusing on explicitly modeling the physical impacts of climate change [72]. This work builds on a long tradition of modeling climate impacts in the higher resolution IAM community (see, e.g., [26, 73–75]). However, to date higher resolution IAMs have examined climate impacts using a sequential methodology, that is, they start with emissions, which are assumed to be given by climate models, and then analyzed the consequences of the ensuing climate change.

New model development is increasingly focused on methods and tools that will allow higher resolution IAMs to examine impacts simultaneously with mitigation and therefore to allow the two to interact. For example, there are on-going research efforts that utilize the higher resolution IAMs to study scenarios in which interactions between policies to mitigate emissions through changes in land-use and land cover – e.g., afforestation policies – and adaptive responses to climate change in agricultural sectors are simultaneously examined. Two complementary model development directions are also worthy of note. First, the higher resolution models are beginning to couple with state-of-the-art natural Earth system models (discussed later in this section) and second, they are beginning to move to finer spatial and temporal resolutions.

The increasing attention to climate impacts implies that the higher resolution IAMs will produce new results that will also contribute to the impacts, adaptation, and vulnerability (IAV) research. For nonmarket impacts of climate change, higher resolution IAMs will compute physical consequences, but not necessarily economic damage estimates, as it has generally been the case with climate impacts that the higher resolution IAMs have examined to date. For climate impacts associated with marketable goods and services, economic costs can also be estimated. But, the nonlinear nature of the human and natural Earth system means that separating out the impact of emissions mitigation from the impact of climate change will be nontrivial.

A good example of new work on the interactions between mitigation and impacts within higher resolution IAMs is land use and land cover. Land use will be affected both by a changing climate and by emissions mitigation effort. Mitigation effects will take the form of forest expansion to reduce land-use change emissions along with the use of bioenergy crops for energy production. A changing climate will bring about many changes in the nature of terrestrial systems, including changes in crop yields. All of these dynamics will interact.

To illustrate these interactions, the effects of climate change on crops were modeled as a response function derived from data reported in IPCC [76]. Figure 9 shows the distribution of estimates of crop yields for maize and wheat for low and other latitudes.



Integrated Assessment Modeling. Figure 9
(Continued)

Both a reference scenario and a policy scenario in which CO_2 concentrations were limited to stay below 500 ppm were presented. Land-use change emissions of CO_2 were recorded for the two scenarios, with and without consideration of climate feedbacks through agricultural crops. These results are displayed in Fig. 10.

Note that cumulative land-use change emissions vary significantly when climate change effects are considered in the reference scenario, with land-use change emissions significantly higher as a consequence of crop yield reductions in the face of climate change.

Results for the scenario in which CO_2 concentrations were not allowed to exceed 500 ppm exhibit lower emissions than either of the reference scenarios. This is because the mitigation scenario valued terrestrial carbon emissions equally with fossil fuel emissions (results would have been very different had terrestrial carbon not been valued; see also [33, 77]). Equally as interesting, land-use change emissions with and without consideration of climate change effects on crop yields are not significantly different between the two scenarios. This result follows directly from the fact that limiting CO_2 concentrations to 500 ppm would also limit the magnitude of climate change, which in turn moderates the effects on crop yields. The purpose of this example is not so much to showcase results, but rather to motivate the joint consideration of impacts, adaptation, and vulnerability with integrated assessment of emissions mitigation.

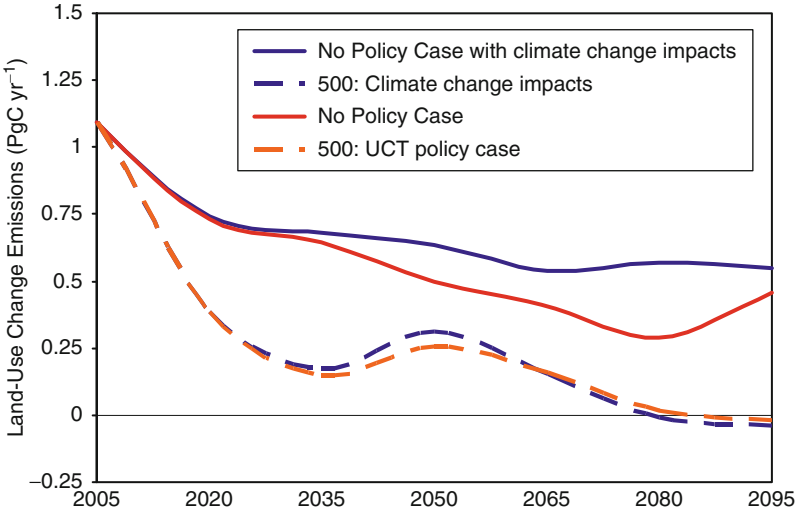
Linking Higher Resolution IAMs into integrated Earth System Models (iESMs)

Several research teams have undertaken joint work with the climate modeling community. The IGSM team has developed a relationship with climate researchers at the US National Center for Atmospheric Research (NCAR). The IMAGE team has developed several collaborative relationships including those with the Oak Ridge National Laboratory (ORNL), the Centre National de Recherches Météorologiques Coupled global climate Model (CNRM-CM3) team of France, and other European climate modeling teams to develop coupled scenarios. The MESSAGE integrated assessment modeling team has developed a collaboration with the NASA Goddard Institute for Space Studies climate modeling team. The GCAM team has developed a collaboration with ORNL and the Lawrence Berkeley National Laboratory (LBNL) in the development of a modeling system that joins the Community Earth System Model (CESM) representation of natural Earth systems with the GCAM representation of human Earth systems. To date, the collaborations have produced one-way coupling models, where emission scenarios from IAMs affect the climate, while the resulting climate change does not feedback to emissions. However, current effort is focused around developing a two-way coupled system.

The goal of the joint collaborations is to create a first-generation integrated Earth System Model (iESM) by fully integrating the human dimension from an IAM and a natural dimension from a climate model, that is, to create the capability of simultaneously estimating human system impacts on climate change and climate change impacts on human systems. After creating the capacity to examine the coupled natural and human Earth systems, the project could apply the model to the examination of feedbacks between human systems, the climate systems, and land-use systems. For instance, the policy response

Integrated Assessment Modeling. Figure 9

The modeled effects of climate change on crops. Sensitivity of cereal yield to climate change for maize, wheat and rice, as derived from the results of 69 published studies at multiple simulation sites, against mean local temperature change used as a proxy to indicate magnitude of climate change in each study. Responses include cases without adaptation (red dots) and with adaptation (dark green dots). Adaptations represented in these studies include changes in planting, changes in cultivar, and shifts from rain-fed to irrigated conditions. Lines are best-fit polynomials and are used here as a way to summarise results across studies rather than as a predictive tool. The studies span a range of precipitation changes and CO_2 concentrations, and vary in how they represent future changes in climate variability. For instance, lighter-coloured dots in (**b**) and (**c**) represent responses of rain-fed crops under climate scenarios with decreased precipitation. (Source: Parry et al. [76], P. 286)



Integrated Assessment Modeling. Figure 10
Land-use change emissions of CO_2 under the scenarios with and without consideration of climate feedbacks through agricultural crops

of land-use change presented in [33] could be revisited to estimate the magnitude of feedbacks in the system.

Significant effort is required before such research becomes routine. Nonetheless, as the research potential this collaboration opens up is virtually limitless, the importance of integrating human Earth systems with natural Earth systems is sufficiently compelling to drive future collaborations between ESMs and IAMs.

Bibliography

1. Wigley TML, Richels R, Edmonds JA (1996) Economic and environmental choices in the stabilization of atmospheric CO_2 concentrations. Nature 379:240–243
2. Nordhaus WD, Yohe GW (1983) Future carbon dioxide emissions from fossil fuels. Changing climate: report of the carbon dioxide assessment committee. National Academy Press, Washington DC, pp 87–153
3. Nordhaus WD (1993) Optimal greenhouse-gas reductions and tax policy in the "DICE" model. Am Econ Rev 83:313–317
4. Nordhaus WD, Yang Z (1996) A regional dynamic general-equilibrium model of alternative climate-change strategies. Am Econ Rev 86:741–765
5. Dowlatabadi H, Morgan MG (1993) A model framework for integrated studies of the climate problem. Energy Policy 21:209–221
6. Hope C, Anderson J, Wenman P (1993) Policy analysis of the greenhouse effect: an application of the PAGE model. Energy Policy 21:327–338
7. Tol RSJ (1997) On the optimal control of carbon dioxide emissions: an application of FUND. Environ Model Assess 2:151–163
8. Weyant J, Davidson O, Dowlatabadi H, Edmonds J, Grubb M, Parson EA, Richels R, Rotmans J, Shukla PR, Tol RSJ (1996) Integrated assessment of climate change: an overview and comparison of approaches and results. In: Bruce JP, H-sŏng Yi, Haites EF (eds) Climate change 1995: economic and social dimensions of climate change. The contribution of working group III to the second assessment report of the intergovernmental panel on climate change. Cambridge University Press, UK/New York, pp 367–396
9. Parson EA, Fisher-Vanden K (1997) Integrated assessment models of global climate change. Annu Rev Energy Env 22:589–628
10. Nordhaus W (2007) Critical assumptions in the Stern Review on climate change. Science 317:201–202
11. Stern NH (2007) The economics of climate change: the Stern Review. Cambridge University Press, Cambridge, UK/New York
12. Portney PR, Weyant JP (eds) (1999) Discounting and intergenerational equity. Resources for the Future, Washington, DC
13. Dasgupta P, Mäler KG, Barrett S (1999) Intergenerational equity, social discount rates and global warming. In: Portney PR, Weyant JP (eds) Discounting and intergenerational equity. Resources for the Future, Washington, DC
14. Clarke L, Edmonds J, Jacoby H, Pitcher H, Reilly J, Richels R (2007) Scenarios of greenhouse gas emissions and atmospheric concentrations. Synthesis and assessment product 2.1a, report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research. U.S. Government Printing Office, Washington, DC

15. Edmonds J, Reilly J (1983) A long-term global energy-economic model of carbon dioxide release from fossil fuel use. Energy Econ 5:74–88
16. Edmonds J, Reilly J (1983) Global energy and CO_2 to the year 2050. Energy J 4:21–48
17. Edmonds J, Reilly J (1983) Global energy production and use to the year 2050. Energy 8:419–432
18. Edmonds J, Reilly JM (1985) Global energy: assessing the future. Oxford University Press, New York
19. Brenkert A, Smith S, Kim S, Pitcher H (2003) Model documentation for the MiniCAM. Pacific Northwest National Laboratory. Technical report PNNL-14337
20. Kim SH, Edmonds JA, Smith SJ, Wise M, Lurz J (2006) The object-oriented energy climate technology systems (ObjECTS) framework and hybrid modeling of transportation in the MiniCAM long-term, global integrated assessment model. Energy J 27:63–91
21. Clarke L, Wise M, Kim S, Smith S, Lurz J, Edmonds J, Pitcher H (2007) Model documentation for the minicam climate change science program stabilization scenarios: CCSP product 2.1 a. Pacific Northwest National Laboratory, PNNL-16735
22. Wise MA, Calvin KV, Thomson AM, Clarke LE, Bond-Lamberty B, Sands RD, Smith SJ, Janetos AC, Edmonds JA (2009) The implications of limiting CO_2 concentrations for agriculture, land use, land-use change emissions and bioenergy. Pacific Northwest National Laboratory
23. Sokolov AP, Schlosser CA, Dutkiewicz S, Paltsev S, Kicklighter DW, Jacoby HD, Prinn RG, Forest CE, Reilly JM, Wang C, et al (2005) MIT integrated global system model (IGSM) version 2: model description and baseline evaluation, MIT Joint Program for the Science and Policy of Global Change. Report 124, Cambridge, MA
24. Manne A, Mendelsohn R, Richels R (1995) MERGE: a model for evaluating regional and global effects of GHG reduction policies. Energy Policy 23:17–34
25. Blanford GJ, Richels RG, Rutherford TF (2009) Feasible climate targets: the roles of economic growth, coalition development and expectations. Energy Econ 31:S82–S93
26. Kainuma M, Matsuoka Y, Morita T (eds) (2003) Climate policy assessment: Asia-Pacific integrated modeling. Springer-Verlag, Tokyo
27. Wigley TML, Raper SCB (1992) Implications for climate and sea level of revised IPCC emissions scenarios. Nature 357:293–300
28. Wigley TML, Raper SCB (2002) Reasons for larger warming projections in the IPCC third assessment report. J Climate 15:2945–2952
29. Raper SCB, Wigley TML, Warrick RA (1996) Global sea-level rise: past and future. In: Milliman JD Haq BU (eds) Sea-level rise and coastal subsidence: causes, consequences, and strategies. Kluwer Academic Publishers, Dordrecht, Netherlands
30. Clarke JF, Edmonds JA (1993) Modeling energy technologies in a competitive market. Energy Econ 15:123–129
31. McFadden D (1974) Conditional logit analysis of qualitative choice behavior. In: Zarambka P (ed) Frontiers of econometrics. Academic, New York, pp 105–142
32. McFadden D (1981) Econometric models for probabilistic choice among products. In: Manski C, McFadden D (eds) Structural analysis of discrete data with econometric applications. MIT Press, Cambridge, MA, pp 198–272
33. Wise M, Calvin K, Thomson A, Clarke L, Bond-Lamberty B, Sands R, Smith SJ, Janetos A, Edmonds J (2009) Implications of limiting CO_2 concentrations for land use and energy. Science 324:1183–1186
34. Bouwman AF, Kram T, Goldewijk KK (2006) Integrated modeling of global environmental change: an overview of Image 2.4. Netherlands Environmental Assessment Agency
35. Searchinger T, Heimlich R, Houghton RA, Dong F, Elobeid A, Fabiosa J, Tokgoz S, Hayes D, Yu T-H (2008) Use of U.S. Croplands for biofuels increases greenhouse gases through emissions from land-use change. Science 319:1238–1240
36. Lal R (2004) Soil carbon sequestration impacts on global climate change and food security. Science 304:1623–1627
37. Fargione J, Hill J, Tilman D, Polasky S, Hawthorne P (2008) Land clearing and the biofuel carbon debt. Science 319:1235–1238
38. Luckow P, Wise MA, Dooley JJ, Kim SH (2010) Large-scale utilization of biomass energy and carbon dioxide capture and storage in the transport and electricity sectors under stringent CO_2 concentration limit scenarios. Int J Greenh Gas Control 4:865–877
39. Edmonds JA, Scott MJ, Roop JM, MacCracken CN (1999) International emission trading and the cost of greenhouse gas emissions mitigation. The Pew Center on Global Climate Change, Arlington
40. Kyoto Protocol: the kyoto protocol to the united nations framework convention on climate change. UNEP/WMO, Kyoto
41. Peck SC, Wan YS (1996) Analytic solutions of simple optimal greenhouse gas emission models. In: van Ierland E, Górka K (eds) Economics of atmospheric pollution. Springer, Berlin, pp 113–121
42. Edmonds JA, Wise MA, Dooley JJ, Kim SH, Smith SJ, Runci PJ, Clarke LE, Malone EL, Stokes GM (2007) Global energy technology strategy: addressing climate change phase 2 findings from an international public-private sponsored research program. Pacific Northwest National Laboratory (PNNL), Richland
43. Metz B, Davidson O, Bosch P, Dave R, Meyer L (eds) (2007) Climate change 2007: mitigation of climate change; contribution of working group III to the 4th assessment report of the intergovenmental panel on climate change. Cambridge University Press, Cambridge, UK/New York
44. Weyant JP, Francisco C, Blanford GJ (2006) Overview of EMF-21: multigas mitigation and climate policy. Energy J 27:1–32
45. National Research Council (U.S.) (2005) Climate research committee: radiative forcing of climate change: expanding the concept and addressing uncertainties. Academic, Washington, DC
46. Wuebbles DJ, Edmonds J (1991) Primer on greenhouse gases. Lewis Publishers, Chelsea, Michigan
47. Solomon S, Qin D, Manning M, Marquis M, Averyt K, Tignor M, LeRoy Miller H, Chen Z (eds) (2007) Climate change 2007. The physical science basis: contribution of working group I to the

fourth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge, UK/New York
48. Houghton JT, Meiro Filho LG, Callander BA, Harris N, Kattenburg A, Maskell K (eds) (1996) Climate change 1995: the science of climate change. Cambridge University Press, Cambridge, UK
49. Joos F, Prentice IC, Sitch S, Meyer R, Hooss G, Plattner G-K, Gerber S, Hasselmann K (2001) Global warming feedbacks on terrestrial carbon uptake under the Intergovernmental Panel on Climate Change (IPCC) emission scenarios. Global Biogeochem Cy 15:891–908
50. Houghton JT, Jenkins GJ, Ephraums JJ (eds) (1990) Climate change: the IPCC scientific assessment. Cambridge University Press, Cambridge, UK
51. Houghton JT, Jenkins GJ, Ephraums JJ (eds) (1990) Climate change: the IPCC response strategies. Cambridge University Press, Cambridge, UK
52. Manne AS, Richels RG (2001) An alternative approach to establishing trade-offs among greenhouse gases. Nature 410:675–677
53. Arrow KJ (1950) A difficulty in the concept of social welfare. J Polit Econ 58:328–346
54. Milliman SR, Prince R (1989) Firm incentives to promote technological change in pollution control. J Environ Econ Manag 17:247–265
55. McJeon HC, Clarke L, Kyle P, Wise M, Hackbarth A, Bryant BP, Lempert RJ (2011) Technology interactions among low-carbon energy technologies: what can we learn from a large number of scenarios? Energy Econ 33:619–631
56. Böhringer C, Rutherford TF, Tol RSJ (2009) THE EU 20/20/2020 targets: an overview of the EMF22 assessment. Energy Econ 31:S268–S273
57. Edmonds J, Clarke L, Lurz J, Wise M (2008) Stabilizing CO_2 concentrations with incomplete international cooperation. Clim Policy 8:355–376
58. Clarke L, Edmonds J, Krey V, Richels R, Rose S, Tavoni M (2009) International climate policy architectures: overview of the EMF 22 international scenarios. Energy Econ 31: S64–S81
59. Edmonds JA, Reilly J, Trabalka JR, Reichle DE (1984) An analysis of possible future atmospheric retention of fossil fuel CO_2, TR013, US Department of Energy Carbon Dioxide Research Division, Washington DC
60. Trabalka JR, Reichle DE (eds) (1986) The changing carbon cycle: a global analysis. Springer, New York
61. Moss RH, Edmonds JA, Hibbard KA, Manning MR, Rose SK, van Vuuren DP, Carter TR, Emori S, Kainuma M, Kram T, Meehl GA, Mitchell JFB, Nakicenovic N, Riahi K, Smith SJ, Stouffer RJ, Thomson AM, Weyant JP, Wilbanks TJ (2010) The next generation of scenarios for climate change research and assessment. Nature 463:747–756
62. Leggett J, Pepper WJ, Swart RJ, Edmonds J, Meira Filho LG, Mintzer I, Wang MX, Wasson J (1992) Emissions scenarios for the IPCC: an update. Climate change 1992: The supplementary report to the IPCC scientific assessment. Cambridge University Press, Cambridge, UK/New York
63. Nakicenovic N, Alcamo J, Davis G, de Vries B, Fenhann J, Gaffin S, Gregory K, Grubler A, Jung TY, Kram T (2000) Special report on emissions scenarios: a special report of working group III of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK/New York
64. Riahi K, Grübler A, Nakicenovic N (2007) Scenarios of long-term socio-economic and environmental development under climate stabilization. Technol Forecast Soc 74:887–935
65. Fujino J, Nair R, Kainuma M, Masui T, Matsuoka Y (2006) Multi-gas mitigation analysis on stabilization scenarios using AIM global model. Energy J SI3:343–354
66. Hijioka Y, Matsuoka Y, Nishimoto H, Masui M, Kainuma M (2008) Global GHG emissions scenarios under GHG concentration stabilization targets. J Global Environ Eng 13:97–108
67. Smith SJ, Wigley TML (2006) Multi-gas forcing stabilization with MiniCAM. Energy J SI3:373–392
68. van Vuuren DP, Eickhout B, Lucas PL, den Elzen MGJ (2006) Long-term multi-gas scenarios to stabilize radiative forcing-exploring costs and benefits within an integrated assessment framework. Energy J SI3:201–234
69. van Vuuren DP, Elzen MGJ, Lucas PL, Eickhout B, Strengers BJ, Ruijven B, Wonink S, Houdt R (2007) Stabilizing greenhouse gas concentrations at low levels: an assessment of reduction strategies and costs. Clim Change 81:119–159
70. van Vuuren DP, Riahi K, Moss R, Edmonds J, Thomson A, Nakicenovic N, Kram T, Berkhout F, Swart R, Janetos A, Rose SK, Arnell N (2011) A proposal for a new scenario framework to support research and assessment in different climate research communities. Global Environ Chang (In Press)
71. Kriegler E, O'Neill BC, Hallegatte S, Kram T, Lempert R, Moss RH, Wilbanks TJ (2010) Socio-economic scenario development for climate change analysis. CIRED working paper. Centre International de Recherche sur l'Environnement et le Développement, Paris, France
72. Janetos AC, Clarke L, Collins W, Ebi K, Edmonds J, Foster I, Jacoby HJ, Judd K, Leung L, Newell R, Ojima D, Pugh G, Sanstad A, Schultz P, Stevens R, Weyant J, Wilbanks T (2008) Science challenges and future directions: climate change integrated assessment research. U.S. Department of Energy, Office of Science. http://science.energy.gov/~/media/ber/pdf/Ia_workshop_low_res_06_25_09.pdf. Accessed 5 Dec 2011
73. Leemans R, Eickhout B (2004) Another reason for concern: regional and global impacts on ecosystems for different levels of climate change. Global Environ Change Part A 14:219–228
74. Reilly J, Paltsev S, Felzer B, Wang X, Kicklighter D, Melillo J, Prinn R, Sarofim M, Sokolov A, Wang C (2007) Global economic effects of changes in crops, pasture, and forests due to changing climate, carbon dioxide, and ozone. Energy Policy 35:5370–5383
75. Edmonds JA, Rosenberg NJ (2005) Climate change impacts for the conterminous USA: an integrated assessment summary. Clim Change 69:151–162
76. Parry M, Canzaiani O, Palutikof J, Van der Linden P, Hanson C (eds) (2007) Climate change 2007: impacts, adaptation and

vulnerability; Contribution of working group II to the fourth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge, UK/New York
77. Melillo JM, Reilly JM, Kicklighter DW, Gurgel AC, Cronin TW, Paltsev S, Felzer BS, Wang X, Sokolov AP, Schlosser CA (2009) Indirect emissions from biofuels: how important? Science 326:1397–1399
78. Weyant JP, Hill J (1999) The costs of the Kyoto protocol: a multi-model evaluation; introduction and overview. Energy J 20(Special Issue):vii–xliv
79. Kyle P, Clarke L, Rong F, Smith SJ (2010) Climate policy and the long-term evolution of the US buildings sector. Energy J 31:145–172

Integrated Pest Management

Ramon Albajes
Department of Plant Production and Forest Sciences, Universitat de Lleida, Centre UdL-IRTA, Lleida, Spain

Article outline

Glossary

Agroecosystem, agrosystem – agricultural ecosystem An ecosystem that is managed to optimize the production of a crop plant or part of it.

Biological control Use or manipulation of natural enemies (predators, parasitoids, or diseases of pests) to suppress pest populations.

Crop resistance to pests One or more qualities that some crop plant varieties have resulting in less damage by a number of pest individuals in comparison with a variety without those qualities when it is exposed to the same pest numbers.

Damage caused by a pest Damage is the monetary value lost to the commodity as a result of injury by the pest, for instance by yield reduction.

Economic injury level of a pest The pest population density at which the cost to control the pest equals the amount of damage it inflicts.

Economic threshold of a pest The pest population density at which a control measure has to be taken to prevent population from reaching the economic injury level.

Genetically modified (GM) crop, transgenic crop A crop whose genetic material has been modified by genetic engineering techniques, through which novel genes have been introduced into the crop.

Integrated pest management (IPM) A system for controlling pests in an economically, ecologically, and sociologically sound manner by the use of multiple tactics in a compatible manner. The term has many (if not all) common elements with integrated control.

Metapopulation A set of populations occupying different patches among which individuals can occasionally move.

Pest Any herbivore that feeds and causes damages on crop plants in an agroecosystem resulting in crop damages if control measures are not taken. In this entry the term "pest" includes only insects and mites but for other authors the same term would additionally include plant diseases and weeds.

Pheromones Chemicals that are released in the environment by one individual and trigger a behavioral response in other individuals of the same species.

Precision agriculture Precision agriculture aims to apply inputs only when and where they are needed and at optimal amounts according to variable field or environment characteristics.

Definition of the Subject

In addition to producing food and fiber to satisfy an increasing world population, agriculture is being asked to supply energy at reasonable prices thus contributing to the acceleration of the demand for agricultural commodities. Increase of crop yields may be achieved by maximizing the proportion of sunlight energy that is fixed by the crop plant or by reducing the amount of energy that is lost by insect pests, diseases, and weeds. More than 50% of the potential yield of agricultural crops is lost by the three causes. To diminish losses caused by insect pests in agriculture in an economically, ecologically, and sociologically acceptable manner is the goal of integrated pest management (IPM). Given the complexity of agricultural ecosystems, IPM has to consider and manage all the elements and relationships involved in agriculture, including those related to nonagricultural ecosystems. Providing a scientific approach to better understand processes in agroecosystems in order to implement more rapidly sustainable IPM systems is a major challenge for ecology.

Introduction: Agriculture and Insect Pests

Agriculture and Productivity

Providing stable food for human subsistence has been the main goal of agriculture throughout the centuries. Probably because more than 50% of world population lives in urban areas and because developed countries do not feel threatened by hunger, society is not aware that agriculture needs to secure the world's food supplies. Increase of both world population and its income lead to increasing rates of food demand and to remarkable changes in the commodity composition of food consumption, particularly a bigger demand for livestock products.

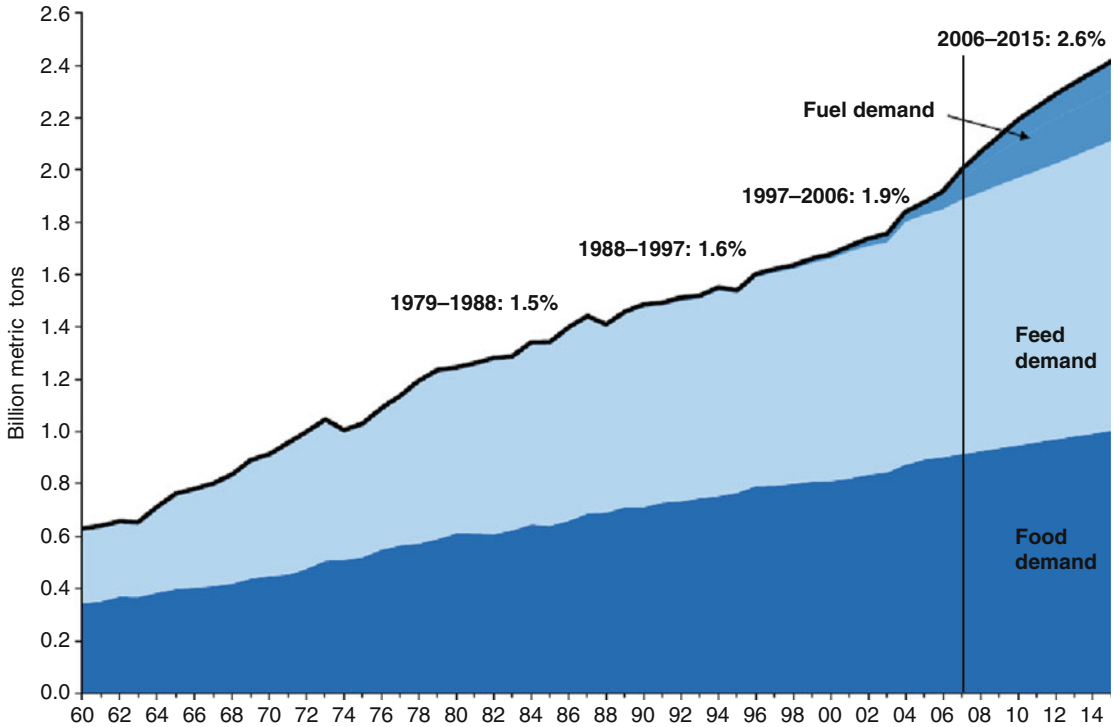
In addition to producing food and fiber to satisfy an increasing world population, agriculture has been asked to supply energy at reasonable prices, mostly for replacing gasoline and petrol for motor vehicles. Although nowadays, as in 2007 and 2008, biofuel demand is linked to rising petrol prices, production of agricultural commodities for bioethanol and biodiesel conversion is expected to increase in the coming years, particularly if their price is competitive with petroleum. Biofuel production has accelerated the demand growth rate for agricultural commodities, which may largely exceed an annual rate of 2% in the coming years (Fig. 1). Reduction of basic food stocks, higher food prices, and increased land use may follow the stronger demand for agricultural commodities. Complementarily, agriculture in developed countries faces increasing environmental, human health, traceability, and competition challenges. Only a significant increase of agricultural productivity and sustainable improvement of protection techniques may respond to these challenges.

Agricultural Ecosystems: Ecological Basis for an Integrated Approach to Pest Control

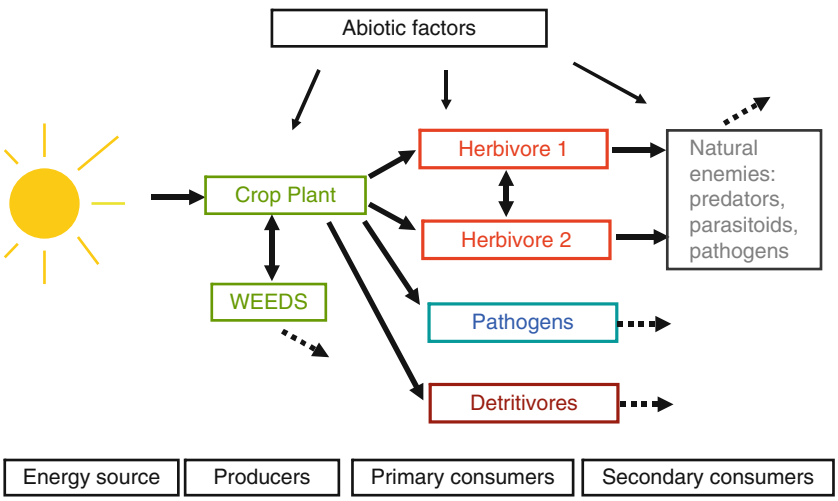
It cannot be forgotten that agricultural ecosystems (also called agroecosystems) have their origin in natural ecosystems which human activity has transformed along millennia to better achieve the goal of agriculture: to provide us with sufficient food. Therefore, objectives of agroecosystems are substantially different to those of natural ecosystems; whereas, in the first case, growers manipulate the ecosystem to obtain an optimal amount of one species or part of one species (grain and not plant biomass, e.g., in cereal systems), natural ecosystems tend to perpetuate the system by themselves and do not favor one particular species or group of species. Agroecosystems differ from natural ecosystems in several ways among which two are particularly relevant in this chapter: (a) agroecosystems introduce a certain amount of energy previously processed (e.g., fuel) in comparison with the second which uses almost exclusively the energy provided by the sun; (b) agroecosystems are generally manipulated to deal with a unique plant species (the crop) and this leads to a drastic simplification of biodiversity in producer and ulterior consumer food web levels.

Composition and relationships in an agroecosystem are of crucial importance to the dynamics of its components. Figure 2 represents a simplified scheme of the main components and relationships concerning crop loss agents.

This chapter only deals with the control of those herbivores that cause damage to crops, the so-called pests. In the literature the term "pest" sometimes refers to pathogens and weeds but here the term will be restricted to herbivore species, mainly insects and



Integrated Pest Management. Figure 1
Evolution of demand of agricultural commodities for food, feed, and fuel use in the last 50 years (http://www.gceholdings.com/pdf/GoldmanReportFoodFeedFuel.pdf. Accessed 17 October 2009)



Integrated Pest Management. Figure 2
Main components and relationships related to insect pests and their control in an agroecosystem food web. *Arrows* show the direction of energy flow between components of adjacent trophic levels. *Double arrows* show competition between components within a trophic level. *Broken arrows* show that energy flow continues but the further components involved are not represented

mites, causing damages to crop yield if preventive measures are not adopted. In order to avoid constant repetition of words such as insects and arthropods the term "pest" will be preferred along all the text.

Maximization of crop yield in Fig. 2 may be achieved by increasing the proportion of sunlight energy that is fixed as biomass by the crop plant or by reducing the amount of energy that is lost by competition of weeds, by damages caused by pests, or by diseases caused by plant pathogens. The process of accumulating biomass in a crop plant to reach a yield may be compared to the process of filling a water deposit that loses water through three holes (pests, diseases, and weeds). More water may be stocked in the deposit by opening the tap (more photosynthetic production) or by repairing holes. The complex nature of the agroecosystem demands an integrated approach to manage all the components and relationships taking into account that any change in a part of them may lead to undesirable consequences in the whole system. Considering the entire complexity of the agroecosystem is crucial to develop sound integrated pest management programs.

Continuously Increasing New Insect Pests: Pest Invasion

Elements and links in the picture of Fig. 2 may change in space and time. Introduction and establishment of invasive alien species and climatic change are among the most decisive factors influencing the composition and relationships in an agroecosystem. Many of the insect pests in agriculture originated as exotic herbivores introduced in the past; this process continues to accelerate, mainly as a consequence of globalization of agricultural trading. The introduction, establishment, and spread in Europe of the Colorado potato beetle may exemplify the impact of insect pest invasions on agriculture. It originally restricted its distribution to the Rocky Mountains (USA) where it fed on wild plants. In the second half of the nineteenth century, the cultivation of potatoes close to that area allowed the beetle to multiply its populations, spread out of its original area, and distribute around the world where it has become the most harmful insect pest of potatoes in general terms. Hundreds of exotic arthropods have been documented as invading agricultural areas, establishing themselves and becoming important pests. The objective of eliminating or limiting the spread of pests has led several national and international institutions to develop legislation and tools to restrict the movement of insect species with a high potential to become agricultural pests. Regional sections of the International Plant Protection Convention have been particularly active in this direction and its Web site (www.ippc.int) is an important source of information. To categorize the risk of potential invasive alien species is a first step to adopt correct and ad hoc measures. Then, measures of exclusion, early detection, containment, or control of the most risky species may be adopted at the national or international level. Prevention of the introduction and spread of crop pests is an important first step of integrated pest management although it is sometimes difficult to implement due to the ease of international travel of both people and plants, and the increasing trade of agricultural commodities.

Injuries and Losses Caused by Pests

Pest control, and particularly integrated pest management, is therefore necessary to safeguard crop productivity against losses caused by herbivore insects and assure human nutrition. Losses caused by insect and mite pests derive mostly from their feeding activity on plants. Insects consume plant materials with their chewing or sucking mouthparts; in addition to lowering plant vigor, some insects are also able to transmit plant pathogens, with additional losses due to disease development, or injection of toxic substances that interfere with plant physiology. Consequently, potential crop yield is not attained and a variable level of losses results according to the amount of insects feeding on plants, the type of injury caused, and the susceptibility of the plant to the amount of injury infringed.

Potential and actual losses due to animal (mostly insect and mite) pests have been estimated by Oerke and Dehne in major crops [1]. Those were estimated as quite variable according to the crop and geographical area but on average they were 17.6% and 10.1% of the attainable yield in the world that give a control efficacy (percentage of losses prevented) of 42.4%. Efficacy to prevent losses from animal pests was higher than from

bacterial and fungal diseases (33.8%) but considerably lower than from weeds (70.1%). These moderate values of efficacy in crop protection, which averaged 52.5% in the world, were reached in spite of particularly high pesticide consumption in Western Europe and North America.

Strategy for Integrated Pest Management

What IPM Is

The necessity to change the strategy of controlling insect pests derived from both theoretical considerations and practical collapse of control systems. Although humans were soon aware of weaknesses in relying on control systems based solely on single tactics, control failure of pesticides due to pesticide resistance in important insect pest species and the consciousness of pesticide transfer to environment might have been the detonators of strong criticism about the mass use of pesticides in the 1950s. Soon after that, scientists, especially applied entomologists, defined new concepts and terms among which were integrated control and integrated pest management (IPM) [2]. These two terms share basically the same underlying concepts although some authors have justified the different use of these two expressions [3]. A broad definition of integrated pest management is this: a pest control system in an economically, ecologically, and sociologically sound manner by the use of multiple tactics in a compatible manner.

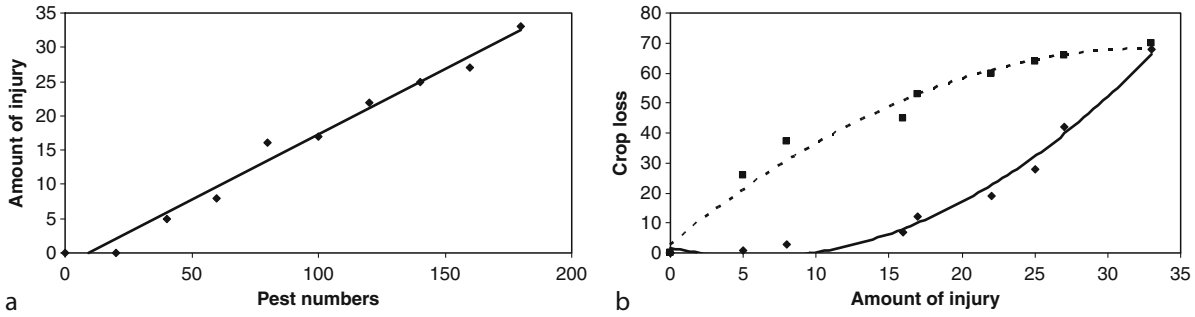
Sustainability should be an inherent qualification of IPM systems. However, development of novel insect pest control methods has often focused more on replacing chemical pesticides than implementing low input tactics, two objectives that do not necessarily progress in parallel ways. Classical biological control provides us with some examples of reduced sustainability when mass-reared natural enemies are repeatedly released to control insect pests in short-lived annual crops as practiced in some Mediterranean greenhouses. Paradoxically, many of pest natural enemies that are mass produced with energetically costly procedures and released into isolated Mediterranean greenhouses are native to the same area and could be managed to enhance their entrance into greenhouses by conservation biological control practices. Early IPM aimed to decrease the use of pesticides and, fortunately, it succeeded in many cases. The novel era of IPM should put emphasis on the agroecosystem management and various entries in the encyclopedia will give some indications on how to proceed in such direction. Nonetheless, some fundamental concepts of IPM as enounced decades ago have kept most of their validity in theory and practice. This is the case of economic threshold which is still a key concept that, under several formulations and criticisms, occupies a central point in IPM development.

Concept of Economic Threshold

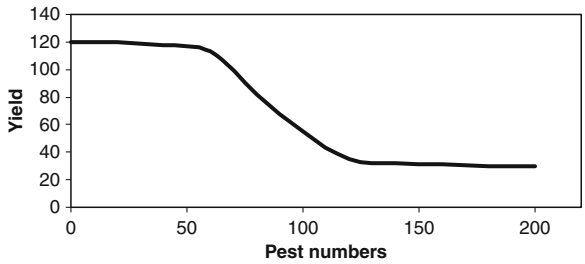
A first question that an IPM practitioner has to face when he/she detects a certain number of pests on the crop is whether intervention is justified. Some criteria to make decisions based on economic, social, ecological, and toxicological needs are necessary. Economic injury level (EIL) and economic threshold (ET) are concepts that primarily were developed to make decisions founded on economic cost-benefit analysis to include later a richer set of the so-called bioeconomic elements that relate pest numbers, crop plants responses to pest injury, and resulting crop losses. With minor modifications these two concepts of EIL and ET still form the basis of the current IPM programs providing the potential to improve economic profits but also reduce environmental impact [4].

Relationships between pest numbers and resulting crop losses are the result of two other subrelationships succeeding each other very closely in time. Pest numbers are grossly related to total injury in a linear manner, whereas the latter relates to crop loss (i.e., lowered yield) by nonlinear figures (Fig. 3) so that the result of combining both curves gives a nonlinear relationship between pest numbers and crop loss (or yield reduction), as that represented in Fig. 4, if crop loss is converted into yield reduction.

As seen in Fig. 4, yield is reduced at low pest numbers very slightly, if at all, but as pest numbers become higher yield is increasingly reduced in a near linear manner until the declining yield per pest number increase decelerates and finally becomes insensitive to pest population growth. Of course the shape and values of this relationship is a function of several factors linked with the plant and pest species. The following are the most significant: yield based on fruit is more



Integrated Pest Management. Figure 3
Relationship between pest numbers and amount of injury (*left*) and two common relationships between the amount of injury and losses caused on the crop (*right*)



Integrated Pest Management. Figure 4
The relationship between pest numbers and yield as a result of combining the relationships of Fig. 3 and converting crop loss to yield reduction

sensitive than is biomass, injury on plant tissues contributing directly to yield are more yield reducing, dissimilar host-plant growth stages are differentially sensitive to injury in terms of yield loss in a manner that young plants have a higher capacity to recover or compensate injuries, and finally biotic and abiotic stressors other than the pest may magnify the consequences of injuries.

Knowledge of the curve relating pest numbers and yield is the basis for the calculation of EIL and ET. Simply defined, EIL is the lowest pest population numbers that will cause economic damage this being the amount of damage that equals control cost. EIL shows the amount of pests (and therefore the amount of injury) that can be tolerated by a crop. Operational consequences of the EIL are summarized by the ET, which determines if management action against a pest is needed. As the time needed to make a decision and to take the action may cause pest population to surpass the EIL, decisions have to be made before the population reaches the EIL and thereby prevent economic damage; this value is called economic threshold, ET. When decision making is delayed, pest population continues to grow, control methods are less efficient, and ET is lower in relation to EIL. On the contrary, if pest population is expected to decrease after reaching the EIL, ET may be higher than EIL. There are still many gaps in the knowledge of the mechanisms of herbivorous insect and host–plant relationships and this may be the cause of the insufficient development of ETs in practice; in addition, a rather confusing literature and proliferation of nomenclatures, often with no novel concepts provided, do not help in the implementation of tools for economically, ecologically, and toxicologically sound pest control.

Limitations of Economic Thresholds

In addition to the complexity of determining ET values already stressed, a number of other limitations may constrain the application of criteria for decision making in IPM. One of the limitations derives from the fact that many factors affect the amount of injury (or pest numbers) and yield (crop losses) relationship, and therefore it is quite unrealistic to use deterministic models – as those based on ET – in insect pest control. If different levels of crop loss may be associated to probabilities of occurrence, IPM practitioners may decide on a risk assumption basis. Risk is common in economic activities and growers have to compare the

risk of crop losses with control costs. Furthermore, growers may have to choose among a number of control methods, each with a different cost and a different expected efficacy. To further complicate the grower's decision, the benefits of different control options – including no action – may be considered in the short or long term, which are often different. For example, application of an insecticide may be more effective at short time if the chemical has a good knockdown but less effective later in the season if the insecticide kills most of the pest natural enemies and pest population outbreaks occur soon after chemical application. Finally, decisions concerning pest control may be altered if derived actions are adopted at farm or regional levels; in the first case, market price is unlikely to change as a consequence of actions undertaken; but when pest control actions are adopted at a regional level, resulting yields may affect the price.

Tools for IPM

Precision Agriculture and IPM

The so-called precision agriculture applied to pest control aims to apply control inputs only when and where they are needed and at optimal amounts according to variable field characteristics. Whereas it is understood that decisions must account for temporal changes in population densities, much less attention has been given to spatial heterogeneity probably because pest control actions are mostly decided and implemented at the farm level. For very mobile pests, when control implementation needs to operate over areas larger than just the farm, or simply because pest dynamics is decisively influenced by the spatial structure of the habitat, area wide pest management strategies have to be developed [5]. This has been practiced in the past, such as with human insect-borne diseases or in a few agricultural problems (i.e., locust plagues). Larger and systematic use of area wide control applied to agricultural pests has been implemented only in the last decades.

One of the most common causes for control failures is pest migration into managed areas from unmanaged areas or from areas managed at different times, particularly when the pest is able to move through several kilometers in a short time. Disposal of fields with different crops or different crop phenologies, or those submitted to different abiotic conditions in a patchwork landscape facilitates the movement of flying pests searching for the best environments to develop and reproduce. An area wide IPM approach differs from local pest management in some important characteristics. Whereas local IPM implementation focuses the control in specific parts of the habitat, the area wide approach considers the control in all potential niches. On the other hand, as area wide strategies need to consider and implement the control on a multiyear basis, this is an incentive for building a more permanent organization for such a purpose and thus leading to more professional management tools. These include geographical information systems (GIS), satellite imagery and remote sensing, online processing of climate data and weather forecasting, and kits to detect potential insecticide-resistant populations that can be useful for prevention of spreading resistance genes across the regional population. Finally, area wide implementation of IPM programs may allow the use of methods – sterile insect techniques, mating disruption, and inundative biological control – that are effective only when applied on big surfaces. Even if these and other methods may be used at individual farm level, economies of scale may derive from acquiring large amounts of materials. On the negative side of area wide IPM, there is the increased complexity of decision making if the area covered by the program is not uniform and many inputs are needed to respond with control measures adapted to each condition.

Area wide IPM application is based on a large amount of data referenced to each geographical position which needs to be stored and organized in a systematic and recoverable way. Once data are stored and organized they can be used to monitor some variables, manage resources, and develop forecasting models. GIS technology serves such a purpose. In recent years, GIS techniques have monitored the influence of crop plant, crop phenology, topographical variables, or climatic variability on insect distribution; movement of herbivore insects or predators at landscape scale; impact of crop rotations or cropping system on pest damage; agricultural production

techniques, insect community structure and impact of damages caused by insect pests; and invasion and establishment processes by alien invasive insects.

Acquisition of data to feed GIS technology may be a long-lasting and tedious process and it is a major limitation for GIS application in IPM. Remote sensing – acquisition of information on an object without being in contact with it – may reduce the efforts for data acquisition. For instance, remote sensing techniques have been developed to map several types of vegetation or plants that are stressed by damage caused by different amounts of herbivore insects and which display changes in absorption and reflectance in the visible and near infrared light due to chlorophyll content decrease, alteration of other pigments, or some other changes in the internal plant anatomy. Plants stressed by insects or diseases also may alter their temperature and thermal imagery may be used to detect these plants. Images may be field-based or taken from planes or satellites and thus cover big surfaces. Before using remote sensing to make maps of insects or insect-damaged plant distribution, accuracy of sensing images should be verified by means of ground surveys. Signals from damaged plants may not be very specific and insect populations or damage caused by one species may be overestimated.

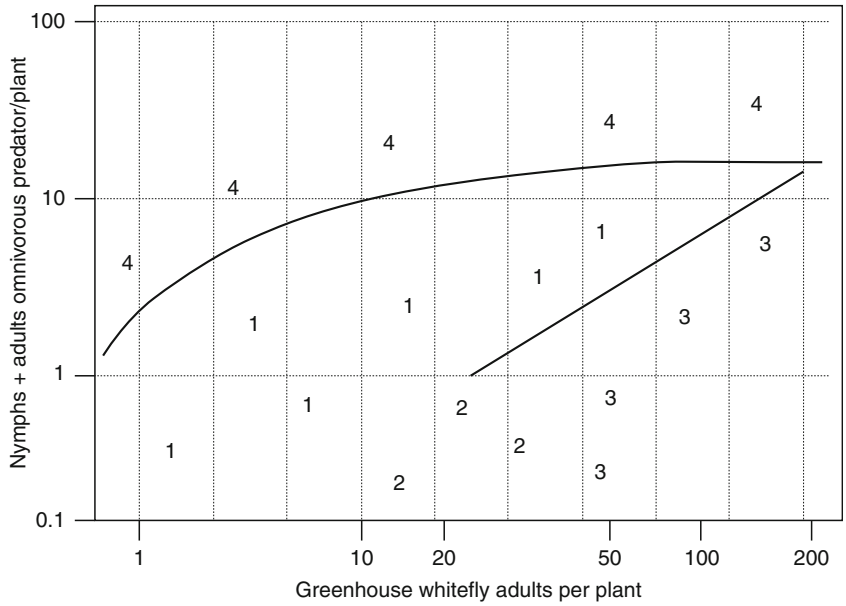
As more data are needed and available for decision making, more sophisticated are the algorithms used to process the information. Processes relating input data with conclusions leading to actions may be performed by expert personnel or by more automatic procedures as computer software; the latter are called expert systems. To consider the role that expert systems may play in IPM, let us recall the steps of any decision in pest control [6]: pest identification, assessment of injury level (density), estimation of likely crop losses, identification of control options, cost/benefit analysis, identification of constraints, integration in the larger framework of crop production [6]. Ideally, expert systems are linked with data bases that allow input and review of historical series of data to analyze past decisions in the light of real posterior events.

Expert systems not only serve for decision making in IPM but may be used as a common base for discussion among scientists or IPM practitioners in order to identify incorrect algorithms or gaps in the current knowledge or incorrect recommendations made in the past. Expert systems can also be used as training tools. Via simulation, students can examine and check situation rules to sample, make calculations, and convert the conclusions into recommendations. When expert systems are used for training or when they have to be presented for demonstration purposes, input variables, algorithms, and logical rules for decision making should be shown in a more explicit format than just computer software, for example, a set of matrices or as a decision chart such as in Fig. 5.

The decision chart in Fig. 5 was designed for field technicians to decide if insecticide treatments are needed in tomato crops. Decision algorithms take into consideration not only the amount of pest (greenhouse whitefly) but also the amount of predators (two mirid bugs) that can keep pest populations under control when the predator–prey ratio is high enough. As mirid bugs are omnivorous predators that may feed on and damage tomato fruit when lacking prey, the strategy adopted was to manage insecticide sprays. This maintained a sufficient number of whiteflies to provide predators with food and prevent tomato damage but was not too high to avoid having honeydew and sooty mold on plants and fruits. Technicians have to sample the field by choosing a number of plants at random and taking seven terminal well-developed leaves to count the number of whitefly adults and predators (adults + nymphs). According to the position in the graph of the values recorded in the field for whiteflies or predators, the decision made is: (a) doing nothing, (b) or spraying against whiteflies, (c) or predators; (d) a four region in the chart does not lead to specific recommendations. Chart is accompanied by keys to identify the targeted whitefly and the predators, some rules to take samples (including a table of random numbers), and an updated list of eligible insecticides with recommendations about their application.

Estimating Insect Population Densities

To make decisions in pest control, pest density needs to be known. As it is normally impossible to count all the insects in a habitat, it is necessary to estimate



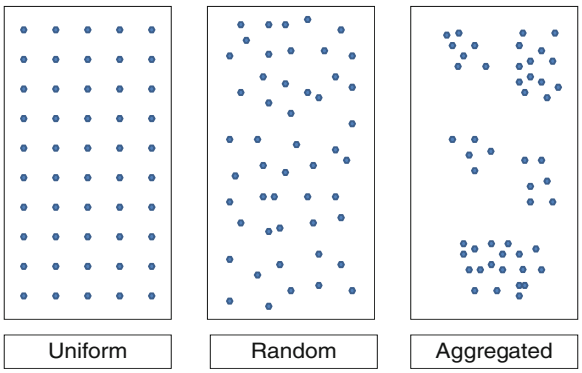
Integrated Pest Management. Figure 5
Decision chart to make recommendations for insecticide spraying in tomato crops against the greenhouse whitefly and the omnivorous predators according to the number of whiteflies and predators recorded on tomato plants. Figure has been simplified to show the most significant elements. Numbers indicate the action to be undertaken after recording densities of the omnivorous predator and the greenhouse whitefly on tomatoes in the field: (1) no action, (2) sample again in one week, (3) spray against whitefly if the recorded numbers remain in zone 2 for two consecutive weeks, and (4) spray against mirids

the population density by sampling. An estimation of absolute or relative population may be needed. In absolute estimations, the estimated number of individuals in a certain surface or volume may be known, whereas in relative estimation a certain and unknown proportion of the population changes in the time or space is determined. The first type of estimate is useful when we want to know if a population has reached the economic threshold or not. The second type serves to compare population numbers in time and space for monitoring purposes. If population numbers may be related to their products – for example, total amount of honeydew excreted by an aphid population – a population index may be more useful than determining the real number of individuals.

An accurate estimate of the density of pests or natural enemies is therefore a major necessity for IPM. To estimate population densities it is necessary to sample the habitat where the population may occur and this may require knowing various aspects. Sampling universe, sampling unit, number of samples to be taken for a certain precision, and data analysis are some of the aspects to be considered in a sampling plan. Readers interested in this concept should consult the Southwood and Henderson book for wider development [7].

How individuals of an insect population are disposed in space – dispersion or distribution of the population – greatly affects the sampling program. Theoretically, three different types of population distributions may be found (Fig. 6). In *uniform distributions* individuals are evenly distributed in space whereas in *random distributions* any point in space has the same probability for occupation by an individual. Finally, most commonly, individuals are clumped on relatively few foci in *aggregated distributions*.

Population dispersion has been described by means of aggregation indices or by fitting mathematical distributions to experimental data. Use of aggregation indices is based on the relationship between variance



Integrated Pest Management. Figure 6
Three kinds of insect population distributions

and mean in the three types of distribution models. In uniform distributions variance is null independently of the mean; in random models mean equals variance; in aggregated distributions variance is higher than mean. The simplest aggregation index uses the variance/mean ratio so that when the ratio is 1 it means that distribution is at random, it is more uniform as the ratio decreases below 1 and, inversely, aggregated models show variance/mean ratios increasing higher than 1 as individuals are more clumped. Other aggregation indices relating variance and mean in different ways have been used – for example Lloyd', Morisita' indices – but most of them have shown a high dependency from sampling unit size and population density, a disturbing inconvenience to characterize a population distribution pattern. Relationship between estimated variance (s^2) and mean (m) has been empirically fitted to a power law: $s^2 = am^b$, where a and b are constants; a is largely a sampling factor while b appears to depend on the degree of aggregation of individuals and thus may be used as an aggregation index. Taylor's power law has shown to overcome some of the limitations mentioned for aggregation indices but usually needs many experimental data to fit a sound power function.

Several mathematical models have been proposed to describe population distributions; for insects many populations have been adequately fitted to the negative binomial distribution which is basically described by the parameter k, which is a measure of the degree of clumping. Although the approach of mathematical models to study population distribution patterns has given more accurate descriptions than aggregation indexes, it often depends on population density and sampling technique. In summary, none of these methods describing population distributions is free from limitations and this can explain why insect ecologists have used so large a variety of approaches to characterize patterns of insect disposition in space.

When populations are very clumped and most individuals occupy a few dense patches among empty patches a question arises: are individuals in one patch moving to another patch? If yes, how often and to which extent does this movement occur? Such a perspective deals with metapopulation ecology. Ecology has classically considered that individuals of a population are capable of unrestricted interaction with each other within a habitat. However, habitats usually occur in a patchwork within a landscape – as islands in the ocean – and populations inhabiting those habitats are consequently patchy. Populations occupying each patch may become extinct but individuals coming from other patches may recolonize that patch and rebuild a new population. Between-patch movement also may be caused by factors other than just extinction and recolonization. In any case, individuals initially belonging to a population in a certain patch occasionally may interact with individuals of other patches. The set of all populations initially occupying different patches is called a metapopulation. There has been a lot of interest in metapopulation ecology in the last decades because of its practical application. In the field of IPM, metapopulation ecology may bring new tools for analyzing the dynamics of insect populations that occur in partially isolated patches (fields or group of fields). Understanding how populations work in a metapopulation may help to increase environmental resistance to pest population development or to favor populations of pest natural enemies. Management of metapopulations and not just local populations may be a valuable approach not only for unstable annual crop systems but also for more permanent crops that are periodically disturbed (i.e., cut, pruned, sprayed with insecticides) so insects are regularly obliged to leave and recolonize fields. In the context of area wide IPM programs, metapopulation ecology may be more predictive than just considering individual populations.

Taxonomic Adscription of Major Pests

Main Animal Taxa with Damaging Species

Most agents causing injuries to crop plants are species belonging to the Phylum Arthropoda and among these the class Insecta includes the majority of arthropods that are crop pests. That is why the discipline dealing with pests and their control very often is called agricultural entomology, the science about insects in relation to agriculture. The class Arachnida, particularly the order Acari (mites), also contains some important pest species. Notice that some herbivorous insects and mites cause injuries to crop plants whereas pathogens are responsible for crop plant diseases.

The class Insecta comprises more than one million identified species and probably even more that have not been identified yet. Insects are grouped in 28 orders – this varies according to the authority– a common taxonomical level to refer to insects. Among those, seven major orders include most of the important insect pests: Orthoptera (grasshoppers and others), Hemiptera (true bugs), Homoptera (hoppers, psyllids, whiteflies, aphids, scale insects), Coleoptera (beetles), Lepidoptera (moths and butterflies), Diptera (mosquitoes, flies), Hymenoptera (ants, wasps, insect parasitoids, and others). Table 1 shows the main families of mites and insects including economically important pests and their most significant pest characteristics.

Other taxa of the animal kingdom include species that can potentially become serious pests. Mollusks (snails and slugs), fishes, reptiles, birds, and mammals may include species that in some circumstances are very damaging.

Identification of Insects and Mites

When a technician observes insects or mites on a crop, the first question that arises is "what are they?" To precisely answer this question is a key starting point for finding an efficient solution if the population grows to become a pest. A wrong answer may lead us to make incorrect decisions as many solutions are specific for each insect pest and crop. Sometimes, even specific identification is needed to adopt correct measures.

Identification of insects and mites is usually done by means of taxonomic keys. Keys are arrangements of related taxa put in clusters. Morphological characters – complemented sometimes with anatomy, appearance, or behavioral features – are mainly used to segregate individuals into clusters. The process goes from more general characteristics (i.e., with wings or wingless) and taxa to more particular characteristics until reaching species-level determination. Even for easy identifications, a certain expertise is needed. For routine identifications field technicians may perform quite well in recognizing common species; when dealing with a new species, correct identification may require sending a sample to a family-level specialist usually working in universities or musea. Availability of good insect taxonomists is therefore critical for developing sound IPM programs, not only to identify insect pests but also their natural enemies (predators and parasitoids).

Molecular techniques are still insufficiently developed to identify insects but rapid progress has been made in recent years. Molecular tools are usually developed to distinguish between two or more taxonomically close species. Molecular identification keys based on a targeted DNA sequence or marker may be useful for such purposes and have some advantages in relation to classical morphological keys. These advantages include: applicability to all developmental stages; less variation than for morphological characters; they may be applied to fragments of the individual to be identified; the technique may be applied for a variety of insects if appropriate specific material and personnel trained in molecular tools are available, in contrast with morphological-based keys that usually require family-level specialists.

Pesticides

Use of Pesticides

In the coming sections, major *control methods* are reviewed in the light of how they can contribute to the sustainability of agriculture. Ecological bases, common applications, and how they can be integrated into IPM systems are presented in each method.

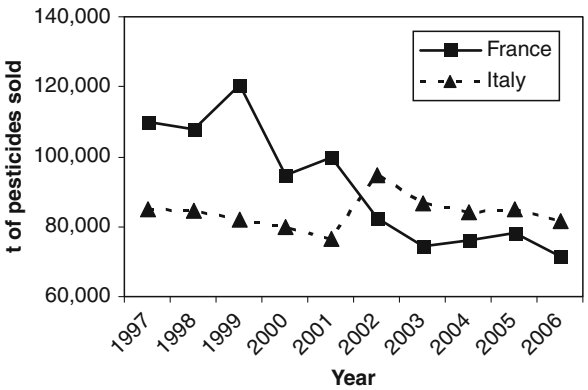
Use of pesticides (mainly herbicides, insecticides, fungicides, and nematocides) in western world agriculture has decreased or been maintained in general.

Integrated Pest Management. Table 1 Main orders and families of arthropods including economically important pest species

Class	Order	Family	Main characteristics and features as pests
Arachnida	Acari	Tetranychidae	Plant-feeding spider mites. They feed on several aerial plant parts but mainly on leaf undersides with loss of photosynthetic products and water
		Eriophyidae	Microscopic mites with only two legs; they feed on plant parts often causing galls
Insecta	Orthoptera	Acrididae	Insects that at high density may aggregate in groups and migrate long distances and become very destructive on many crop and forest plants
	Hemiptera	Pentatomidae	They suck plant sap from several tissues resulting plant wilt, abortion of fruits, or tissue malformations
		Miridae	Numerous, but not only, herbivore species that feed on plant sap causing foliar chlorosis, cankers, abnormal growth, and many kinds of lesions
	Homoptera	Cicadellidae	Usually they feed on leaves where they suck juices and reduce chlorophyll contents and produce small white spots. Plant vigor decrease and disease transmission are common injuries
		Psyllidae	They feed on the phloem causing plant stunting or poor plant growth and sometimes gall forming
		Aleyrodidae	They feed on the phloem and reduce plant vigor, exude honeydew where sooty mold may develop causing fruit depreciation and some species are active plant disease vectors
		Aphididae	As mentioned for Aleyrodidae with special importance for plant virus transmission
		Coccoidea	In addition to the injuries mentioned for the other Homoptera, scale insects may inject saliva into the host causing discoloration, malformations, galls, and also esthetic damages in ornamental plants
		Diaspididae	
		Asterolecaniidae	
		Coccidae	
		Margarodidae	
		Pseudococcidae	
	Coleoptera	Scarabaeidae	As pests they mainly feed on plant roots in larval stages causing plant vigor decrease and even plant death
		Elateridae	Larvae feed below ground on roots and tiller base and kill the plant when it is young. Crops harvested for roots or tubers are more easily injured
		Curculionidae	Many species whose larvae and adults feed on several plant tissues including roots, tillers, leaves, flowers, and fruits. A very damaging family
		Chrysomelidae	Adults and larvae feed on foliage and fruit; in some other species larvae feed on roots
		Scolytidae	Larvae feed internally in tree tissues, below the bark; particularly harmful in forest trees but also in orchard and ornamental trees
	Lepidoptera	Tortricidae	Larvae feed on leaves, fruit, buds, and stems
		Pyralidae	Larvae are leaf-rollers, borers, and detritivorous attacking a large number of crops including stored products
		Crambidae	Larvae are mostly grass stem borers

Integrated Pest Management. Table 1 (Continued)

Class	Order	Family	Main characteristics and features as pests
		Noctuidae	Larvae feed on leaves, stem, and fruits devastating many crops. This is one of the most important families with many economically important species
	Diptera	Cecidomyiidae	Main injury comes from their capacity to cause galls in several plant tissues
		Tephritidae	Larvae feed internally in fruits. They have a high destructive potential
		Agromyzidae	Larvae mine leaves by feeding parenchyma cells between up and down epidermis
	Hymenoptera	Tenthredinidae	Most damage is caused by larvae feeding on leaves that reduce photosynthesis activity and thus plant vigor. They commonly defoliate forest trees but also some agricultural crop plants
		Cephidae	Larvae bore into stems of grass plants and cause their breakage



Integrated Pest Management. Figure 7
Evolution of sales of active ingredients with pesticide activity in two significant consumers in Western Europe (From http://epp.eurostat.ec.europa.eu/tgm.Accessed 6 February 2010)

In Fig. 7, the amount of pesticides sold in two main consumer European countries is shown. Whereas in France, the highest consumer of pesticides in Europe, the amount of pesticides sold (mostly applied in France) has decreased significantly, in Italy the amount of pesticides is more or less stable or shows a slight increase taking into account that modern pesticides are used at considerably lower doses than classical active ingredients. Beyond differences due to variable economic and climatic conditions, an important part of the decrease in many countries has been achieved by the progress of application of IPM systems to control insect pests, diseases, and weeds.

Pesticides are chemicals aimed to kill any kind of plant pest or otherwise lower their populations to prevent their reaching economic thresholds. Pesticides include four main groups of substances according to their target: herbicides against weeds, insecticides against insects or other pests, fungicides against in general disease-causing agents, and nematocides against plant pathogenic nematodes. A number of characteristics of pesticide use may explain the success of pesticides for pest control in the last decades although probably the amount applied has started to decrease in the developed world in recent years due to the restrictions posed by the legislation and also by the progress experimented by the R&D in implementing IPM systems.

Pesticides are easy to use; growers may fill the tank and spray many hectares while sitting in the tractor and listening to the radio. When used correctly they are effective to lower pest populations and frequently cheaper than other control alternatives. In spite of the better selectivity and lower permanence of modern pesticide active ingredients, at least one pesticide is usually available in the market for each pest. Until the discovery of insecticide properties of DDT in the 1940s, most pesticides were inorganic or extracted from plants. Since the 1940s until the end of the century, the amount of pesticides applied in the world multiplied dramatically per 20 or 30 times according to the country and several chemical families were available to

control insect pests. Since the 1950s, however, scientists were aware of the problems derived from the excessive confidence in the efficacy of pesticides. Fewer than 20 years of mass application of pesticides in western agriculture were sufficient to display some of their negative effects. In ulterior years, problems became harder and many field data confirmed first concerns. Development of alternatives to chemical pesticides therefore became the goal of R&D programs in most of those countries.

Problems Associated to Pesticide Use

Problems derived from inadequate and excess pesticide use include (a) risks to public health and environment (e.g., wildlife and groundwater), (b) disturbance within agrosystems due to the common toxicity to natural enemies and secondary pest resurgence, (c) development of pesticide resistance in the targeted pests (more than 600 pest species related to agriculture and human and livestock health are nowadays confirmed to be resistant to one of more pesticides) (www.pesticideresistance.org accessed on February 10, 2010), and (d) shorter and shorter shelf life and increasing costs to innovate by producing more selective and environmentally friendly new active ingredients. Industry has tried to develop new more compatible chemicals in order to integrate selective chemicals in IPM strategies but innovation is increasingly slow and expensive. Legislation is becoming very strict for registration of new pesticides and obliges repeated registration of old active ingredients for health and environmental safety. As a consequence the number of active ingredients available for chemical pest control is decreasing constantly. It is expected that the number of active ingredients registered as insecticides in the coming years in the EU will be less than a third of those allowed at the end of the twentieth century. Lack of effective insecticides is pressing research in and the development of new and efficient IPM systems.

Controversy on pesticide use in modern agriculture cannot lead us to forget that pesticides, at the moment, have still an important role in IPM systems. Some important pests lack sufficiently effective control methods so that no-chemical methods have to be combined with chemical ones. In other, although few, cases, pests have only chemical insecticides to control them. Invasive exotic insect pests, due to short experience in their control and novelty, may be contained only by the regional application of chemical pesticides. A rigorous analysis of how sustainable is the use of each insecticide for each pest should permit detection of those pest problems in which insecticides are irreplaceable at least at short time and those other pests in which one insecticide is superfluous – therefore that pesticide can be banned – because at least one efficient nonchemical method is available. Unfortunately the control of unnecessary use of chemicals is frequently difficult but should be implemented to speed up the adoption of IPM technology.

Crop Resistance

Ecological Bases of Crop Resistance

Most pests cause plant damage when feeding. However, it is well known by interested observers of nature that not all herbivore insects may feed on any plant. Usually some insects may feed on a few close plant species (monophagous insects), or on plant species belonging to one family or only a few families (oligophagous insects), and finally some others may feed on a broad range of plants (polyphagous insects). Even polyphagous species feeding is usually restricted to a relatively small number of plants available in the habitat. These associations between herbivore insects and host plants are the result of the coevolution of the two components; in such coevolution, plants develop mechanisms to defend themselves from herbivore insects and insects try to develop mechanisms to overcome plant defenses. That a plant possesses some characteristics that diminish its access for one insect pest and that such a trait may be introduced into a host plant for pest control is an old idea. Still, until well into the twentieth century crop resistance to insects was not systematically considered as a universal tool for pest control in spite of early successes in the use of plant resistance. Probably, the most famous early case of using crop plant resistance for pest control was the control of grape phylloxera in European grapevines. Practically all the European wine industry was ruined by the entrance into Europe of the North American phylloxera aphid in the 1860s. Successful control of the pest was achieved at the very end of that century by grafting European vineyards on resistant American rootstocks.

Insect–Plant Relationships and Plant Characteristics for Crop Resistance

Many physical and chemical factors are involved in insect–plant interactions. Plant stimuli and elicited insect responses are usually studied in a sequence of five behavioral steps: (a) host habitat searching, (b) host searching within the habitat, (c) recognition of a host as suitable for feeding and ovipositing, (d) host acceptance, and (e) host suitability. Step (a) is important for species that migrate or disperse over long distances and it is only occasional for pests staying in the crop habitat. Preferred abiotic conditions of the habitat are usually the most involved signals in habitat selection whereas in host selection (step b) visual and olfactory stimuli have a major relevance to bring the herbivore close to the host plant and not only olfactory but also tactile inputs stimulate the herbivore to remain on the plant. The host is recognized through gustatory receptors that identify particular host substances when bitten by the herbivore or when the female starts ovipositing. Similar mechanisms but different substances in the host plant cause the herbivore to continue or stop feeding or ovipositing after the recognition phase. Host adequacy for the herbivore and descendants is determined by the nutritional value of the plant and the absence of toxic compounds.

Physical and chemical plant characteristics that confer resistance to the host plant against the exploitation by the herbivore may be found in each of the behavioral steps and mechanisms and those traits may be incorporated into the crop plant genotype for pest control. Mechanisms involved in host-plant resistance to herbivores may be grouped into two main categories:

- *Antixenotic* mechanisms prevent herbivores from approaching or establishing on a plant to feed or oviposit on it. There is a varied array of chemical and physical deterrents in plants to prevent or modulate preference of herbivores for feeding/ovipositing. Chemicals may be plant volatiles that act at long distances or nonvolatiles that intervene once the herbivore has landed on the plant or after it has probed the host. Physical characteristics in plants with antixenotic properties include morphological and structural features that interfere with normal feeding or oviposition. For example, plant epidermis hairs and trichomes are common morphological features that impede normal feeding in many herbivore insects and confer resistance to hairy cultivars.
- *Antibiotic* mechanisms cause deleterious effects, including mortality, for the herbivore once it has ingested a certain amount of the host plant. Antibiotic effects on herbivores that have fed on resistant plants may be expressed in a variety of consequences, and not only mortality: lowered development rates, failure of development features like pupation or adult emergence, reduced fecundity or fertility, and irregular behavior. Antibiosis is caused by toxic plant compounds or by other nontoxic plant characteristics like low nutritional quality, unbalanced composition in nutrients, or presence of enzymes interfering with normal insect digestive physiology.

There are mechanisms of plant resistance that are not related to host-plant constitution. These are "ecological resistance," "induced resistance," and "tolerance." Although they have been often neglected in plant breeding programs, their consideration when crop cycling and crop management practices are planned may contribute greatly to reduce crop losses by pests. *Ecological resistance* derives from the phenological asynchrony of crop and pest populations that prevents or diminishes the coincidence of the most susceptible crop growth stages with the most pest stages that are able to attack the host; sowing date may thus be planned for enhancing ecological resistance. *Induced resistance* is the response of a host plant to an environmental stress that reduces herbivore insect fitness or the plant availability for the insect. Once again, many agricultural techniques, like fertilization or irrigation, may alter plant physiology to enhance or decrease induced resistance. *Tolerance* is the capacity of some crop plants or crop cultivars to recover from injuries caused by the pest so that tolerant plants may attain yields that a similar amount of pest on a nontolerant plant would reduce. Several physiological processes have been identified in plants that may compensate for the injury caused by a herbivore insect and those processes may be facilitated by good agricultural practices.

Crop Resistance and IPM

Host crop resistance has been used profusely for disease control but it is increasingly considered and applied for integrated pest management. It is largely *compatible* with other IPM tactics, particularly biological control, although interferences of the host plant in predator–prey relationships have been reported recently. Its effectiveness is cumulative as its effects on pest population is exerted on several pest generations and usually persists after a pest's long exposure to the resistance traits; however, some cases of pests that have overcome resistance barriers at midterm have been described. One of the major advantages of crop resistance as a pest control method in the framework of IPM programs is that growers may easily adopt it as no particular expertise is needed. From a point of view of environment, crop resistance is in general safe, a required trait of any IPM method.

There are also some important limitations for a wider adoption of crop resistance in IPM. Frequently a long time is needed to develop a resistant cultivar, particularly in tree crops, and reaction to new and urgent problems is very slow. Modern biotechnology techniques, particularly genetic engineering (see below), are contributing to mitigate this handicap. Additionally, resistance traits are not always identified and available or they are very difficult to introduce into crop plants (also genetic engineering may favor the transfer of resistance genes between nonsexually compatible organisms). Another difficulty to implement crop resistance for pest control is the incompatibility of the resistance trait and commercial requirements; bad taste for consumers, for example, has been found in some resistant cultivars. In spite of the persistence of crop resistance mentioned as an advantage of the method, and as also mentioned earlier, local biotypes of the pest that are able to overcome or avoid resistance characters in the plant may develop and rapidly multiply as a consequence of their supremacy to exploit the resistant cultivar.

Biological Control

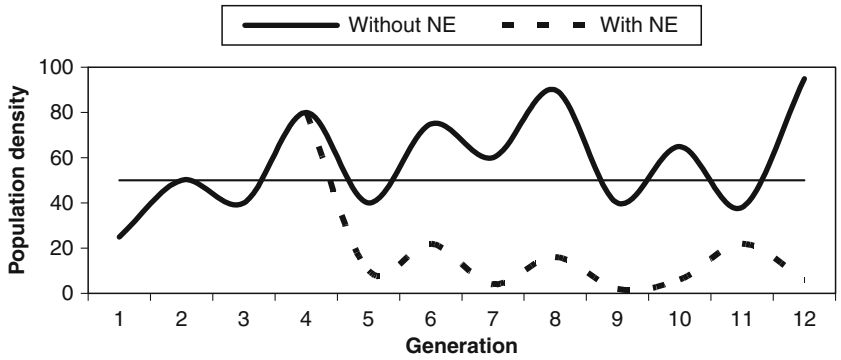
Ecological Basis of Biological Control

Biological control may be defined in general terms as the use or manipulation of natural enemies to suppress pest populations. Some authors include in the term any kind of nonchemical method that is biology-based. However, this is generally considered as incorrect and the narrower meaning given above is preferable. Although biological control may be practiced under several modalities, its principle is unique and responds to the scientific background of predator–prey ecology.

Natural enemies have been signaled as major components of natural control keeping populations within cyclic oscillations between maximal and minimal bounds in the framework of a "spontaneous balance of nature." Therefore the idea that exotic pests greatly increase their densities mainly due to lack of natural enemies in the new habitat led to attempts of reconstituting that balance by importation and release of exotic natural enemies. The ecological basis of biological control may be represented as shown in Fig. 8. A pest population that oscillated around a mean density very often above economic threshold is reduced to oscillations below that threshold after a release of an effective natural enemy. An important part of recent population ecology developments has dealt with prey-predator models that should provide the theoretical basis of biological control allowing it to progress from a rather empirical practice to a scientifically based technology. Unfortunately, the contribution of theoretical developments on predator–prey relationships has not been as fruitful as expected for biological control applications although they have clearly helped to progress beyond the "trial and error" stage of the first half of the twentieth century.

Understanding relationships between predators and prey and between parasitoids and hosts is important to optimize biological control practices. Prey consumption or host parasitization is the successful result of several behavioral steps of predator/parasitoid including:

- Selection of a *suitable habitat* where prey/host is more likely to occur. Predators and parasitoids may respond to biotic and abiotic characteristics of habitats where prey may be found. Several long-distance visual and olfactive stimuli from habitat components, including the proper prey/host, may be involved in habitat suitability recognition by searching predators and parasitoids.
- Once in the suitable habitat the predator/parasitoid has to *find a prey/host*. Within the habitat vision and



Integrated Pest Management. Figure 8
Fluctuations of a pest population before and after releasing an effective natural enemy. *Broken line* shows the fluctuation under the action of the natural enemy whereas *solid line* shows the dynamics of the pest without the natural enemy. *Horizontal line* represents the value of economic threshold

olfaction also play a major role, but short distance stimuli may be decisive as adults sometimes look first for a prey in a random way until they come in contact with a potential prey.

- *Acceptance of a prey/host* when it is attacked by the predator/parasitoid is influenced by physical and chemical characteristics of the prey/host but also by hunger of the predator. Abundance may also determine if a prey is attacked or not.
- Prior to the final decision to consume/parasitize requires checking that a *prey/host is suitable* for the predator/parasitoid. Internal composition of prey largely determines if the predator rejects or continues to feed on it.

A relatively low amount of predatory arthropods are quite specific (they feed only on a particular family of prey) whereas generalist predators (they may feed on a wide range of prey taxa) are more common among both insects and spiders. In some groups both juvenile and adult stages are predatory but in some others only juveniles or adults prey. Parasitoids are usually more specific than predators and in general they can parasitize species belonging to one family or to a narrow range of families.

Taxonomic Adscription of Insect Natural Enemies

Natural enemies include three kinds of organisms: predators, parasitoids, and entomopathogens. Biological control only deals with the former two types whereas the third is the subject of microbial control. Predators are organisms that kill and consume a number of other organisms, called prey, along their lifespan from which they obtain the energy needed to grow, develop, and reproduce. Parasites are organisms that usually need to consume only one organism, the host, to develop and reproduce. When the host dies as a result of the action of one parasite this is called parasitoid.

Predaceous habits are relatively common in classes Insecta and Arachnida. Among insects, five orders include particularly important predators in agrosystems: Hemiptera (true bugs), Neuroptera (nerve-winged insects), Coleoptera (beetles), Diptera (flies), and Hymenoptera (wasps and ants). Among Arachnida, the order Acari includes some predatory families, particularly Phytoseiidae, and many predatory species grouped in several families belong to Araneae (spiders). Whereas predators are located in many insect and mite families, parasitoids mainly belong to certain families of Hymenoptera and a few to Diptera. Early on biological control almost exclusively used entomophagous insects but currently a wide range of organisms are being applied or manipulated to control pests. Table 2 shows those predators and parasitoids that are or have been commercially used in Europe and Mediterranean countries (http://archives.eppo.org/EPPOStandards/biocontrol_web) (accessed in December 2009).

Integrated Pest Management. Table 2 Main predators and parasitoids that are or have been commercially provided for biological control in Europe and non-European Mediterranean countries

Phylum/order	Family	Species	Main target pests
Predators			
Insecta/Hemiptera	Pentatomidae	*Podisus maculiventris*	Lepidoptera larvae, Colorado potato beetle
		Picromerus bidens	Lepidoptera larvae
	Anthocoridae	*Orius albidipennis*	Thrips
		O. laevigatus	Thrips
		O. majusculus	Thrips
		Anthocoris nemoralis	Psyllidae in orchards
		A. nemorum	Pear psylla
	Miridae	*Macrolophus caliginosus*	Whiteflies
Thysanoptera	Aeolothripidae	*Franklinothrips megalops*	Thrips
		Franklinothrips vespiformis	Thrips
		Karniothrips melaleucus	Scales
Neuroptera	Chrysopidae	*Chrysoperla carnea*	Aphids
Coleoptera	Staphylinidae	*Aleochara bilineata*	Larvae of *Delia* sp. flies in soil
	Coccinellidae	*Adalia 2-punctata*	Aphids
		Chilocorus baileyii	Armored scale insects
		C. bipustulatus	Armored scale insects, olive black scale
		C. circumdatus	Armored scale insects
		C. nigrita	Armored scale insects, pit scales
		Coccinella septempunctata	Aphids
		Cryptolaemus montrouzieri	Mealybugs
		Delphastus catalinae	Whiteflies
		Rhyzobius lophanthae	Armored scale insects
		Rodolia cardinalis	Cottony cushion scale
		Scymnus rubromaculatus	Aphids
		Stethorus punctillum	Red spider mite
Diptera	Cecidomyiidae	*Aphidoletes aphidimyza*	Aphids
		Feltiella acarisuga	Red spider mite
	Syrphidae	*Episyrphus balteatus*	Aphids
Arachnida/Acari	Phytoseiidae	*Amblyseius barkeri*	Thrips, tarsonemid mites
		A. degenerans	Thrips
		Hypoaspis aculeifer	Sciarid flies in soil substrates
		Metaseiulus occidentalis	Spider mites
		Neoseiulus californicus	Spider mites

Integrated Pest Management. Table 2 (Continued)

Phylum/order	Family	Species	Main target pests
		N. cucumeris	Thrips
		Phytoseiulus persimilis	Red spider mite
		Typhlodromus pyri	Some spider and eriophyoid mites
	Laelapidae	*Stratiolaelaps miles*	Sciarid flies in soil substrates
	Cheyletidae	*Cheyletus eruditus*	Storage mites
Parasitoids			
Insecta/Hymenoptera	Mymaridae	*Anagrus atomus*	Leafhoppers
	Encyrtidae	*Anagyrus fusciventris*	Mealybugs
		A. pseudococci	Mealybugs
		Comperiella bifasciata	Armored scale insects
		Encyrtus aurantii	Scales
		E. infelix	Scales
		Gyranusoidea litura	*Pseudococcus longispinus* (mealybug)
		Leptomastidea abnormis	Mealybugs
		Leptomastix dactylopii	*Pseudococcus citri* (mealybug)
		L. epona	Mealybugs
		Metaphycus flavus	Soft scales
		M. helvolus	Soft scales
		M. lounsburyi	Soft scales
		M. swirskii	Soft scales
		Microterys nietneri	Soft scales
		Pseudaphycus maculipennis	Mealybugs
		Tetracnemoidea peregrina	Mealybugs
	Aphelinidae	*Aphelinus abdominalis*	Some aphids
		Aphytis diaspidis	Some armored scales
		A. holoxanthus	Armored scales
		A. lingnanensis	Some armored scales
		A. melinus	*Aonidiella aurantii* (armored scale)
		Coccophagus lycimnia	Soft scales
		C. rusti	Soft scales
		C. scutellaris	Soft scales
		Encarsia citrina	Armored scales
		E. formosa	Greenhosue whitefly
		Eretmocerus eremicus	*Bemisia tabaci* (whitefly)
		E. mundus	*Bemisia tabaci* (whitefly)
		Cales noacki	*Aleurothrixus floccosus* (whitefly)

Integrated Pest Management. Table 2 (Continued)

Phylum/order	Family	Species	Main target pests
	Aphidiidae	*Aphidius ervi*	Aphids
		A. colemani	Aphids
		A. matricariae	Aphids
	Braconidae	*Bracon hebetor*	Lepidoptera
		Cotesia marginiventris	Lepidoptera
		Dacnusa sybirica	*Liriomyza* spp. (leafminers)
		Opius pallipes	*Liriomyza* spp. (leafminers)
		Praon volucre	Aphids
	Eulophidae	*Aprostocetus hagenowii*	Cockroaches
		Diglyphus isaea	*Liriomyza* spp. (leafminers)
		Thripobius javae	Thrips
	Trichogrammatidae	*Trichogramma brassicae*	Lepidoptera
		Trichogramma cacoeciae	Lepidoptera
		Trichogramma dendrolimi	Lepidoptera
		Trichogramma evanescens	Lepidoptera
	Pteromalidae	*Scutellista caerulea*	Soft scales

The dozens of natural enemies included in Table 2 are only a part of those used in biological control in Europe. At least the same amount can be added when established agents managed for biological control by conservation (conservation and enhancement of natural enemies by crop and habitat management, see biological control by conservation below) are considered. Many other arthropod (e.g., dermapterans, carabids, staphylinids, coccinellids, syrphids, mirids, nabids, lygaeids, spiders, phytoseiids, and stigmaeids) and non-arthropod groups include predatory species that are or have been directly or indirectly managed to suppress pest populations in agriculture in Europe. Similarly, the same or close families to those cited in Table 2 include other species of parasitoids that have been used in biological control by conservation.

Strategies of Biological Control

To achieve pest suppression biological control may follow three main strategies: release of new natural enemies in a habitat (classical biological control); augmentation of natural enemies in the habitat (augmentative biological control); and conservation of those natural enemies already established in the habitat (conservation biological control) [8].

Classical biological control. Modern biological pest control began at the end of the nineteenth century when an exotic pest, the cottony cushion scale, invaded Californian citrus orchards and caused severe damages. Among other species, a predatory coccinellid beetle, *Rodolia cardinalis*, was imported from its Australian origin and released in some orchards of California. In a few years the predators spread throughout the citrus growing area and greatly reduced scale densities. Success of the strategy created much enthusiasm and many other cases of introduction of exotic natural enemies to control exotic pests followed the release of the Rodolia beetle. Some of them resulted in successful pest suppression but some other failed due to several causes. In spite of this, several benefits derived from successful and failed attempts. First, biological control was shown as a feasible technique to suppress pests. Second, a network of entomology laboratories – mainly

American – spread around the world to support expeditions to look for natural enemies and some of them were the embryo of future biological control institutes. Third, successes and failures have pushed entomology science to develop the taxonomy of many families of predators and parasitoids, to furnish an ecological basis of predator–prey relationships for a more theoretically founded biological control and convince governments and policy makers that nonchemical methods may be as efficient as pesticides in controlling insect pests (if not more so). Although there are no reliable records of all introductions of new natural enemies for biological control, more than 1,000 cases have been reported in the last 100 years and about 60% provided a complete control or a substantial reduction of pest damage. A classical reference on the history of biological control in the twentieth century is DeBach [9].

A major concern and limitation of classical biological control is the potential impact that introduced exotic natural enemies may cause on native fauna. Criticism of biological control by introduction of exotic natural enemies was soon theoretically enounced and several authors have advocated for limiting the trade of commercial biological control agents. In spite of little field data supporting the idea of strong impacts on native fauna, several countries have prepared regulations to forbid the importation of the most risky species of natural enemies. Although some principles have been established to assess risks of classical biological control agents much more knowledge on natural enemy interactions is needed. Two references may be useful for readers interested in environmental impact of biological control agents [10, 11].

Conservation biological control. As shown in Fig. 2, pest population development is the result of interactions among many biotic and abiotic components among which are herbivores and their natural enemies. In fact, pest population outbreaks are often due to agricultural practices which interfere with natural enemies that exert a certain control on herbivore populations in agrosystems. The goal of conservation biological control is to restore or enhance conditions for natural enemy survival, reproduction, and activity.

Conservation biological control may be implemented by manipulating the crop or the habitat. Main concerns of this kind of biological control are the identification of the natural enemies that can play a major role in keeping pest population at tolerable levels and what practices interfere with their functioning. Of course, pesticides are one of the principal interferences with natural enemies either by directly causing mortality or by indirectly influencing their biology, behavior, and movement at sublethal concentrations. Nowadays, data on the effects of pesticides on common natural enemies are needed before a pesticide is authorized; today, selectivity and low persistence in the environment, contrarily to that of years ago, are positive characteristics of pesticides used in pest control. Innovation in pesticide application techniques also tries to reduce the amount of pesticide and concentrates the application on certain sites, two goals that may increase selectivity in relation to natural enemies.

Certain practices to manage soil, water, and crop residues may contribute to natural enemy enhancement. Many insect pests live in the soil or have a part of their life cycle in the soil where they may be attacked by natural enemies; soil tillage has to be practiced in a safe and timely manner for such natural enemies. This includes crop residue management. After harvesting, crop plants may host a high variety of pest natural enemies; destroying crop residues also destroys pest individuals but may inflict a higher damage on natural enemy populations. Water management must provide relative humidity values that are optimal for natural enemies while damaging for pests.

Generally, the more stable is the agrosystem, the more chance natural enemies have to play their role as biological control agents. Whereas permanent crops allow population establishment and increase year after year, annual crops have to be colonized each year by natural enemies. However, cropping patterns may be designed to promote earlier and more abundant field colonization by natural enemies or enhance their survival and reproduction once the crop is established. Manipulation of sowing and harvesting dates may facilitate earlier colonization in the season and maintenance of natural enemies after season. Some other practices may also contribute to stabilize agrosystems. Some examples of crop manipulation to benefit

natural enemy survival and activity include strip harvesting, variable crop phenology, and inclusion of banker plants to keep prey and predators in the field between seasons.

Crop diversification seeks to delay crop colonization by insects, both pests and natural enemies, or reduce their retention in the crop. Intercropping – growing two or more crops in the same field at the same time – can cause a more abundant colonization by natural enemies although there are records of the contrary phenomenon. In non-crop-diversified agrosystems, herbivore colonizers arrive earlier and in higher numbers and consequently higher populations of natural enemies may be built up. Crop diversification may be performed by different spatial and temporal patterns of crops in the landscape where one field is the source for colonization of neighboring fields. Diversification of landscape with nonagricultural vegetation is another way to enhance natural enemy presence and activity. Non-crop plants in inter-rows, margins, roads, "fallow" plots, etc., may provide shelter and even food for natural enemies when conditions on crop plants are not suitable for natural enemies and may facilitate the movement of predators and parasitoids among fields in a patchwork landscape.

Biological control by conservation has shown to be very efficient when practiced after a sound knowledge of the agrosystem. Furthermore it is safe as managed natural enemies are already established in the habitat and no negative effects on the environment may be expected. Unfortunately, biological control by conservation is not always possible due to a lack of effective natural enemies in the habitat or a lack of knowledge about how crop or habitat may be managed to effectively enhance predator or parasitoid action. Research needed to implement a program of conservation biological control has to be closely linked with local conditions, and solutions are not universal but related to each particular location.

Augmentative biological control. Although natural enemies exist, sometimes conservation and enhancement practices are not sufficient to increase their populations until reaching levels which are capable to suppress pest populations at desirable levels. In these cases the release of reared natural enemies is needed. Augmentative releases may be necessary when the natural enemy is not present at the place and time needed. There are two types of augmentative releases:

- A relatively low number of individuals of the natural enemy is released and the suppression is expected to be achieved by the first or ulterior descendants: *inoculative augmentation.*
- A high number of individuals are released and control is expected to be exerted by them: *inundative augmentation.*

As in the previously described biological control strategies, augmentative releases use natural enemies that are able to search, locate, recognize as suitable, and attack the prey, but supplementary characteristics in the natural enemy are required in augmentative releases. Biocontrol agents have to be reared, and in inundative releases, they have to be mass reared. Rearing techniques have to meet several economical and quality requirements. The high costs of rearing are mainly due to a lack of true artificial diets but also by the necessity of labor to manipulate materials and individuals and the necessity of producing natural enemies for a rather narrow interval of time in the year. Predators and parasitoids need to be reared on their natural herbivore prey/host or, in some cases, an easily reared alternative prey/host, as in the case of flour moth eggs that can be produced daily for millions at relatively low cost and supplied as food to rear many generalist predators. In addition, herbivores have to be reared on their natural host plant, or similar alternatives, that usually need specific temperature, humidity, and light conditions, sufficient space, and a lot of manpower. Automatic processes to produce plants and natural enemies in biofactories have been designed but much more still has to be done to lower production costs. The extra cost of producing for a few demand peaks in the year may be mitigated if an effective storage method is available. For instance, the egg parasitoids *Trichogramma* spp. may be produced throughout the year and stored for months as diapaused larvae that are reactivated in their development some weeks before they have to be applied in the field.

Continuous rearing of natural enemies in nonnatural conditions and ulterior transport to the field may alter their quality to perform as biological control agent. Genetic uniformity and inbreeding,

negative selection for desirable characteristics to perform in the field, high impact of diseases in mass rearing colonies, and lack of learning opportunities of natural features may be some of the problems derived from natural enemy mass rearing. Procedures for quality control are being developed for the most common natural enemies [12]. Dispersal and search capacities, fecundity, health, correct species, and biotype are some of the characteristics checked in natural enemies for quality control.

Biological Control and IPM

Biological control is an important component of many successful IPM programs. The success of biological control has pushed pesticide companies to design new active ingredients with less impact on natural enemies. Whereas a few decades ago long persistence and broad action spectrum were two positive characteristics of new pesticides, current innovation of chemicals for integrated pest management looks for high selectivity and short persistence thus helping pesticides to become more compatible with biological control. The introduction of a new natural enemy in agrosystems to control a certain pest may allow reduced pesticide application and express the full potential of native natural enemies to control other pests.

Predator–prey interactions may be mediated by the host plant so that a natural enemy that is able to successfully control a pest on a certain crop may fail to do so on another crop. Hairy cucumber cultivars, which are more resistant to greenhouse whitefly than glabrous ones, were preferred to control the whitefly before *Encarsia formosa* was profusely used for biological control of this pest. Once the biological control was generally adopted in Dutch greenhouses, objectives of plant breeding shifted 180º and less hairy cultivars were again cultivated because these facilitated the inspection of leaves by the parasitoid to select and parasitize a host, thus improving the efficacy of greenhouse whitefly biological control. Tritrophic relationships – host plant, herbivore, and natural enemy – have to be considered before planning a biological control program. For IPM programs based on biological control, crop plant management and general cultural practices also have to be adapted to enhance natural enemy activity. Detection of those cultural practices that injure natural enemies may allow cultural modification in order to make predators or parasitoids more compatible with agricultural environments. For instance, tomato deleafing to produce more colored fruits has been found to interfere with the establishment of whitefly parasitoids that are inside the host in lower leaves. Tomatoes may be equally deleafed but leaves should be left on the soil between rows for some days until adult parasitoids emerge and fly to upper leaves to parasitize the host there.

Microbial Control

Entomopathogenic Organisms as a Natural Mortality Factor on Insect Populations

Insects are naturally affected in nature and also in agroecosystems by a varied array of pathogenic organisms, called entomopathogens, that cause diseases on insect pests. When naturally occurring epizootics are not efficient enough to lower pest populations under economic thresholds or they occur too late (natural epizootics are very dependent on natural abiotic conditions), the entomopathogen can be released into the environment at the time needed or manipulate the habitat to enhance the impact of the disease on the pest population. Entomopathogens may be mass produced and formulated for applying as a chemical pesticide. Many aspects described and discussed in biological control may be applied to microbial control. For instance, microbial control can be employed with classical, conservation, and augmentation techniques. In classical microbial control, a pathogen is isolated in a foreign insect and inoculated into a population that previously has never been exposed to that pathogen, whereas in conservation microbial control the impact of an already established pathogen is enhanced by manipulating crop or habitat conditions. Finally, augmentative microbial control inoculates a nonexotic pathogen at the time needed, or the crop field is inundated by microbial pesticides whose efficacy relies on the primary infection and not the secondary one as in the other techniques.

Main Insect Pathogens Used for Microbial Control

There are several kinds of microbes that cause diseases in insects and are used to control insect pests. Groups

with more species and microbial control uses are: bacteria, viruses, fungi, and nematodes.

- Bacteria. The most known and used bacterium is *Bacillus thuringiensis* (Bt). This bacterium has some interesting properties that make its use very compatible with other IPM methods. Entomopathogenic action of Bt originates in some insecticidal toxins – the so-called delta endotoxins – that are produced by the bacterium during its sporulation and once ingested by the insect is activated in the digestive system and causes gut paralysis, feeding cessation, and later larvae show general paralysis. One of the most valuable characteristics of Bt toxins is their selectivity according to the Bt strain. Historically Bt has been used to control specific Lepidoptera pests and, to less extent, some Diptera and Coleoptera. Nowadays, the range of pests targeted by Bt toxins include species of some other insect groups. Also importantly, Bt can be produced in large scale and at a reasonable price by fermentation and then formulated like an insecticide to easily spray or powder the crop. Genes encoding the production of Bt toxins may be transferred by genetic engineering to crop plants which become resistant to insects that are susceptible to the Bt toxins introduced into the plant (see section "Crop Resistance"). Biopesticides based on Bt are widely used in world agriculture, particularly in IPM programs and organic farming. Their relatively high prices and low permanence on the crop plant, mainly reduced by UV radiation when the crop is grown in the open air, are major limitations. Bt biopesticides are less than 1% of the pesticide world market in economic terms.
- Viruses. There are many types of insect pathogenic viruses but only baculoviruses are commercially available for insect pest control, mainly for Lepidoptera and Hymenoptera pests. Baculoviruses have a characteristic that gives them some advantages in comparison with other entomopathogens: they are able to create secondary inoculums and ulterior epizootics so that permanence of control efficacy may be longer than for other microbial control agents, although, like Bt, they are very sensible to deactivation by UV radiation. Furthermore, viruses are very selective and thus especially good for integrating their use in IPM programs. Genetic engineering techniques have shown a high potential to overcome some of the limitations of baculoviruses; higher knockdown effects and host range are two of the traits introduced in genetically modified baculoviruses. Probably their high price is the principal constraint for repeated field applications.
- Fungi. They comprise microorganisms that cause diseases on a variety of insect groups. In comparison with the two above, fungi are able to act in topical applications and do not need to be ingested to be active. This is why they are preferably used against sucking insects like whiteflies, scales, or aphids. Several fungal entomopathogens occur naturally and reduce insect pest incidence in agroecosystems but they are also mass produced and applied as biopesticides. Efficacy of fungal biopesticides is largely limited by low relative humidity although suitable formulations may attenuate this constraint; also UV-radiation protectors in fungal biopesticides are being tried to prolong their persistence on crop plants.
- Nematodes. Nematodes are a quite large group including many species that have several kinds of associations with insects. Facultative and obligate parasitism is among these. Some insect parasitic species have established a symbiotic relationship with entomopathogenic bacteria that confer to these kinds of nematodes a more virulent action against insect pests. In this case the bacterium is responsible for killing the host and, once dead, to preserve the host insect from being invaded by other microorganisms and therefore available for the nematode. These entomopathogenic nematodes are the most valuable for pest control, especially when part of their life cycle is in the soil, where nematodes are more effective. For pests that do not inhabit the soil, nematode desiccation at low humidity conditions is a major limitation of these agents for use in microbial control.

Use of Entomopathogens for Insect Control in IPM: Advantages, Disadvantages, and Techniques

Microbial control is one of the most promising methods to control insect pests within IPM programs. It is highly specific and selective for nontarget

arthropods like natural enemies, it is harmless for vertebrates including man, it has very little risk of environmental pollution, it is easy to apply as most of them are formulated to be sprayed with conventional machinery, and there are techniques for engineered modified entomopathogens with improved performance. Disadvantages include short permanence in the environment for pest control, slow efficacy, moderate probability of generating resistance to the active ingredient in targeted pests, host production cost, and poor acceptance of biotechnology products by consumers in certain countries. In summary, microbial control is an easy method to integrate in IPM and should substantially replace chemical treatments if efficacy and cost problems are solved and biopesticides are perceived by public opinion as a safe replacement for chemicals.

Behavioral Control

Pheromones and Other Semiochemicals

It is well known that insects, and other animals, communicate within the same species and with individuals of other species by chemical signals; chemicals involved in communication are called semiochemicals. Intrapopulation semiochemicals are called pheromones and those devoted to communicating among species are allelochemicals. These are divided into two categories depending if they benefit the chemical releaser (allomones) or to the receiver (kairomones). As more is known about insect behavior, more relevance is given to the role of semiochemicals in the communication governing crucial insect functions and more applications of semiochemicals are envisaged for insect control. Several kinds of semiochemicals have been investigated for both scientific and practical purposes but most attention has focused on pheromones.

Although chemical signals in insect communication had been observed some centuries ago, the chemical identification, synthesis, and demonstration of the sex attraction capacity of a pheromone was performed in the 1950s. A general enthusiasm on potentialities to govern insect behavior and suppress insect pests followed that pioneering work with considerable practical achievements but also with some limitations. The number of insect functions governed or mediated by pheromones is large; in addition to courtship behavior – the most studied for practical applications – social, physiology, trail, defense, finding, discriminating, and aggregating on the host plant are among insect biology features with pheromone involvement. The chemical nature of pheromones is quite diverse according to insect taxon and function. Pheromone composition usually has to meet two main requirements: be highly specific and be easily transportable by air currents. This double requirement is in part contradictory as volatile compounds need to be short molecules, with low molecular weights, but long enough to make possible several combinations of atoms for specificity. As pheromones are typically blends of several organic components synthesized by the insect or less commonly sequestered from the plant, specificity is achieved not only by chemical structure of components but also by the exact composition of the mixture.

Main Application of Pheromones for IPM

The main applications of pheromones for insect pest control may be included in one of the three following groups:

- Pheromones for *detection and monitoring* pest populations. One first application of pheromones is to know when and where a pest population is present. The pheromone is put in a trapping device and number of trapped individuals recorded. There are many kinds of traps that have been used; optimal trap design depends on pheromone composition and insect species. Generally, trap catch numbers will give relative estimates from which absolute numbers of pest population cannot be derived. Early warning, determination of timing for control intervention according to the pest population phenology, early detection for quarantine actions, and dispersal studies are some of the purposes that may be reached with pheromone trapping. However, decisions that need to know population numbers or densities rarely may be based solely on trap catch records unless a sound relationship between relative and absolute estimates had been previously established.
- Pheromones may be used to directly control insect pests by *mass trapping*, that is, by trapping

a sufficient number of individuals from a pest population. Usually a huge amount of traps is needed to remove a significant number of individuals to lower pest damage. The rice stem borer – as other Lepidoptera – is controlled nowadays by mass trapping in the Mediterranean area. A first interesting variant of mass trapping with pheromone traps is practiced with *trap trees* in bark beetles. When a pioneer bark beetle recognizes a tree as a suitable host, it starts releasing an aggregation pheromone. This is soon followed by another attractant emitted by the tree that complements aggregation of many bark beetles on the same tree that may be destroyed, burnt, or sprayed with insecticides to kill many beetles. A second variant of mass trapping is when the attracted insect is not glued on the trap but killed or sterilized and in any case eliminated. Some of the most harmful fruit flies are caught in traps, sterilized, and released again to the environment for control as in the sterile technique programs (see below in the section "Genetic Control"). Principles of mass trapping are very simple but many limitations have prevented successful application. One major constraint of the method is its low efficacy when pest population densities are high; in these circumstances a reducing treatment is needed before applying mass trapping.

- Species in which orientation of one sex to the other for mating is performed by sex pheromones are sensitive to the application of *mating disruption* techniques. The principle, as in the other cases where pheromones are involved, is rather simple. The permeation of the air with synthetic pheromone components – or the whole pheromone blend – interferes with the orientation of the searching sex that usually is unable to meet the other sex; oviposition is thus prevented and population rapidly declines. Despite many studies conducted on the mechanisms underlying mating disruption techniques, much remains still unknown. A number of factors influence feasibility and efficacy of these techniques. Systems for pheromone delivery have to assure air permeation for all the period during which males and females may meet and mate, a requirement difficult to satisfy for very volatile molecules. Synthetic components to be delivered have to be produced and released at reasonable prices; otherwise, the technique is not competitive with insecticides that are cheaper and relatively easy to apply. Efficacy of mating disruption has repeatedly shown to be drastically limited by high pest population densities and this means that mating disruption has to be applied in early generations when populations are still low and forecasting whether they will reach economic thresholds is difficult. In many countries authorization to sell pheromones for mating disruption purposes has to follow hard administrative processes too close to chemical insecticide registration; this sometimes deters companies from investing money in research and development. Dozens of pests are controlled nowadays by mating disruption techniques with acceptable efficacies, especially to manage Lepidoptera in vineyards and fruit orchards.

I

Compatibility of Pheromones in IPM Systems

Pheromones are easy to integrate into IPM systems because they are very compatible with other control systems such as biological control. Selectivity of the method is a major advantage. As environmentally friendly substances they do not cause pollution problems, generally have no toxic effects on nontarget species, and permanence in the environment is low. When used for monitoring pheromones are easy to work with and field technicians do not need particular skills.

Pheromones are an elegant tool for IPM and used more and more profusely sometimes without sufficient rigor. On the other side, scientific research is not sensible enough for solving real problems that would lead to a faster adoption of pheromones in the field. As stated by Millar [13] more pragmatism is necessary in research on pheromones [13]. The deeper knowledge of how pheromones work acquired in the last decades should allow focusing our efforts on those systems that can be most effective. Furthermore, globalization of insect pests should push local and national efforts to more international cooperation because "my pest today may be your pest tomorrow" and vice versa.

Genetic Control

Genetic pest control comprises those techniques that use the insect pest for its own destruction. Genetic

control consists of the release of sterilized, sterile, or incompatible individuals of one species into its wild population to cause a high proportion of sterile matings and hence reduce or eliminate the wild population. Three main methods causing sterility by different mechanisms are available for agricultural pests: sterile insects, incompatible insects, and hybrid sterility. Some other mechanisms have been exploited to control pests by genetic methods. Note that use of resistant/tolerant host crops is not considered within this section of genetic control.

In the sterile insect technique, insects are mass reared in controlled conditions, sterilized, and then released into the field. Insects can be sterilized by irradiation or by chemosterilization. Gamma irradiation, in which dominant lethal mutations arise as a result of chromosome break in treated cells, has been the most used technique in field programs. One of the key aspects to be determined before applying the technique in the field is the optimal irradiation dose to produce enough degree of sterility without causing somatic damages in the sterilized individual. The optimal dose is a function of several factors including insect species, its sex and physiological age, and the level of sterility required. Males are usually sterilized and released. Efficacy of the method relies on the capacity of released males to compete with wild males to mate with wild females. Progeny of females mated with sterile males will die soon after egg oviposition. Ideally insect mass rearing procedures should target males for sterilizing and releasing because females may need a higher dose for sterilization and once released they may be harmful for the crop as a consequence of ovipositing.

With no doubt the successful eradication of the parasitic screwworm fly – the larvae of which eats the living tissue of warm-blooded animals including humans in some circumstances – in wide areas of North and Central America contributed decisively to an increase in the amount of funds devoted to the research and applications of the sterile insect technique. In agricultural pests there have also been a number of successful male sterile programs, particularly in fruit flies. The Mediterranean fruit fly, *Ceratitis capitata*, has been eradicated in some countries of Latin America and they have been declared medfly-free regions; those countries can export fruits to countries which have fruit flies as quarantine pests. In contrast, attempts to use sterile male techniques in several moth pests have failed mainly due to the difficulty or mass rearing the moth at reasonable prices.

Insect incompatibility has been also used in field conditions but less extensively than the sterile insect technique. One of the methods of insect incompatibility used is based on the effect of crossing sexes in two conspecific populations resulting in only partial embryonation and thus population decrease. Incompatibility may be caused by microorganisms that are present in one population and not in the other. For practical pest control purposes, the release of one sex of one geographic population into another location may result in nonviable progeny and therefore population suppression.

Finally, sterility also may be achieved by crossing individuals of two species that produce apparently normal but completely or partially sterile hybrids. If the hybrid mates with at least one of the parent species, it can be mass reared and released in the field for genetic control. An advantage of this system comes from the fact that it does not need a sterilization method and the quality of released individuals is better than in the case of irradiation.

Genetic control has been shown as successfully practicable in commercial conditions in several cases. However, some failures have taught us about the constraints of the method. Reinvasion of treated areas by gravid females is one of the common causes of failure. Real knowledge of the dispersal capacity of the targeted species should prevent applying the sterile male technique in too small areas. Area wide programs are particularly needed for sterile insect technique application to prevent early recolonization of the treated area from neighboring zones.

One of the key factors for successful pest control with the sterile male technique is the ability of treated individuals to compete with wild males for wild females. Lack of adequate information on how to assure competitive treated males and about mating behavior could have been the main failure cause of many genetic control programs. More attention was then devoted to developing quality control of

mass-reared insects that are now routinely applied in many programs. A major conclusion after several years of applying genetic control programs is that the method by itself may be insufficiently effective to be uniquely applied and in many situations it may be useful when integrated in IPM systems to control the target and other interacting species.

Another perspective about the future of genetic control relates to potential contributions of insect molecular biology, more developed in model insects like the fly Drosophila but increasingly interesting for insect pests and pest natural enemies. As genetic control was soon seen as very limited by the lack of appropriate genetic material or by too reduced fitness in irradiated individuals, transgenesis may contribute to renewing the interest for genetic pest control. Introduction of desired transgenes into target insect species or the release of insects infected with engineered microorganisms – mainly rickettsia-like, Wolbachia – which mate with no infected individuals cause reduced fitness progeny. The reader curious about potential contributions of molecular biology to genetic control should consult the review by [14].

Cultural Control

As for any organism, insect populations – and thus pests – have a variable rate of increase depending on, among other factors, biotic and abiotic factors (see Fig. 2). An insect population becomes a pest in the agroecosystems when the insect environment is favorable enough to enhance population increase until economic threshold is reached and control measures have to be adopted. To manipulate that environment to make it less favorable for pests is called cultural control. Many kinds of crop or habitat manipulations devoted to constrain pest population development or enhance natural enemy numbers and activity may be included within the term.

It is quite difficult to list the practices that may be applied to reduce pest populations or enhance natural enemies. Some of the practices routinely applied are apparently unrelated to pests and natural enemies and only when they are eliminated or modified their implication in pest control is discovered. In other cases growers abandoned cultural practices that had been used to control pests due to the efficacy and reliability of cheap pesticides and they had to rediscover their usefulness. (a) Crop rotation, for example, prevented populations of non-generalist herbivorous insects from building up high populations without moving from the same field; crop monoculture provoked a high resource concentration on the same place for a long time and facilitated insect development and reproduction.

In addition to crop rotation, some general cultural practices are used to lower pest populations: (b) removal of crop residues between two successive seasons may reduce insect survival in unfavorable seasons (e.g., cold winters or dry summers); (c) management of planting or harvesting can avoid pest populations peaking at particularly susceptible crop growth stages; (d) unbalanced fertilization use to favor some herbivore insects like aphids that can multiply per several units their rate of increase when fed on plants with an excess of nitrogen; (e) the same as in the previous point could be said for other agricultural inputs like water, mulching, or plant hormones that alter physical environment and crop plant physiology in favor of or detriment to pests.

Many cultural practices have an important impact on pest populations by enhancement of *natural enemies*. Weed management, beyond preventing damages to the crop should take into account their role in insect biology. Nonagricultural vegetation in margins and hedgerows may offer shelter, refuge, or food sources for predators and parasitoids, but also for herbivorous insects that colonize crops early in the season. This double role of margins has to be carefully studied before margin management practices are recommended. The behavioral manipulation of the insect pest and their natural enemies may allow making the protected resource (e.g., the crop) unattractive for the pest and attract it to an unprotected resource (e.g., nonagricultural plants). The opposite done for natural enemies in this kind of strategy is called "push and pull" [15]. *Intercropping* – the cultivation of two or more crops simultaneously on the same field – is another potential way to manipulate the environment for making it unfavorable to the pest. It is frequently observed that there are fewer pests in fields with intercropping than in monoculture; more

attractiveness for natural enemies and less for herbivorous insects are two hypotheses to explain such a phenomenon.

Biotechnology and IPM

Emerging Biotechnological Techniques for IPM

Developments in plant biotechnology have contributed decisively to progress in agriculture in general and IPM in particular in the last decades of the twentieth century [16]. Biotechnology has been defined by the United Nations Convention on Biological Diversity as any technological application that uses biological systems, living organisms, or derivatives thereof, to make or modify products or processes for specific use (http://www.cbd.int/convention/ accessed on January 27, 2010). Probably, genetically engineered crops are the most socially known products of biotechnology applied to plant breeding. However, there are many other achievements and tools issued from plant biotechnology that have allowed progress in the scientific basis of IPM and multiple applications in the last decades and may allow even faster progress in the future. Following are some achievements of plant biotechnology that are relevant for IPM: (a) incorporation of insect resistance genes into commercial crop varieties; (b) the design of chemical and biological novel insecticides; (c) genetic modification of insect pests with lethal characters for genetic control or with beneficial traits to improve activity and efficacy of biological control agents; (d) rapid and reliable detection of insecticide resistance before genes responsible of resistance are widespread in the pest population and control fails in the field; and (e) identification of arthropod species and biotypes for pest diagnostics or trophic studies.

Host-Plant Resistance: Integrating GM Crops into IPM Systems

Host-plant resistance has been used in IPM to a rather limited degree (see above). Host resistance to herbivore insects is generally a quantitative trait that is difficult to manage and long to be incorporated into crop plants in conventional plant breeding programs. Techniques of genetic engineering have allowed some of those problems to be overcome by faster identification of insect resistance sources by means of molecular markers associated to resistance traits and by speeding up gene transfer from those sources to crop varieties to produce the genetically modified (GM) crops. The capacity to produce entomopathogenic toxins and digestive enzymes inhibitors have been the most used characters to confer insect resistance to crop plants by means of genetic engineering techniques. However, a varied array of other characters has also been successfully introduced into crop plants for insect pest control purposes. Insecticidal capacity of an entomopathogenic bacterium, *B. thuringiensis* (Bt) is caused by some of the toxins that the microorganism produces when it sporulates. Truncated genes expressing Bt toxins (dozens of Bt toxins and corresponding genes have been identified) have been transferred to several crop and forest plants to give the so-called Bt crops and plants. More than 35 million ha of maize and 15 million ha of cotton with Bt-expressing genes (alone or stacked with other transgenes) were grown in the world in 2009 [17] to control mainly Lepidopteran pests [17]. In Spain – the European country with the largest surface of Bt maize, the only GM crop allowed for cultivation in Europe – a survey conducted among more than 400 growers on the economic, social, and environmental impact of Bt maize found this: a mean increase of 12% in the gross margin in Bt growers, more than a 50% decrease in the number of insecticides applied due to the cultivation of Bt maize; this reduction was not very high because only a minority of growers used to spray against the pest targeted by the GM crop due to low efficacy [18].

In spite of the potential contribution of Bt crops to the sustainability of IPM through the selective control of key pests and the important savings of insecticide sprays, the deployment of GM crops has been very controversial in some areas like the European Union. Major risks concern potential development of resistance to Bt toxins in targeted pests and negative effects that Bt crops could have on nontarget organisms (NTO). The development of resistance to Bt toxins in targeted pests would cause a significant loss of an important tool for selective control of certain pests in the framework of IPM programs and organic agriculture and could drastically reduce the lifetime of Bt crops. Most countries that grow Bt crops have specific programs to monitor targeted pest populations for the

evolution of resistance and have implemented strategies to prevent resistance development. Until now, no Bt-resistance has been reported even in areas with almost 15 years of cultivation of Bt crops. More public attention has been paid to potential negative effects of GM crops on nontarget organisms. Most of the work conducted in this area has been devoted to Bt maize and practically no negative effects on biological control functions have been reported [19]. Consequently, Bt crops may prevent a substantial part of current insecticide usage and they can be integrated with biological control into more sustainable IPM programs.

Introduction of Traits into Insects

The introduction of deleterious or beneficial traits into insects has been achieved in several cases for varied purposes. Lethal gene inoculation into wild pest populations for genetic control may overcome some of the problems of conventional sterile insect techniques in which insects may be harmed when they are sterilized with irradiation or chemosterilization techniques and their competitive ability decreased. However, introduction of desirable traits for enhanced efficacy of biological control agents is a remarkable contribution of biotechnology to IPM. Unfortunately, most programs in that last respect have focused on insecticide-resistant genes that increase the possibility of using chemical insecticides and biological control in a compatible way. Despite this innovative approach, which may contribute to increasing the effective application of biological control, it also may lead to increased use of chemical insecticides.

Detection and Monitoring of Insecticide Resistance

Insects, like any other organism, may become resistant to insecticide active ingredients if they are submitted to frequent pressure of that ingredient. More than 600 insect species are nowadays resistant to one or more insecticides (http://www.pesticideresistance.org/search/1/). As fewer insecticide active ingredients are available in world agriculture, more is needed to increase the lifetime of registered substances and to implement strategies for resistance development prevention.

Several tactics to prevent resistance development in the pest population may be implemented and may succeed at least to delay the wide spread of resistance genes in the targeted population. Commonly, insecticide resistance in a pest is first detected when typical doses fail to control it and usually it is too late. Even at low frequency, earlier detection of the presence of resistance genes in the targeted population before they become common is crucial for the successful implementation of any antiresistance strategy application. As resistance is caused by a varied set of mechanisms, there are also several methods to detect resistance genes. Biochemical, immunological, and molecular methods are available now for the most common insecticides that allow screening a large amount of individuals for resistance gene presence. Improved comprehension of resistance mechanisms should lead to developing more specific, faster, and cheaper methods in order to monitor more populations at reasonable prices.

Identification of Arthropods

IPM needs the correct identification of insect species (or even biotype) in multiple situations: to apply a specific and selective method to control a certain species, to release a biological control agent that needs particular characteristics only available at biotype level, to detect quarantine pests in border inspections, and to study predator diet in trophic ecological studies. Morphological features that classically have been used for insect identification are frequently unknown or they are difficult to observe for nonspecialists. Biotechnological tools also may be valuable for various ecological studies that need markers for distinguishing target individuals from nontarget ones. Some of the current applications of biotechnology for identifying insects and their functions include: markers for dispersal measurements or to estimate insect densities by capture, release, and recapture of marked individuals, silencing genes for investigation of the function of certain proteins, invasion and spreading processes, and phylogenetic relationships between taxonomical groups.

Novel Bioinsecticides and Tension Actives

As described in microbial control, entomopathogenic microorganisms are used to control pests although

some factors linked with their costly production and narrow host range are limiting applicability. Entomopathogens may be genetically engineered to incorporate genes expressing foreign proteins with new insecticidal capacities, including larger host range, or proteins that negatively interfere with insect metabolism and physiology like insect hormones or juvenile hormone esterase that are involved in insect metamorphosis. A major concern about the bioengineered entomopathogens deals with the fate and permanence of foreign genes in the environment. Another promising line of biotechnological research in relation to IPM includes biosurfactants. A considerable amount of pest control agents is applied as sprays that need surfactants for correct application and spreading on target surfaces. Classical chemical surfactants are increasingly rejected by consumers and legislation due to their environmental impact and this stimulates the research on alternative biosurfactants. These are a group of heterogeneous secondary metabolites produced by a variety of microorganisms during their growth with significantly improved characteristics in comparison with homologous chemicals: those are biodegradable, have a lower per se toxicity, have more environment-friendly characteristics, require cheaper fermentative processes to produce them, are efficient at more variable conditions and at lower quantities, and have the potential of tailoring to suit specific applications. In addition to their activity as surface tension reducers and other chemical functions, biosurfactants have shown considerable biological antifungal and antiviral activity although in many cases their mode of action is poorly understood.

Implementation of IPM: Incentives and Constraints

More than 20 years ago, Wearing [20] wrote an article reporting results of a survey conducted among researchers and extensionists of three regions of world agriculture (Australia–New Zealand, Europe, and USA) on the IPM implementation process [20]. Still in 2010, many of the conclusions remain valid and following is a summary of those aspects of IPM implementation that enhance or constrain the adoption of IPM systems in western agriculture. There is general agreement among scientists about the accelerated progress of research on pest knowledge and control as well as the slow adoption of the new methods in practice. Knowing the key elements in the technology transfer and implementation of IPM may accelerate its application.

Among incentives to adopt IPM systems in Wearing's work, most surveys first perceived the cost advantage, followed by the development of pesticide resistance in local pest populations, the hazard to the grower from using pesticides, and environmental issues. The surveys also examined the major constraints for IPM implementation. The importance of obstacles for faster and wider application of IPM varied from one region to another. Whereas in USA over the 50% of respondents ranked social/market obstacles first, in Australia–New Zealand organizational obstacles were the first ranked, and technical obstacles were signaled in Europe as the least important to implement IPM.

In Europe much has been done in labeling agricultural products issued from IPM technology, particularly from regional administrations: IPM labels and integrated production (IP) labels have proliferated in the second half of the twentieth century in northern and southern Europe as well. The International Organization for Biological and Integrated Control (IOBC/WPRS) soon established the bases of IP and later developed guidelines in specific crop groups (see the organization Web site www.iobc-wprs.org). Food retailers and marketing food chains more recently have developed auditing procedures, in part inspirited in IPM principles which have pressed growers to progress more rapidly to more integrated production techniques. The impact of labeling and certification initiatives on IPM adoption has been quite variable in each country and commodity but the initiatives probably have contributed to publicize IP and IPM techniques among consumers.

A common question among policy makers is how to measure IPM adoption. Most likely, data on acquisition and use of tools for implementing IPM in the field (for instance, monitoring devices, meteorological receptors, users of warning systems and significant Web sites, and varied software for decision making among others) in combination with data related to pesticide usage (amount and economic value of pesticides sold,

commodity rejection because of high residue levels, inspection of pesticide residues in food trading, and significant habitat and ecosystem elements) may give an approximate idea of the progress made by the implementation of IPM. The number and size of companies (including small local firms) devoted to selling IPM products (e.g., natural enemies for biological control) may be another realistic way to monitor IPM adoption in world agriculture. Data on all these indicators are very disperse but perhaps they allow moderate optimism for faster IPM adoption in developed and emerging countries. Additionally, more strict legislation on pesticide use and the "disappearance" of most insecticides in the coming years may accelerate this process.

Future Directions

Most elements of IPM are among the factors that contribute greatly to the sustainability of agriculture. Beyond IPM, integrated control of insect pests, diseases, and weeds should progress toward the integration of all elements of agroecosystems to produce a balanced and harmonized growth of plants through true integrated production techniques. Exploitation of local natural resources and energy saving are common benefits of implementing novel integrated management systems. Additionally, most insecticides are being prohibited by western legislations. This should be the main focus of IPM progress to accelerate the transition of classical pest control based on chemicals to more integrated approaches.

Review of the vast literature on IPM confirms that success has come from a fundamental understanding of the processes acting in agroecosystems, rarely from a revolutionarily new control tactic. However, faster adoption of IPM strategies and tactics should also come from a more intense linkage among research, development, education, extension, and production. It is well documented that much of the scientific progress issued from R&D is not applied in practice or it takes too long to do it. Analysis of surveys conducted to identify the major incentives for the adoption of IPM systems by growers shows that an innovation is not adopted unless it contributes to producers' economic goals and meets the requisites for acceptance by the whole society.

Acknowledgments

Mr. Thomas Holland, MA from Stanford University in California, and now working at the Centro de Linguas at the Universidade da Coruña, proofread the English for this entry. His patience and professionalism has, I hope, helped ensure a comprehensible reading.

Bibliography

1. Oerke EC, Dehne HW (2004) Safeguarding production-losses in major crops and the role of crop protection. Crop Prot 23:275–285
2. Stern VM, Smith RF, van den Bosch R, Hagen KS (1959) The integrated control concept. Hilgardia 29:81–101
3. Kogan M (1998) Integrated pest management: historical perspectives and contemporary developments. Annu Rev Entomol 43:243–270
4. Higley LG, Pedigo LP (1996) Economic thresholds for integrated pest management. University of Nebraska Press, Lincoln
5. Klassen W (2008) Area-wide insect pest management. In: Capinera JL (ed) Encyclopaedia of entomology, vol 2. Springer, Dordrecht, pp 266–282
6. Mumford JD, Norton GA (1993) Expert systems. In: Norton GA, Mumford JD (eds) Decision tools for pest management. CAB International, Wallingford, pp 167–179
7. Southwood TRE, Henderson PA (2000) Ecological methods, 3rd edn. Blackwell Science, Oxford
8. Van Driesche RG, Bellows TS (1996) Biological control. Chapman & Hall, New York
9. DeBach P (1974) Biological control by natural enemies. Cambridge University Press, London
10. Bigler F, Babendreier D, Kuhlmann U (eds) (2006) Environmental impact of invertebrates for biological control of arthropods: methods and risk assessment. CAB International, Wallingford
11. Follett PA, Duan JJ (eds) (2000) Nontarget effects of biological control. Kluwer, Boston
12. Van Lenteren JC (ed) (2003) Quality control and production of biological control agents. CAB International, Wallingford
13. Millar JG (2007) Insect pheromones for integrated pest management: promise *versus* reality. Redia 90:51–55
14. Gould F, Schliekelman P (2004) Population genetics of autocidal control and strain replacement. Annu Rev Entomol 49:193–217
15. Cook SM, Khan ZR, Pickett JA (2007) Push and pull strategies for insect pest management. Annu Rev Entomol 52:375–400
16. Shelton AM, Bellinder RR (2007) Role of biotechnological advances in shaping the future of integrated pest management. In: Koul O, Cuperus GW (eds) Ecologically based integrated pest management. CAB International, Wallingford, pp 269–288

17. James C (2009) Global status of commercialized biotech/GM crops: 2009. ISAAA brief n. 41. Ithaca, NY
18. Gomez-Barbero M, Berbel J, Rodríguez-Cerezo E (2008) Bt corn in Spain – the performance of the EU's first GM crop. Nat Biotechnol 26:384–386
19. Romeis J et al (2008) Assessment of risk of insect-resistant transgenic crops to non target arthropods. Nat Biotechnol 26:203–208
20. Wearing CH (1988) Evaluating the IPM implementation process. Annu Rev Entomol 33:17–38

Intelligent Vehicles Technology, Introduction

Fei-Yue Wang
Institute of Automation, Chinese Academy of Sciences, Beijing, China

The basic goal of intelligent vehicle research is to provide a safe and comfortable ride for drivers and passengers. To reach this goal, an intelligent vehicle needs to dynamically sense the surrounding environment, receive the correct command from drivers, and make appropriate movements in real time.

In the 15 chapters in this book, 11 address how to sense the environment via vision sensors, 3 study how to monitor drivers' status and understand their commands, and 1 discusses vehicle motion control. The basis and content of these chapters are described below.

To obtain richer and more accurate visual information, researchers are trying to design more powerful vision sensors. One important direction combines different vision sensors to achieve better image quality. For example, how to obtain true color night vision video was studied in the chapter "► True Color Night Vision Video Systems in Intelligent Vehicles." The proposed night vision cameras sense the near-infrared radiation of the car's own headlights to extend the range of forward vision. The near-infrared and visible spectrum can then be combined into natural-looking color images for the driver to read.

Adjusting the focus of high-resolution video cameras and directing them to the regions of interest can provide better image data of certain objects. "► Active Multifocal Vision System, Adaptive Control of" presents methods for obtaining active pointing and gaze stabilization camera systems via an adaptive multi-focal vision system. Sliding mode control has been shown to be an effective control strategy to reach this goal.

Stereo vision sensors have become more popular because they can recover more three-dimensional information about nearby objects via stereo video data. In "► 3D Pose Estimation of Vehicles Using Stereo Camera," neighboring vehicles pose a detection problem and are discussed as a special application of the 3D stereo vision technique. The common characteristic points clustering problem is discussed and a model-based scene flow method is applied to determine the positions and instantaneous motion states of the vehicles.

"► Dynamic Environment Sensing Using an Intelligent Vehicle" studies the 3D environment reconstruction problem. Not only moving vehicles but also stationary objects (e.g., trees, walls, etc.) are recognized and modeled in such applications. Similarly, "► Driver Assistance Systems, Automatic Detection and Site Mapping" discusses how to reconstruct objects in the environment, particularly when driving in constructing sites.

LIDAR (Light Detection And Ranging) sensor is another frequently used vision sensor and often serves as a supplement to ordinary vision sensors. It can measure the distance between objects and the vehicle and provides depth information that cannot be directly collected via ordinary vision sensors. In "► Vehicle Detection, Tightly Coupled LIDAR and Computer Vision Integration for," a calibration algorithm was developed to map the geometry transformation collected via a multi-plane LIDAR system into camera vision data. Both simulations and real-world experiments have been carried out to show its reliability. "► Active Pedestrian Protection System, Scenario-Driven Search Method for" studies pedestrian detection problem via the fusion of LIDAR and camera systems.

Another type of approach combines camera vision systems and radar systems. Some related discussion is presented in "► Night Vision Pedestrian Warning in Intelligent Vehicles." The pedestrian detection problem is studied as an example to prove the merit of radar/camera fusion systems.

Since occlusion is generally unavoidable, the vision capability of a single vehicle is always limited. One straightforward way to solve this difficulty is to employ vision information that is collected from different vehicles in a small area. Clearly, this objective requires us to share the understandings of surrounding environments in an efficient way. As an early attempt, "► Cooperative Group of Vehicles and Dangerous Situations, Recognition of" discusses how to define the objects around several vehicles and share their position information.

The final chapter in vision sensory field, "► Driving under Reduced Visibility Conditions for Older Adults," discusses how to enhance the gathered vision information in reduced visible conditions and transform it sufficiently for older adults. This chapter can be viewed as an interdisciplinary study that crosses the boundaries of vision sensor system design and driver assistance systems.

The first chapter on driver assistance, "► Driver Assistance System, Biologically inspired," is hard to categorize. It discusses how drivers pay attention to certain changes in vision and how we can learn from drivers' response. It points out that advanced driver assistance systems may be designed to alert drivers in way that fit their biological characteristics. However, a detailed notification method is yet to be developed.

"► Driver Behavior at Intersections" provides a relatively mature approach that identifies drivers' turning features at intersections based on the sampled vehicle motion records. Following a stylized procedure, it first introduces how to gather the required driving data. Then a decision model of drivers' side turn behavior is proposed and the gathered driving data are used to calibrate the model. This chapter can be viewed as a representative of related reports.

The final chapter on driver assistance is "► Driver Characteristics Based on Driver Behavior." It presents a classical method to model driver characteristics, the driver behavior questionnaire. If well designed, a driver behavior questionnaire can tell us how drivers make their decisions under certain conditions. This method is not novel but may still useful to driver modeling.

The only chapter on vehicle motion control is "► Unscented Kalman filter in Intelligent Vehicles." It covers vehicle tire friction monitoring. Because a vehicle's motion is primarily determined by the friction forces transferred from the road via the tires, tire friction features receive attention. This chapter illustrates how researchers consider the connection between vehicle motion dynamics and tire friction features. It also presents application of the Unscented Kalman Filter (UKF) to realize tire feature monitoring.

There is an additional chapter related to intelligent vehicles. As mentioned above, sharing information between vehicles and infrastructures is a hot topic in recent studies. In "► Vehicular ad hoc networks, enhanced GPSR and Beacon-assist Geographic Forwarding in," utilization of real-time vehicle location information to accelerate vehicle ad hoc communication is discussed.

Internal Combustion Engines, Alternative Fuels for

Thomas Wallner[1], Scott A. Miers[2]
[1]Energy Systems Division, Argonne National Laboratory, Argonne, IL, USA
[2]Mechanical Engineering, Michigan Technological University, Houghton, MI, USA

Article Outline

Glossary

Advanced biofuel Renewable fuel not produced from food crops such as cellulosic biofuel and biomass-based diesel.

Alternative fuel A fuel that can serve or be used in place of another or an unconventional fuel choice.

Diesel gallon equivalent (DGE) The amount of alternative fuel required to match the energy content of 1 gal of diesel fuel.

Gasoline gallon equivalent (GGE) The amount of alternative fuel required to match the energy content of 1 gal of gasoline.

Mixture calorific value The amount of energy contained per volume of fresh charge typically at stoichiometric conditions that can be introduced into the cylinders of an internal combustion engine.

Renewable fuel A fuel created from resources that are never used up or can be replaced by new growth.

Definition of the Subject and Its Importance

Alternative fuels for internal combustion engines include a wide range of liquid and gaseous chemicals for both spark-ignition (SI) and compression-ignition (CI) applications. *Alternative* in this context describes a fuel that can be used in place of a conventional fuel such as gasoline or diesel. Alternative fuels are not equivalent but to a wide extent include renewable fuels since *renewable* only suggests a fuel can be created from resources that are never used up or can be replaced by new growth. The most common alternative fuels for SI engines are ethanol in blends with gasoline as well as compressed natural gas (CNG), liquefied natural gas (LNG), and liquefied petroleum gas (LPG). Biodiesel, synthetic diesel, green diesel, and DME are the most common alternative fuels for CI engines. In light of increasing energy demand worldwide, limited fossil fuel reserves partially located in politically unstable regions, growing environmental awareness and concern related to air pollutants as well as greenhouse gas emissions, alternative fuels have transitioned into social and political focus. Legislation and mandates were put in place specifying required amounts of alternative fuels in the fuel mix in Europe, Asia, and the United States. This article highlights some of the dominant alternative fuels, their market share, properties as well as their implication on engine and vehicle design, engine efficiency, and emissions.

Introduction

Although gasoline and diesel are nowadays considered conventional fuels, early engine concepts were designed to use other fuels. Combustion engines date back as early as 1807 with François Isaac de Rivaz of Switzerland and his combustion engine that used a mixture of hydrogen and oxygen for fuel [1]. The idea of alternative fuels is as old as engines themselves with Rudolf Diesel including considerations for use of gaseous, liquid and solid fuels in "Theory and Construction of a Rational Heat Motor" in 1894 [2]. The first mass produced vehicle, Ford Model T, was designed to run on ethanol.

Economic growth and prosperity are closely linked with availability and cost of energy. A large portion of worldwide energy demand and the majority of energy used for transportation are covered through fossil resources. The two most dominant fuels for transportation applications are gasoline in spark-ignition engines and diesel in compression-ignition engines. Announcements of newly discovered crude oil reserves and uncertainty on size of known reserves result in a large error band in prediction of the time span in which fossil fuels will be depleted. However, the fact that fossil fuels are finite is indisputable and with increasing worldwide energy demand the need for viable and sustainable alternative fuels becomes ever more pressing.

Today more than ever, society is concerned with the availability of energy which has become a matter of national security. The transportation sector alone accounted for 70% of the total US oil consumption in 2007. Foreign sources continue to outpace domestically produced fuel in the United States as well as Europe and Asia, leading to increased dependence on oil from politically unstable regions. Alternatives, either from domestic reserves or locally produced have been identified as a critical factor in ensuring national security and also as a means to stabilizing fuel prices in times of fluctuating crude oil prices.

As third-world countries develop and further growth of other countries continues, the amount of transportation-borne pollution increases. This development poses a threat due to local pollution resulting from emissions of regulated constituents and air toxics as well as the global impact of increased greenhouse gas emissions. Alternative fuels can play a major role in reducing regulated emissions through cleaner combustion, and renewable fuels in particular can help reduce greenhouse gas emissions.

There are several potential solutions to these issues from reduced consumption, increased production, and/or alternative sources of energy. It is this last

option that is the focus of this article and how alternative fuels and their properties impact engine as well as vehicle operation.

Limitations

This article is focused on technical implications of alternative fuels on engines and vehicles, their performance and emissions characteristics with particular focus on automotive applications. The characteristics of the most dominant alternative fuels are defined and compared to conventional gasoline and diesel fuel. The discussion on engine and fuel efficiency is strictly limited to the effects of alternative fuels on engine performance. Differences of alternative fuels in production pathways, production efficiency and challenges for their distribution are not included in this article. The interested reader is referred to [3] for detailed well-to-wheel analysis of alternative fuel pathways.

Before an attempt is made to classify alternative fuels it is important to define alternative fuels. The word alternative in the context of fuels suggests that a fuel can serve or be used in place of another and might also suggest an unconventional choice. However, alternative fuels are not to be confused with renewable fuels since renewable suggests that the resources used to create the fuel are never used up or can be replaced by new growth. Alternative fuels and renewable fuels are neither mutually exclusive nor mutually inclusive; although the vast majority of renewable fuels are also considered alternative fuels, by far not all alternative fuels are renewable. Methanol is just one example for a fuel that is renewable if produced, e.g., from biomass, and since not widely used also an alternative fuel. Ethanol is a renewable fuel if produced from corn or sugarcane, but since it is commonly used it may or may not be considered an alternative fuel in the sense of an unconventional choice. For our discussion ethanol is still considered an alternative since it is used in place of other fuels. Natural gas is just one example for an alternative fuel that, if fossil reserves are used, is not a renewable fuel.

Legislation

Recently introduced legislation worldwide mandates the increased use of alternative fuels. Directive 2009/28/EC of the European Parliament and of the Council of 23 April 2009 on the promotion of the use of energy from renewable sources and amending and subsequently repealing Directives 2001/77/EC and 2003/30/EC states the need to increase energy efficiency as part of the triple goal of the "20-20-20" initiative for 2020, which means a saving of 20% of the European Union's primary energy consumption and greenhouse gas emissions, as well as the inclusion of 20% of renewable energies in energy consumption.

Japan issued a New National Energy Strategy in 2007 targeting to expand the domestic biofuel production to cover approximately 10% of the fuel consumption by 2030 mainly through increased bioethanol and biodiesel utilization [4].

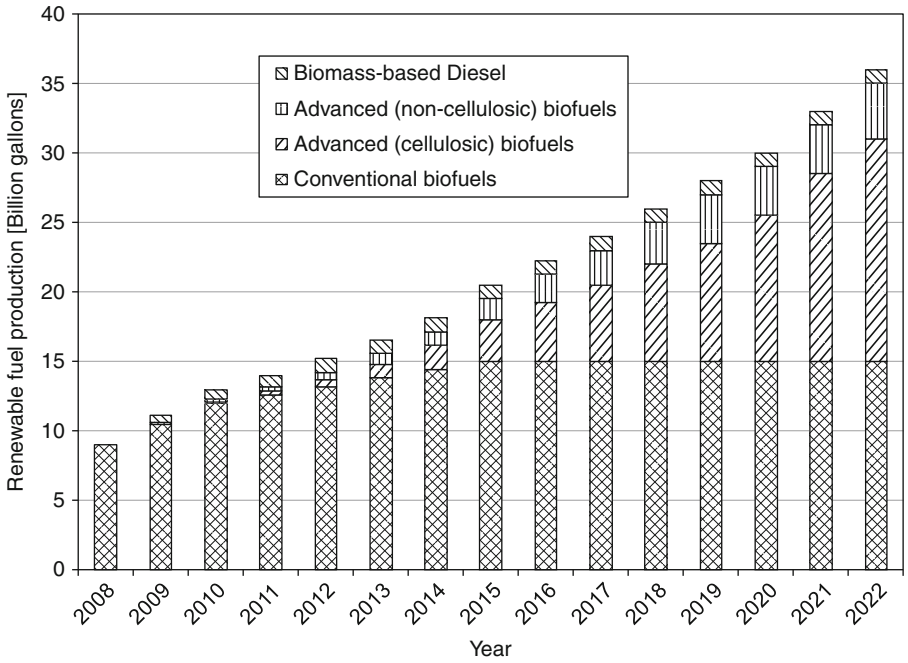
US legislation through the Renewable Fuel Standard (RFS) mandates a fourfold increase in renewable fuel production from approximately 9 billion gallons in 2008 to 36 billion gallons in 2022. The RFS also limits the amount of conventional biofuels to 15 billion gallons [5]. Figure 1 shows the mandated biofuels production in the United States which regulates contribution by conventional biofuels, advanced cellulosic and non-cellulosic and biomass-based diesel. Despite the rapid growth, US biofuels consumption remains a small contributor to US motor fuels, comprising about 4.3% of total transportation fuel consumption (on an energy-equivalent basis) in 2009 which is expected to grow to about 7% with the mandated 36 billion gallons by 2022 [6].

Overview of Alternative Fuels

Alternative fuels for internal combustion engines include fuels used in blends with conventional fuels such as ethanol which is used in low level gasoline/ethanol blends (e.g., E10) or biodiesel in low level blends with diesel (e.g., B5) as well as fuels requiring a dedicated fuel system design such as compressed natural gas (CNG) or liquefied petroleum gas (LPG). The following chapters focus on these mainstream alternative fuels, including their market share, properties, and impact on engine efficiency and emissions as well as other renewable fuel options with considerable market share either globally or in local markets.

Classification of Alternative Fuels

As mentioned earlier, different alternative fuels require more or less significant changes to vehicle fuel system



Internal Combustion Engines, Alternative Fuels for. Figure 1
Mandated biofuel production in the United States

design, engine calibration, and safety equipment. Some of these changes will be addressed in detail when discussing individual alternative fuels. In general, a classification can include dedicated vehicles, flexible fuel vehicles, and multi-fuel vehicles. Dedicated vehicles only operate on a single (alternative) fuel and are therefore typically purpose-built and optimized for this specific fuel. Dedicated vehicles are typically built for applications with fuels that cannot be stored in conventional fuel tanks, such as compressed gaseous fuels or liquefied fuels. On the other hand, blended operation describes usage of two or more fuels of the same type (liquid or gaseous) at varying mixture ratios. Vehicles capable of operating on varying blends of gasoline and ethanol are commonly referred to as FlexFuel vehicles. Finally, multi-fuel vehicles are defined here as vehicles that can operate on two or more fuels and feature multiple fuel storage and delivery systems.

According to their most common use, alternative fuels can also be classified as alternatives for spark-ignition engines or compression-ignition engines. This very generic classification is used for the following discussion although there are examples of alternative fuels that can be used in both applications. Gasoline is used as the baseline fuel for spark-ignition engine applications. For compression-ignition engine applications diesel is used as the reference.

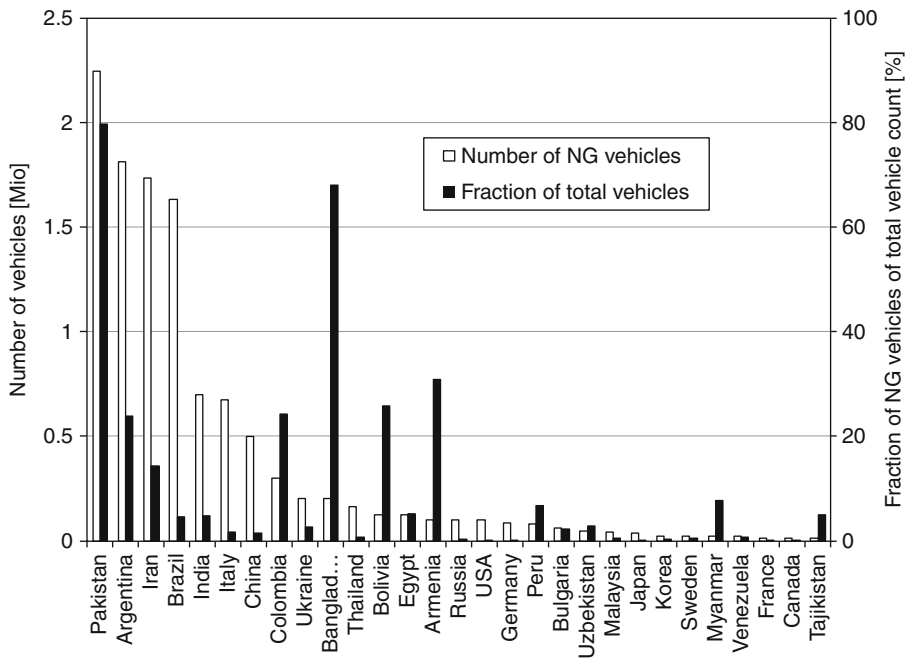
Market Share of Alternative Fuels

Worldwide approximately 5% of the automobile fleet are alternative fuel or advanced technology vehicles [7]. These vehicles can be grouped into dedicated vehicles that are designed and built for a certain type of fuel on the one hand and vehicles that are capable of operating on alternative or conventional fuels on the other hand. Total vehicle population is an estimated 884 million vehicles which includes all types of vehicles such as light-duty, medium-duty, and heavy-duty vehicles as well as busses.

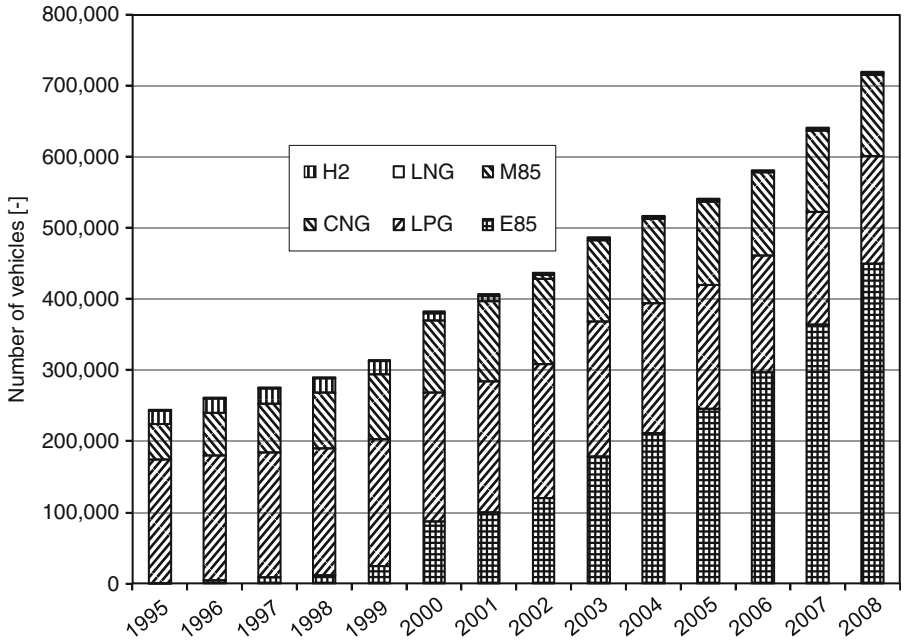
Considering all vehicle types CNG is the dominant alternative fuel for dedicated vehicles with a share of 1.27% which amounts to a total of more than 11 million NG vehicles worldwide. However, distribution and share of CNG vehicles is not uniform and strongly

depends on local fuel supplies. Figure 2 shows the absolute number of vehicles for all countries with a total CNG vehicle population in excess of 10,000 units. In addition the fraction of natural gas vehicles compared to the countries' total vehicle population is presented. The leading nations in terms of total CNG population are Pakistan (2.25 million units), Argentina (1.8 million units), Iran (1.7 million units), and Brazil (1.6 million units). Following India with approximately 700,000 units is Italy with 677,000 units as the largest European CNG vehicle fleet. The CNG vehicle fleet in the United States is approximately 100,000 vehicles, which accounts for only 0.04% of the entire vehicle population. The number of CNG vehicles available in the United States dropped from eight models in MY2004, to five in MY2005, and one in MY2006 [8].

In terms of market share of alternative fuel technology, CNG is only exceeded by ethanol and gasoline/ethanol flexible fuel (FlexFuel) vehicles. FlexFuel vehicles are vehicles that can run on gasoline and any gasoline/ethanol blend up to 85 vol % (E85). Since their introduction in 1979, more than 12.5 million ethanol and FlexFuel vehicles have been sold in Brazil [9]. In the United States, the FlexFuel fleet in 2009 accounted for approximately 8.35 million vehicles [10]. Although the number of FlexFuel vehicles exceeds that of natural gas–fueled vehicles, the total number of vehicles likely operated on ethanol is lower than that of CNG. In the United States, the number of vehicles actually operated on E85 is only 450,000 [11]. Figure 3 shows the development of alternative fuel vehicles in use in the United States since 1995. While LPG was the dominant alternative fuel in 1995 with more than 170,000 units, this number dropped to approximately 150,000 units in 2008. In the same time frame, the number of CNG vehicles increased from approximately 50,000 units in 1995 to more than 110,000 units in 2008. As mentioned earlier, the number of E85 vehicles that are actually fueled with gasoline/ethanol blends was around 450,000 units in 2008, still making it the dominant alternative fuel in the United States. Methanol as an alternative fuel was fairly popular in the 1990s with a peak of more than 21,000 units in 1997. Although the number of hydrogen powered vehicles has been steadily increasing since 2003, their market share with slightly over 300 vehicles in 2008 was less than 0.05% of all alternative fuel vehicles.



Internal Combustion Engines, Alternative Fuels for. Figure 2
NG vehicles worldwide (Based on data supplied via NGVA Europe and the GVR)



Internal Combustion Engines, Alternative Fuels for. Figure 3
Alternative fuel vehicles in use in the USA [11]

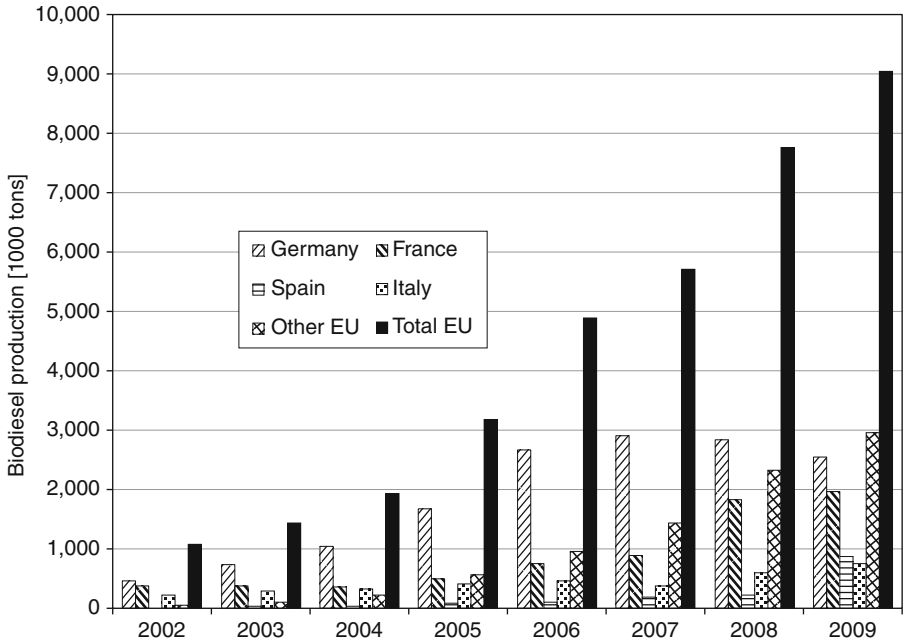
The United States consumed approximately 43 billion gallons of petroleum diesel fuel in 2005. Europe consumed 59.4 billion gallons, and Asia (excluding the middle east) consumed 48.5 billion gallons in 2005. In 2009, Europe produced the largest percentage of world-wide biodiesel followed by the United States and Asia Pacific. Figure 4 shows the distribution of biodiesel production for EU member states from 1998 to 2010 [12]. Overall biodiesel production has been declining in Europe since 2001, while it is increasing in the United States and Asia Pacific.

Since 2000, US production of biodiesel has increased by a factor of 350, as shown in Fig. 5. Government-imposed incentives combined with the requirements of the Renewable Fuel Standard have been the primary reasons behind the expansion of biodiesel production and associated utilization in the United States.

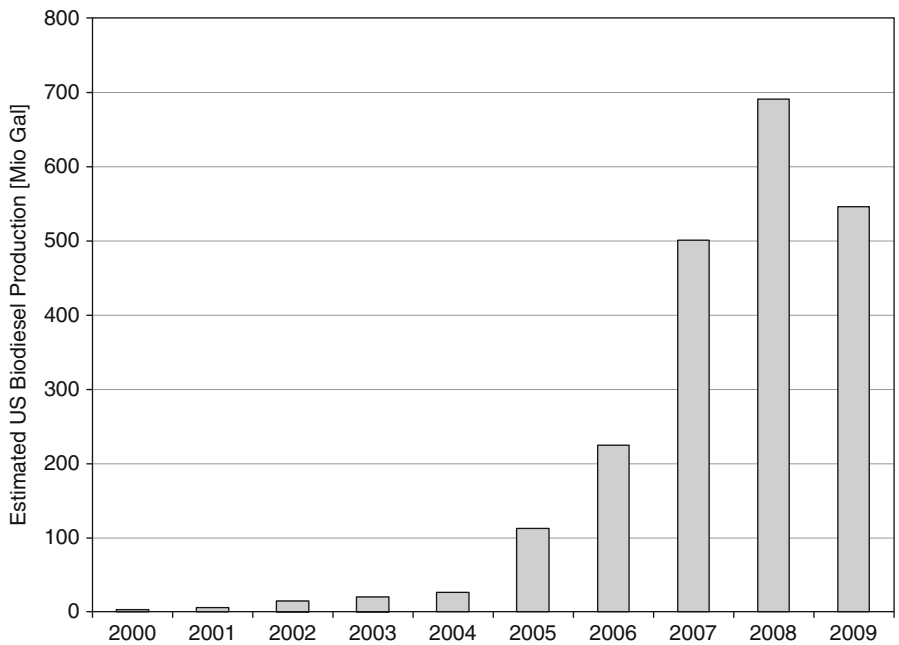
In 2009, the United States produced 17.7% of the world's biodiesel, making it the second largest producer behind Europe. The biodiesel market is expected to grow in the United States over the next several years and is predicted to reach ~6,500 million liters (1,717 million gallons) by 2020. Biodiesel is considered a biomass-based diesel source, and a minimum of 3.79 billion liters (1 billion gallons) must be produced by 2012. This is a 50% increase over 2010 production. The biodiesel market in Germany is not anticipating growth but rather a decline in production. Taxes have made the biodiesel too expensive for consumers combined with export subsidies for US producers. The European Union has imposed import tariffs on US biodiesel to promote and protect European production.

Fuel Properties

An understanding of alternative fuels and their impact on vehicle design as well as engine performance and emissions characteristics requires knowledge of fuel properties. For this purpose the fuel properties of the most common liquid as well as gaseous fuels are discussed in detail and compared to conventional fuels as a baseline. The alternative fuels are grouped into alcohol fuels including methanol, ethanol as well as *iso*-butanol as the most promising butanol isomer;



Internal Combustion Engines, Alternative Fuels for. Figure 4
EU biodiesel production [12, 13]



Internal Combustion Engines, Alternative Fuels for. Figure 5
Estimated US Biodiesel Production by year [14]

liquefied fuels including liquefied natural gas (LNG) and liquefied petroleum gas (LPG), and gaseous fuels including methane/CNG and hydrogen. Since these fuels are mainly used in spark-ignition (SI) engines, gasoline as the most common SI engine fuel is used as the baseline for comparison. Biodiesel and synthetic diesel fuel produced through the Fischer–Tropsch process, green diesel as well as dimethyl ether (DME) are compared to diesel fuel for CI engine applications.

In order to compare the energy content of alternative fuels to a conventional fuel, gasoline gallon equivalent (GGE) is introduced. GGE refers to the amount of alternative fuel required to match the energy content of 1 gal of gasoline. Although this measure is somewhat controversial due to the variability in the energy content of gasoline as well as alternative fuels [15], it is used here to simplify the discussion regarding energy content of alternative fuels. In a similar way, diesel gallon equivalent (DGE) is defined as the amount of alternative fuel required to match the energy content of 1 gal of diesel fuel. GGE is used for alternative fuels that are mainly used in spark-ignition engines and therefore displace gasoline whereas DGE is used for alternative fuels used in compression-ignition engines with the potential to substitute diesel fuel. The conversion factor between GGE and DGE is approximately 0.88.

Alcohol Fuels

Table 1 summarizes the fuel properties of the most common alcohol fuels and compares them to gasoline as a baseline. Characterized by a hydroxyl functional group (-OH) bound to a carbon atom the most common alcohols are methanol, with only a single carbon atom, ethanol (C2) as well as butanol, a four carbon alcohol. The negative impact of the oxygen in the hydroxyl group on the energy content decreases with increasing carbon count. While the lower heating value of methanol is less than 47% of gasoline, it increases with carbon count to approximately 63% for ethanol and 78% for butanol. Since the density of these fuels is in a similar range, the ratio of volumetric energy

Internal Combustion Engines, Alternative Fuels for. Table 1 Comparison of fuel properties – gasoline versus common alcohol fuels [16, 18, 19]

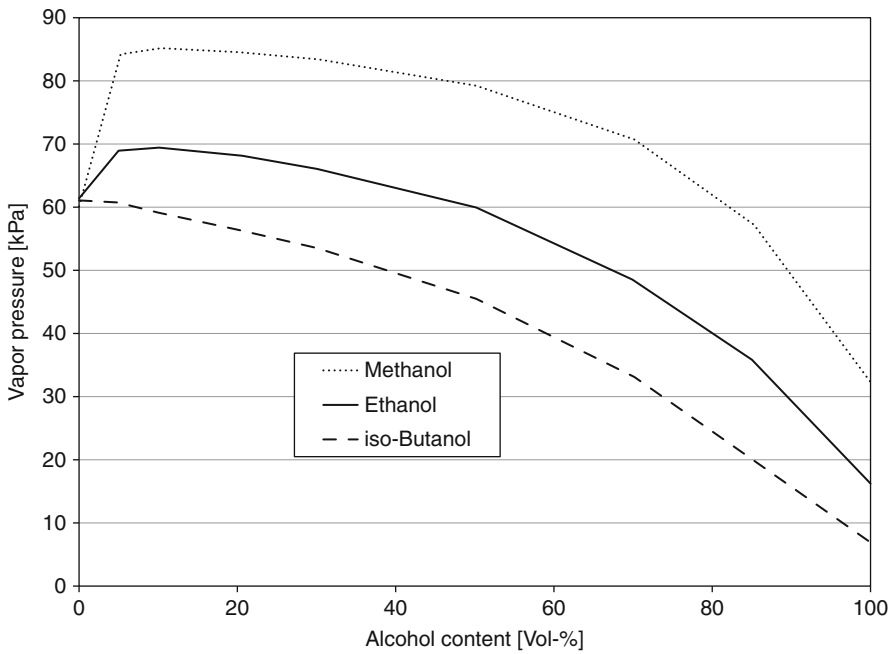
Parameter	Unit	Gasoline	Methanol	Ethanol	*iso*-Butanol
Chemical formula		$C_4 - C_{12}$	CH_3OH	C_2H_5OH	C_4H_9OH
Composition (C, H, O)	wt.%	86, 14, 0	37.5, 12.6, 49.9	52, 13, 35	65, 13.5, 21.5
Lower heating value	MJ/kg	42.7	20	26.9	33.1
Density	kg/m^3	720–780	792	789	805
Volumetric energy content	MJ/l	31.6	15.8	21.2	26.5
Vol. energy content relative to gasoline	%	100	50	67	84
GGE	gal	1	2.01	1.5	1.19
Research octane number (RON)	–	92–98	106–114	107–111	105
Motor octane number (MON)	–	80–90	91–95	89–94	91
AKI		90	100	100	98
Stoichiometric air/fuel ratio	–	14.7	6.45	9.0	11.2
Solubility in water @ 20°C	ml/100 ml H_2O	<0.1	Miscible	Miscible	7.6
Boiling temperature	°C	25–215	64.7	78.4	107–108
Flashpoint	°C	–43	12	17	28
Ignition limits	vol % λ	1.0–7.6 0.4–1.4	4–75	3.5–15	1.2–10.9
Latent heat of vaporization	MJ/kg	0.305	1.103	0.84	0.69

content is similar to the lower heating value with a significant advantage of gasoline compared to the alcohols in general and the short chain alcohols in particular. The lower volumetric energy content is also reflected in the GGE values for alcohol fuels. Approximately 2 gal of methanol contain the same energy as 1.5 gal of ethanol, 1.2 gal of butanol, or 1 gal of gasoline. A major advantage of these most common alcohol fuels compared to gasoline is their higher knock resistance expressed as octane number, which is around 100 for methanol as well as ethanol and 98 for *iso*-butanol compared to approximately 90 for gasoline. Although concerns have been voiced that laboratory engine Research and Motor octane rating procedures are not suitable for use with neat oxygenates [16], the presented values are still indicators to evaluate knock resistance. The difference in oxygen content is once again apparent when comparing the stoichiometric air demand. It increases from 6.45 for methanol to 9 for ethanol and 11.2 for butanol compared to 14.7 for gasoline. Finally, the solubility of ethanol in water is known to present a challenge for transporting gasoline/ethanol blends in pipelines [17].

A critical factor for liquid fuels is their evaporation characteristics. These characteristics affect engine behavior, in particular at cold start, as well as evaporative vehicle emissions. Vapor pressure curves for alcohol fuels as a function of blend level are shown in Fig. 6 based on data from Anderson et al. [20]. The vapor pressure of alcohols is lower than that of gasoline. When alcohols are mixed with gasoline, they form an azeotropic mixture with some components of the gasoline which results in an asymmetric relation between vapor pressure of the blended fuel and the volumetric blend ratio [21]. Evaporation of ethanol produces 4.6 times as much cooling of the charge as *iso*-octane, a major component of gasoline, while methanol produces 8.4 times as much cooling. Charge cooling reduces the tendency of the end gas to autoignite and therefore, further decreases the tendency for combustion knock [22].

Liquefied Fuels

Table 2 compares the properties of liquefied natural gas (LNG) and liquefied petroleum gas (LPG) to those of

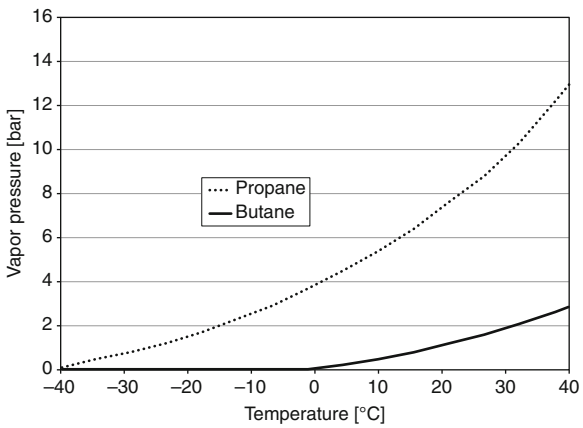


Internal Combustion Engines, Alternative Fuels for. Figure 6
Vapor pressure of alcohol fuels blends based on data from [20]

Internal Combustion Engines, Alternative Fuels for. Table 2 Comparison of fuel properties of gasoline with LNG and LPG [18]

Parameter	Unit	Gasoline	LNG	LPG (USA)
Chemical formula		$C_4 - C_{12}$	CH_4 (typ. > 95%)	C_3H_8 (maj) C_4H_{10} (min.)
Lower heating value	MJ/kg	42.7	46.7	46
Density	kg/m^3	715 – 765	430–470	508
Volumetric energy content	MJ/l	31.6	21	23.4
Vol. energy content relative to gasoline	%	100	66	74
GGE	gal	1	1.5	1.35
Octane number ((R + M)/2)	–	90	120	105
Stoichiometric air/fuel ratio	–	14.7	17.2	15.8
Boiling temperature	°C	25–215	–162	–42
Physical state for storage	–	Liquid	Cryogenic liquid	Pressurized liquid
Ignition limits	vol % λ	1.0–7.6 0.4–1.4	5.3–15 0.7–2.1	2.1–9.5 0.4–2

gasoline. Natural gas, after it has been cooled down to its liquid state at a temperature of –161°C, is referred to as liquefied natural gas (LNG). Since the liquefaction requires the removal of certain components that could form solids during the process, LNG typically contains approximately 95% methane [23]. Liquefied petroleum gas (LPG), also called autogas, is a mixture of propane (C_3H_8) and butane (C_4H_{10}) with changing composition depending on the region as well as seasonal differences. In the United States, Canada, and Japan, LPG consists mainly of propane, whereas the butane content in other areas of the world can be significant and typically increases for countries with warmer climate (butane content in Greece is 80%). The actual composition of LPG significantly affects its properties. Figure 7 shows the vapor pressure of propane and butane, the two main constituents of LPG as a function of temperature. Although the lower heating value of both LNG and LPG is higher than that of gasoline, the volumetric energy content of both fuels is significantly lower. This is due to the lower density of the two alternative fuels compared to gasoline and results in a volumetric energy content of LNG of only 66% and 74% for LPG compared to gasoline. The resulting GGE values of LNG and LPG are 1.5 and 1.35, respectively. The octane rating of LNG is approximately 120 and significantly higher than gasoline (90) and LPG with an octane rating of 105.



Internal Combustion Engines, Alternative Fuels for. Figure 7
Vapor pressure as a function of temperature for propane and butane

Gaseous Fuels

As mentioned earlier, LNG typically contains more than 95% methane (CH_4) since other impurities are removed for the liquefaction process. Although compressed natural gas (CNG) also mainly contains methane, its range of composition is wider than that of LNG and varies geographically. In addition to methane it

typically includes significant amounts of ethane, propane, and butane as well as other constituents in smaller amounts, as shown in Table 3.

Given that methane is by far the dominating constituent of CNG, the comparison of fuel properties of gaseous fuels shown in Table 4 uses the properties of methane as representative for CNG. Although not widely used for transportation applications at this point, the fuel properties of hydrogen as another promising alternative fuel are included in this comparison. The mass-specific lower heating value increases from gasoline to methane to hydrogen starting at 43.5–50 MJ/kg and 120 MJ/kg respectively. It is interesting to note that the lower heating value increases with increasing hydrogen content in the mass-specific fuel composition. However, since both methane and hydrogen are gaseous at ambient conditions with a density that is several orders of magnitude lower than that of gasoline, it takes 912 gal of methane or 3,075 gal of hydrogen to achieve the energy equivalent of 1 gal of gasoline. The comparison is more favorable on a weight basis, suggesting that 1 kg of hydrogen contains the same energy as 2.5 kg of methane and 2.86 kg of gasoline.

Internal Combustion Engines, Alternative Fuels for.

Table 3 Typical composition of natural gas [24]

Methane	CH_4	70–90%
Ethane	C_2H_6	0–20%
Propane	C_3H_8	
Butane	C_4H_{10}	
Carbon dioxide	CO_2	0–8%
Oxygen	O_2	0–0.2%
Nitrogen	N_2	0–5%
Hydrogen sulfide	H_2S	0–5%
Rare gases	A, He, Ne, Xe	Trace

Biodiesel and Synthetic Diesel

The average properties of petroleum diesel fuel and biodiesel are shown in Table 5. In addition, to accentuate the differences in properties between biodiesel

Internal Combustion Engines, Alternative Fuels for. Table 4 Comparison of fuel properties of gasoline and gaseous fuels (methane and hydrogen) [18, 19]

Parameter	Unit	Gasoline	Methane	Hydrogen
Chemical formula		$C_4 - C_{12}$	CH_4	H_2
Composition (C, H, O)	wt.%	86, 14, 0	75, 25, 0	0, 100, 0
Lower heating value	MJ/kg	43.5	50	120
Density gaseous liquid	kg/m^3	– 730–780	0.72 430–470	0.089 71
GGE (by volume)	gal	1	912	3075
GGE (by weight)	kg	2.86	2.5	1
Stoichiometric air demand	–	14.7	17.2	34.3
Boiling temperature	°C	25–215	–162	–253
Ignition limits	vol % λ	1.0–7.6 0.4–1.4	5–15 0.7–2.1	4–75 0.2–10
Minimum ignition energy	mJ	0.24	0.29	0.02
Self-ignition temperature	°C	approximately 350	595	585
Diffusion coefficient	mm/s	–	1.9×10^{-6}	8.5×10^{-6}
Quenching distance	Mm	2	2.03	0.64
Laminar flame speed	cm/s	40–80	40	200

Internal Combustion Engines, Alternative Fuels for. Table 5 Fuel properties for diesel, biodiesel, and synthetic diesel [25, 27]

Parameter	Unit	Diesel	Biodiesel	Synthetic diesel
Chemical formula		C_8–C_{25}	C_{12} – C_{22}	
Composition (C, H, O)	wt%	87, 13, 0	77, 12, 11	85–87, 13–15, 0
Lower heating value	MJ/kg	42–44	36–38	43.9
Density	kg/m^3	810–860	870–895	770–780
Volumetric energy content	MJ/l	34–37.8	31–34	33.8–34.2
Vol. energy content relative to diesel	%	100	93	94.4–95.5
DGE	gal	1	1.074	1.05–1.06
Cetane number	–	40–55	45–65	73–80
Viscosity @ 40°C	mm^2/s	2.8–5.0	2–6	2.0–3.5
Flash point	°C	60–78	100–170	59–109
Cloud point	°C	–24 – –10	–5–5	–19–0
Stoichiometric air/fuel ratio	–	14.7	13.8	
Boiling range/temperature	°C	180–340	315–350	160–350
Ignition limits	vol %	0.6–6.5	0.3–10	
Lubricity		Baseline	Higher	Lower

and other alternative diesel fuels, the properties of a typical synthetic diesel fuel produced through the Fischer–Tropsch process are included in Table 5 as well.

Diesel fuel properties are controlled by ASTM D975. Hundred percent biodiesel has been specified as an alternative fuel to diesel and has its own standard, ASTM 6751. For blends between 6% and 20% biodiesel, a third standard is used, ASTM D7467. Before neat biodiesel can be blended with petroleum diesel, both fuels must meet their respective ASTM standards and then be analyzed as a blend using D7467 if the biodiesel content is between 6 and 20 vol %.

For example, if the cetane number falls too low, it can lead to increased ignition delay, combustion pressure, and noise, along with increased smoke during cold start.

Increased viscosity has been shown to increase fuel spray penetration resulting in cylinder wall wetting and impingement. Biodiesel has a slightly higher viscosity than petroleum diesel, while synthetic diesel is similar to petroleum diesel. In addition, the atomization of the fuel is reduced as fuel viscosity is increased. Too low of viscosity will increase pump wear due to reduced lubricity. Biodiesel has been utilized as a lubrication additive recently.

Sulfur content has previously been utilized for lubrication in fuel injection systems but increased emissions controls have forced the reduction of sulfur in diesel fuel. The sulfur can be converted into sulfur dioxide and sulfur trioxide, poisoning the catalyst and sensors in the exhaust stream. In addition, small droplets of sulfuric acid and other sulfates can become nucleation sites for particulate matter emissions. In the United States, the sulfur content of on-road diesel fuel was mandated to not exceed 15 ppm sulfur as of 2006. Biodiesel and synthetic diesel do not contain sulfur.

The cloud point is the temperature at which crystals begin to form in the fuel and prevent the flow of fuel through filters and pumps. A significant challenge for biodiesel fuels is overcoming the significant increase in cloud point temperature compared to petroleum diesel fuel. Additives, tank heaters, and various blends of petroleum and biodiesel have all been experimented with to improve the cold-weather operation of biodiesel.

Synthetic diesel fuel whether produced from biomass (BTL), coal (CTL), or natural gas (GTL) does not vary widely in properties based on the feedstock due to the consistent processing techniques employed. If the Fischer–Tropsch process is employed, the feedstock is first converted to a synthesis gas (syngas) consisting of hydrogen and carbon monoxide. This process involves combustion of the feedstock in a reduced oxygen environment. Once the hydrogen and carbon monoxide are produced, the effect of the feedstock composition is removed from the final fuel properties. The remaining process to produce liquid fuel (Fischer–Tropsch) basically receives the same input regardless of feedstock. The other method of producing synthetic diesel fuel is to first convert the biomass to bio-oil through pyrolysis (zero oxygen environment) and then convert the bio-oil to fuel. Again, the pyrolysis step removes the effect of feedstock composition, thus producing a consistent fuel. Both processes produce a high quality alternative diesel fuel with high cetane, good cold-flow and oxidative stability properties, and ultra low sulfur content. The Fischer–Tropsch process was first commercialized by Sasol in South Africa. They produce over 165,000 barrels/day of transportation fuel from coal. However, the process is on the order of three times more expensive to produce fuel than conventional petroleum production processes [28].

Green Diesel and DME

Hydrogenated-derived renewable diesel fuel or "green" diesel is an alternative compression-ignition engine fuel that is produced by utilizing the conventional processing method for petroleum diesel fuel, fractional distillation. The process is feedstock tolerant and unlike transesterification, the properties of the fuel do not change significantly when the feedstock varies from vegetable oil to waste grease. However, there are significant differences in properties compared to biodiesel and even petroleum diesel. Green diesel has a higher cetane number, typically 75–90 compared to 40–50 for petroleum diesel. The pour point and energy content can be nearly the same as petroleum diesel, alleviating the negative aspects of biodiesel. Green diesel is not oxygenated so the oxidative stability is very good. The process of producing green diesel is more energy intensive and requires higher temperatures and pressures compared to transesterification. The majority of research is currently focused on increasing the output of fuel and scaling up production facilities.

Dimethyl ether (DME) is a two-carbon, oxygenated molecule that has unique properties in the liquid state as an alternative diesel fuel. DME is primarily produced by converting hydrocarbons such as natural gas to synthesis gas (syngas). The syngas is then converted to methanol in the presence of a copper-based catalyst. An additional dehydration process of methanol using a silica-alumina catalyst results in the production of DME. This two-step process is the most widely utilized currently, but a one-step production process is currently being researched. Because DME is in the vapor state at standard temperature and pressure, it must be compressed, like liquid propane, and stored in tanks for transportation and consumption. Table 6 shows some of the primary differences between green diesel and DME compared to petroleum diesel fuel.

Effect of Feedstock Composition on Biodiesel Properties Unlike ethanol whose properties are relatively unaffected by the feedstock due to the processing technique employed (distillation), biodiesel properties have been shown to vary widely depending on the feedstock properties (fatty acid profile) and catalyst employed (typically methanol or ethanol) during the fuel production process.

The fatty acid profile of the feedstock directly influences the properties of the biodiesel. The primary features of interest for the fatty acid include chain length, degree of unsaturation, and branching of the chain. The properties of the biodiesel that are most directly impacted by changes in the fatty acid profile include cetane number, heat of combustion, cold-flow, oxidative stability, viscosity, and lubricity. Table 7 shows the fatty acid profiles for the primary feedstock options in biodiesel and the wt% of each fatty acid contained in the feedstock. Note that C18:1 refers to an 18 carbon chain molecule with one double bond. Increasing numbers of double bonds reduce the saturation of the molecule. If no double bonds exist, the molecule is said to be fully saturated.

It is well known that cetane number decreases with decreasing chain length and increasing branching. As well, oxidative stability increases with degree of saturation. Table 8 shows the range of cetane number,

Internal Combustion Engines, Alternative Fuels for. Table 6 Green diesel and DME fuel properties compared to petroleum diesel [26–29]

	Unit	Petroleum diesel	Green diesel	DME
Chemical formula		C_8–C_{25}		C_2H_6O
Lower heating value	MJ/kg	42–44	44	28
Density (+20°C)	kg/m^3	810–860	780	675
Volumetric energy content	MJ/l	34–37.8	34.3	18.9
Vol. energy content relative to diesel	%	100	94	52
DGE	–	1	1.06	1.92
Cetane number	–	40–55	98–99	>65
Viscosity @ 40°C	mm^2/s	2.8–5.0	3.0–3.5	0.21
Flash point	°C	52		–41
Auto ignition temperature	°C	250		350
Cloud point	°C	–24 – –10	–15	n/a
Lubricity		Baseline	Similar to baseline	Poor
Stoichiometric air/fuel ratio	–	14.7		8.99
Ignition limits	vol %	0.6–6.5		3.4–28

Internal Combustion Engines, Alternative Fuels for. Table 7 Fatty acid profile for the most common biodiesel feedstocks [30, 31]

Vegetable oil	C16:0	C18:0	C18:1	C18:2	C18:3
Castor	1	1	3	4	
Coconut	7.5–10.5	1–3.5	5–8	1–2.6	
Cottonseed	22–26	2–3	15–22	47–58	
Palm	40–47	3–6	36–44	6–12	
Peanut	6–14	2–6	36.4–67.1	13–43	
Rapeseed (Canola)	2–6	4–6	52–65	18–25	10–11
Soybean	10–12	3–5	18–26	49–57	6–9
Sunflower	5–7	3–6	14–40	48–74	

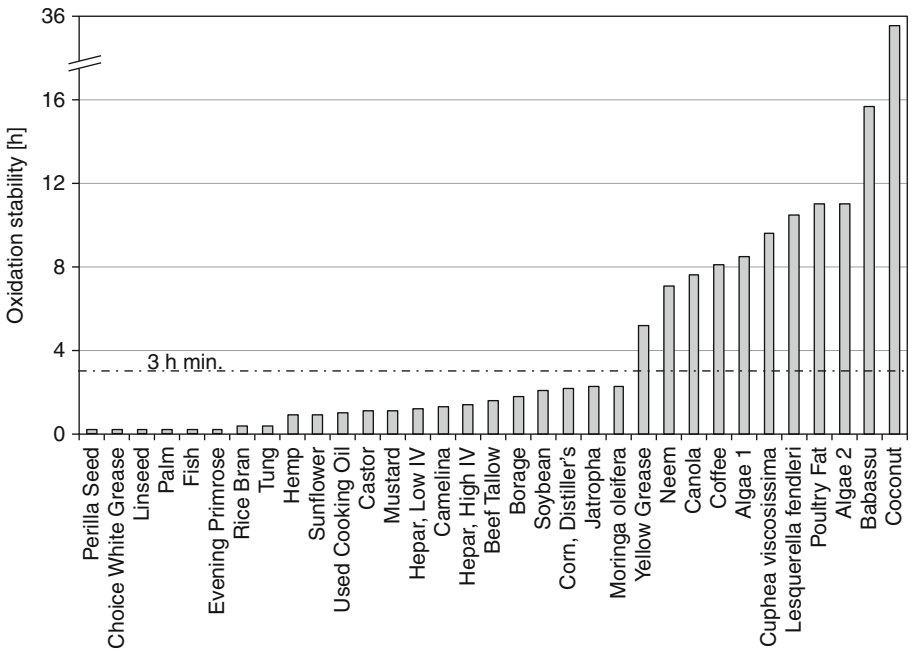
heat of combustion, cloud point or cold filter plugging point (CFPP), and kinematic viscosity for several biodiesel fuels.

Figure 8 shows the effect of 33 different feedstocks on fuel oxidative stability, measured in hours. This data shows that as the number of double bonds increases (increased degree of unsaturation), the oxidative stability decreases significantly. A test of iodine number of the biodiesel provides a measure of the degree of unsaturation of the fuel.

One of the significant advantages of biodiesel is the ability to improve the lubricity of the fuel with only a small quantity of ester. As shown in Fig. 9, the lubricity improves significantly for the first 1% of methyl ester added to the baseline diesel fuel. Lubricity is measured using the ASTM D6078 Scuffing Load

Internal Combustion Engines, Alternative Fuels for. Table 8 Fuel properties of biodiesel fuels from various feedstocks [31]

Vegetable oil	Cetane number	Heat of combustion [KJ/kg]	Cloud point or CFPP [°C]	Kinematic viscosity [40°C, mm^2/s]
Coconut ethyl	67.4	38,158	5	3.08
Cottonseed	51.2		–5 (pour point)	6.8 (21°C)
Palm ethyl	56.2	39,070	8	4.5 (37.8°C)
Rapeseed (Canola)	48–56	37,300–39,870	–3; CFPP –6	4.53
Soybean	48–56	39,720–40,080	from –2 to 3	4.0– 4.3
Sunflower	54–58	38,100–38,472	0–1.5	4.39



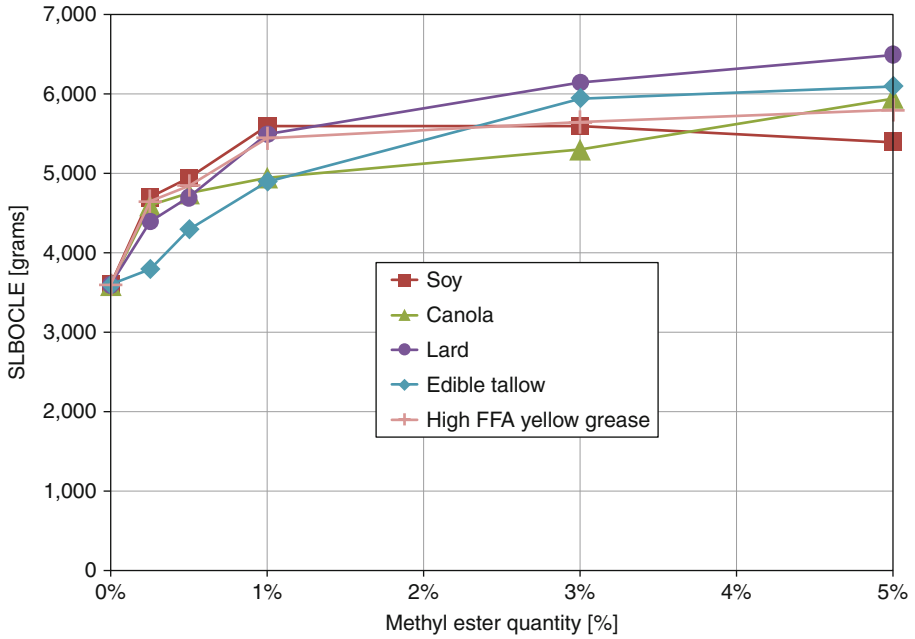
Internal Combustion Engines, Alternative Fuels for. Figure 8
Dependence of biodiesel oxidative stability on feedstock [32]

Ball-on-Cylinder Lubricity Evaluator (SLBOCLE) test rig. The higher the value, the higher the load before scuffing occurs which translates to a fuel with improved lubricity.

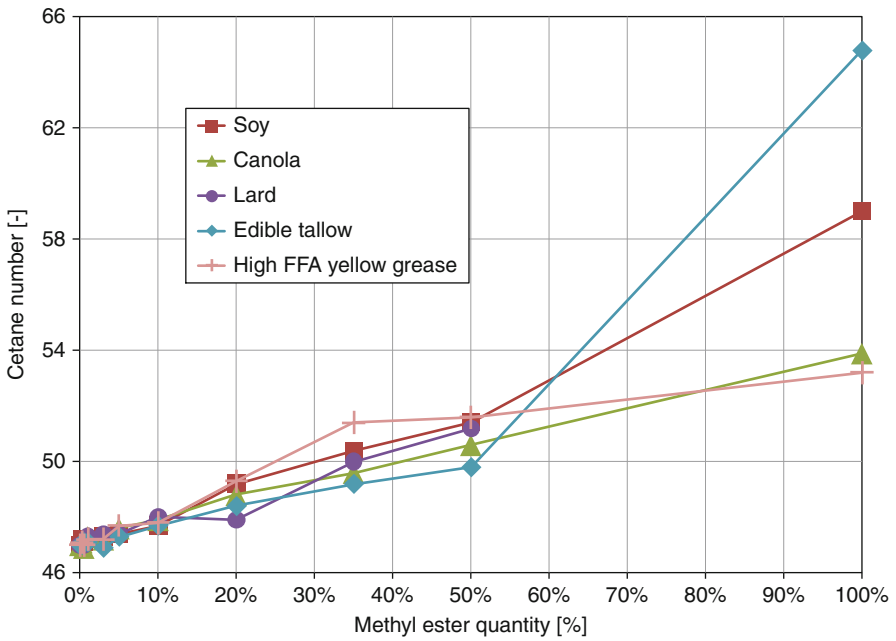
The impact of biodiesel on cetane number can be clearly identified in Fig. 10. As a general rule regardless of feedstock, increasing quantity of biodiesel volume increases the cetane number and improves the overall combustion quality. The data shown in Fig. 10 utilized a baseline petroleum diesel fuel meeting ASTM D975 specifications.

Comparison of Fuels Based on Mixture Calorific Value

In addition to the fuel properties highlighted in Tables 1–9, the mixture calorific value for different fuels and mixture formation strategies can be calculated



Internal Combustion Engines, Alternative Fuels for. Figure 9
Lubricity improvement with methyl ester quantity [33]



Internal Combustion Engines, Alternative Fuels for. Figure 10
Effect of methyl ester on cetane number [33]

Internal Combustion Engines, Alternative Fuels for. Table 9 Comparison of methyl ester and ethyl ester properties with a soybean feedstock

Ester	Acid No.	Iodine No.	Peroxide No.	Glycerol- free/ Glycerol-bound	Water/ Sediment	Cetane No.	Density g/cm^3	Oxygen wt%	Kinematic viscosity	
									40°C	100°C
Methyl soy	0.15	121	340	0.007/0.223	0	52.3	0.8836	11.44	4.03	1.64
Ethyl soy	3.02	122	123	0.003/0.031	0	47.3	0.8817	11.55	4.33	1.74

and presents a measure for the volume-specific energy content of fuel/air mixtures. It specifies the amount of energy contained per volume of fresh charge typically at stoichiometric conditions that can be introduced into the cylinders of an internal combustion engine. The mixture calorific value is not only used to describe the fuel properties of a specific fuel and mixture formation strategy, but also allows estimation of the theoretical power density of a specific alternative fuel. The brake mean effective pressure (BMEP) or engine torque can be calculated based on the volumetric efficiency (VE) of a specific engine, the brake thermal efficiency (BTE), and the mixture calorific value (H_G). Under the (theoretical) assumption that volumetric efficiency and brake thermal efficiency remain unchanged when switching from one fuel to another, any change in mixture calorific value is directly reflected in corresponding change in (maximum) brake mean effective pressure, in other words, maximum torque.

$$\mathrm{BMEP} = \mathrm{VE} \times \mathrm{BTE} \times \mathrm{H_G} \tag{1}$$

For mixture-aspirating engines (carbureted or port injected/external mixture formation) the mixture calorific value is defined relative to 1 m^3 of mixture and for air-aspirating engines (direct injection/internal mixture formation) to 1 m^3 of air. The mixture calorific value for mixture-aspirating operation is calculated based on lower heating value (LHV), density of the fuel/air mixture ($\rho_{Mixture}$), relative air/fuel ratio (λ), and stoichiometric air demand (L_{St}) and is defined as:

$$\mathrm{H_G} = \frac{\mathrm{LHV}\rho_{\mathrm{Mixture}}}{\lambda \mathrm{L_{St}} + 1} \tag{2}$$

For air-aspirating operation the density of air (ρ_{Air}) is used instead of the fuel/air mixture density and the value is defined as:

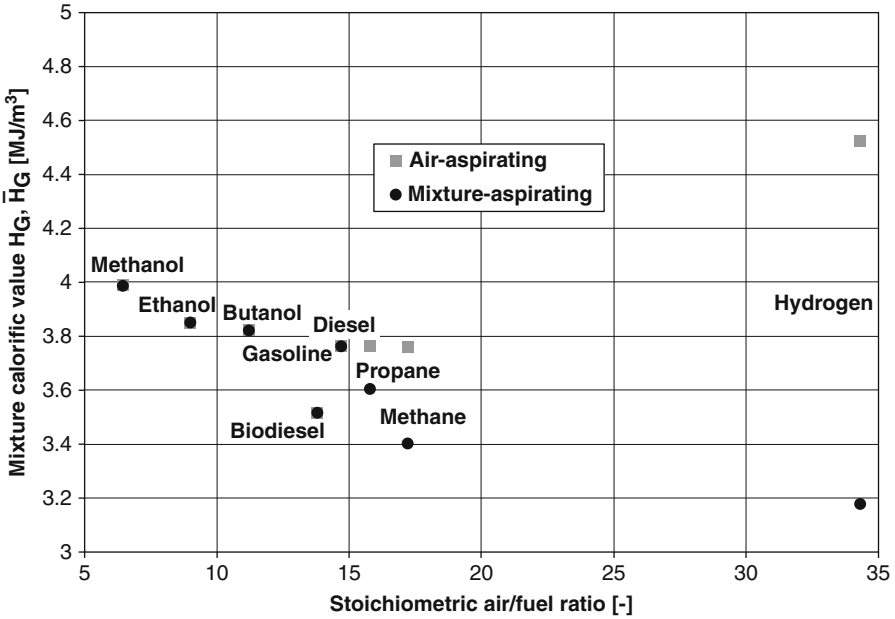
$$\bar{H}_G = \frac{\mathrm{LHV}\rho_{\mathrm{Air}}}{\lambda \mathrm{L_{St}}} \tag{3}$$

Figure 11 shows the mixture calorific value for various alternative fuels both for air-aspirating and mixture-aspirating operation as a function of stoichiometric air/fuel ratio. To achieve high power density, the mixture calorific value should be as high as possible. With liquid fuels the difference in the mixture calorific value between air- and mixture-aspirating operation is minimal since the liquid fuel due to its high density takes up only a small volume. However, for gaseous fuels the difference between the mixture calorific value in air-aspirating and mixture-aspirating operation increases significantly with decreasing gas density ($\rho_{Propane}$ = 1.83 kg/m^3, $\rho_{Methane}$ = 0.72 kg/m^3, $\rho_{Hydrogen}$ = 0.089 kg/m^3). The impact is most pronounced for hydrogen as a fuel, because it displaces approximately 30% of the aspirated air at stoichiometric conditions with external mixture formation. Even for liquid fuels there is noticeable difference in mixture calorific values that should be considered when comparing alternatives.

Alternative Fuels for Spark-Ignition (SI) Engines

Alcohol Fuels (Ethanol, Butanol, Methanol)

Methanol As per the Energy Policy Act of 1992, Methanol (CH_3OH), also known as wood alcohol, is considered an alternative fuel. The main production routes for methanol all use natural gas as a feedstock. Methanol can be used to make methyl tertiary-butyl ether (MTBE), an oxygenate that was used in blends with gasoline as an octane booster. MTBE production and use has declined in recent years because it has been found to contaminate ground water. As an engine



Internal Combustion Engines, Alternative Fuels for. Figure 11
Mixture calorific value for conventional and alternative fuel options

fuel, methanol has similar chemical and physical characteristics to ethanol.

A limited study of low level methanol gasoline blends up to 15 vol % on a four-stroke, single-cylinder, variable compression ratio engine with a displacement volume of 582 cm^3 suggests an approximately 10% increase in torque at the highest blend level which is attributed to the improved volumetric efficiency due to the higher heat of vaporization [34]. An experimental evaluation of engine cold start behavior with gasoline as well as 10 vol % and 30 vol % blends of methanol in gasoline on a small three-cylinder engine came to the conclusion that the addition of methanol improves combustion stability, indicated mean effective pressure (torque), and reduces misfires. In addition a 70% reduction in CO emissions as well as a 40% reduction in HC emissions were also observed. These positive effects of methanol addition were attributed to higher vapor pressure, lower boiling temperature as well as the high oxygen content of methanol [35]. A detailed thermodynamic evaluation of alcohol fuels including their effect on engine efficiency and NO_x emissions concluded that the use of methanol could result in increase in engine brake thermal efficiency of up to 1% regardless of engine load and speed. This increase was attributed to slightly reduced pumping losses because of the increased amount of fuel vapor at constant load compared to the *iso*-octane baseline. The study further found that a 25% NO_x emissions reduction can be expected with methanol compared to the baseline fuel which is mainly caused by reduced process temperatures [36]. The high octane rating of methanol also makes it well suited for knock-free operation at high compression ratios. A study that was performed on single-cylinder, four-stroke, naturally aspirated, high-compression direct-injection stratified charge spark-ignition engine revealed the potential of methanol for high engine efficiencies. Performing optimization of injection timing and spark timing at the research engine with a compression ratio of 16:1 at full load conditions at 1,600 RPM resulted in a maximum indicated efficiency of approximately 51% [37].

Due to challenges with methanol such as toxicity as well as safety related issues due to invisible methanol flames widespread use of methanol as a transportation fuel is rather unlikely.

Ethanol Ethanol is probably the most widely used liquid alternative fuel and is typically used in blends

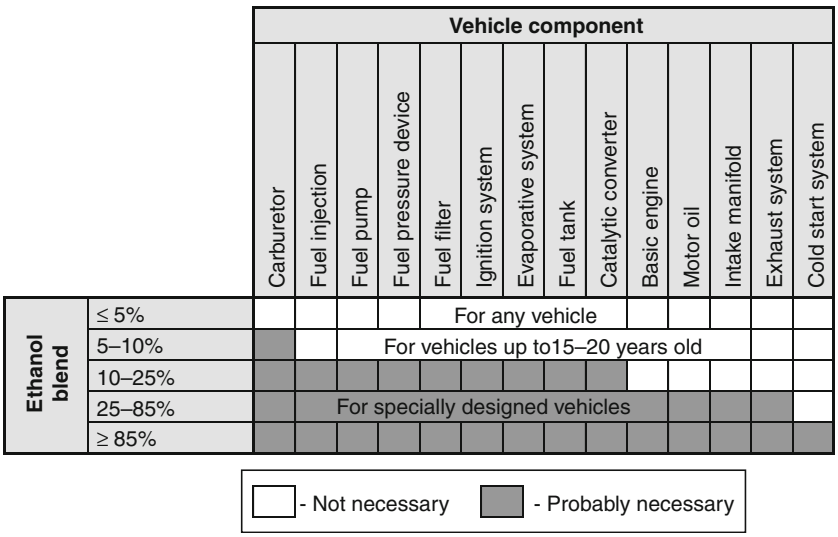
with gasoline. The two main variables when using gasoline/ethanol blends are the blend level and whether hydrous or anhydrous ethanol is used. Typically gasoline is blended with anhydrous ethanol whereas neat ethanol when used in vehicles is hydrous (E100). Although attempts are made to get higher blends approved, the most commonly used blend throughout the world is E10, a blend of 90 vol % gasoline with 10 vol % anhydrous ethanol. US legislation limits the amount of oxygen content in blends of gasoline with alcohol fuels. In 1991, the maximum oxygen content was increased from 2 wt.% to 2.7 wt.% for blends of aliphatic alcohols and/or ethers excluding methanol [38]. To ensure sufficient gasoline base was available for ethanol blending, the EPA also ruled that gasoline containing up to 2 vol % of MTBE could subsequently be blended with 10 vol % of ethanol [39]. Only recently EPA granted a partial waiver to allow gasoline that contains greater than 10 vol % ethanol and up to 15 vol % ethanol (E15) for use in model year 2007 and newer light-duty motor vehicles, which includes passenger cars, light-duty trucks, and sport utility vehicles (SUV) [40].

Several markets also promote higher ethanol blends, with E85 being the most prominent one with existing distribution networks in several countries including the United States and Sweden.

Due to its large domestic production capabilities of ethanol from sugarcane Brazil has a special role in the worldwide ethanol market with a distribution network for both, Type C gasoline as well as neat hydrous ethanol (~5 vol % water content). Type C Gasoline describes a 25 ± 1 vol % anhydrous ethanol blend with gasoline [41] and is referred to as gasohol in Brazil.

A critical aspect potentially limiting the maximum blend level of any alcohol fuel with gasoline is the vapor pressure of the fuel blend since it is critical for cold startability. Figure 6 shows the vapor pressure for several alcohol fuels as a function of volumetric blend level. Vapor pressures below 45 kPa [42] are known to potentially cause cold start problems which effectively limits the maximum ethanol blend level without taking additional measures to approximately 75% during winter months. Other limitations when increasing the amount of ethanol in the fuel might be inferred due to incompatibility of the blend with metals, plastics, and elastomers in the fuel system. Despite these direct effects, higher ethanol blends potentially also limit the use of existing fuel infrastructure, in particular pipelines [17]; however, these limitations are beyond the focus of this article.

A more detailed overview of vehicle components that are affected by ethanol addition as a function of blend level are shown in Fig. 12 [43]. Blend levels below

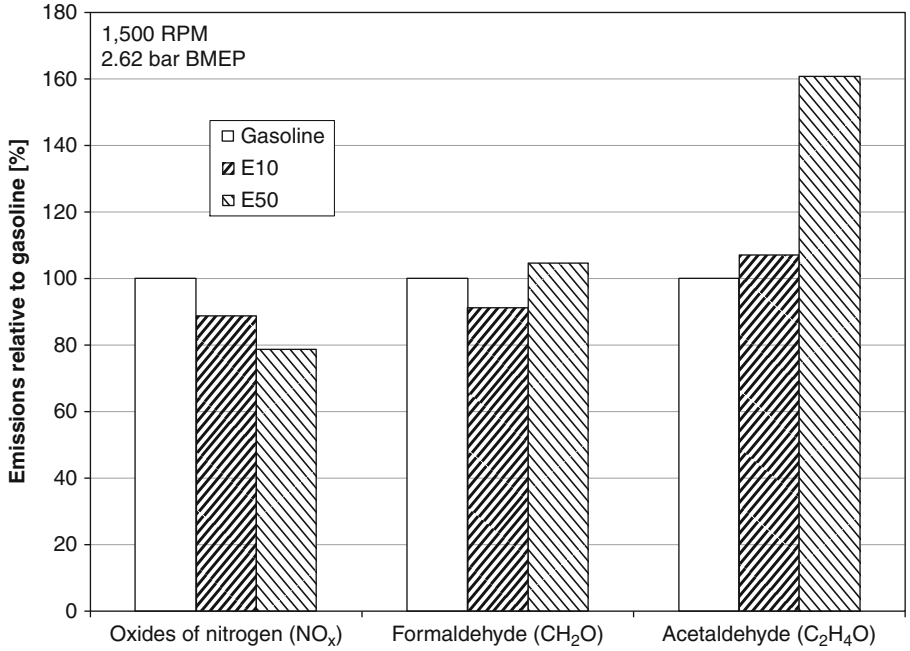


Internal Combustion Engines, Alternative Fuels for. Figure 12
Required vehicle modifications with ethanol blend level (Based on [43])

5 vol % do not require any changes to the vehicle regardless of vehicle age. Typically, modifications are also not required at blend levels up to 10 vol % for most vehicles less than 15–20 years of age. Only older vehicles might require changes to the engine carburetors. Although studies suggest that most fuel system components are not negatively affected at a blend level of 20 vol % [44–46], modifications to the fuel system including injectors, pumps, regulators and filters, the ignition system as well as the catalytic converters are common for blend levels up to 25 vol %. At higher blend levels up to 85 vol % additional modifications include the base engine, engine oil as well as intake and exhaust system. Only for blend ratios beyond 85 vol % and neat ethanol, a separate cold start system might be required.

As mentioned earlier, specific engine designs are typically only found for engines that are intended to run on higher ethanol blends in excess of 25 vol %. Experimental results with port fuel injection on a 1.8 L four-cylinder engine converted to single cylinder operation revealed the efficiency and emissions potential of high ethanol blends. In particular at moderate engine speed of 1,500 RPM torque could be improved significantly due to the increased knock resistance resulting in a 20% increase with E100 compared to gasoline, while improving indicated thermal efficiency from 32% to 40%. At the same time a reduction in oxides of nitrogen emissions by approximately 125 ppm together with a 2°C reduction in exhaust temperature per 10% increase in liquid volume fraction of ethanol in the fuel could be observed which was attributed to lower adiabatic flame temperatures [47]. A theoretical as well as experimental study comparing performance, efficiency, and emissions of gasoline and E85 with direct fuel injection concluded that at full load at 2,000 RPM a 9% improvement in efficiency as well as maximum torque could be achieved while reducing exhaust gas temperatures by 60°C [48]. The Saab Biopower, a turbocharged sedan built to operate on gasoline as well as E85 utilizes the advantageous fuel properties of the ethanol blend. Rated at a maximum power of 110 kW at 5,500 RPM and a peak torque of 240 Nm at 1,800 RPM with gasoline, the same engine achieves a maximum power of 132 kW at 5,500 RPM and a peak torque of 280 Nm at 2,400 RPM with E85 [49]. With a compression ratio of 9.3:1 the 2.0 L engine takes advantage of the higher octane rating of the ethanol blend by operating the turbo engine at higher combustion pressures, producing more power without risk of engine knock. Independent vehicle dynamometer testing of the Saab Biopower showed that the vehicle acceleration time with E85 was 1 s faster than with gasoline, and that the car also met US Tier 2, Bin 5 emissions levels on both gasoline and E85, which is significant since Europe, where the car is certified, does not require emissions certification on E85 and applicable Euro 4 emissions standards are less stringent. In addition, a detailed exhaust speciation revealed that ethanol and aldehyde emissions were higher on E85 while hydrocarbon-based hazardous air pollutants were higher on gasoline [50]. A detailed study of the impact of alcohol blends on regulated emissions and air toxics with direct injection and ethanol blend levels of up to 50 vol % also concluded that ethanol addition resulted in an increase in aldehyde emissions, in particular acetaldehyde, whereas formaldehyde emissions remained almost constant with ethanol addition. The same study also concluded that with E50 oxide of nitrogen emissions were reduced by approximately 15% compared to gasoline operation, which was attributed to a temperature reduction as a result of direct injection and the increased in-cylinder cooling effect of ethanol compared to gasoline. Specific emissions trends for oxides of nitrogen as well as relevant air toxics at a typical part load engine operating point are summarized in Fig. 13 [51]. These trends hold true independent of engine load, the magnitude changes slightly depending on engine speed and load conditions.

A study targeted at designing an engine specifically for neat ethanol application suggested that a combination of direct injection and turbocharging at increased compression ratios is ideally suited for operation on E100. The proposed direct injection, turbo charged engine differs in the direct injector operating pressure, piston shape and compression ratio to efficiently run with E100 and gasoline. The E100 version operates at an increased injection pressure (300 bar vs 200 bar) and an increased compression ratio (13:1 vs 9:1). Improvements in power and torque vary from 20% to 28% over the range of engine speed, while the fuel conversion efficiency increases 17% to 23% with a peak efficiency of approximately 40% [52]. The report acknowledges



Internal Combustion Engines, Alternative Fuels for. Figure 13
Influence of ethanol content on oxides of nitrogen and air toxics [51]

that cold start issues exist for E100 engines and indicates that fuel is injected after intake valve closing which helps to insure complete vaporization of the ethanol and prevents or minimizes wall wetting, however, no specific measures to ensure proper cold start behavior are discussed.

Measures to mitigate the cold start issue with E100 include fuel choice, e.g., E85 where the 15 vol % of highly volatile gasoline are typically sufficient to improve cold startability, sometimes combined with intentional overfueling to further improve the cold start behavior. However, it would be preferable to use neat ethanol to fully utilize its fuel properties and also avoid overfueling in order to reduce cold start emissions. Technical solutions that have been developed to mitigate the cold start problem include heated fuel injectors [53], heated fuel-rails as well as installation of a second tank, filled with highly volatile gasoline to run the engine for a few seconds before switching to ethanol [54].

Ethanol currently is one of the most dominant alternative fuels which is mainly due to its favorable properties and production pathways. The properties and challenges of ethanol as a fuel are well understood and most technical hurdles for use of ethanol in internal combustion engines have been overcome. It is likely that ethanol will remain a dominant alternative fuel for both, low and mid-level blends with gasoline for the majority of applications as well as high level blends and neat ethanol for local markets. Further development is expected in dedicated engines for high level blends and neat ethanol, however, the main area for advancement will have to be in renewable production pathways which create ethanol from biomass rather than food crops.

Butanol Butanol, a four carbon alcohol that exists in several isomers, is currently under investigation as a potential alternative fuel. Butanol is particularly attractive compared to ethanol due to its higher volumetric energy content and its lower affinity to water (see Table 1). In a joint venture BP and DuPont founded Butamax™ Advanced Biofuels LLC to market and promote biobutanol, which consists mainly of *iso*-butanol. In addition to 135 cars, including 1995 – 2009 model years, that have been tested to date with 1.5 million miles driven, a retail demonstration was completed in 2009 in the UK. A total of 10 million liters of

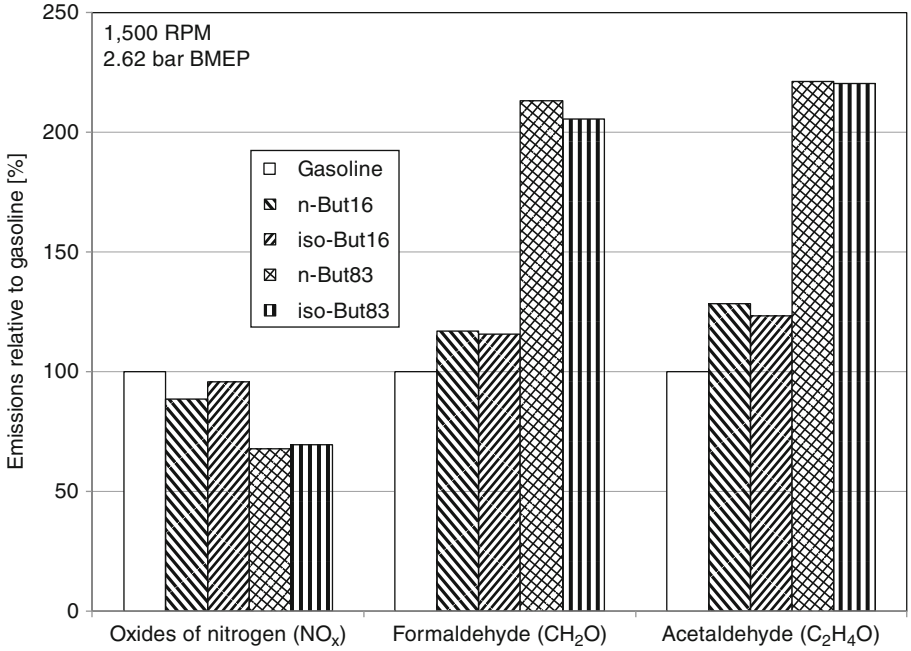
biobutanol were blended and supplied to ten retail sites as EN228 compliant gasoline. Over the course of the demonstration approximately 250,000 vehicle fills were performed and 80 million miles driven suggesting biobutanol can be treated as a normal fuel component and used without special procedures [55]. Since butanol has a lower oxygen content than ethanol, a 16 vol % blend of butanol with gasoline contains the same amount of oxygen, approximately 3.7 wt.%, than a 10 vol % blend of ethanol with gasoline (E10).

In addition to vehicle tests performed and publicized by BP and DuPont, few engine performance and efficiency tests have been performed on *iso*-butanol as well as other butanol isomers. A comparative study performed with gasoline as well as 10 vol % ethanol in gasoline (E10) and 10 vol % *n*-butanol in gasoline on a 2.2 L Direct-injection engine suggests that engine efficiency and emissions characteristics are almost identical throughout a majority of the engine operating regime. Apparent differences were only found at the highest speed and load conditions which could be attributed to differences in the knock resistance of the different fuels [56]. Consequently the study was extended to include *iso*-butanol in addition to *n*-butanol and blend levels were increased up to 85 vol % with similar findings as for the low level blends. Even at the highest blend levels for ethanol, *n*-butanol, and *iso*-butanol, the engine efficiency and emissions characteristics are almost identical to the gasoline baseline. However, at high load conditions the engine could be run with significantly earlier spark timing with ethanol and *iso*-butanol compared to *n*-butanol and gasoline. The earlier combustion phasing due to the increased knock resistance of ethanol and *iso*-butanol compared to gasoline resulted in an increase in brake thermal efficiency in excess of 1% [57]. Due to its lower octane rating *n*-butanol does not exhibit the advantages in terms of knock resistance that ethanol and *iso*-butanol show compared to gasoline. In fact, *n*-butanol was found to behave very similar to regular gasoline in terms of combustion knock [58]. A study of gasoline-butanol blends with up to 80 vol % alcohol content on a port injected spark-ignition engine suggests that the emissions characteristics remain similar with increasing blend ratio, however, improved combustion stability with increasing butanol was observed which could be attributed to shorter ignition delays [59]. An issue that was raised specifically with the use of alcohol fuels is the amount of aldehydes and other air toxics as a result of the combustion. A study of regulated and non-regulated emissions as a result of combustion of various blends of gasoline with ethanol, *n*-butanol, and *iso*-butanol suggests that aside from a consistent reduction in engine-out emissions of oxides of nitrogen regardless of blending agent, both formaldehyde and acetaldehyde emissions increased with addition of butanol to the fuel blend, whereas formaldehyde did not increase significantly with addition of ethanol. Figure 14 shows oxides of nitrogen as well as formaldehyde and acetaldehyde emissions for *n*-butanol as well as *iso*-Butanol gasoline blends at blend levels of 16 and 83 vol % [51]. Since the load point is identical and the butanol blends with 16, and 83 vol % contain the same amount of oxygen as ethanol gasoline blends at 10 and 50 vol %, these results are directly comparable to those for ethanol blends shown in Fig. 13. Although the increase in air toxics in engine-out emissions is noticeable, further studies will have to clarify the effect on tailpipe emissions.

Butanol appears to be a promising alternative fuel with certain fuel properties superior to ethanol such as volumetric energy content and affinity to water. Several studies suggest that butanol is well suited as a fuel for spark-ignition internal combustion engines. However, butanol nowadays is produced mainly as an industrial chemical rather than a transportation fuel. Since feedstock and production pathways for butanol are similar to ethanol, the higher alcohol also faces the same challenges in terms of economy of production from celluloses.

Liquefied Fuels (LPG, LNG)

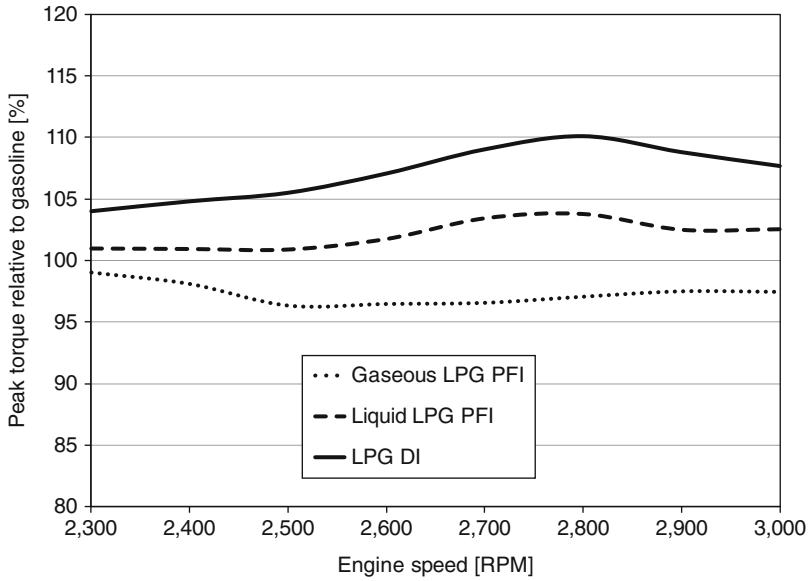
Liquefied Petroleum Gas At normal ambient pressure the boiling point of methane is –162°C, propane is at –42°C, isobutane –10°C, and *n*-butane 1°C. Special measures have to be taken to keep liquefied fuels from evaporating at ambient conditions. As mentioned earlier, LPG composition differs depending on regional supplies which affect vapor pressure and storage conditions (see Fig. 7). Typically LPG is stored in a steel or composite tank at moderate pressure around 10 bar. The two most common mixture formation strategies for LPG are vapor fumigation and LPG injection into the intake. More recently work has also been



Internal Combustion Engines, Alternative Fuels for. Figure 14
Influence of butanol content on oxides of nitrogen and air toxics [51]

reported on LPG direct injection. For vapor fumigation the LPG supply to the engine is typically controlled by a regulator or vaporizer, followed by a mixer in the intake manifold. For a typical vaporizer system a power density that is up to 15% lower than a comparable gasoline engine can be expected [60, 61]. LPG injection systems supply the liquid phase to the intake manifold, where a phase change occurs once the fuel is injected. Since this phase change results in increased charge density due to the cooling effect of the evaporation, LPG injection systems are superior compared to fumigation systems in terms of power density and also reduce the tendency for backfiring [62]. An increase in power density of LPG injection systems of approximately 5–10% compared to the gasoline baseline has been reported [60, 61]. Most recently LPG engines have also been operated with direct injection, and a performance advantage in excess of 10% compared to a gasoline baseline has been reported [61]. Figure 15 shows the maximum engine torque as a function of engine speed for gaseous LPG port injection, liquid LPG port injection, and LPG direct injection relative to gasoline.

A comparative emissions study of a bi-fuel vehicle capable of operating on gasoline or LPG using a vaporizer system showed significant emissions advantages for LPG. Depending on the drive cycle a reduction in CO emissions of 10–30%, 30–50% in HC, and 40–77% in NO_x was reported [63]. The same study also concluded that CO_2 emissions were reduced by approximately 10% when operated on LPG compared to gasoline regardless of drive cycle. Given that the CO_2 emissions factor per unit energy for LPG is approximately 12% lower than that of gasoline [64], the previously mentioned reduction in CO_2 emissions suggests that the two fuels are equivalent in fuel economy. Due to the rapid vaporization of LPG a significant advantage also in terms of particulate emissions can be realized. A study of combustion and emissions characteristics with direct injection of LPG and gasoline concluded that the particulate emissions in LPG operation were lower by a factor of 100 compared to gasoline [65]. Finally, the higher octane rating of LPG compared to gasoline indicates that a dedicated LPG engine could possibly run at a higher, more efficient compression ratio. A computational study that focused on developing an engine design for direct injection of CNG and LPG proposed 14.2:1 as a suitable compression ratio for LPG operation [66]. Since the fuel properties of



Internal Combustion Engines, Alternative Fuels for. Figure 15
Relative peak torque of LPG mixture formation concepts compared to gasoline (Based on [61])

propane were used for the study, the actual compression ratio for real world LPG with a higher butane content might be lower. Nonetheless the proposed compression ratio is still significantly higher than that of typical gasoline engines with compression ratios below 10.5:1 for port injection and less than 12:1 for gasoline direct injection [67].

Liquefied petrol gas (LPG) is well established as an alternative fuel in certain markets and the combustion properties of LPG are well understood. However, the market share compared to mainstream fuels is small in most markets limiting dedicated vehicle development. Besides the challenges due to the liquefied state of the fuel, additional issues with infrastructure limitations and safety concerns appear to further limit the interest in large scale deployment.

Liquefied Natural Gas LNG is stored on-board of vehicles as a cryogenic liquid in super-insulated storage tanks. The typical operating pressure of these tanks is in the range of approximately 5 bar but can reach up to 15 bar. Although LNG is stored in its liquid form, it is typically evaporated before use. The quality of the evaporated gas differs depending on whether natural boil-off gas or forced boil-off gas is used. The gaseous phase in the top portion of LNG tanks is called natural boil-off gas, consists mainly of methane and some nitrogen, and has a high knock resistance. If LNG is extracted from the liquid phase and evaporated separately, it is considered forced boil-off gas which could have a different knock resistance than the natural boil-off gas. In addition to the potential difference in knock resistance the lower heating value of natural boil-off gas is approximately 33–35 MJ/m^3 and significantly lower than that of forced boil-off gas with an LHV of 38–39 MJ/m^3 [68].

Regardless of which type of gas from an LNG tank is used, the fuel is delivered to the engine in its gaseous state. Therefore, an engine combustion system designed and marketed for operation on LNG is in fact designed to run on natural gas or methane and does not necessarily differ from CNG engines [69, 70]. There are, as outlined above, significant differences in the fuel storage, supply, and conditioning systems. For engine design, efficiency, performance, and emissions the same considerations as for compressed natural gas apply.

Gaseous Fuels (Natural Gas, Hydrogen)

Compressed Natural Gas Compressed natural gas is typically stored on-board a vehicle in pressure cylinders, with single steel cylinders mainly used in the past

which are more and more replaced by several smaller composite cylinders. The gas is typically stored at pressures around 200–270 bar with the additional cylinders adding around 60 kg of extra weight to the average vehicle [71].

The first CNG model of the Honda Civic released in 1998 had a natural gas storage capacity of approximately 8 GGE at a fuel economy of 24 MPG City and 34 MPG Highway [72]. The 2011 model still has the same storage capacity (7.8 GGE) at reported fuel economy numbers of 24/36 MPG allowing for a vehicle range of up to 450 km (280 miles) [73].

Vehicle range and weight of the fuel storage system are challenges for natural gas vehicles. At constant range a gaseous storage system operating at 200 bar requires approximately four times the storage volume and outweighs the gasoline system by a factor of 3. Lightweight materials, increased storage pressures, and improved engine efficiencies are expected to improve the volumetric ratio to 1:2.9 and the gravimetric ratio to 1:1.4 with a vehicle range of up to 600 km [74].

The dominant mixture formation strategy for natural gas in automotive applications is intake manifold injection. The higher knock resistance of natural gas compared to gasoline allows for dedicated NG engines to operate at higher compression ratios resulting in higher thermal efficiencies on the one hand and higher NO_x emissions on the other. The lower mixture calorific value of natural gas with external mixture formation (see Fig. 11) compared to gasoline causes reduced power density with the gaseous fuel which can be further compromised when a lean burn strategy is employed. The increased NO_x levels with NG also warrant the use of exhaust gas recirculation (EGR), however, EGR is limited due to combustion instabilities and misfires which occur at increased rates. The reduced power density with injection into the intake manifold can be compensated with supercharging or turbocharging with intercooling. More recently, attempts have also been made to employ direct injection (DI) of natural gas but efforts are hampered since high pressure gaseous injection hardware is not available on the open market [75].

A comparative study of gasoline and CNG on a four-cylinder spark-ignition engine showed significant differences between the two fuels at identical operating conditions. At wide open throttle operation at speeds between 1,500 and 5,000 RPM the engine produced between 8% and 16% less torque with sequential CNG injection relative to gasoline port fuel injection which was mainly attributed to the reduced mixture calorific value. The increased reduction in engine torque at higher speeds was further attributed to the lower flame speeds of CNG compared to gasoline. Similar results in terms of reduction in maximum torque had been reported on the same engine when operated with a mixer type CNG system instead of the sequential fuel injection. However, the fuel conversion efficiency with intake manifold injection was reported to be an average of 13% higher compared to gasoline whereas the carburetor system only showed an average 3% improvement over the gasoline baseline. Both the carburetor and the sequential injection system showed emissions reductions of up to 80% for CO, 8–20% for CO_2, and an average reduction in hydrocarbons of approximately 50%, while NO_x emissions were increased by around 33% with the carburetor system compared to gasoline and no NO_x emissions were reported for the sequential injection system [76, 77].

The previously mentioned results are representative of stoichiometric engine operation just like conventional gasoline engines. However, the relatively wide ignition limits of natural gas suggest that a lean burn strategy which is beneficial for engine efficiency be applied.

Experiments on a 1.3 L four-cylinder engine with one cylinder operated with port fuel injection of natural gas showed that lean operation with relative air/fuel ratios up to 1.3 is feasible without a dramatic decrease in combustion stability. The lean burn strategy also resulted in an up to 75% reduction in NO_x emissions compared to the stoichiometric case with only a moderate increase in HC and CO emissions [78]. Modern stationary natural gas engines operate at relative air/fuel ratios around 1.7 which is close to the misfire region which results in a further increase in engine efficiency while minimizing NO_x emissions [79].

Further extension of the lean combustion limits can be achieved with stratification of the fuel/air mixture which is frequently accomplished by employing direct injection. Research on a single-cylinder 0.9 L engine with direct injection during the compression stroke

suggests that the lean limit could be expanded to approximately 1.8 also identifying the injection timing as a critical parameter to influence combustion and emissions characteristics [80]. A comparison of early injection, late injection, and split injection on a 1 L single-cylinder research engine further showed that stratified mixtures resulting from late and split injection can extend the lean limit up to relative air/fuel ratios in excess of 2.5, eliminating NO_x emissions. However, decreased combustion stabilities and exponential increase in hydrocarbon emissions were reported for these conditions [81]. Finally, hydrogen addition has also been examined as a tool to extend the lean limit of natural gas engines. Results on a 1.6 L single cylinder engine with hydrogen addition rates of up to 19% showed an extension of the lean limit from a relative air/fuel ratio around 1.8 without hydrogen addition to almost 2 with 19% hydrogen. A simultaneous increase in indicated efficiency which can be attributed to the increased burn rate with hydrogen addition was also reported [82].

Energy security and domestic production in addition to natural gas being considered a clean-burning fuel are the main drivers for renewed interest in natural gas in the United States [83]. Natural gas has been widely used for stationary power supply at outstanding engine efficiencies and emissions levels. However, most natural gas vehicle applications rely on conversion from conventional gasoline engines for light-duty applications or diesel natural gas dual fuel applications for heavy duty applications. Future development for natural gas applications is expected to result in dedicated engines optimized to cater to the specific properties of the gaseous fuels. In terms of storage technology both compressed natural gas as well as liquefied natural gas storage will have their market share since either one is well suited for certain applications.

Hydrogen Although the situation for hydrogen in terms of fuel storage is similar to that of natural gas with compressed and cryogenic storage as the two dominant techniques, hydrogen vehicles are not typically classified based on their storage system. Since, as for natural gas, the fuel supply to the engine is almost exclusively in gaseous form, engine technology does not depend on the storage system. Attempts have been made to utilize the low temperature of gaseous hydrogen from liquid storage to increase the power density of hydrogen port injection engines [84].

Hydrogen storage is a significant challenge for the development and viability of hydrogen-powered vehicles. On-board hydrogen storage in the range of approximately 5–13 kg is required to enable a driving range of greater than 300 miles for the full platform of light-duty automotive vehicles using fuel cell power plants or hydrogen internal combustion engines. Current on-board hydrogen storage approaches involve compressed hydrogen gas tanks, liquid hydrogen tanks, cryogenic compressed hydrogen, metal hydrides, high-surface-area adsorbents, and chemical hydrogen storage materials. Storage as a gas or liquid or storage in metal hydrides or high-surface-area adsorbents constitutes "reversible" on-board hydrogen storage systems because hydrogen regeneration or refill can take place on-board the vehicle. For chemical hydrogen storage approaches (such as a chemical reaction on-board the vehicle to produce hydrogen), hydrogen regeneration is not possible on-board the vehicle and thus, these spent materials must be removed from the vehicle and regenerated off-board. On-board hydrogen storage system performance targets were developed through the FreedomCAR and Fuel Partnership, a collaboration among DOE and the US Council for Automotive Research, and the major energy and utility companies. The targets developed are system-level targets and are customer-driven based on achieving similar performance and cost levels as competitive vehicles. The storage system includes the tank, storage media, safety system, valves, regulators, piping, mounting brackets, insulation, added cooling capacity, and any other balance-of-plant components. In order to achieve system-level capacities of 1.8 kWh/kg system (5.5 wt.% hydrogen) and 1.3 kWh/L (0.040 kg hydrogen/L) in 2015 and the ultimate targets of 2.5 kWh/kg system (7.5 wt.% hydrogen) and 2.3 kWh/L (0.070 kg hydrogen/L), the gravimetric and volumetric capacities of the material alone must clearly be higher than the system-level targets [85, 86].

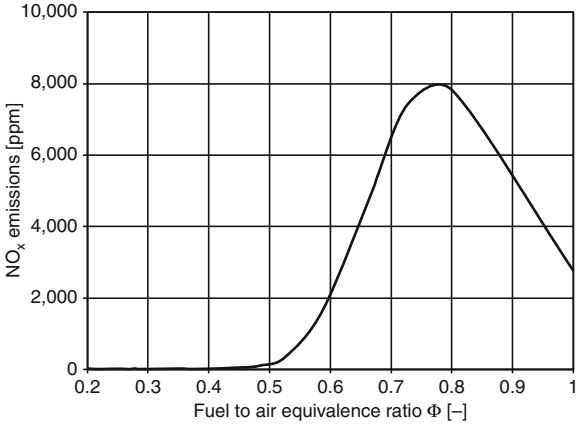
A primary classification of mixture formation strategies of hydrogen engines can be done based on the location of mixture formation or the location of the hydrogen dosing devices. External mixture formation refers to concepts in which hydrogen and air are mixed

outside the combustion chamber, whereas internal mixture formation refers to concepts with hydrogen being introduced directly into the combustion chamber. Some researchers have also proposed combined concepts with a combination of external and internal mixture formation [87]. As indicated in Fig. 11, the mixture formation strategy, especially in hydrogen operation, has significant impact on the theoretical power output of the engine. The dramatic difference in theoretical power output is mainly caused by the low density of hydrogen, resulting in a significant decrease in mixture density when external mixture formation is employed [88].

A study of hydrogen mixture formation concepts in comparison to gasoline performed on an automotive single-cylinder engine confirmed the superior power density with hydrogen direct injection. A 15% improvement in torque at 2,800 RPM with hydrogen direct injection compared to gasoline port injection could be achieved; the relative torque improvement of direct injection compared to hydrogen port injection exceeded 75% [89].

In terms of engine efficiency hydrogen internal combustion engines have been shown to be superior to other, more conventional fuels. Peak efficiency numbers reported by several research groups around the globe suggest that brake thermal efficiencies in the range of 45% can be achieved with hydrogen direct injection [90–93].

Since hydrogen is the only fuel that does not contain any carbon, its combustion does not create any carbon-based emissions constituents except for traces that are expected to result from lube oil combustion. The major challenge for hydrogen combustion engines is their NO_x emissions which are a result of high temperatures during the combustion process (Fig. 16). A trade-off between relative air/fuel ratio and NO_x emissions for homogeneous hydrogen combustion resulting from port injection or early direct injection has been reported by several researchers [89, 94]. Combustion of lean hydrogen–air mixtures with fuel-to-air equivalence ratios of less than 0.5 ($\lambda > 2$) results in extremely low NO_x emissions. Due to the excess air available in the combustion chamber, the combustion temperatures do not exceed the NO_x critical value of approximately 1,800 K [94]. Exceeding the NO_x critical equivalence ratio results in an exponential increase in



Internal Combustion Engines, Alternative Fuels for. Figure 16
NO_x emissions trends for hydrogen fuelled engines

oxides of nitrogen emissions, which peaks around a fuel-to-air equivalence ratio of 0.75 ($\lambda \sim 1.3$). At stoichiometric conditions, the NO_x emissions are at around 1/3 of the peak value. The highest burned gas temperatures in hydrogen operation occur around a fuel-to-air equivalence ratio near 1.1, but at this equivalence ratio oxygen concentration is low so the NO_x concentration does not peak there [95]. As the mixture gets leaner, increasing oxygen concentrations initially offset the falling gas temperatures, and NO_x emissions peak around a fuel-to-air equivalence ratio of 0.75 (λ 1.3). Since hydrogen engines, due to the efficiency advantages, are generally operated with fuel lean conditions, the above mentioned NO_x emissions trade-off with air/fuel ratio also limits the power output.

If NO_x emissions have to be avoided, hydrogen engine operation with port injection is frequently limited to air/fuel ratios below 0.5 and supercharging is employed to increase power density [96, 97]. With hydrogen direct injection the mixture homogeneity can be influenced through tailored design of injection strategy including injector location and nozzle design as well as injection strategy. Parameter studies with hydrogen direct injection revealed that the NO_x emissions characteristics significantly depend on the injection strategy. A consequent optimization of injection timing with conventional hydrogen direct injection allowed a NO_x tailpipe emissions reduction below the

level of a comparable gasoline engine; with a multiple injection approach NO_x emissions could be further reduced to a level of less than 25% compared to single-injection strategies [98].

A recently released hydrogen internal combustion engine vehicle has been pushing the envelope in terms of emissions levels achievable with combustion engines. A BMW Hydrogen 7 Mono-Fuel demonstration vehicle that was tested for fuel economy as well as emissions on the Federal Test Procedure FTP-75 cold-start test as well as the highway test achieved emissions levels that were only 3.9% of the Super Ultra Low Emissions Vehicle (SULEV) standard for nitric oxide (NO_x) and 0.3% for carbon monoxide (CO) emissions. For non-methane hydrocarbon (NMHC) emissions the cycle-averaged emissions are actually 0 g/mile, which require the car to actively reduce emissions compared to the ambient concentration [99].

The properties of hydrogen make it well suited for internal combustion applications with excellent engine efficiency potential and inherently low emissions signatures for carbon based emissions. Further development of injection equipment will be required for production applications of advanced mixture formation concepts. The main challenges for a widespread utilization of hydrogen in transportation remain the lack of a fuel infrastructure as well as sufficient on-board storage density.

Alternative Fuels for Compression-Ignition (CI) Engines

Operation of compression-ignition engines on non-petroleum fuel is not a new concept. In fact, Rudolf Diesel, one of the pioneers of the diesel engine, originally designed the compression-ignition engine to run on a variety of fuels, including straight vegetable oil. In 1900, at the World Exhibition in Paris, he demonstrated his engine running on pure peanut oil.

Operation of today's modern, highly sophisticated compression-ignition engines on straight vegetable oil is not recommended due to the higher viscosity and increased contaminants in the oil. While short-term operation may be possible, long-term effects include build-up of engine deposits, ring sticking, lube oil gelling, and other significant maintenance problems that can reduce engine life.

The primary alternative fuels for compression-ignition engines include biodiesel and synthetic diesel. Emerging fuel technologies include hydrogenated-derived diesel fuel ("green" diesel) and dimethyl ether (DME).

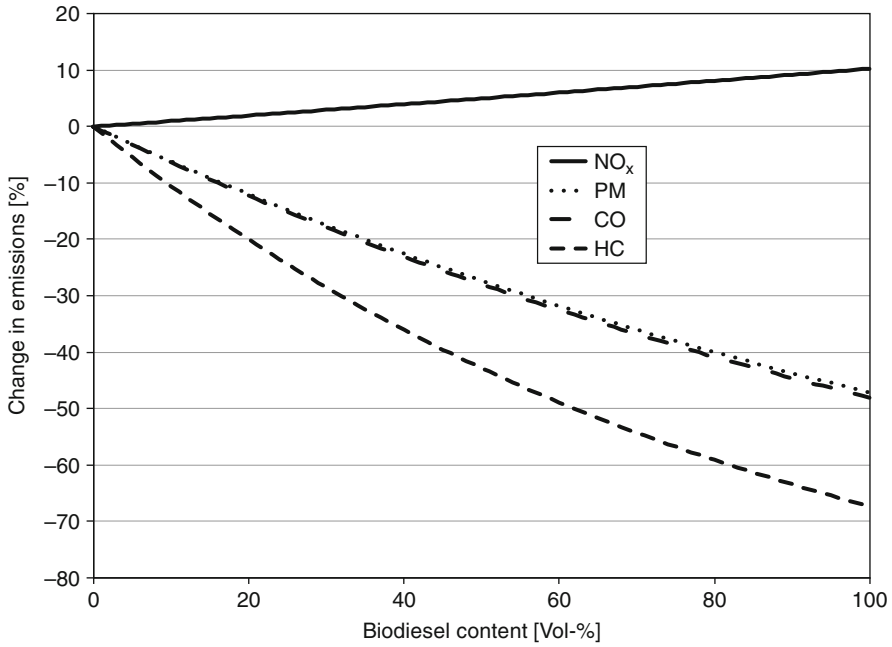
Biodiesel

Biodiesel is the most widely used alternative diesel fuel in the world. Biodiesel is defined as, "a fuel comprised of mono-alkyl esters of long chain fatty acids derived from vegetable oils or animal fats, designated B100, and meeting the requirements of ASTM D 6751" [100]. The properties of biodiesel can vary widely, depending on the feedstock and processing technique employed.

Biodiesel is produced from a fat or oil by reacting it with an alcohol like methanol or ethanol in the presence of a catalyst such as sodium or potassium hydroxide. The transesterification process produces esters (biodiesel) and glycerin. Typically, the alcohol is supplied in excess to ensure a quick reaction and it can be reused. The processing temperatures are around 65°C and the pressures rarely exceed 140 kPa. The conversion efficiency, from oil to methyl ester exceeds 98%. If the alcohol is methanol, then the biodiesel is called a fatty acid methyl ester (FAME). If the alcohol is ethanol, then the biodiesel is called a fatty acid ethyl ester (FAEE). The two "biodiesels" have significantly different properties, as shown in Table 9 for a biodiesel derived from a soybean feedstock. As the properties of biodiesel change, there is a direct impact on fuel economy, emissions, performance, and engine efficiency.

Experimental testing has shown that particulate matter, CO, and HC emissions can be significantly reduced with relatively low levels of biodiesel blend, while NO_x emissions have been shown to increase. Figure 17 shows a typical trend in emissions as the quantity of biodiesel is increased.

One theory as to why NO_x increases with increasing biodiesel content relates to the reduction in PM. As the combustion event occurs, the formation of soot and particulate matter is reduced due to the presence of oxygen in the fuel. Typically, this mass in the cylinder would increase the specific heat of the burned-gas mixture and lower the overall temperature. With the reduction in particulate matter mass, the temperature rises and the production of NO_x increases [102].



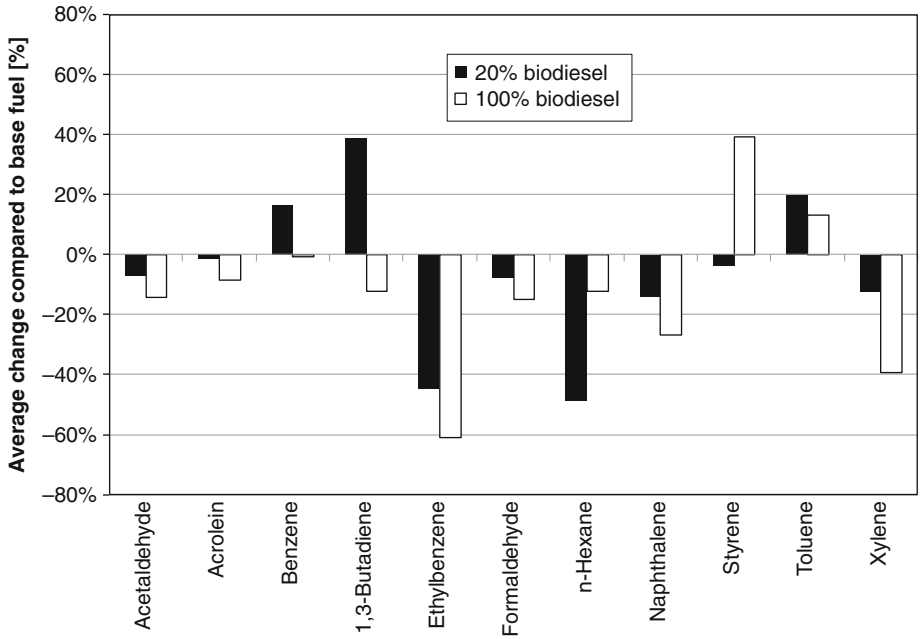
Internal Combustion Engines, Alternative Fuels for. Figure 17
Impact of emissions on biodiesel percentage [101]

A comparison of unregulated, gaseous toxic emissions for two biodiesel blend ratios is shown in Fig. 18. Except for Styrene and Toluene, all toxic emissions levels reduced as the biodiesel blend quantity increased from 20% to 100%. Even for the 20% blend compared to a base petroleum diesel fuel, all but three toxic emissions constituents decreased measurably.

Because of the lower energy density compared to petroleum diesel fuel, biodiesel is associated with a decrease in fuel economy. As the blend volume increases, the impact on fuel economy becomes greater. One would expect up to a 15% reduction in fuel economy due to the reduction in lower heating value. To maintain equivalent power, additional fuel must be injected when operating on biodiesel. This is achieved through an increase in injection pressure or injection duration. It is advantageous to increase injection pressure to not only increase the fuel delivery rate but also improve atomization of the liquid fuel. However, field data has not shown the significant reduction in fuel economy as expected. The oxygenated fuel is also associated with an increase in combustion efficiency up to 5% at blend levels below 25% [104]. The positive impact on combustion efficiency reduces as the blend level increases beyond 30%. The higher cetane number of biodiesel reduces the ignition delay, resulting in reduced noise and vibration from the engine [104]. This also reduces the premixed combustion event of diesel operation, known to have a significant influence on the production of NO_x emissions.

Hydrogenated-Derived Renewable Fuel, Hydrogenated Vegetable Oil (HVO), or Green Diesel

Green diesel refers to the diesel-like fuel product produced from renewable feedstocks utilizing the conventional distillation process for petroleum fuel. Because conventional distillation methods are used to produce this fuel, the properties are very similar to petroleum diesel and it blends well with conventional fuel. The fuels are straight chain paraffinic hydrocarbons that do not contain aromatics, oxygen, or sulfur. Neste claims that their green diesel, named NExBTL results in an 18% reduction in NO_x and almost a 30% reduction in PM, compared to petroleum diesel fuel [105]. This data was the result of a delayed injection event from a lower bulk modulus compared to petroleum diesel. It has been found that the bulk modulus on



Internal Combustion Engines, Alternative Fuels for. Figure 18
Comparison of gaseous, toxic emissions at two biodiesel blend ratios [103]

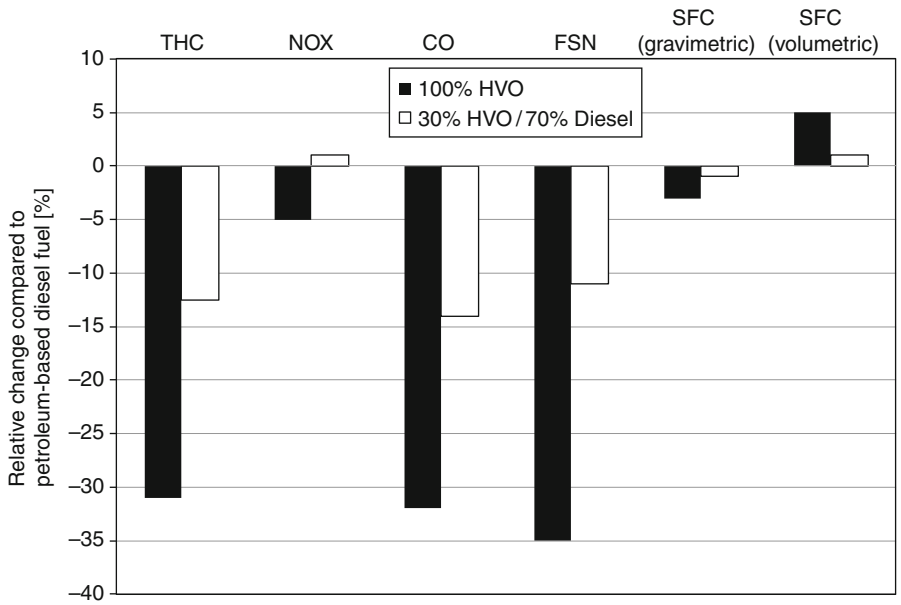
pump-line-nozzle injection systems is a sensitive parameter in affecting the start of injection. Therefore, each fuel system needs to be custom tuned for different alternative fuels to ensure optimum performance and minimal emissions. This requires advanced sensors and significantly more involved calibration to enable flex-fuel operation in the field.

In another study using a common-rail injection system, the impact of 30% and 100% HVO on emissions and fuel consumption was studied. Figure 19 shows the significant reduction in emissions and even gravimetric fuel consumption when no changes were made to the injection system. Additional benefits were realized by modifying the injection timing depending on the parameter of interest. For example, when specific fuel consumption (SFC) was held constant for all three fuels, the 100% HVO reduced NO_x by 16% and smoke by 23% compared to petroleum diesel fuel [106].

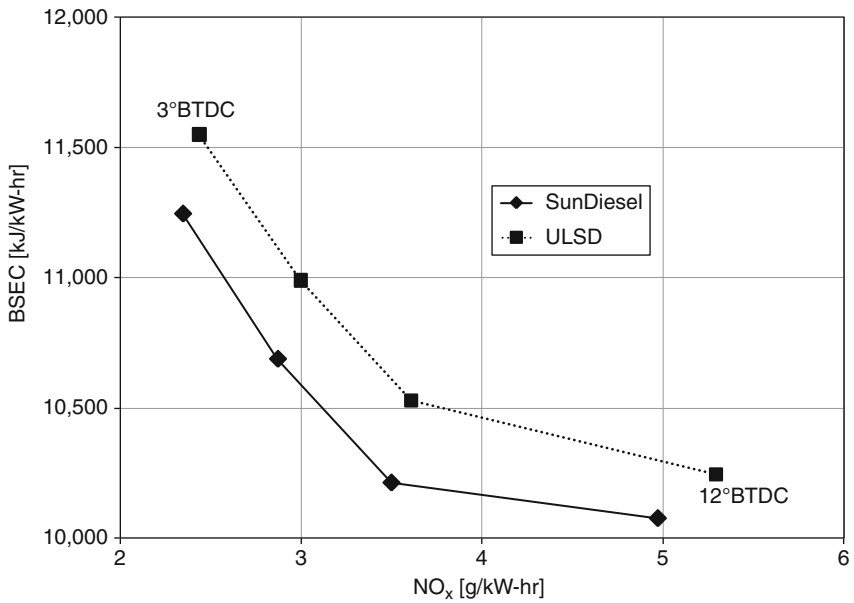
Synthetic Diesel

Synthetic diesel fuel is characterized by high cetane number, low pour point, low sulfur content, and high energy density. These characteristics provide a high quality alternative diesel fuel that is capable of blending with petroleum diesel at any ratio. In one study utilizing a synthetic diesel fuel derived from biomass feedstock (BTL), emissions and fuel consumption were positively impacted when operating the engine on the BTL fuel. The engine was a 1.7 L, four-cylinder, direct injected diesel (Euro II) with intercooled turbocharging. A single speed/load point was utilized for all the testing, which simulated a typical cruise point for this engine. The fuel was produced by Choren Industries under the brand name SunDiesel. The impact on NO_x and BSEC is shown in Fig. 20. Brake specific energy consumption (BSEC) was utilized instead of brake specific fuel consumption (BSFC) due to the varying energy densities between the baseline D2 diesel fuel and SunDiesel. Depending on the parameter of concern, a significant reduction in NO_x emissions at the same BSEC level was achievable with the synthetic diesel fuel [107].

A significant reduction in the number of soot particles in the exhaust stream was measured when the engine was operated on SunDiesel, as shown in Fig. 21. There is also a slight shift toward smaller particles for



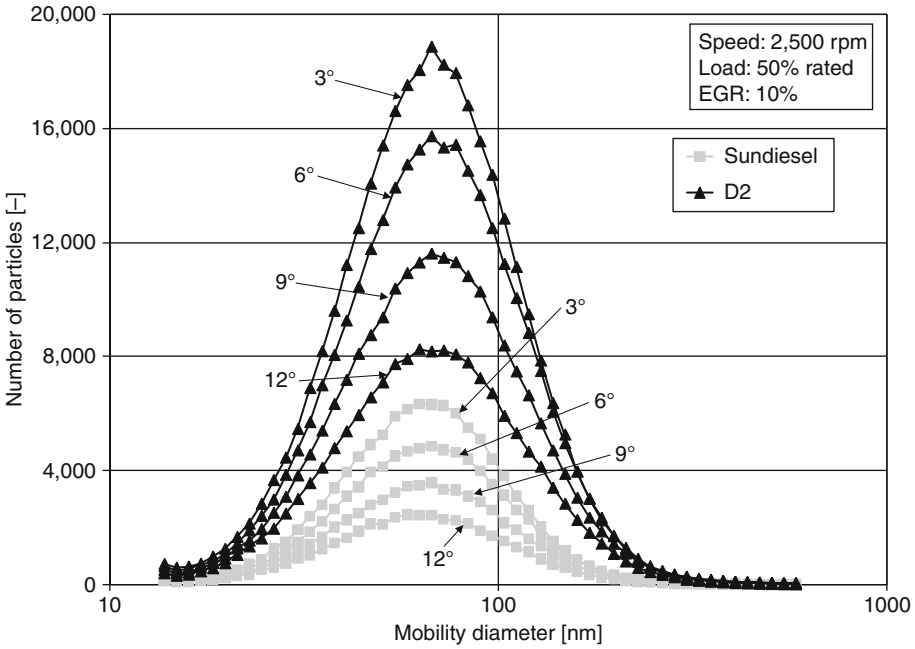
Internal Combustion Engines, Alternative Fuels for. Figure 19
Impact of green diesel (HVO) on emissions and fuel consumption [106]



Internal Combustion Engines, Alternative Fuels for. Figure 20
Brake-specific energy consumption versus specific NO_x emissions for D2 and SunDiesel [107]

the synthetic diesel fuel compared to the petroleum based fuel. This modification of the particulate matter size and quantity can have implications on the particulate filter efficiency and regeneration.

In a comprehensive look at GTL, BTL, and CTL fuels produced using the Fischer–Tropsch process, Gill et al. [108] concluded that the higher cetane number and lower aromatic content were the primary fuel



Internal Combustion Engines, Alternative Fuels for. Figure 21
Soot particle size and distribution for D2 and SunDiesel [107]

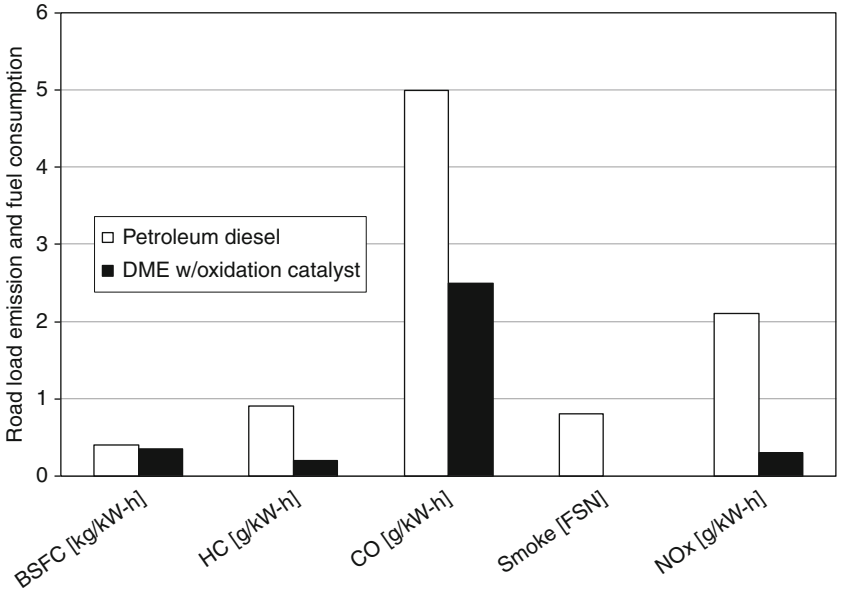
properties that directly affected the emissions. The higher cetane number reduced the ignition delay, leading to a reduction in NO_x emissions. Typically, injection timing is retarded to reduce NO_x emissions but at the expense of fuel consumption (higher BSFC). With a higher cetane number and thus lower ignition delay, the injection timing can be delayed with almost no impact on the BSFC. The shorter ignition delay also provides additional time for the oxidation of soot particles and subsequently reduced the engine-out particulate matter. The reduced aromatic content of the XTL fuels reduced the formation of soot by removing soot nucleation sites [108]. EGR is another effective method of reducing NO_x emissions but typically increases the PM production. XTL fuels with their lower aromatic content have a higher tolerance to EGR, producing less PM emission for the same EGR rate as petroleum diesel fuel.

Dimethyl Ether (DME)

DME is characterized by high cetane (>55) and low emissions characteristics. The oxygenated fuel has been shown to reduce PM, NO_x, and THC emissions compared to petroleum diesel fuel and engine noise is typically reduced as well. The absence of carbon–carbon bonds combined with high oxygen content (~35 wt.%) and rapid evaporation are the primary reasons for the emissions reduction with DME. This makes it an attractive alternative fuel for compression-ignition engines and provides additional means to reduce NO_x emissions through injection timing without a significant increase in PM emissions. However, DME has poor lubricity and high compressibility compared to diesel fuel and thus requires modification to the fuel delivery and storage system. To achieve comparable mileage to a petroleum diesel vehicle, the fuel storage capacity for a DME-powered vehicle would have to be doubled [110].

As shown in Fig. 22, DME has shown significant reductions in HC, CO, and NO_x emissions compared to petroleum diesel. Smoke emissions for DME were too low to be measured. Of interest is the associated reduction in BSFC, though perhaps a BSEC comparison would be more meaningful due to differences in energy density.

DME can be produced from the conversion of natural gas, coal, oil residues, and bio-mass [109]. DME is



Internal Combustion Engines, Alternative Fuels for. Figure 22
Road load emissions and fuel consumption for DME and conventional petroleum diesel [110]

an effective energy carrier and reforming of DME to produce hydrogen-rich fuel-cell streams is currently being researched. In fact, hydrogen yields have been found to be equivalent to methanol at comparable operating temperatures [110].

When comparing alternative fuels on a well-to-wheels analysis for efficiency and GHG production, DME combined with existing engine technology is quite favorable. For example, compression-ignition engines operating on DME has the highest well-to-wheel efficiency of all non-petroleum–based fuel utilizing conventional, hybrid-electric, and fuel processor fuel cell technologies (excluding natural gas) [110]. When compared against existing engine technology, DME produces the lowest amount of well-to-wheel GHG emissions compared to FT diesel, biodiesel, methanol, and ethanol [110].

Alcohol/Diesel Blends

Various drivers including energy security and potential for emissions reduction of diesel engines have triggered attempts to blend diesel with alcohol fuels. A variety of factors including blend stability, viscosity and lubricity, materials compatibility, energy content, cetane number, safety and biodegradability have to be considered if alcohol is to be added to conventional diesel fuel for engine applications. Although all factors require attention, the safety aspect is particularly emphasized at this point. Diesel fuel has a flashpoint of 64°C and is therefore classified as a Class II or combustible liquid according to NFPA guidelines. Alcohol fuels as well as gasoline with flashpoints below 37.8°C (see Table 1) are classified as Class I or flammable liquids requiring more stringent storage requirements such as greater distance in location of storage tanks from property lines, buildings, and other tanks. Studies have shown that blends of 10, 15, and 20 vol % of ethanol in diesel exhibit combustion safety characteristics essentially identical to those for pure ethanol. Essentially any blend of alcohol with diesel fuel has to be stored and handled like gasoline rather than diesel [111, 112].

Blends of diesel and ethanol, also referred to as E-Diesel, generally show increased specific fuel consumption due to the lower energy content of ethanol compared to diesel. However, engine efficiencies remain constant or improve slightly with addition of ethanol. It is also accepted that the addition of ethanol to diesel fuel has a beneficial effect in reducing the PM emissions at least. The amount of improvement varies

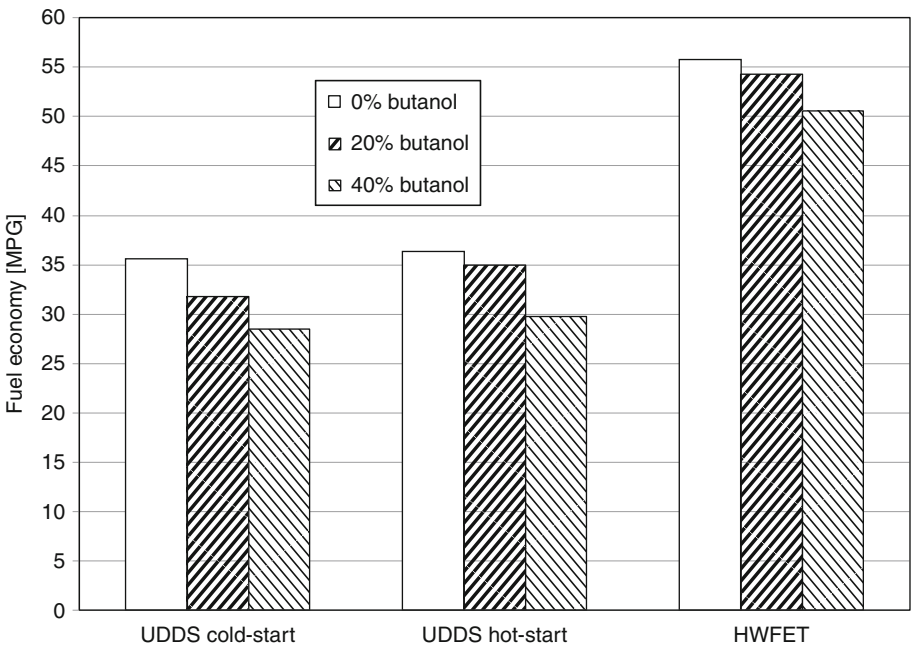
from engine to engine and also within the working range of the engine itself [113]. A heat release and emissions study of 5, 10, and 15 vol % ethanol/diesel blends also concluded that soot, NO_x, and CO emissions decreased with increasing blend levels while hydrocarbon emissions showed an opposite trend [114].

A comparative study of methanol diesel and ethanol diesel blends at blend levels of 5 and 10 vol % on a single-cylinder engine concluded that trends for both alcohols were similar. The results suggest a consistent reduction in soot and CO emissions as well as increased specific fuel consumption with alcohol addition. However, the reported trends of decreasing HC emissions and increasing NO_x emissions with addition of ethanol and methanol are inconsistent with earlier studies. All reported trends did not show any significant difference between addition of ethanol and methanol [115].

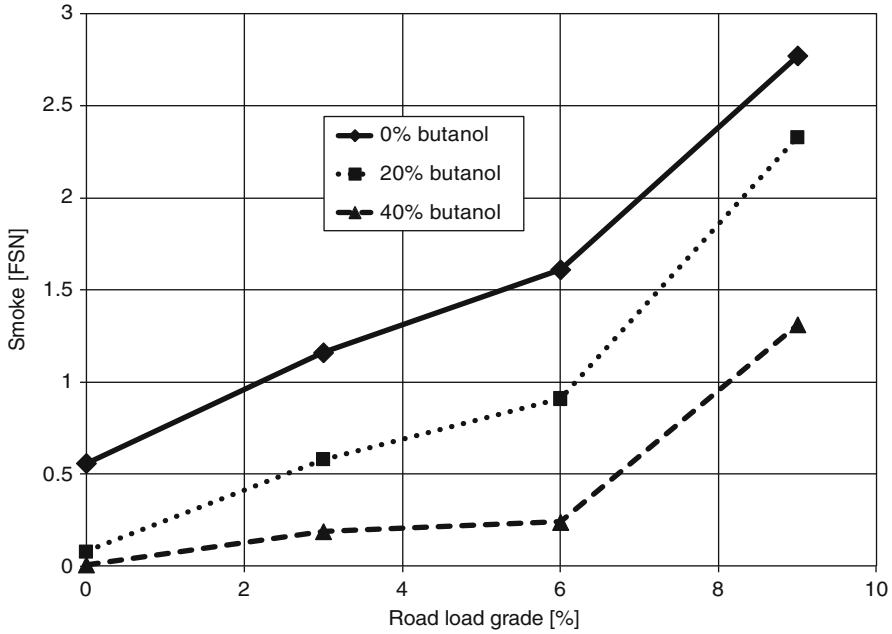
A study was performed using a four-cylinder Mercedes Benz C220 turbo-diesel vehicle and comparing the emissions for conventional diesel fuel, 20 vol % and 40 vol % *n*-butanol. Tests were performed for the cold-start UDDS, hot-start UDDS, and HWFET drive cycles as well as several steady-state points. The results showed that for the urban drive cycle, both total hydrocarbon (THC) and carbon monoxide (CO) emissions increased as larger quantities of butanol were added to the diesel fuel. Oxides of nitrogen (NO_x) were not significantly affected by the 20% butanol blend and decreased with the 40% butanol blend. Drivability of the vehicle decreased noticeably for the 40% butanol blend, especially for the cold-start urban drive cycle. Fuel consumption increased and thus fuel economy decreased (mpg) as the blend ratio of butanol increased, as shown in Fig. 23 due primarily to the lower energy density of butanol compared to diesel.

For the steady state tests, a significant reduction in filter smoke number (FSN) with increasing butanol quantity was observed, as shown in Fig. 24 [116].

A performance and emissions study performed on a 6.0 L bus engine at several steady state operating points reached similar conclusions. At the tested blend levels of 8 and 16 vol % of *n*-butanol with diesel, smoke, NO_x, and CO emissions were consistently reduced with increasing butanol content. At the same time increasing butanol content also resulted in increased HC emissions as well as increased specific



Internal Combustion Engines, Alternative Fuels for. Figure 23
Impact on fuel economy for butanol/diesel blends over three drive cycles [116]



Internal Combustion Engines, Alternative Fuels for. Figure 24
Effect of *n*-butanol in diesel fuel on smoke emissions [116]

fuel consumption despite a simultaneous increase in engine efficiency [117].

Although the technical advantages of alcohol addition to diesel have been demonstrated other limiting factors such as blend stability, materials compatibility and, especially, safety have to be critically evaluated. Given that even addition of small amounts of alcohol changes the flammability dramatically raises the question on whether a transition to alcohol diesel blends is economically viable.

Future Directions

Although alternative fuels can be grouped according to their fuel properties or suitability for certain engine types, their diversity requires dedicated engines to fully utilize their unique characteristics. Due to the wide range of engine applications with their specific requirements such as emissions limits or efficiency targets and the variety of alternative fuel options with specific advantages and limitations it is likely that a future transportation scenario will rely on an increased range of fuels. Meeting current and future engine and vehicle efficiency as well as emissions targets will likely also require advanced combustion concepts to be implemented which in turn might benefit from introduction of alternative fuels. To this extent a stronger link between engine development and fuel development is desirable which should result in optimized utilization of certain fuel properties in dedicated engines. This would put an end to the currently applied approach of developing fuels that best mimic the properties of gasoline or diesel which themselves are far from optimal for certain engine concepts.

Further, an isolated evaluation of fuels based on their efficiency potential or emissions characteristics will likely yield a suboptimal solution. Assessment of alternative fuel options must include technical merits and downsides as well as economical aspects. While technical considerations have been discussed here, an economical assessment including the entire vehicle and fuel life cycle known as "well-to-wheel" analysis has to be performed. These assessments have to include factors such as fuel production, storage and distribution options as well as secondary factors such as water use, recycling, or land usage. Furthermore, findings and recommendations of such studies might vary dramatically depending on regional factors such as climate as

well as typical driver behavior. To this extent a link between engine and fuel related aspects and vehicle related aspects will have to be established which will clarify the impact of engine and fuel choices for different powertrain configurations such as hybrid vehicles.

Overall improving energy security while reducing harmful pollutants as well as greenhouse gas emissions will require a combination of improved transportation efficiency and increased use of alternative fuels. Any efficiency benefit a new engine technology can provide is also desirable when burning alternative fuels along the lines of "Using alternative fuels is desirable, having to use less even more so."

Bibliography

Primary Literature

1. Dutton K (2006) A brief history of the car. New Ideas 1
2. Diesel R (1894) Theory and construction of a rotational heat motor. Spon & Chamberlain, London
3. MacLean H, Lave LB (2003) Evaluating automobile fuel/propulsion system technologies. Prog Energy Combust Sci 29:1–69
4. Tanaka R (2007) Biofuels in Japan. Report by the British Embassy in Tokyo
5. Section 201-202 Renewable Fuel Standard (RFS) Energy Independence and Security Act of 2007 (Pub.L. 110-140, originally named the CLEAN Energy Act of 2007)
6. Schnepf R, Yacobucci B (2010) Renewable Fuel Standard (RFS): Overview and issues. CRS Report for Congress, Order Code R40155
7. Sperling D, Gordon D (2009) Two billion cars: driving toward sustainability. Oxford University Press, New York, pp 42–43. ISBN 978-0-19-537664-7
8. Yacobucci BD (2008) Natural gas passenger vehicles: availability, cost, and performance. CRS Report for Congress. Order Code RS22971
9. Brazilian Automotive Industry Association – ANFAVEA, Brazilian Automotive Industry Yearbook 2010. Available online at http://anfavea2010.virapagina.com.br/anfavea2010/
10. Energy Information Administration (2008) Annual survey of alternative fuel vehicle suppliers and users, "as reported in" Alternatives to traditional Transportation fuels 1998–2008 reports (Tables 14 or S1 depending on year of report). www.eia.doe.gov/cneaf/alternate/page/atftables/afv-atf2008.pdf
11. US Energy Information Administration (EIA) (Apr 2010) EIA's Alternatives to traditional transportation fuels, Table V1. www.eia.doe.gov/cneaf/alternate/page/atftables/afv-atf2008.pdf
12. European Biodiesel Board (2009) http://www.ebb-eu.org/stats.php
13. International Energy Agency (IEA) Statistics division (2007) Energy balances of OECD countries (2008 edition)–Extended balances and energy balances of Non-OECD countries (2007 edition)–Extended balances. IEA, Paris. http://data.iea.org/ieastore/default.asp
14. http://www.biodiesel.org/buyingbiodiesel/plants/
15. Boyd RA (2009) Proposed method of sale and quality specification for hydrogen vehicle fuel. Summary of current information. Standards Fuel Specifications Subcommittee (FSS). U.S. National Work Group for the Development of Commercial Hydrogen Measurement
16. American Petroleum Institute (API) (2001) Alcohols and ethers, Publication No. 4261, 3rd edn. API, Washington, DC
17. Whims J (2002) Pipelines considerations for ethanol, Agricultural marketing resource center. Kansas State University
18. Perry RH, Green DW (1999) Perry's chemical engineers' handbook. McGraw Hill, Malaysia
19. Heywood JB (1988) Internal combustion engine fundamentals. McGraw-Hill, New York
20. Andersen V, Anderson JE, Wallington TJ, Mueller SA, Nielsen OJ (2010) Vapor pressures of alcohol-gasoline blends. Energy Fuels 24:3647–3654
21. The Royal Society (2008) Sustainable biofuels: prospects and challenges. The Royal Society, London. ISBN 978 0 85403 662 2
22. Bradley D (2009) Combustion and the design of future engine fuels. Proc. IMechE Part C: J. Mech Eng Sci 223:JMES1519. doi: 10.1243/09544062JMES1519
23. Foss M (2007) Introduction to LNG. An overview on liquefied natural gas (LNG), its properties, organization of the LNG industry and safety considerations
24. http://www.naturalgas.org/overview/background.asp
25. Lapuerta M, Armas O, Rodrıguez-Fernandez J (2008) Effect of biodiesel fuels on diesel engine emissions. Prog Energy Combust Sci 34:198–223
26. Teng H, McCandless JC, Schneyer JB (2004) Thermodynamic properties of dimethyl ether – an alternative fuel for compression-ignition engines. SAE Technical Paper 2004-01-0093
27. Tilli A, Kaario O, Imperato M, Larmi M (2009) Fuel injection system simulation with renewable diesel fuels. SAE Technical paper 2009-24-0105
28. Teng H, McCandless JC, Schneyer JB (2001) Thermochemical characteristics of dimethyl ether — an alternative fuel for compression-ignition engines. SAE Technical Paper 2001-01-0154
29. Teng H, McCandless JC, Schneyer JB (2002) Viscosity and lubricity of (Liquid) dimethyl ether – an alternative fuel for compression-ignition engines. SAE Technical Paper 2002-01-0862
30. Knothe G (2008) "Designer" biodiesel: optimizing fatty ester composition to improve fuel properties. Energy Fuels 22:1358–1364
31. Knothe G (2005) Dependence of biodiesel fuel properties on the structure of fatty acid alkyl esters. Fuel Process Technol 86:1059–1070
32. Sanford SD et al (2009) Feedstock and biodiesel characteristics report, Renewable Energy Group, Inc. www.regfuel.com

33. Kinast JA (2003) Production of biodiesel from multiple feedstocks and properties of biodiesel and biodiesel/diesel blends, NREL/SR-510-31460, March 2003
34. Abu-Zaid M, Badran O, Yamin J (2004) Effect of methanol addition on the performance of spark ignition engines. Energy Fuels 18:312–315
35. Song R, Hu T, Liu S, Liang X (2008) Combustion characteristics of SI engine fueled with methanol gasoline blends during cold start. Front Energy Power Eng China 2(4):395–400
36. Caton JA (2009) A thermodynamic evaluation of the use of alcohol fuels in a spark-ignition engine. SAE Paper No. 2009-01-2621
37. Li J, Gong C, Su Y, Dou H, Liu X (2010) Effect of injection and ignition timings on performance and emissions from a spark-ignition engine fueled with methanol. Fuel 89:3919–3925
38. Environmental Protection Agency. Regulation of fuels and fuel additives; definition of substantially similar. [FRL-3856-9]
39. http://www.faqs.org/faqs/autos/gasoline-faq/part1/section-4.html
40. http://www.epa.gov/otaq/regs/fuels/additive/e15/420f10054.htm
41. Ré-Poppi N, Almeida FFP, Cardoso CAL, Raposo JL Jr, Viana LH, Silva TQ, Souza JLC, Ferreira VS (2009) Screening analysis of type C Brazilian gasoline by gas chromatography – Flame ionization detector. Fuel 88:418–423
42. Bennett J (2007) Bioethanol in road transport fuels. Presentation at the Royal Society 'International Biofuels Opportunities' conference and workshop, London
43. Joseph Jr H (2007) The vehicle adaptation to ethanol fuel. Presentation at the Royal Society 'International Biofuels Opportunities' conference and workshop, London
44. Jones B, Mead G, Steevens P, Timanus M (2008) The Effects of E20 on metals used in automotive fuel system components. Report by the Minnesota Center for automotive research at Minnesota State University, Mankato
45. Jones B, Mead G, Steevens P (2008) The Effects of E20 on plastic automotive fuel system components. Report by the Minnesota center for automotive research at Minnesota State University, Mankato
46. Jones B, Mead G, Steevens P, Connors C (2008) The Effects of E20 on elastomers used in automotive fuel system components. Report by the Minnesota center for automotive research at Minnesota State University, Mankat
47. Kar K, Cheng W, Ishii K (2009) Effects of ethanol content on gasohol PFI engine wide-open-throttle operation. SAE Paper No. 2009-01-1907
48. Grabner P, Eichlseder H, Eckhard G (2010) Potential of E85 direct injection for passenger car application. SAE Paper No. 2010-01-2086
49. Saab engine specifications (2008) http://dyc.saab-web.com/main/GLOBAL/en/model/95/techspecs.shtml. Accessed 31 May 2011
50. West B, López A, Theiss T, Graves R, Storey J, Lewis S (2007) Fuel economy and emissions of the ethanol-optimized Saab 9-5 biopower. SAE Paper No. 2007-01-3994
51. Wallner T, Frazee R (2010) Study of regulated and non-regulated emissions from combustion of gasoline, alcohol fuels and their blends in a DI-SI engine. SAE Paper No. 2010-01-1571
52. Boretti A (2010) Analysis of design of pure ethanol engines. SAE Paper No. 2010-01-1453
53. Kabasin D, Hoyer K, Kazour J, Lamers R, Hurter T (2009) Heated injectors for ethanol cold starts. SAE Paper No. 2009-01-0615
54. Jeuland N, Montagne X, Gautrot X (2004) Potentiality of ethanol as a fuel for dedicated engine. Oil & Gas Sci Technol Rev IFP 59(6):559–570
55. Schubert A (2010) Expanding the role of biofuels through development of advanced BioButanol. Biofuels Workshop. SAE International Fuels & Lubricants Meeting. Rio de Janeiro
56. Wallner T, Miers S, McConnell S (2009) A comparison of ethanol and butanol as an oxygenate and their effect on efficiency, combustion performance and emissions of a direct-injected 4-cylinder engine. J Engin Gas Turbines Power 131(3):032802–032809
57. Cooney C, Wallner T, McConnell S, Gillen J, Abell C, Miers S (2009) Effects of blending gasoline with ethanol and butanol on engine efficiency and emissions. In: ASME Spring technical conference. Milwaukee/WI. ASME Paper No. ICES2009-76155
58. Szwaja S, Naber JD (2010) Combustion of *n*-butanol in a spark-ignition IC engine. Fuel 89:1573–1582
59. Dernotte J, Mounaim-Rousselle C, Halter F, Seers P (2010) Evaluation of Butanol–gasoline blends in a port fuel-injection, Spark-ignition engine. Oil & Gas Science and Technology – Rev IFP 65(2):345–351
60. http://www.lpgli.com/features.html
61. Li X, Yang L, Pang M, Liang X (2010) Effect of LPG injection methods on engine performance. Adv Mater Res 97–101:2279–2282
62. LPG Autogas (2005) Expansion Issue 1
63. Saraf RR, Thipse SS, Saxena PK (2009) Comparative emission analysis of gasoline/LPG automotive bifuel engine. Int J Environ Sci Eng 1:4
64. Energy Information Administration, Documentation for Emissions of Greenhouse Gases in the U.S. 2005, DOE/EIA-0638 (2005), Oct 2007
65. Oh S, Lee S, Choi Y, Kang K, Cho J, Cha K (2010) Combustion and emission characteristics in a direct injection LPG/Gasoline spark ignition engine. SAE Paper No. 2010-01-1461
66. Boretti AA, Watson HC (2009) Development of a direct injection high flexibility CNG/LPG spark ignition engine. SAE Paper No. 2009-01-1969
67. Heywood JB, Welling OZ (2009) Trends in performance characteristics of modern automobile SI and diesel engines. SAE Paper No. 2009-01-1892
68. CIMAC Working Group "Gas Engines" (2008) Information about the use of LNG as engine fuel. CIMAC position paper
69. Westport Innovations Inc (2010) Direct injection natural gas demonstration project. Westport GX Geavy-Duty LNG engine, Canada

70. Mattas G, Thijssen B (2004) Dual-fuel diesel engines for LNG carriers. Marine News 1
71. ETSAP (2010) Energy Technology Systems Analysis Programme. Automotive LPG and Natural gas engines. http://www.etsap.org
72. NREL (1999) Honda civic dedicated CNG Sedan. Fact Sheet
73. Honda website. http://automobiles.honda.com
74. Thien U (2008) Widening the driving range of NGV 's. Presentation 6th European Forum Gas, Bratislava
75. Korakianitis T, Namasivayam AM, Crookes RJ (2011) Natural-gas fueled spark-ignition (SI) and compression-ignition (CI) engine performance and emissions. Prog Energy Combust Sci 37(1):89–112
76. Aslam MU, Masjuki HH, Kalam MA, Abdesselam H, Mahlia TMI, Amalina MA (2006) An experimental investigation of CNG as an alternative fuel for a retrofitted gasoline vehicle. Fuel 85:717–724
77. Geok HH, Mohamad TI, Abdullah S, Ali Y, Shamsudeen A, Adril E (2009) Experimental investigation of performance and emission of a sequential port injection natural gas engine. Eur J Sci Res 30(2):204–214, ISSN: 1450-216X
78. Cho HM, He B-Q (2008) Combustion and emissions characteristics of a lean burn natural gas engine. Int J Automot Technol 9(4):415–422
79. Packham K (2007) Lean-burn engine technology increases efficiency, reduces NO_x emissions. Power topic #7009. Technical information from cummins power generation
80. Zeng K, Huang Z, Liu B, Liu L, Jiang D, Ren Y, Wang J (2006) Combustion characteristics of a direct-injection natural gas engine under various fuel injection timings. Appl Therm Eng 26:806–813
81. Goto Y, Sato Y, Narusawa K (1998) Combustion and emissions characteristics in a direct injection natural gas engine using multiple stage injection. Proceedings of the International Symposium on Diagnostics and Modeling of Combustion in Internal Combustion Engines (COMODIA), 1998. Japan Society of Mechanical Engineers, pp 543–548
82. Tunestål P, Christensen M, Einewall P, Andersson T, Johansson B (2002) Hydrogen addition for improved lean burn capability of slow and fast burning natural gas combustion chambers. SAE Paper No 2002-01-2686
83. ASTM Standardization News (2010) The renewed promise of natural gas. ASTM Committee D03 and natural gas standards by Adele Bassett. May/June 2010
84. Hallmannsegger M, Fickel H-C (2004) The mixture formation process of an internal combustion engine for zero CO_2-emission vehicles fueled with cryogenic hydrogen. In: IFP International Conference, Rueil-Malmaison, France
85. DOE EERE website. http://www1.eere.energy.gov/hydrogenandfuelcells/storage/current_technology.html?m=1&
86. Wallner T (2009) Opportunities and risks for hydrogen internal combustion engines in the United States. In: Conference proceedings "Der Arbeitsprozess der Verbrennungskraftmaschine." Verlag d. Technischen Universität Graz. ISBN 978-3-85125-068-8
87. Rottengruber H, Berckmueller M, Elsaesser G, Brehm N, Schwarz C (2004) Operation strategies for hydrogen engines with high power density and high efficiency. In: 15th annual U.S. hydrogen conference, Los Angeles, CA
88. Verhelst S, Wallner T (2009) Hydrogen-fueled internal combustion engines (H2ICEs). Prog Energy Combust Sci 35:490–527
89. Eichlseder H, Wallner T, Freymann R, Ringler J (2003) The potential of hydrogen internal combustion engines in a future mobility scenario. SAE Paper No. 2003-01-2267
90. Welch A, Mumford D, Munshi S, Holbery J, Boyer B, Younkins M, Jung H (2008) Challenges in developing hydrogen direct injection technology for internal combustion engines. SAE Paper No. 2008-01-2379
91. Heindl R, Eichlseder H, Spuller C, Gerbig F, Heller K (2009) New and innovative combustion systems for the H2-ICE: compression ignition and combined processes. SAE Paper No. 2009-01-1421
92. Obermair H, Scarcelli R, Waller T (2010) Efficiency improved combustion system for hydrogen direct injection operation. SAE Paper No. 2010-01-2170
93. Tanno S, Ito Y, Michikawauchi R, Nakamura M, Tomita H (2010) High-efficiency and Low-NO_x hydrogen combustion by high pressure direct injection. SAE Paper No. 2010-01-2173
94. Tang X, Stockhausen WF, Kabat DM, Natkin RJ, Heffel JW (2002) FordP hydrogen engine dynamometer development. SAE Paper No. 2002-01-0242
95. Salimi F, Shamekhi AH, Pourkhesalian AM (2009) Role of mixture richness, spark and valve timing in hydrogen-fuelled engine performance and emission. Int J Hydrogen Energy 34:3922–3929
96. ETEC hydrogen internal combustion engine full-size pickup truck conversion. Hydrogen ICE truck brochure
97. Wallner T, Lohse-Busch H, Shidore N (2008) Operating strategy for a hydrogen engine for improved drive-cycle efficiency and emissions behavior. Int J Hydrogen Energy 34:4617–4625
98. Wimmer A, Wallner T, Ringler J, Gerbig F (2005) H2-direct injection – a highly promising combustion concept. SAE Paper No. 2005-01-0108
99. Wallner T, Lohse-Busch H, Gurski S, Duoba M, Thiel W, Martin D, Korn T (2008) Fuel economy and emissions evaluation of a BMW hydrogen 7 mono-fuel demonstration vehicle. Int J Hydrogen Energy 33:7607–7618
100. http://www.biodiesel.org/resources/biodiesel_basics/
101. United States Environmental Agency (2002) A comprehensive analysis of biodiesel impacts on exhaust emissions. EPA420-P-02-001, Oct 2002
102. Mueller CJ, Boehman AL, Martin GC (2009) An experimental investigation of the origin of increased NO_x emissions when fueling a heavy-duty compression-ignition engine with soy biodiesel. SAE Int J Fuels Lubr 2:789–816
103. Graboski et al (2003) The effect of biodiesel composition on engine emissions from a DDC series 60 engine. NREL/SR-510-31461

104. Agarwal AK (2006) Biofuels (alcohols and biodiesel) applications as fuels for internal combustion engines. Prog Energy Combust Sci 33:233–271
105. Kuronen M, Mikkonen S, Aakko P, Murtonen T (2007) Hydrotreated vegetable oil as fuel for heavy duty diesel engines. SAE Technical Paper 2007-01-4031
106. Aatola H, Larmi M, Sarjovaara T, Mikkonen S (2008) Hydrotreated Vegetable Oil (HVO) as a renewable diesel fuel: trade-off between NO_x, particulate emission, and fuel consumption of a heavy duty engine. SAE Technical Paper 2008-01-2500
107. Miers SA, Ng H, Ciatti SA, Stork K (2005) Emissions, performance, and In-cylinder combustion analysis in a light-duty diesel engine operating on a Fischer-Tropsch, Biomass-to-Liquid Fuel. SAE Technical Paper 2005-01-367
108. Gill SS et al (2010) Combustion characteristics and emissions of Fischer Tropsch diesel fuels in IC engines. Progr Energy Combust Sci 31:466–487. doi:10.1016/j.pecs.2010.09.001
109. Arcoumanis C, Bae C, Crookes R, Kinoshita E (2008) The potential of di-methyl ether (DME) as an alternative fuel for compression-ignition engines: a review. Fuel 87:1014–1030
110. Semelsberger TA, Borup RL, Greene HL (2006) Dimethyl ether (DME) as an alternative fuel. J Power Sour 156:497–511
111. McCormick RL, Parish R (2001) Advanced petroleum based fuels Program and renewable diesel program milestone report: technical barriers to the use of ethanol in diesel fuel. NREL/MP-540-32674
112. Waterland LR, Venkatesh S, Unnasch S (2003) Safety and performance assessment of ethanol/diesel blends (E-diesel). Subcontractor Report NREL/SR-540-34817
113. Hansen AC, Zhang Q, Lyne PWL (2005) Ethanol - diesel fuel blends – a review. Bioresour Technol 96:277–285
114. Rakopoulos CD, Antonopoulos KA, Rakopoulos DC (2007) Experimental heat release analysis and emissions of a HSDI diesel engine fueled with ethanol–diesel fuel blends. Energy 32:1791–1808
115. Sayin C (2010) Engine performance and exhaust gas emissions of methanol and ethanol–diesel blends. Fuel 89:3410–3415
116. Miers S, Carlson R, Ng H, McConnell S, Wallner T, LeFeber J (2008) Drive cycle analysis of butanol/diesel blends in a light-duty vehicle. SAE Paper No. 2008-01-2381
117. Rakopoulos DC, Rakopoulos CD, Hountalas DT, Kakaras EC, Giakoumis EG, Papagiannakis RG (2010) Investigation of the performance and emissions of bus engine operating on butanol/diesel fuel blends. Fuel 89:2781–2790

Books and Reviews

Agarwal AK (2007) Biofuels (alcohols and biodiesel) applications as fuels for internal combustion engines. Prog Energy Combust Sci 33:233–271
Eichlseder H, Klell M (2008) Hydrogen in automotive engineering (in German, Wasserstoff in der Fahrzeugtechnik). Vieweg + Teubner, Wiesbaden. ISBN 978-9-8348-0478-5
Huo H, Wu Y, Wang M (2009) Total versus urban: well-to-wheels assessment of criteria pollutant emissions from various vehicle/fuel systems. Atmos Environ 43:1796–1804
MacLeana HL, Lave LB (2003) Evaluating automobile fuel/propulsion system technologies. Prog Energy Combust Sci 29:1–69
Pischinger R, Klell M, Sams T (2002) Thermodynamics of internal combustion engines (in German: Thermodynamik der Verbrennungskraftmaschine). Springer, Wien. ISBN 3-211-83679-9

Internal Combustion Engines, Developments in

Timothy J. Jacobs
Department of Mechanical Engineering,
Texas A&M University, College Station, TX, USA

I

Article Outline

Greek Symbols

η_c Combustion efficiency
η_f Fuel conversion efficiency
$\eta_{f,b}$ Brake fuel conversion efficiency
$\eta_{f,i}$, Indicated fuel conversion efficiency
η_m Mechanical efficiency
η_{th} Thermal efficiency
$\eta_{th,Carnot}$ Thermal efficiency of the ideal Carnot cycle
$\eta_{th,Otto}$ Thermal efficiency of the ideal heat engine Otto cycle
η_v Volumetric efficiency
γ Ratio of specific heats
γ_b Ratio of specific heats of the burned mixture

φ Fuel–air equivalence ratio. For $\varphi < 1$ mixture is lean. For $\varphi = 1$, mixture is stoichiometric. For $\varphi > 1$, mixture is rich. Note that φ is the inverse of the often-used air–fuel equivalence ratio, or λ.

$\rho_{a,i}$ Inlet air density

τ Engine torque

Symbols

BMEP Brake mean effective pressure

$C_{p,b}$ Constant pressure specific heat of the burned mixture

$C_{v,b}$ Constant volume specific heat of the burned mixture

f Residual fraction

f_{final} Final calculated residual fraction

$f_{final\ -\ 1}$ Previous iteration residual fraction to final calculated residual fraction

(F/A) Fuel–air ratio

FMEP Friction mean effective pressure

h_1 Specific enthalpy at state 1

h_2 Specific enthalpy at state 2

h_3 Specific enthalpy at state 3

h_{3a} Specific enthalpy at state 3a

h_5 Specific enthalpy at state 5

h_6 Specific enthalpy at state 6

h_e Specific enthalpy of exhaust mixture

h_i Specific enthalpy of inlet mixture

$IMEP_g$ Gross indicated mean effective pressure

$IMEP_n$ Net indicated mean effective pressure

m Mass

m_1 Mass at state 1

m_2 Mass at state 2

m_3 Mass at state 3

m_4 Mass at state 4

m_6 Mass at state 6

m_a Mass of air

m_f Mass of fuel

m_r Residual mass

m_{total} Total mass

$\dot{m}_a$ Mass flow rate of air

$\dot{m}_f$ Mass flow rate of fuel

MEP Mean effective pressure

M_b Molecular weight of the burned mixture

n_R Number of revolutions per engine cycle

N Engine speed

$\overline{R}$ Universal gas constant

R Gas constant

R_5 Gas constant of mixture at state 5

R_6 Gas constant of mixture at state 6

R_e Gas constant of exhaust mixture

P Cylinder pressure or power

P_1 Pressure at state 1

P_2 Pressure at state 2

P_3 Pressure at state 3

P_{3a} Pressure at state 3a

P_4 Pressure at state 4

P_5 Pressure at state 5

P_6 Pressure at state 6

P_7 Pressure at state 7

P_b Brake power

P_e Exhaust pressure

P_i Inlet (initial) pressure

P_{in} Net indicated power

P_{limit} Limit pressure

PMEP Pumping mean effective pressure

$_1Q_2$ Heat transfer of process 1–2

$_6Q_1$ Heat transfer for process 6–1

Q_{HV} Heating value of fuel

$Q_{HV,f}$ Heating value of fuel

$Q_{HV,i}$ Heating value of specie i

r_c Compression ratio

s_1 Entropy at state 1

s_2 Entropy at state 2

s_3 Entropy at state 3

s_4 Entropy at state 4

s_5 Entropy at state 5

T_1 Temperature at state 1

T_4 Temperature at state 4

T_5 Temperature at state 5

T_6 Temperature at state 6

$T_{cv,adiabatic}$ Constant volume adiabatic flame temperature

T_e Exhaust temperature

T_H Temperature of a source reservoir

T_i Inlet temperature

T_L Temperature of a sink reservoir

T_r Residual fraction temperature

u_1 Specific internal energy at state 1

u_2 Specific internal energy at state 2

u_3 Specific internal energy at state 3

u_{3a} Specific internal energy at state 3a

u_4 Specific internal energy at state 4

U_1 Internal energy at state 1

U_2 Internal energy at state 2
U_3 Internal energy at state 3
U_4 Internal energy at state 4
v_1 Specific volume at state 1
v_2 Specific volume at state 2
v_3 Specific volume at state 3
v_{3a} Specific volume at state 3a
v_4 Specific volume at state 4
V Cylinder volume
V_1 Volume at state 1
V_2 Volume at state 2
V_3 Volume at state 3
V_5 Volume at state 5
V_6 Volume at state 6
V_d Displaced volume
V_{max} Maximum cylinder volume
V_{min} Minimum cylinder volume
W Thermodynamic work
${}_1W_2$ Work for process 1–2
${}_2W_3$ Work for process 2–3
${}_3W_4$ Work for process 3–4
${}_4W_5$ Work for process 4–5
${}_5W_6$ Work for process 5–6
${}_6W_1$ Work for process 6–1
W_b Brake work
W_f Friction work
W_{gross} Gross work
W_{ig} Gross indicated work
W_{in} Net indicated work
W_{ip} Pump work
W_{net} Net work
x_i Mole fraction of specie i
y_i Mass fraction of specie i

Glossary

Combustion Rapid oxidation of a fuel–air mixture (reactants) converting reactants to products and in the process releasing thermal energy.

Lean Air–fuel mixture is such that there is more air than is chemically necessary to oxidize the available fuel.

Products Species formed as the result of a chemical reaction (in the context of this article species formed as the result of a combustion reaction).

Reactants Species that are to be involved in a chemical reaction (in the context of this article species that are to be involved in a combustion reaction).

Rich Air–fuel mixture is such that there is less air than is chemically necessary to oxidize the available fuel.

Stoichiometric Air–fuel mixture is chemically balanced such that there is the correct amount of air to fully oxidize the available fuel.

Definition of the Subject

An internal combustion (IC) engine is a thermodynamic work conversion device that converts chemical energy (typically delivered to the engine in the form of a liquid or gaseous fuel) to work energy (typically in the form of shaft work issuing from a rotating crankshaft). The internal combustion engine markedly distinguishes itself from other types of power-producing equipment – most notably, the heat engine (e.g., steam engine or steam-cycle plant). In the former, chemical energy is released, via combustion (i.e., rapid oxidation) mechanism, internal the same device that converts the released energy to work energy. In the latter, thermal energy is passed into the device via heat transfer; the device thereby converts the thermal energy to work energy. Because of this subtle difference, the thermodynamic limits of maximum efficiency of the internal combustion engine are constrained differently than the thermodynamic limit of maximum efficiency of the heat engine.

There are several types of internal combustion engines; the two most common being the piston/cylinder reciprocating engine and the gas turbine engine. This article constrains its discussion to just piston/cylinder reciprocating engines. As this article is meant to be brief, readers seeking additional and thorough information are referred to the associated cited articles and the authorities listed in "Books and Reviews".

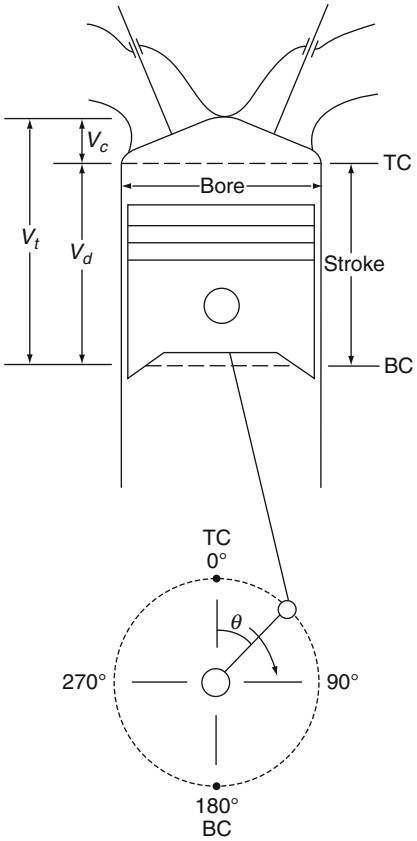
Introduction

Internal combustion engines are pervasive to our daily activities. They are the primary powerhouse of the transportation industry. They serve as neighborhood and small municipality backup power generators or primary power stations. They provide convenience by powering lawnmowers, leaf blowers, and weed trimmers. They deliver excitement and entertainment in pleasure boats, race cars, and motor bikes. Their scale of usability nearly spans the scales of classical physics,

from as small as "micro engines" that fit in the palm of your hand to as large as marine engines that scale three stories and use human-sized doorways for entry into the cylinder block.

The story of the internal combustion engine dates back around 150 years ago, when J.J.E. Lenoir built an engine that combusted coal–air mixtures in a cylinder outfitted with a piston with a two-stroke type fashion (without compression). Soon thereafter, Nicolaus A. Otto and his colleague Eugen Langen expanded on the Lenoir concept, creating an engine that had 11% efficiency compared to Lenoir's 5%. Determined to improve efficiencies of internal combustion engines (which at 11%, were not much better than the steam engine), Otto built the first engine operating on the four-stroke principle which today serves as the primary operating cycle of engines [1]. It is here noted that, although Otto built the first working four-stroke engine, Alphonse Beau de Rochas described in theory the principles of the four-stroke cycle. Further, Beau de Rochas outlined several conditions to achieve maximum efficiency of the internal combustion engine [1].

From this point, and with the aid of several important developers including Rudolf Diesel, Sir Harry Ricardo, Robert Bosch, and Charles Kettering, the internal combustion engine has become one of the most highly efficient, power dense, cost-effective, easily maintained, and versatile power machinery available to consumers. This article will provide some of the important basic information about internal combustion engines, and indicate some of the more recent developments that continue to make internal combustion engines competitive as the preferred power-producing technology.



Internal Combustion Engines, Developments in. Figure 1
Illustration of a piston/cylinder arrangement, as often employed in a reciprocating-piston internal combustion engine (Used with permission from [2])

The Basics of Internal Combustion Engines

Basic Operating Cycle

As this article concentrates its discussions of IC engines on those of the piston/cylinder reciprocating type, Fig. 1 [2] shows the basic geometrical considerations of the piston/cylinder/crankshaft arrangement, which kinematically is described as a crank-slider mechanism. The chemical energy to work energy conversion occurs inside the cylinder, usually bound on the sides by the cylinder walls, on the top by the cylinder head (which typically houses the gas exchange valves, such as the intake and exhaust valves, and other important hardware such as spark plugs and/or fuel injectors), and on the bottom by the piston which reciprocates within the cylinder. A rigid connecting rod fastens the piston to an eccentric location on the rotating crankshaft. The eccentric placement of the connecting rod converts the reciprocating motion of the piston (i.e., boundary motion work) to the rotating motion of the crankshaft (i.e., shaft work). The eccentric placement of the connecting rod to the crankshaft also dictates the important geometrical parameter of the piston engine called the "stroke." The stroke and "bore," or diameter of the cylinder, create the displaced volume, V_d, of the piston engine. The maximum volume of the cylinder, V_{max}, is attained when the piston is at its bottom-most

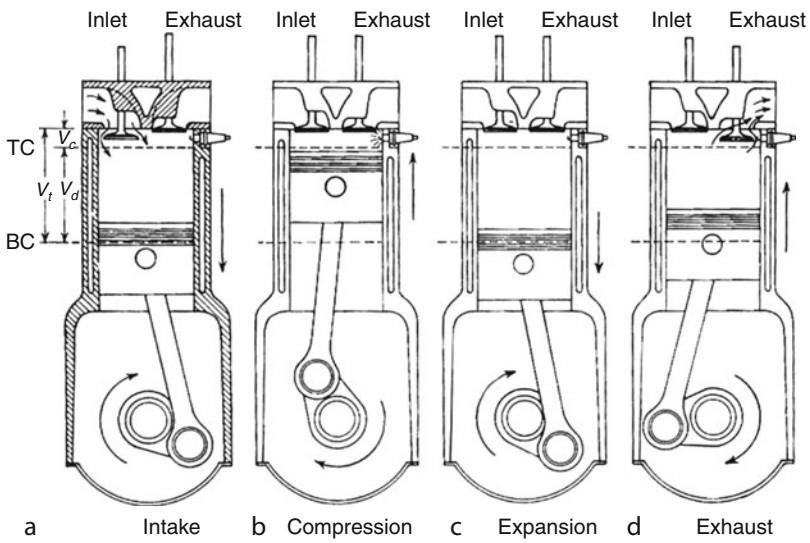
position; a position referred to as "bottom dead center," or BDC. The minimum volume, or clearance volume, V_c, is attained when the piston is at its top-most position; a position referred to as "top dead center," or TDC. The ratio between V_{max} and V_{min} is called the compression ratio, r_c, and is given as Eq. 1:

$$r_c = \frac{V_{max}}{V_{min}} \quad (1)$$

The compression ratio, as will be described in section on "Thermodynamic Analysis of Internal Combustion Engines", is a fundamentally critical parameter for controlling the efficiency (i.e., the ratio of work energy out to chemical energy in) of an internal combustion piston engine.

There are two major cycles used to exploit the piston engine's conversion of chemical energy to work energy: a "two-stoke" cycle and a "four-stroke" cycle. The earliest prototype engines were of the two-stroke variety (e.g., Lenoir and Otto/Langen engines) [1]. In pursuit of achieving higher efficiency, Otto (for whom the thermodynamic ideal "Otto Cycle" is named) built the four-stroke version of his engine [1]. Today, four-stroke cycle engines are the dominant form; thus, most of the article will center on the details of the four-stroke cycle.

Four-stroke cycle engines require four strokes of the piston to complete one power-producing cycle, as shown in Fig. 2 [3]; the reader is also referred to Fig. 4a to aid the discussion. Consider first the cycle starting with the piston at TDC and the intake valve open. The piston moves from TDC to BDC, inducting fresh mixture (conventionally, fuel and air in a gasoline engine and only air in a diesel engine) through the open intake valve in what is called the "intake stroke." At some location near BDC, the intake valve closes, and the piston reverses its motion at BDC. Once the valve closes, the piston/cylinder arrangement creates a closed system. As the piston moves from BDC to TDC, the trapped mixture is compressed in what is called the "compression stroke," increasing the mixture's temperature, pressure, and decreasing its specific volume (i.e., increasing its density). At a point near TDC combustion is expected to commence. In the case of a spark ignition engine (e.g., a conventional gasoline engine), combustion is initiated by the release of spark at a point near (usually advanced of) TDC. In the case of a compression ignition engine (e.g., a conventional diesel engine), combustion is initiated by injecting liquid fuel directly into the cylinder; the compressed air at elevated temperature and pressure atomizes, vaporizes, and mixes with the injected fuel. After a period of time,



Internal Combustion Engines, Developments in. Figure 2
Illustration of the four-stroke operating principle (Used with permission from [3])

the high temperature environment causes chemical reaction and start of combustion. Around start of combustion, the piston reaches TDC, reverses direction, and expands the cylinder volume as combustion converts chemical energy into work energy. This stroke takes on many names, including "power stroke," "combustion stroke," and "expansion stroke." As the piston approaches BDC, the exhaust valve opens, allowing the products of combustion to escape the cylinder. At BDC, the piston reverses direction and motions toward TDC with the exhaust valve open, in what is called the "exhaust stroke." Depending on the engine's crankshaft rotational speed – which can vary from as low as 100 rev/min for large marine-application engines to as high as 15,000 rev/min for race car engines – the four-stroke cycle requires as much as about 1.2 s to as little as 8 ms to complete.

Up to this point, the discussion has centered on the processes occurring in a single cylinder. Most engines, however, are composed of many cylinders and take on various forms (e.g., in-line four-cylinder, "V6", "V8", and "W" engine). In such cases, each cylinder undergoes the same processes but usually out of phase. For example, in an in-line four-cylinder engine (i.e., an engine that has four cylinders oriented sequentially in a single-line bank, an example of which is shown in Fig. 3), the cylinder processes are typically out of phase by 180° (720°/4). The firing order is usually not linear, however, in order to ensure smooth and continuous operation. For example, a four-cylinder firing order may go 1-3-2-4; that is, as the engine crankshaft rotates cylinder 1 produces a power stroke first, followed by cylinder 3, then cylinder 2, and finally by cylinder 4 (which is then followed again by cylinder 1). Finally, when describing an engine's displacement, it refers to a summation of each cylinder's displacement; thus, each cylinder displacement is on average the total engine displacement divided by the number of cylinders.

Internal Combustion Engines, Developments in. Figure 3
Illustration of the Audi 2.0-L TSI, an example of an in-line four-cylinder engine (Used with permission from [4])

How an Engine Makes Power

The reciprocating motion of the piston in a cylinder is the means by which the chemical energy released during combustion is converted into useful work. Work is transferred when a force acts through a displacement; in the case of the piston/cylinder engine, the in-cylinder pressure, P, is interpreted as the force and the changing cylinder volume, dV, during piston strokes is interpreted as the displacement. Thus, the thermodynamic work of an engine cycle is given by Eq. 2:

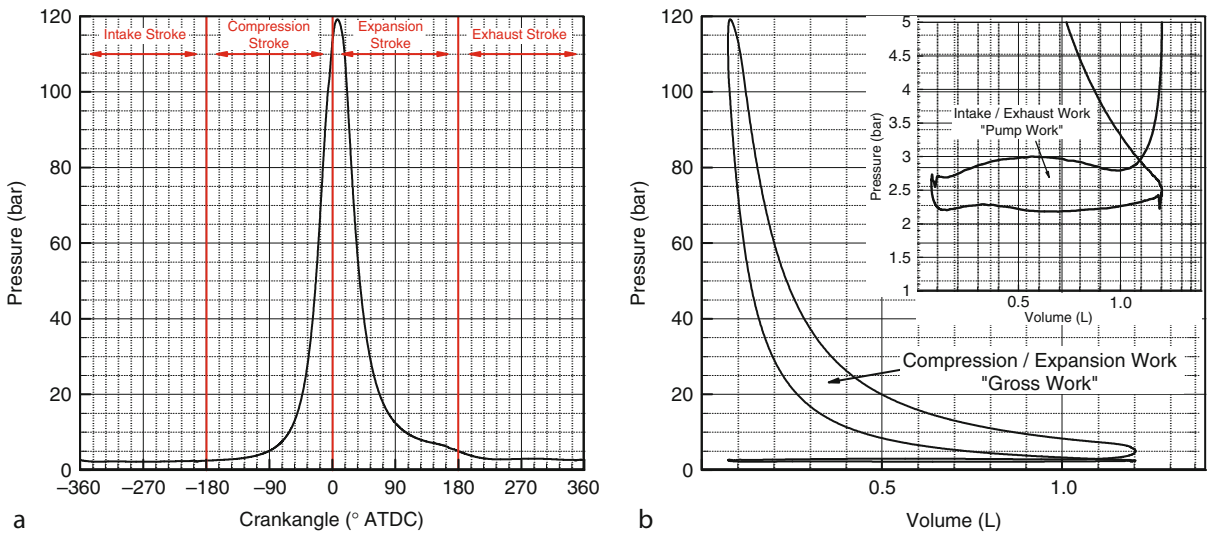
$$W = \oint P\,dV \qquad [\text{kJ/cycle}] \tag{2}$$

The work of the cycle given by (Eq. 2) is considered *boundary work*, since it results from the changing boundary of the control system (in this context, the control system is that enclosed by the piston/cylinder arrangement, and the moving boundary is manifested by the moving piston). Conventional engines convert this boundary work of the piston to *shaft work* through the piston connecting rod's eccentrically located connection to the crankshaft. The shaft work issuing from the crankshaft is often best interpreted as *torque*, τ; the work given by Eq. 2 is related to the torque of the crankshaft via Eq. 3:

$$W = 2\pi n_R \tau \qquad [\text{kJ/cycle}] \tag{3}$$

where n_R is the number of crankshaft revolutions per power cycle (i.e., $n_R = 2$ for a four-stroke cycle and $n_R = 1$ for a two-stroke cycle). The units for τ in Eq. 3 are kN-m.

The torque of an engine is routinely measured with a dynamometer; in this way, the torque is considered "brake torque," or the amount of resistance torque the dynamometer must apply to "brake" the engine to a certain speed condition. Via application of Eq. 3, the "brake work," W_b, is determined. Only determining brake work, however, reveals no insight into the in-cylinder work processes. The in-cylinder work processes can be calculated using in-cylinder pressure measurement that is precisely coupled to the in-cylinder volume via a crankshaft encoder and detailed knowledge of the cylinder's and crank-slider's geometries. An example "pressure–crankangle" diagram is shown Fig. 4a, where crankangle is reported in degrees after top dead center (° ATDC) relative to "combustion TDC." Also shown in Fig. 4a are the four strokes of the four-stroke cycle as described in section "Basic Operating Cycle". In-cylinder pressure is typically measured with a piezo-electric pressure transducer which is able to provide a fast-response indication during the engine cycle [5]. It should be noted that measuring in-cylinder pressure is not a trivial task and great care must be taken to do it properly [6–9]. When using digital equipment (i.e., an analog–digital converter) to electronically record in-cylinder pressure, it becomes necessary to determine the sample rate, which is usually determined by the crankshaft encoder. Varying crankangle resolutions can be used, depending on the level of precision needed of the analysis. For calculating in-cylinder work (described next), a crankangle resolution of up to 10° [6] can be used; for detailed combustion analysis much finer resolution must be used (e.g., about 1° for gasoline engine combustion and 0.25° for diesel engine combustion). The data shown in Fig. 4 is recorded every 0.2°. In addition to the crankangle resolution, the engine speed also determines the needed sample rate of the data acquisition system. Finally, it is equally important to know the cylinder volume at each record of pressure when calculating in-cylinder work. This requires precise phasing between the piston's location and the crankshaft encoder; it also requires knowing the precise geometries of the cylinder and crank-slide mechanism. Specifically, the minimum volume (i.e., clearance volume), the piston stroke, and the cylinder bore must be precisely known. It is often best to determine these using precise measuring instruments, rather than rely on manufacturer specifications (which, although have



Internal Combustion Engines, Developments in. Figure 4

(**a**) Pressure as a function of crankshaft rotation, or crankangle, in degrees after top dead center (° ATDC), and (**b**) pressure as a function of cylinder volume, also illustrating the areas of the *P–v* plane that render gross work and pump work (Data from author's laboratory, Texas A&M University)

tight tolerances, are nominal values). An example "pressure–volume," or "P–V," diagram is shown in Fig. 4b.

Once a precise P–V diagram is determined, the in-cylinder work associated with each process can be determined. The area between the P–V curves, as suggested by Eq. 2, represents the in-cylinder work, or "indicated" work (named for the antiquated use of a mechanical stylus-indicator device to record pressure [10]). There are two portions of the typical four-stroke engine cycle, as shown in Fig. 4: (1) compression and expansion strokes which in combination result in the "gross work," W_{ig}, and (2) intake and exhaust strokes which in combination result in the "pump work," W_{ip}. It is important to note that gross work and pump work correspond to the respective strokes of the piston, not necessarily the valve events (i.e., not necessarily the closed portion of the cycle versus open portion of the cycle). In combination, i.e., the sum of gross work and pump work result in the "net work," as given by Eq. 4:

$$W_{in} = W_{ig} + W_{ip} \tag{4}$$

Note that the subscript "*i*" on the terms in Eq. 4 represents "indicated"; since terms "gross," "pump," and "net" only have relevance from indicated data (i.e., in-cylinder pressure data), it is often dropped as a designator on the work terms.

Pump work, which will be nonzero when intake pressure is different from exhaust pressure (nearly all situations), often decreases net work relative to gross work (i.e., the gas exchange process requires the piston to do work on the gas, or, pump work is negative). There are a few situations, with the use of a turbocharger or supercharger for example, when intake pressure is greater than exhaust pressure and pump work is positive. In such situations, net work will be greater than gross work.

The difference between net work and brake work, as shown in Eq. 5, is the friction work, W_f, of the engine. Friction, of course, always requires work from the system; thus, brake work will always be less than net work. In Eq. 5, W_f captures all forms of mechanical friction of the engine, including friction among crank-slider components, in bearings, in valve springs, and in various accessories (e.g., water and oil pumps):

$$W_f = W_{in} - W_b \tag{5}$$

Engine researchers typically quantify all of the above-described work parameters on volume-normalized parameters, which in general represent a "mean effective pressure." The general definition for mean effective pressure, MEP, is given as Eq. 6:

$$\text{MEP} = \frac{W}{V_d} \tag{6}$$

The various mean effective pressures and their definitions are summarized in Table 1.

One of the major benefits of describing the work of an engine in terms of mean effective pressure is that the "size" of the cylinder is removed from consideration. In other words, it is possible to produce more work from an engine that has a larger displaced volume; however, the mean effective pressure may be lower (relative to a lower displaced volume engine) due to a number of other possible influencing parameters that affect and engine's ability to make power (i.e., fuel conversion and volumetric efficiencies, fuel–air ratio, inlet air density, and fuel heating value). To make these types of assessments, refer to Eq. 7, which is developed from the basic definition of power, P (i.e., work per unit time), and respective definitions of other involved parameters as shown by Heywood [11]. An example of the use of Eq. 7 to assess factors affecting power in a technology comparison is provided in [12].

$$P = \frac{\eta_f \eta_v \rho_{a,i} (F/A) Q_{HV} V_d N}{n_R} \tag{7}$$

where

Internal Combustion Engines, Developments in.

Table 1 Summary of various mean effective pressures describing the various work transfers defined for a reciprocating-piston internal combustion engine

Name	Definition
Gross indicated mean effective pressure (gross IMEP)	$\text{IMEP}_g = W_{ig}/V_d$
Net indicated mean effective pressure (net IMEP)	$\text{IMEP}_n = W_{in}/V_d$
Pump mean effective pressure (PMEP)	$\text{PMEP} = W_{ip}/V_d$
Friction mean effective pressure (FMEP)	$\text{FMEP} = W_f/V_d$
Brake mean effective pressure (BMEP)	$\text{BMEP} = W_b/V_d$

η_f Fuel conversion efficiency (described in section on "Thermodynamic Analysis of Internal Combustion Engines").

η_v Volumetric efficiency, or the engine's effectiveness at "breathing" air. It is defined as $\eta_v = \frac{m_a}{\rho_{a,i} V_d}$, where m_a is the actual mass of air inducted per cycle (kg) and $\rho_{a,i}$ is the density of the intake air at some reference point (kg/m^3) (usually the intake manifold, but also may be atmospheric air upstream of the engine air filter). Note that V_d has units of (m^3) in Eq. 7. It is important to note that volumetric efficiency only quantifies the engine's ability to breath air, i.e., not a fuel–air mixture. Thus, for example, premixing fuel with air prior to induction tends to lower the engine's volumetric efficiency. It is also important to note that volumetric efficiency of the engine depends on the chosen reference point. Thus, if the volumetric efficiency of the whole breathing stream is desired, the reference point is taken as upstream of the intake air filter (for example). If, however, volumetric efficiency of just the intake ports and through the valves is desired, then the reference point is taken as the intake manifold.

$\rho_{a,i}$ As defined above, the density of the intake air at some reference point (kg/m^3).

F/A The mass-based fuel–air ratio of the mixture.

Q_{HV} Heating value of the fuel (typically, lower heating value is used since the products leaving the piston/cylinder system with water as a vapor) (kJ/kg).

N Engine speed (rev/s)

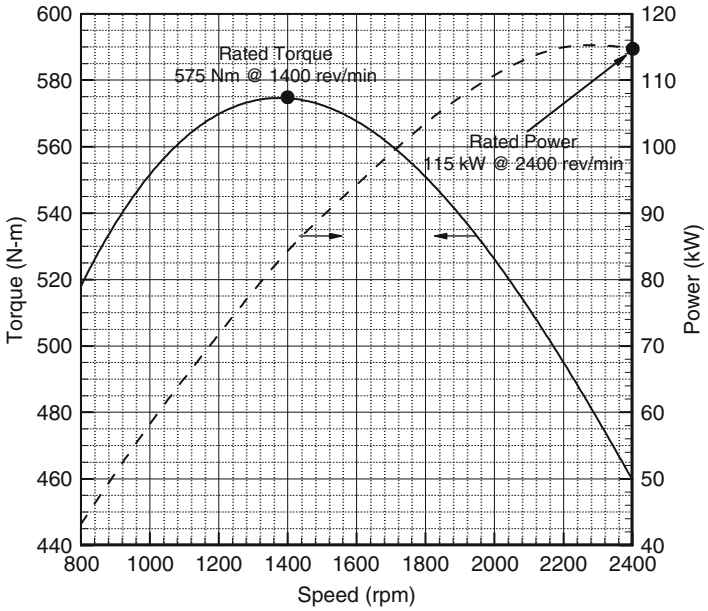
To reveal the form of mean effective pressure, Eq. 8 is given as a modified form of Eq. 7.

$$\text{MEP} = \eta_f \eta_v \rho_{a,i} (F/A) Q_{HV} \tag{8}$$

Thus, it is apparent from Eq. 8 how an engine with a relatively small displacement might have a higher mean effective pressure than an engine with a large displacement, even though the larger-displaced engine may produce more power. It is also clear from Eqs. 7 and 8 how the performance (i.e., power) of an engine may be improved beyond the "easy" action of increasing displaced volume. One parameter, for example that benefits the power and mean effective pressure of the engine is the fuel conversion efficiency. This very important parameter is discussed in the next section.

Finally, this section concludes by illustrating a typical power/torque/speed curve. Apparent from Eq. 3 and the definition of power, there is a relationship among power, torque, and speed of an engine. This relationship for a typical internal combustion engine is shown in Fig. 5. There are a few interesting features to point out in this figure. The first, a practical feature, is the identification of "rated torque" and "rated power". Engines are usually "sized" based on the speed at which they develop maximum torque (also known as rated torque) and the speed at which they develop maximum power (also known as rated power). The second feature, which is apparent from the knowledge that rated torque exists, is the seemingly dependent relationship of torque on engine speed. It is clear from the definition of power that it should have a functional relationship on the speed of the engine, which dictates the amount of work per unit time the engine can deliver. But, based on assessment of Eq. 3, there is not a direct relationship between in-cylinder work per cycle and the speed of the engine. In fact, in an ideal sense, the torque of the engine should be constant with engine speed (and, if constraining to an ideal engine, higher than the maximum attained torque of the real engine), similar to how torque of an electric motor is nearly constant with motor speed. Thus, the behavior seen in Fig. 5 suggests that there exist factors in a real engine that depend on engine speed and also affect the work produced per cycle (thereby affecting the torque of the engine). These factors predominantly consist of heat transfer and friction; both of which have dependencies on the speed at which the engine operates. Heat transfer of course is a time-dependent phenomenon. At low engine speeds there is more time per cycle for thermal energy to be transferred from the cylinder; thus, torque droops at low engine speeds as heat transfer diminishes the energy available for in-cylinder work production. Friction, too, is a speed dependent function due to the mechanical behavior of the engine's interacting components. At high engine speeds, there is increased friction to be overcome on a per cycle basis; thus torque droops at high engine speeds as increased work energy is required to overcome increased friction.

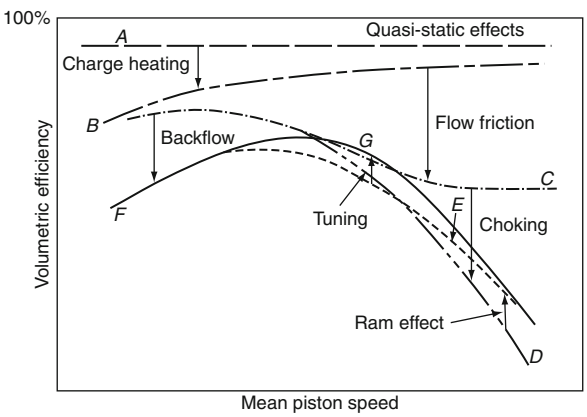
Relating this discussion to Eq. 7, both heat transfer and friction effects tend to decrease η_f. There is, however, another factor of Eq. 7 being influenced by heat transfer and friction and thus serving as a major



Internal Combustion Engines, Developments in. Figure 5
Torque and power as functions of engine speed for a typical internal combustion engine. Data is collected from 4.5 L medium-duty diesel engine with advanced technology such as turbocharging and exhaust gas recirculation (Data from author's laboratory, Texas A&M University)

contributor to the behavior of the torque curve shown in Fig. 5; this factor is the volumetric efficiency. The maximum amount of work that can be developed per cycle is firstly dependent on the amount of air (or oxidizer) the engine can breathe; the amount of air the engine can breathe ultimately dictates the amount of fuel that can be delivered, which of course serves as the energy carrier to be converted in the cylinder.

A representative volumetric efficiency curve as a function of mean piston speed (which is correlated to engine speed) is shown in Fig. 6. There is much important detail in this figure, and the following will describe this in detail; after this discussion, a return to the effect of volumetric efficiency on the torque curve of Fig. 5 will be made. First, notice that under ideal situations the volumetric efficiency curve would be 100% and independent of engine speed (like work, trapped mass per cycle ideally has no functional dependency on the number of cycles per unit time, or, speed of the engine). It is noted that, depending on the reference point chosen for quantifying it, volumetric efficiency could be greater than 100% if the engine is, for example, boosted. In such a situation, the chosen reference point is upstream of the boosting device



Internal Combustion Engines, Developments in. Figure 6
Example of volumetric efficiency, illustrating the many factors that impact the engine's ability to effectively breathe air (Used with permission from [13]). Improvements to the engine's volumetric efficiency can yield substantial improvements to its ability to make power, evident from Eq. 7

(e.g., turbocharger compressor inlet). Although such a chosen reference point is useful to indicating the quantity of trapped mass in the cylinder, it prevents the quantification of the engine's breathing system outside of the boosting device, and masks opportunities for further improvement to the engine's design.

Returning to the volumetric efficiency curve of Fig. 6, and for the specific case of premixed charge engines such as conventional gasoline spark-ignited engines, the premixing of fuel with air prior to induction diminishes the amount of trapped mass of air in the cycle. Such a factor is considered a "quasi-static" effect since it will be present regardless of the speed of the engine. Other quasi-static effects include factors such as, for example, residual fraction due to manifold pressure differences (between exhaust manifold and intake manifold) and exhaust gas recirculation. Thus, notice secondly that the diminishing effect of quasi-static factors is represented as curve "A" in Fig. 6. The next phenomenon to be captured in the volumetric efficiency curve is called "charge heating" and is a heat transfer effect. During normal engine operation, the breathing system of the engine reaches a steady state temperature that is higher than the ambient temperature. Thus, the elevated temperature of the intake port results in heat transfer to the intake air, increasing the air temperature, decreasing its density, and decreasing the amount of mass per unit volume. As described above, heat transfer is a time-dependent phenomenon, therefore its presence is most notable at low engine speeds (low mean piston speeds); notice thirdly in Fig. 6 the effect of charge heating on the volumetric efficiency curve, which in combination with quasi-static effects lowers it to the curve labeled "B." Similar to how friction affects work per cycle, friction also affects air flow and has a dependency on mass flow rate (i.e., has a dependency on engine speed). The fourth factor to notice in Fig. 6 is the effect of flow friction; its effect in concert with quasi-static and charge heating factors result in the curve labeled "C." Notice at very high speeds, the flow friction effect seems to level off and play a speed-independent role as engine speed increases; this results from the flow attaining choke conditions where, regardless of the pressure drop across the intake system, the mass flow rate remains constant as its velocity reaches the speed of sound. Thus, the fifth factor to notice in Fig. 6 is the effect of choking on the volumetric efficiency curve. The net result of all combined effects including choke is the curved labeled "D." Interestingly, some phenomena help to increase the engine's volumetric efficiency; ram effect is one such example. Ram effect occurs at high flow rates where fluid momentum results in continued charging of the cylinder even as the in-cylinder motion of the piston no longer provides the pumping action (i.e., volume increase during intake stroke). The sixth factor to note in Fig. 6 is the increase in volumetric efficiency due to ram effect. The combined effects, including ram effect, result in the curve labeled "E." A similar but diminishing gas exchange phenomenon occurs at low engine speeds, known as backflow. Backflow occurs during the valve overlap period (the period during which both exhaust and intake valves are open as exhaust valves close and intake valves open) where in-cylinder motion becomes quiescent as the piston reaches top dead center and exhaust products of combustion backflow into the intake port. The seventh factor to notice in Fig. 6 is the effect of backflow on the engine, which tends to pervade at low engine speeds where exhaust momentum is low. The combined effects, including backflow, on volumetric efficiency result in the curve labeled "F." Finally, to end on a positive note, the eighth and last factor to observe in Fig. 6 is that of tuning. Tuning is the effort to make use of established resonating sound waves in the intake and exhaust systems to either aid in charging the cylinder (intake tuning) or discharging the cylinder (exhaust tuning). Because this charging or discharging benefits relies upon the timing of compression and rarefaction waves being positioned at precise locations in the intake or exhaust systems, fixed geometry intake and exhaust systems can only be tuned at a narrow speed range. The benefit of tuning is shown in Fig. 6, where for this particular application tuning is designed for the mid-speed range of the engine. Variable tuning systems are becoming available in production, which through the use of creative flow channeling and blocking, can change the "effective" runner length of the flow passage and allow for tuning across broader speed ranges than allowed by fixed geometry runners. With all factors considered, the final volumetric efficiency curve is the solid curve labeled "G."

Now that the details of the volumetric efficiency curve as a function of engine speed, identified as

curve "G" in Fig. 6, have been discussed, a return to its effect on engine torque will be made. In fact, it is interesting to note the similarities between the final volumetric efficiency curve (curve "G") of Fig. 6 and the torque curve of Fig. 5. Such a similarity is not a coincidence, as the maximum amount of work attainable per cycle is a strong function of how well the engine breathes air per cycle. Correspondingly, based on the assessment of Eq. 7, significant improvements to an engine's ability to make power can be realized through creative improvements to the engine's volumetric efficiency.

Thermodynamic Analysis of Internal Combustion Engines

At its core, an internal combustion engine is a thermodynamic device. That is, it takes one form of energy (i.e., chemical energy) and converts it into another form of energy (i.e., mechanical "shaft work" energy). There are several considerations that should be given to the thermodynamics of the internal combustion engine, in pursuit of improving its ability to make power and, more importantly, to make power *efficiently*. Efficiency is an important thermodynamic concept which quantifies the ability of a work conversion device to convert one form of energy (e.g., chemical energy) into work energy (i.e., mechanical shaft work). In the context of internal combustion engines, *fuel conversion efficiency* is used, as given by Eq. 9:

$$\eta_f = \frac{W}{m_f Q_{HV}} \tag{9}$$

where m_f is the mass of fuel per cycle. Typical maximum fuel conversion efficiencies for internal combustion engines vary from ca. 20% to 45%, depending on engine design and type of combustion system. Fuel conversion efficiency, like power, can be designated as brake or indicated fuel conversion efficiency. The relationship between brake and indicated fuel conversion efficiency is given as Eq. 10:

$$\eta_{f,b} = \eta_{f,i}\eta_m \tag{10}$$

where $\eta_{f,b}$ is brake fuel conversion efficiency, $\eta_{f,i}$ is net indicated fuel conversion efficiency, and η_m is mechanical efficiency and given by $\eta_m = \frac{W_b}{W_{in}} = \frac{P_b}{P_{in}} = \frac{BMEP}{IMEP_n}$. In words, fuel conversion efficiency is the ratio of work output to fuel energy delivered. There is another efficiency definition that can be considered, called *thermal efficiency*, which is the ratio of work output to fuel energy released. In other words, since the combustion process converts the fuel's chemical energy to thermal energy, the thermal efficiency indicates the conversion of thermal energy to work energy. It is not possible to precisely measure the energy released by the fuel during the combustion process; it is, however, possible to precisely determine the extent of incomplete combustion, or combustion inefficiency. Thus, *combustion efficiency* is introduced to allow for the determination of the engine's thermal efficiency. Combustion inefficiency is defined as the energy of unreacted fuel in the exhaust to fuel energy delivered to the engine. Combustion efficiency, η_c, is thusly given as Eq. 11 [14]:

$$\eta_c = 1 - \frac{\sum y_i Q_{HV,i}}{\left[\dot{m}_f \Big/ \dot{m}_f + \dot{m}_a\right] Q_{HV,f}} \tag{11}$$

where y_i is an unreacted fuel specie in the exhaust (e.g., carbon monoxide), $Q_{HV,i}$ is the heating value of specie i, $\dot{m}_f$ and $\dot{m}_a$ are the fuel and air mass flow rates, respectively, and $Q_{HV,f}$ is the fuel's heating value. Fuel conversion efficiency, combustion efficiency, and thermal efficiency, η_{th}, are all related by Eq. 12:

$$\eta_f = \eta_c \eta_{th} \tag{12}$$

Note that η_f of Eq. 12 could be either brake or indicated, in which case the corresponding η_{th} will have the same basis.

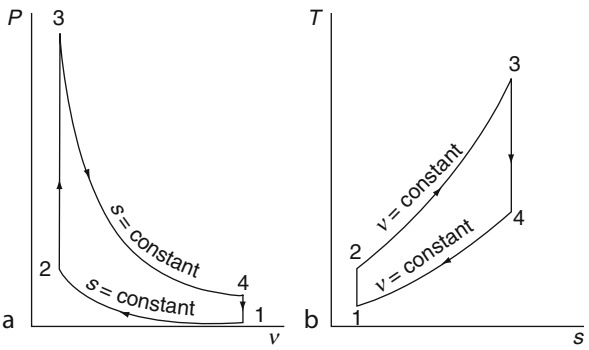
At this point, it is instructive to question what the maximum possible fuel conversion or thermal efficiency of an internal combustion engine could be. To begin to answer this question, it should be established that, contrary to what is sometimes simply conveyed, an internal combustion engine is not, by rigid definition, a *heat engine* [15–17]. A heat engine is a device that collects thermal energy, via heat transfer mechanism, from a high temperature reservoir and converts the *available portion* of this thermal energy to useful work. The portion of thermal energy that is not available for conversion to useful work (i.e., the entropy of the thermal energy) is rejected to a low temperature sink via heat transfer mechanism. A practical example of a heat engine is the steam engine. It is the steam

engine that caused N.L. Sadi Carnot to establish his two postulates on the maximal conditions of the heat engine [18], thus laying the foundation for what would become the second law of thermodynamics and the mathematical axioms for the property entropy. The fully reversible heat engine cycle, i.e., the Carnot cycle, is the most efficient cycle for converting thermal energy into useful work (which, because of its reversible nature, is the most efficient cycle for converting work energy into thermal energy). These details are provided because a common misconception is that the maximum efficiency of an internal combustion engine is limited by a so-called Carnot efficiency; this really has no technical basis because the internal combustion engine does not operate on the same principle (i.e., the heat engine principle) around which the Carnot cycle is created.

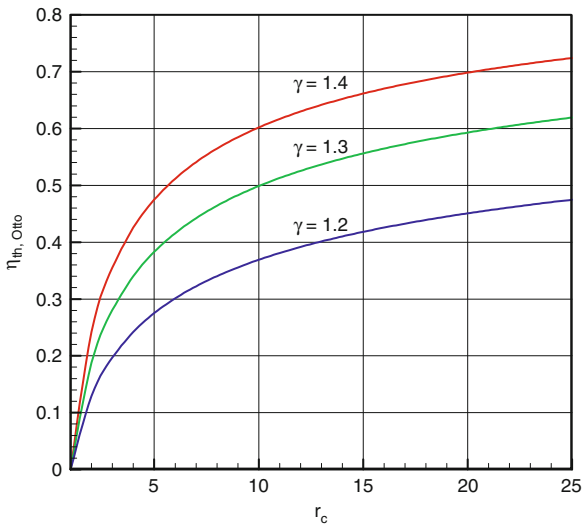
Perhaps much of the misdiagnosis of considering an internal combustion engine a heat engine comes from the use of ideal heat engine cycles to "model" an internal combustion engine; namely the Otto and Diesel cycles. These cycles do serve an important purpose in that certain fundamental parameters affecting the efficiency of the ideal heat engine cycles (i.e., Otto and Diesel cycles) transcend to also affecting the efficiency of the internal combustion engine. An example of such is compression ratio. Thermodynamic analysis of the Otto Cycle, the four processes of which are shown in Fig. 7, on (a) pressure–volume and (b) temperature–entropy diagrams reveal that efficiency of the cycle, $\eta_{th,Otto}$, is given by Eq. 13:

$$\eta_{th,Otto} = 1 - \frac{1}{r_c^{\gamma-1}} \tag{13}$$

where γ is the ratio of specific heats for the working fluid (e.g., γ_{air} = 1.4 at 300 K). This result reveals the asymptotic relationship of increasing thermal efficiency with compression ratio, as shown in Fig. 8. Similar behavior is observed in internal combustion engines (i.e., increasing efficiency with increasing compression ratio) as the capability to expand the working fluid (i.e., combustion products), and thus the ability to convert thermal energy to work energy, increases with increasing compression ratio. Another fundamental behavior is apparent Fig. 8 related to the changing value of γ. Notice that thermal efficiency tends to increase with an increase in γ; similar to the increase in compression ratio, a mixture with a higher γ suggests it requires more energy to increase its temperature during a work-transfer (i.e., constant pressure) process compared to a nonwork (i.e., constant volume) process. The corollary to this is more work energy is transferred out of a mixture with higher γ as the mixture is expanded. In internal combustion engines,



Internal Combustion Engines, Developments in. Figure 7
(**a**) Pressure–volume and (**b**) temperature–entropy diagrams of the air-standard Otto cycle (Used with permission from [19])



Internal Combustion Engines, Developments in. Figure 8
Otto cycle thermal efficiency as a function of compression ratio for several γ values

leaner fuel/air mixtures (i.e., those with less than stoichiometric, or chemically complete, concentrations of fuel in the mixture) have higher γ values, and thus realize an efficiency improvement.

Although the examples of using the Otto cycle to assess the effects of compression ratio and γ value on thermal efficiency are helpful, the Otto cycle efficiency cannot be used to predict maximum attainable efficiency limits of an internal combustion engine. There are several inaccuracies in applying the Otto cycle (or any ideal heat engine cycle) to predict internal combustion engine performance and efficiency. Specifically, the assumptions that go into the development of Eq. 13 are as follows (refer to Fig. 7):

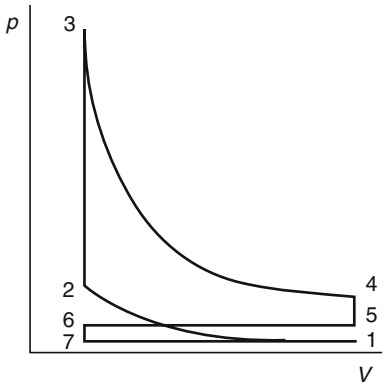
1. Processes 1–2 and 3–4 are *reversible and adiabatic* (i.e., *isentropic*, or constant entropy) processes.
2. Process 2–3 is a constant volume *heat addition* process, where thermal energy of the system increases due to heat transfer with the surroundings at a temperature of no less than T_3.
3. Process 4–1 is a constant volume *heat rejection* process, where thermal energy of the system decreases due to heat transfer with the surroundings at a temperature no greater than T_1.
4. Properties of the working fluid are constant throughout the cycle.

Of the above four assumptions, all four are impractical (meaning, an actual engine's efficiency will be less than the ideal limit because of real effects, which will be discussed in more detail below). Further, assumptions 2–4 are theoretically improper to impose on an internal combustion engine. Assumptions 2–4 imply a heat engine cycle, which has already been established that an internal combustion engine is not. Instead, an internal combustion engine operates as a collection of processes, all which have ideal limits dictating, in aggregate, the ideal maximum efficiency of the engine. Energy is transferred to the control system (i.e., the cylinder of the engine) through mass flow rate, as opposed to heat transfer mechanism (as in the heat engine apparatus). This energy, initially in the form of chemical energy, is converted to thermal energy during the combustion process and the expansion of the piston thereby converts the thermal energy to work energy. Additionally, because of the change from reactants to products during the combustion process and the dependence of species' properties on temperature and pressure, it is not appropriate to assume properties of the working substance remain constant during the engine's operation (i.e., assumption no. 4 is not valid).

Instead, maximum ideal efficiency limits of internal combustion engines could be assessed by improving upon the inappropriate assumptions of the Otto heat engine cycle, i.e., allow energy to be transferred into the system via mass flow (i.e., via the intake process) and allow properties of the mixture to change with species, temperature, and pressure. The resulting analysis, referred to as *Fuel–Air Cycle Analysis* [20], employs the following processes (refer to Fig. 9) and assumptions in modeling an internal combustion engine cylinder (i.e., the control volume):

Process 6–7–1: Ideal intake process of fuel and air and adiabatic mixing with residual gas from the preceding cycle. Residual gas, or often referred to as "residual fraction," is left-over products of combustion not fully exhausted from the cylinder during the previous cycle's exhaust displacement process.

Process 1–2: Reversible and adiabatic compression of the fuel/air/residual mixture (i.e., the reactants). Chemical species are frozen.



Internal Combustion Engines, Developments in. Figure 9
Pressure–volume diagram illustrating the processes assumed in *Fuel–Air Cycle Analysis* of an internal combustion engine. Note that shown cycle is representative of an engine under "throttled" conditions, where intake pressure is less than atmospheric (Used with permission from [21])

Process 2–3: Complete, adiabatic combustion of reactants to products (the latter of which are assumed to exist in equilibrium concentrations). Combustion, in ideal sense, is modeled as constant volume, constant pressure, or limited pressure. Limited pressure is a combination of constant volume and constant pressure processes. Constant volume combustion occurs until a "limit pressure" is reached, when remainder of combustion occurs at constant pressure.

Process 3–4: Reversible and adiabatic expansion of products, which remain in chemical equilibrium throughout the process.

Process 4 – 5: Reversible and adiabatic "blowdown" of products to exhaust pressure of products remaining in the cylinder. Products remain in fixed composition based on concentrations at State 4.

Process 5–6: Ideal, constant pressure, and adiabatic exhaust displacement of products, remaining in fixed composition.

Ideal gas behavior of the mixture and conserved mass are assumed throughout the analysis. Further, mass transfers only occur during intake and exhaust processes (i.e., no leakage, blow-by, or crevice flow during the cycle). By employing first and second laws and various assumptions to each of the processes, respective states can be thermodynamically fixed, thus allowing determination of work transfers. An item to note about *Fuel–Air Cycle Analysis*, because it accommodates changing species, is that *three* independent properties must be known to fix the states (i.e., two thermodynamic properties, such as temperature and pressure, and species composition). Process 6–7–1 is assumed adiabatic. Further, it is assumed that during the stroke from TDC to BDC, the cylinder pressure is constant and equal to the intake pressure, P_i:

$$P_7 = P_1 = P_i$$

$$_6Q_1 = 0$$

With these two statements, and use of first law, the enthalpy at State 1, h_1, is given as:

$$h_1 = f\left[h_e + R_e T_e\left(\frac{P_i}{P_e} - 1\right) - h_i\right] + h_i$$

where f is residual gas fraction, h_e is enthalpy of the exhaust (and equal to $h_6 = h_5$), R_e is the gas constant of the exhaust (and equal to $R_6 = R_5$), T_e is temperature of the exhaust (and equal to $T_6 = T_5$), P_e is pressure of exhaust (and equal to $P_6 = P_5$), and h_i is the enthalpy of the fresh intake mixture. Residual fraction, as explained above, is the residual mass, m_r (which is equal to m_6, the mass at State 6), divided by the total cylinder mass, m_{total} (which is equal to $m_1 = m_2 = m_3 = m_4$, the mass of the system at States 1, 2, 3, and 4, respectively), and is related to r_c, P_e, T_e, and the pressure and temperature at State 4, P_4 and T_4, respectively, by the following:

$$f = \frac{m_r}{m_{total}} = \frac{m_6}{m_1} = \frac{m_6}{m_4} = \frac{1}{r_c}\frac{P_e T_4}{P_4 T_e}$$

It is clear from the equations for h_1 and f that a priori knowledge of the cycle is needed in order to fix State 1 (note that r_c, P_i, P_e, h_i are chosen based on desired cycle compression ratio, manifold conditions, inlet composition, and inlet temperature, T_i, respectively). Thus, initial values for State 1 are typically assumed for the first iteration; State 4 temperature, pressure, and composition from the initial iteration are then used to estimate f and h_1 for the next iteration. This iterative process continues until some convergence criterion is met (e.g., close match between f_{final} and $f_{final\ -\ 1}$). Rather than assume values of f and h_1, it is perhaps more intuitive to assume initial values of f and temperature at State 1, T_1; the following expressions [22] are approximations for establishing initial iteration State 1 properties:

$$f = \left\{1 + \frac{T_e}{T_i}\left[r_c\left(\frac{P_i}{P_e}\right) - \left(\frac{P_i}{P_e}\right)^{(\gamma-1)/\gamma}\right]\right\}^{-1}$$

$$T_1 = T_e r_c f\left(\frac{P_i}{P_e}\right)$$

It is also noted that composition of the residual gas must be assumed to fully fix State 1; it is suggested that ideal products of stoichiometric combustion of the fuel–air mixture be used as the residual gas species.

Work analysis of Process 6–7–1 gives the intake work quantity:

$$_6W_1 = P_i(V_1 - V_6)$$

Process 1–2 is assumed reversible and adiabatic (i.e., $_1Q_2 = 0$); from second law, this gives:

$$s_1 = s_2$$

Geometric relationship gives: $v_1 = r_c v_2$

The assumption that chemical composition remains fixed during compression (i.e., Process 1–2) gives chemical composition at State 2; thus, State 2 is fixed. First law analysis of Process 1–2 gives:

$$ {}_1W_2 = U_2 - -U_1 = m(u_2 - u_1) $$

Process 2–3 is the adiabatic combustion process, where reactants become products (products assumed to be in chemical equilibrium) modeled typically in three simple fashions: (1) constant volume combustion, (2) constant pressure combustion, and (3) limited pressure combustion. First law analysis for each of the three cases results in the following, respectively:

1. Adiabatic constant volume combustion
   $u_3 = u_2$
   $v_3 = v_2$
   ${}_2W_3 = 0$
2. Adiabatic constant pressure combustion
   $h_3 = h_2$
   $P_3 = P_2$
   ${}_2W_3 = P_2(V_3 - V_2) = P_3(V_3 - V_2)$
3. Adiabatic limited pressure combustion
   Constant volume portion
   $u_{3a} = u_2$
   $P_{3a} = P_{\text{limit}}$
   $v_{3a} = v_2$
   Constant pressure portion
   $h_3 = h_{3a}$
   $P_{3a} = P_3 = P_{\text{limit}}$
   ${}_2W_3 = P_3(V_3 - V_2)$

At this point, it becomes necessary to describe a way in which product species of combustion can be modeled. It is clear from the above discussion that combustion is modeled as adiabatic and (a) constant volume, (b) constant pressure, or (c) limited pressure. In order to capture a representative combustion reaction, it is necessary to assume that several product species may exist and will exist in equilibrium. Also, it is important to clarify that this approach assumes initial and final equilibrium states in predicting combustion; i.e., the thermodynamic process goes from one equilibrium state (reactants) to a different equilibrium state (products). This is in contrast to using a more detailed approach that models the progression of combustion and relies upon reaction kinetics to predict intermediate species and their nonequilibrium concentrations. See, for example, Foster and Myers [16] for overview of detailed engine modeling and, for example, Westbrook and Dryer [23] for overview of kinetic modeling of combustion. While perhaps hundreds of species could be included as products, Olikara and Borman [24] propose a combustion reaction, given as Reaction (R1), that might nearly be general for internal combustion engine purposes:

$$ \begin{aligned} x_{13}\Big[\mathrm{C}_n\mathrm{H}_m\mathrm{O}_l\mathrm{N}_k + \frac{n + m/4 - l/2}{\varphi}(O_2 + 3.7274\,\mathrm{N}_2 \\ + 0.0444\,\mathrm{Ar})\Big] \rightarrow x_1\mathrm{H} + x_2\mathrm{O} + x_3\mathrm{N} + x_4\mathrm{H}_2 \\ + x_5\mathrm{OH} + x_6\mathrm{CO} + x_7\mathrm{NO} + x_8\mathrm{O}_2 + x_9\mathrm{H}_2\mathrm{O} \\ + x_{10}\mathrm{CO}_2 + x_{11}\mathrm{N}_2 + x_{12}\mathrm{Ar} \end{aligned} \quad \text{(R1)} $$

where x_1 through x_{12} are the mole fractions of the respective product species. x_{13} represents the moles of fuel per mole of products. Solution of the 13 unknowns of course requires 13 independent equations; five of these equations come from the species balance (i.e., balance of C, H, O, N, and Ar species). The remaining eight equations come from the assumption of equilibrium among various product species [24]. As a final note, it is recognized that NO appears as a product species; this represents the equilibrium concentration of NO at the reaction temperature and pressure. Because both the formation and decomposition of NO in the post-flame gas regions are rate limited [25], it is necessary to take a chemical kinetic approach to modeling IC engine exhaust NO concentrations (see, e.g., [26–29]).

Process 3–4 is reversible and adiabatic expansion with the mixture in chemical equilibrium (resulting in different species at State 4 relative to State 3), resulting in the following thermodynamic state (along with geometrical constraint imposed):

$$ s_4 = s_3 $$

$$ v_4 = v_1 $$

First law analysis gives the following expression, which is the expansion work of the cycle:

$$ {}_3W_4 = U_4 - U_3 = m(u_4 - u_3) $$

Process 4–5 is the ideal blowdown process (i.e., mixture that remains in the cylinder expands isentropically, filling the cylinder volume voided by the products irreversibly escaping past the open exhaust valve).

Species are fixed based on State 4 concentrations. Thus, State 5 is fixed with the following relationships:

$$s_5 = s_4$$

$$P_5 = P_e$$

Although specific volumes change during the blowdown process (due to mass transfer), total volume is constant; thus, work transfer is 0:

$${}_4W_5 = 0$$

Finally, Process 5–6 is the ideal exhaust displacement process, i.e., cylinder contents are transferred out of the adiabatic control volume via piston displacement at constant pressure.

At this point, it is useful to describe the attainment of various property values indicated above (e.g., s_1). Because *Fuel–Air Cycle Analysis* seeks to capture the effects of changing properties on the cycle analysis, the approximations for constant-specific heat ideal gas property equations cannot be used. Instead, it becomes necessary to employ techniques that accommodate changes to mixture properties based on the mixture's temperature, pressure, and species composition. One such technique is to use NASA's Chemical Equilibrium with Applications program [30–34] coupled with the JANAF Thermochemical Tables. By doing so, species at any given temperature and pressure can be predicted (using NASA's Chemical Equilibrium with Applications program) and the corresponding mixture's properties (using JANAF Thermochemical Tables) can be determined. This type of technique, for example, can be programmed into a computer routine.

Based on first law analysis and constant pressure assumption, State 6 is fixed as follows:

$$T_6 = T_5 = T_r$$

$$P_6 = P_5 = P_e$$

The exhaust work is given by the following:

$${}_5W_6 = P_e(V_6 - V_5)$$

Based on the above analysis, the net work and gross work of the cycle are given by:

$$W_{net} = {}_6W_1 + {}_1W_2 + {}_2W_3 + {}_3W_4 + {}_5W_6$$

$$W_{gross} = {}_1W_2 + {}_2W_3 + {}_3W_4$$

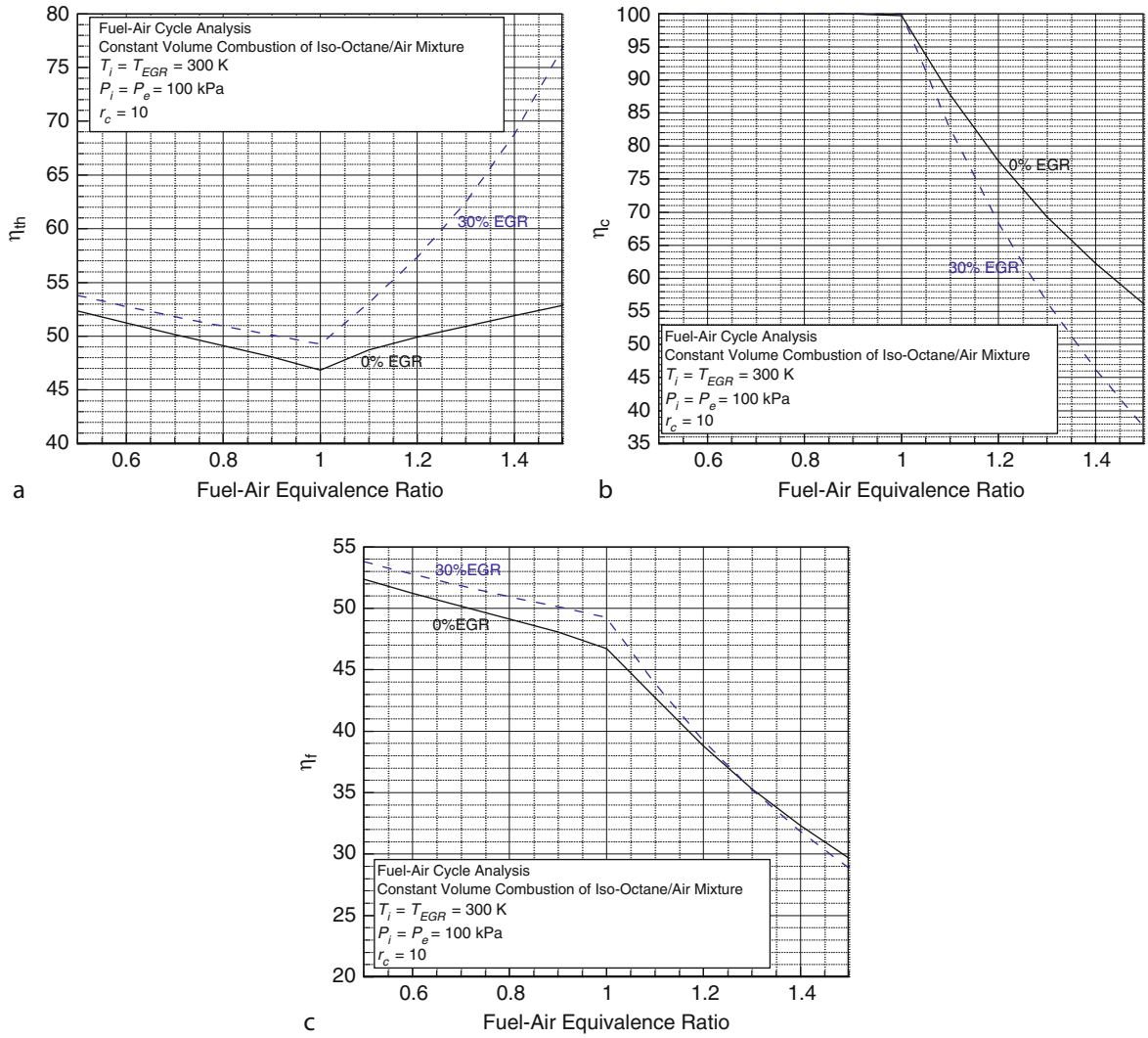
Pump work is as defined by Eq. 4.

Further, various efficiencies can be calculated using Eqs. 9, 11, and 12. Note that although *Fuel–Air Cycle Analysis* is ideal, there may still be incomplete combustion due to the assumption of species existing in chemical equilibrium. Thus, combustion inefficiency will occur at fuel–air mixtures approaching stoichiometric and rich conditions, as dissociation creates concentrations of CO, H_2, and other partially oxidized species. Upon execution of the above analysis in a computer routine, for example, the theoretical limits of maximum efficiency of an internal combustion engine can be determined. Examples of such ideal efficiencies for constant volume combustion of isooctane/air mixtures at two different EGR levels are given in Fig. 10.

Figure 10 reveals interesting behavior of two important internal combustion engine parameters on the various efficiencies, i.e., fuel–air equivalence ratio (often designated as φ) and EGR level. First, notice that as φ approaches stoichiometric (i.e., $\varphi = 1$), η_{th} decreases, and then begins to increase as φ goes rich (i.e., $\varphi > 1$). Further, η_{th} increases as EGR level increases. To understand this behavior, it is necessary to understand how mixture properties are affected by (a) effect of varying φ on species, (b) effect of varying EGR level on species, and (c) effects of φ and EGR level on temperatures and pressures of the cycle. Ultimately, as fundamentally revealed by Eq. 13, a change to mixture properties that results in an increase in γ will cause an increase in efficiency. Figure 11 [35] summarizes how burned gas mixture properties of an isooctane/fuel mixture are affected by various parameters such as φ and temperature. The behaviors of the parameters shown in Fig. 11 can be related to γ by recognizing that γ can be written as given by Eq. 14:

$$\gamma_b = \frac{1}{1 - \frac{\bar{R}}{M_b C_{p,b}}} \tag{14}$$

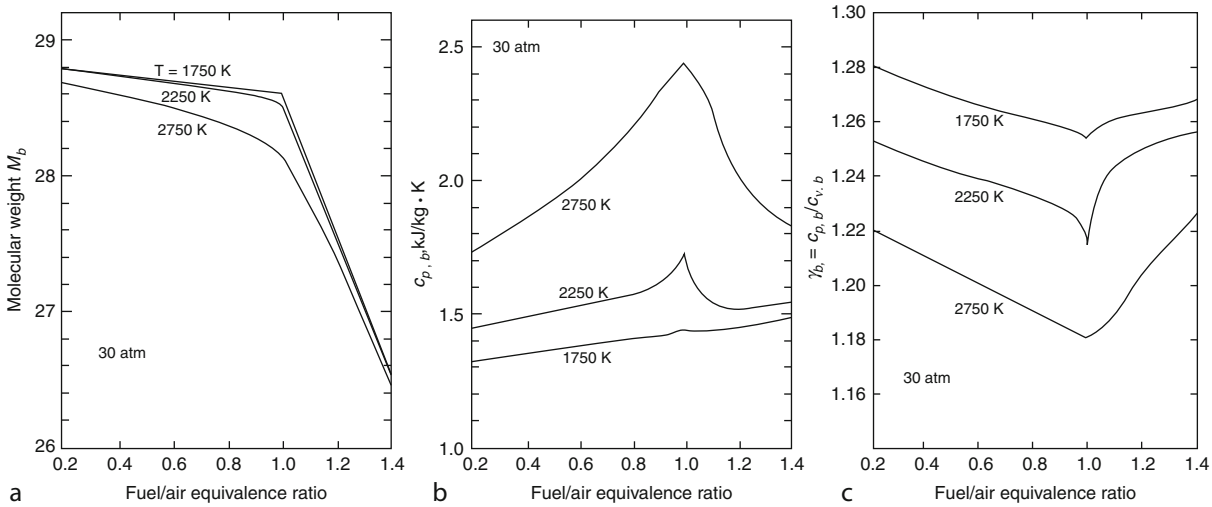
where γ_b is γ for the burned gas mixture, $\bar{R}$ is the universal gas constant (i.e., $\bar{R} \approx 8.314$ kJ/kg-K), M_b is the molecular weight of the burned gas mixture, and $C_{p,b}$ is the constant pressure specific heat of the burned gas mixture. As an example of the type of analysis that might be done on Fig. 11, notice that M_b, (Fig. 11a) decreases as φ increases (for any given temperature of the burned mixture); although concentrations of CO_2 ($M_{CO_2} \approx 44$ g/mol) increase as φ increases, so too do concentrations of H_2O ($M_{H_2O} \approx 18$ g/mol). Since



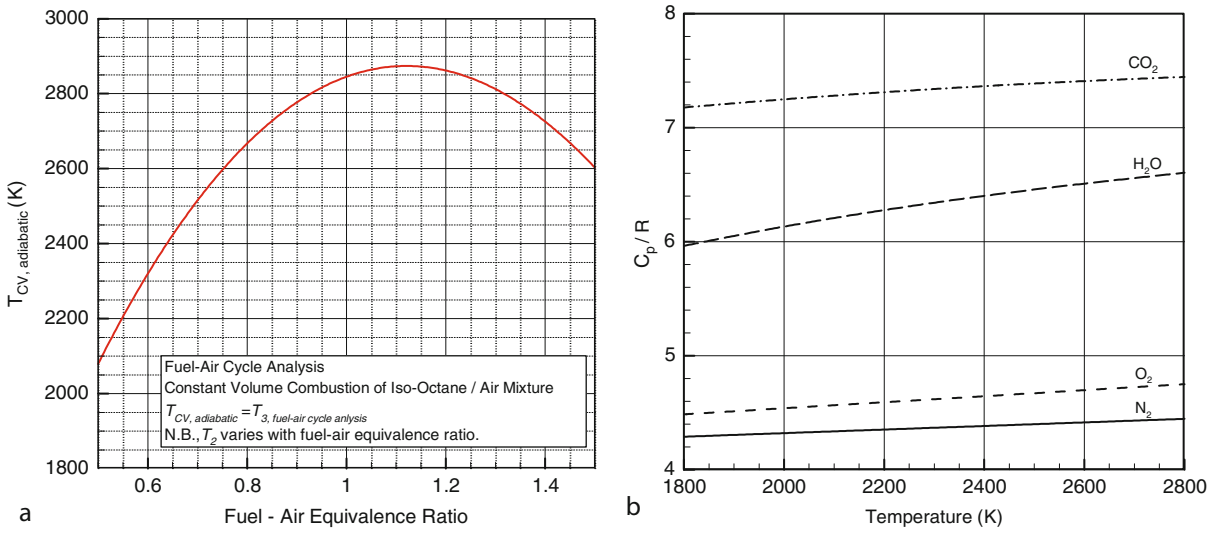
Internal Combustion Engines, Developments in. Figure 10

(**a**) Thermal efficiency, (**b**) combustion efficiency, and (**c**) fuel conversion efficiency of an ideal fuel–air cycle analysis of an internal combustion engine assuming constant volume combustion of isooctane (C_8H_{18}) with T_i = 300 K, $P_i = P_e$ = 100 kPa, T_{EGR} = 300 K, and r_c = 10

isooctane, being a paraffin, produces more H_2O than CO_2 in the products, the lighter, higher concentration of H_2O dominates the net effect of decreasing M_b with increasing φ. As the mixture approaches $\varphi = 1$ and becomes rich ($\varphi > 1$), M_b decreases substantially as concentrations of partially oxidized species (e.g., CO and H_2) with relatively lighter molecular weights increase. Based on analysis of Eq. 14, a decrease in M_b will tend to increase γ_b, which tends to increase η_{th}. Also contributing to γ_b behavior, however, is $C_{p,b}$, which is shown in Fig. 11b. Notice that, up to about $\varphi = 1$, $C_{p,b}$ increases with φ. A portion of this increase is due to the increasing concentration of species with higher constant pressure specific heats as φ increases (e.g., At 1,750 K, $(C_p/R)_{H_2O} \approx 5.94$ and $(C_p/R)_{CO_2} \approx 7.15$ whereas $(C_p/R)_{N_2} \approx 4.28$ and $(C_p/R)_{O_2} \approx 4.46$, as shown in Fig. 12b); another portion, however, comes from the increase in burned gas temperature which causes an increase in constant pressure specific heat for most species. Specifically, as φ increases, adiabatic



Internal Combustion Engines, Developments in. Figure 11
Burned mass (**a**) molecular weight, (**b**) constant pressure specific heat, and (**c**) ratio of specific heats as functions of equivalence ratio of isooctane/air mixtures at various burned mass temperatures and pressure of 30 atm (Used with permission from [35])



Internal Combustion Engines, Developments in. Figure 12
(**a**) Constant volume adiabatic flame temperature as function of fuel–air equivalence ratio of isooctane with initial temperature and pressure as predicted by *Fuel–Air Cycle Analysis* of isooctane/air mixture and 0% EGR, and (**b**) C_P/R for various species typically found in burned mixtures as function of temperature (Data adapted from JANAF *Thermodynamic Tables* [36])

flame temperature and thus burned mixture temperature increases up to a peak that occurs slightly rich of stoichiometric (for isooctane; location of peak adiabatic flame temperature will vary depending on the type of fuel), as shown in Fig. 12a for isooctane/air–fuel cycle analysis with 0% EGR (note that adiabatic flame temperature, or $T_{CV,adiabatic}$ is equal to State 3 temperature, or T_3 of *Fuel–Air Cycle Analysis* and that

T_3 is partially influenced by the changing State 2 temperature, or T_2, as φ varies). The effects of increasing burned gas temperature on the various species' constant pressure specific heats are shown in Fig. 12b. Thus as φ increases $C_{p,b}$ increases as species concentrations change (i.e., increased concentrations of CO_2 and H_2O) and burned gas temperature increases. An increase in $C_{p,b}$, like M_b, will cause γ_b to decrease. Thus, for φ less than stoichiometric, there are two competing effects on γ_b; a decreasing M_b and an increasing $C_{p,b}$ as φ increases. The net result, as shown in Fig. 11c, is a decrease in γ_b which correspondingly explains the decrease in η_{th} as φ increases up to stoichiometric (see Fig. 10a).

The behaviors of M_b and $C_{p,b}$ change as φ increases beyond $\varphi = 1$ (i.e., the mixture becomes rich), causing a corresponding change in γ_b and ultimately in η_{th}. As explained above, M_b decreases more dramatically rich of stoichiometric due to the abundantly increasing concentrations of CO, H_2, and other dissociated species which have lower molecular weights than the fully associated (i.e., CO_2 and H_2O) species. The dramatic decrease in M_b generally dominates any potential increase in $C_{p,b}$, resulting in a net increase in γ_b. Thus, as φ becomes larger than 1, η_{th} increases. Further, at certain temperatures (e.g., 2,750 K) $C_{p,b}$ decreases as φ becomes greater than 1; the higher temperature of the burned mixture causes higher dissociation, which causes increased concentrations of dissociated species (e.g., CO and H_2), which have lower constant pressure specific heats than the fully associated species (e.g., at 2,750 K, $(C_p/R)_{H_2O} \approx 6.6$ whereas $(C_p/R)_{H_2} \approx 4.4$). For such temperatures where $C_{p,b}$ simultaneously decreases along with M_b as φ becomes larger than 1, there is a dramatic increase in γ_b and a corresponding dramatic increase in η_{th}.

Although η_{th} may tend to increase as φ becomes larger than 1, η_c tends to decrease (again, as concentrations of dissociated species such as CO and H_2 increase as φ becomes larger than 1). The result is a general decrease in η_f. Similar analysis can be conducted to understand the effect of EGR on the various efficiencies. It becomes necessary at this point, however, to assess the extent of the utility of such "first law analysis" – i.e., an analysis centered only on energy transfer – in pursuing future developments of internal combustion engines. For example, an engineer may "dream" of a situation where an engine operates rich at a fuel–air equivalence ratio of 1.4 and 30% EGR, yielding a η_{th} of nearly 70%; this certainly outperforms an engine operating lean at a fuel–air equivalence ratio of 0.5 and 30% EGR (see Fig. 10a). If only an energy-analysis is conducted to assess this "dream," the idea will immediately be discounted because the corresponding η_f for the engine running at 1.4 fuel–air equivalence ratio and 30% EGR is only ca. 32% (compared to the lean operation which is close to 54%). But, the products of the rich combustion still have useful energy, that is, energy that can be converted to useful work. While it is true the cycle cannot convert such *available* energy to useful work, it is more than a "dream" that some other device could convert it (e.g., a fuel cell or a thermoelectric generator). Thus, it might be conceivable to design an engine that allows it to operate at maximal η_{th} and couple it to another device that, while not having as high of a η_{th}, has a high enough η_{th} to suitably convert the available chemical energy to useful work, so that the net efficiency of both devices is greater than the η_f of the internal combustion engine operating at the lean condition (for example).

Analysis of such "dreams" can only be completed by (a) introducing the concept of *available energy*, or modernly called *exergy*, by employing second law considerations on the analysis and (b) introducing real effects on the engine analysis that include such factors as heat transfer, friction, and other energy "losses" – but not necessarily exergy losses (friction is, of course, an exergy loss). Thus, a brief discussion will follow describing the major benefit of conducting an exergy analysis of an engine system. The idea of available energy – or, that portion of energy which is available to do useful work – is introduced by Gibbs [37] and, in its basic form, suggests that a system with a given amount of energy can transfer that energy out of the system as useful work until the point of equilibrium between the system and its environment (i.e., its surroundings). Using a simplified example, a piston/cylinder arrangement that contains a gas at 2 atm. will, if allowed to interact with its surroundings at 1 atm., expand and transfer work energy until the system pressure is in equilibrium with the surrounding's pressure (i.e., 1 atm.). In a more theoretical sense, it captures the notion that only an orderly flow of energy (e.g., like that established because of a temperature gradient

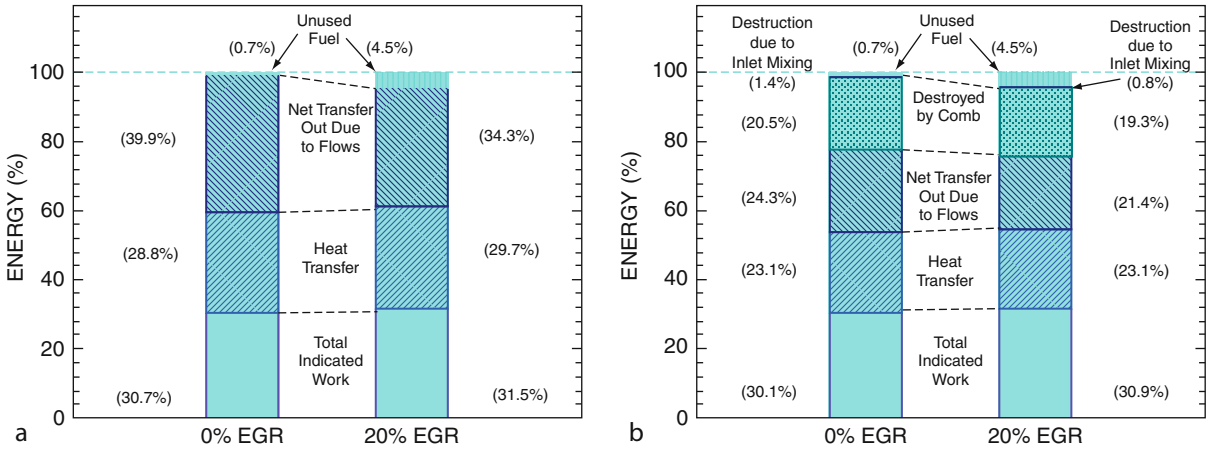
between a system and its surroundings) can be converted into useful work; disorderly energy cannot be converted to useful work. Entropy is the idea of disorder within a system; thus, exergy analysis combines the effect of a system's entropy on its ability to convert its energy into useful work. Because real processes in net increase entropy (i.e., real processes are irreversible and result in entropy generation), systems undergoing real processes destroy exergy; that is, the opportunity to convert energy into useful work is destroyed via some disordering of what had been an orderly flow of energy.

Examples of real effects, or irreversibilities, either in a general thermodynamic system or between a general thermodynamic system and its environment (all of which happen to be present, also, in internal combustion engines) are heat transfer through a finite temperature difference, mixing, unrestrained expansion, combustion, and friction. The general premise of an irreversibility is one which renders a system/environment combination changed as the system undergoes a cycle and returns to its initial state. For example, if a warm body exchanges thermal energy with a cold body (i.e., heat transfer through a finite temperature difference), there is heat transfer, but not work transfer, until the two bodies are in thermal equilibrium. To return the two bodies to their initial states (i.e., transfer thermal energy from the cold body to the warm body), work energy must be transferred into the system to effect the heat transfer. In the forward process (i.e., the bodies attaining thermal equilibrium), heat transfer occurs without work transfer; in the reverse process (i.e., the bodies returning to their initial states), heat transfer is manifested by a necessary work transfer. A net effect has been made to the surroundings (transfer of work energy) as the system is returned to its initial state; thus, the process is irreversible. If, on the other hand, a device had been positioned between the two bodies of dissimilar temperature that was able to extract the orderly sense of energy motion into useful work (i.e., a heat engine), the irreversibility of the heat transfer through finite temperature difference is diminished. Further, if the heat engine is conceived to be ideal itself, then all the orderly sense of energy motion is converted into useful work; hence, the concept of the Carnot heat engine, or, fully reversible heat engine.

On the one hand, the heat transfer is viewed as a loss; on the other hand, it becomes apparent that there may still be an opportunity to extract useful work from the heat transfer. It is the latter supposition that exergy analysis centers on. To begin a discussion of energy and exergy analysis of an internal combustion engine, a revisit to the assumptions made in the *Fuel–Air Cycle Analysis* is necessary. First, the assumption that the engine is adiabatic is not realistic; all real engines have heat transfer. Thus, it becomes necessary to capture an understanding of the effects of heat transfer on the engine cycle. Second, the assumption of how combustion is modeled is inaccurate. Of course, constant volume combustion in an operating engine is not physically possible due to the finite time required by chemical kinetics to decompose the fuel molecule. Even constant pressure combustion is not realistic because it supposes a constant burn rate with instantaneous starts and ends of combustion. Thus, it becomes necessary to include the effects of "real" combustion on the engine cycle. Third, if it is desired to know the actual torque from the engine (as opposed to what is indicated by the pressure–volume relationship), then a sense of friction must be included in the analysis.

The inclusion of "real effects" in engine cycle analysis is well established. For example, comparisons between actual engine cycles and *Fuel–Air Analysis Cycles* are made by Edson and Taylor [38]. Inclusion of heat transfer, finite combustion rates, and mechanical losses is demonstrated by Strange [39]. The beginnings of detailed combustion modeling, including spatial location of start of combustion and flame propagation, are demonstrated by Patterson and Van Wylen [40]. Such early demonstrations, and the corresponding advancement in digital computers, result in advanced models of heat transfer (e.g., [41–43]), friction [44, 45], and combustion; the latter of which are described in the context of doing a complete engine cycle simulation (e.g., [44, 46–51]). The influences of "real effects" on both energy and exergy perspectives of internal combustion engines are also well established (see, e.g., the reviews by Caton [52] and Rakopoulos and Giakoumis [53]). For example, Fig. 13 illustrates both the energy (Fig. 13a) and exergy (Fig. 13b) distributions of using 0% and 20% adiabatic EGR in a spark ignition engine, as computed by a thermodynamic simulation [54]. Note that when real factors, such as heat transfer and

I



Internal Combustion Engines, Developments in. Figure 13

(**a**) Energy distribution and (**b**) exergy (availability) distribution of various transfer and destruction (in the case of exergy) mechanisms for 0% EGR and 20% EGR in a spark ignition engine modeled using a thermodynamic simulation [54] (Used with permission from Professor J. Caton, Texas A&M University)

real combustion, are considered, the maximum indicated fuel conversion efficiency is about 30.7% for non-EGR case (0% EGR) and 31.5% for 20% EGR case. Most of the balance of energy is transferred to the coolant (ca. 29% for non-EGR case) due to nonadiabatic control system and out the exhaust (ca. 40% for non-EGR case) due to underutilized conversion of thermal to work energy. Thus, the importance of considering real effects is clear. It is noted that the data shown in Fig. 13 do not include the effect of friction; the brake fuel conversion efficiency for the studied conditions of Fig. 13 are available in [54]. It is further noted that friction offers no opportunity for conversion to useful work (i.e., energy lost to friction is all destroyed exergy). Finally, it is noted that EGR seems to have a small to negligible effect on the calculated friction of the simulation [54].

Also clear from Fig. 13 is the importance of considering the internal combustion engine from an exergy perspective. If considering only the energy transfers of Fig. 13a, it may appear that ca. 70% of the energy is "lost," or unavailable for conversion to useful work. While it is correct to say that a fraction of the fuel's energy is underutilized in the engine control system's (i.e., piston/cylinder chamber) conversion to useful work, it is not correct to say that all this energy is lost. As explained above, exergy provides insight into the fraction of energy that is available to do useful work. For example, Fig. 13 illustrates that (for non-EGR cases) 28.8% of the fuel's *energy* is transferred out of the control system via heat transfer, whereas only 23.1% of the fuel's *exergy* is transferred out of the control system via heat transfer. That is, for one unit of fuel, 0.288 unit of energy is transferred via heat transfer, but only ca. 80% (0.231/0.288) of that energy is available to do useful work; thus, 0.231 unit of fuel exergy are a missed opportunity for conversion to useful work. This 0.231 unit of missed opportunity could be exploited by configuring an ideal heat engine between, for example, the cylinder walls and the environment (i.e., representative T_H and T_L). Such a conceptualization seems impractical, but exergy is not just passed into the coolant. Exergy is also transferred to the exhaust; ca. 24.3% of the fuel's exergy for the non-EGR case under study in Fig. 13b. The high temperature exhaust, with ease of accessibility, offers an opportunity to exploit the fuel's exergy transferred through the exhaust system. There are several practical technologies that offer an opportunity to extract the fuel's exergy transferred out of the engine control system; these are discussed in more detail in the section "Waste Heat Recovery".

One of the important features of Fig. 13b is the exergy destroyed due to combustion. Combustion, in any application, is an irreversible phenomenon and thus will render entropy generation (i.e., exergy

destruction). Like friction, the ca. 20% of fuel exergy destroyed due to combustion cannot be recovered for conversion to useful work. The major causes for exergy destruction of a typical hydrocarbon-based combustion process include (1) thermal energy exchange among particles within the system, (2) diffusion among fuel/oxidizer particles, and (3) mixing among product species [55]; of these three, the dominant exergy destruction source is the thermal energy exchange among particles. At the onset of combustion, with the system containing reactants, a metastable equilibrium exists as gradients exist between fuel and oxidizer molecules. As combustion proceeds during the process, natural phenomena cause the system to eliminate such gradients and maximize entropy in the pursuit of attaining a more stable equilibrium (i.e., the products state). The maximization of entropy in reducing the gradients is manifested entirely through entropy generation. Using the dominant source of exergy destruction (i.e., entropy generation) during combustion as an example: internal thermal gradients among particles [56] established during the combustion process create microscopic opportunities to do useful work (i.e., thermal gradients can be converted to useful work via a heat engine). Of course, practical implementations of microscopic heat engines do not exist, thus, the thermal gradients are reduced to zero during the combustion process without any conversion to useful work; heat transfer through a finite temperature difference, as described above, generates entropy (destroys exergy).

The general behavior of exergy destruction due to combustion in internal combustion engines is generally well characterized, as summarized by [57]; generally, an increase in combustion temperature decreases exergy destroyed due to combustion. This statement should not be confused with the situation of internal thermal gradients which, as mentioned above, are a major source of combustion-based exergy destruction. While it is tempting to seek ways to minimize (or even eliminate) combustion irreversibilities (and thus minimize or eliminate combustion-based exergy destruction), it should be recognized that such efforts may decrease the conversion of exergy in other ways. Consider, for example, the use of a lean mixture versus the use of a stoichiometric mixture; the former has higher exergy destruction than the later. While a stoichiometric mixture reduces exergy destruction, its mixture composition also causes lower conversion of thermal to work energy during the expansion process (and, coincidentally, increases exergy transfer via heat transfer). Thus, overall efficiency is lower with a stoichiometric mixture. Again, energy- and exergy-based analyses, such as that presented in Fig. 13, are necessary to quantitatively make such assessments. In the case of eliminating combustion irreversibilities, a reversible combustion process [58] can be conceptualized [59, 60] where initially separated reactant species are isentropically compressed to their respective partial pressures and a certain temperature; concentrations of individual species, when allowed to interact with each other, will be in equilibrium. After compression, individual species will be collected forming a mixture that is a priori in equilibrium (thus, there are no irreversibilities due to mixing) and allowed to expand isentropically. Because of the increase in moles of the mixture as it expands isentropically maintaining equilibrium along the path, net work is extracted from the collection of processes. Quantitative analysis [60] of such a concept reveals 0% exergy destruction; because of exergy retention in the exhaust products, however, thermal efficiency of the "reversible combustion engine" is around 28% (for a fuel–air equivalence ratio of 1.0, compression pressure and temperature of 10 MPa and 6,000 K, respectively, and an expansion ratio of 18:1). Of course, such concepts are presently impractical; but the notion of reversible combustion is theoretically possible.

In fact, it is briefly noted that, although not explicitly calling his concept one of "reversible combustion," a close inspection of Diesel's original engine design [61] describes a method of nearly attaining reversible combustion (at least, one which eliminates internal thermal gradients). This is described in more detail in section "A Case Study: Diesel Engines Versus Gasoline Engines".

In closing this section on "Thermodynamic Analysis of Internal Combustion Engine", it is noted that assigning a "maximum possible efficiency" of an internal combustion engine is anything but straightforward. *Fuel–Air Cycle Analysis* reveals that the efficiency of an engine cycle depends on several different characteristics of the control system. A more advanced and appropriate analysis of an engine – that which includes real

effects of heat transfer, real combustion, friction, and other irreversibilities – gives realistic senses of tangible factors that engineers can strive to improve. Further, exergetic-based analysis offers the important insight of what future opportunities might be exploited in improving the overall system efficiency associated with an internal combustion engine. In returning to the engineer's "dream" engine concept described above, such computed numbers as shown in Fig. 13 are more realistic than the ideal *Fuel/Air Cycle Analysis* numbers shown in Fig. 10. Further, the exergy-based analysis of Fig. 13 could be used to make the necessary assessment of the engineer's dream, to quantify if there are real opportunities of "downstream" exergy conversion when the engine is operated close to maximum theoretical thermal efficiency levels.

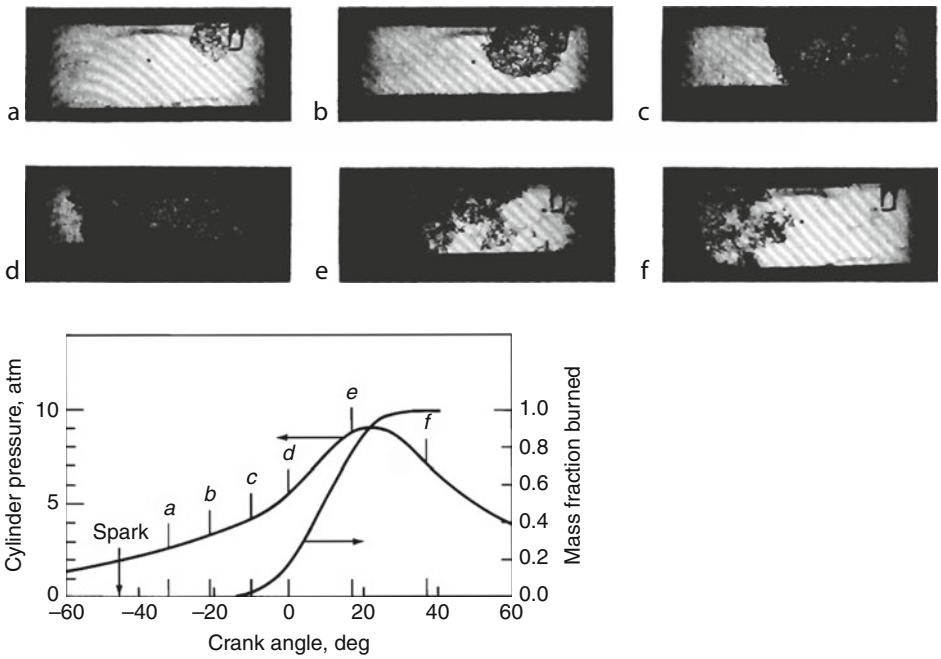
Spark Ignition Combustion

As discussed in section on "Thermodynamic Analysis of Internal Combustion Engines", real combustion in an internal combustion engine is not ideal in the sense that it can be modeled as precisely constant volume, constant pressure, or even the combination of the two (limited pressure). Real combustion, instead, involves several complex and interacting features that, on the one hand, make it difficult to predict and model, but on the other hand create opportunities for further development. Conventional engines are either spark ignited or compression ignited. Since the type of ignition results in substantially different combustion features, the two are separated into respective sections. This section describes spark ignition combustion.

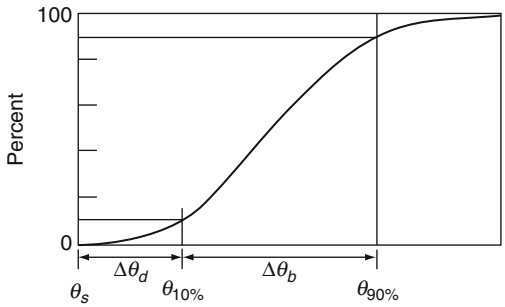
Spark ignition combustion is the typical form of combustion found in the commonly called "gasoline" engine (also commonly called either "petrol" engine or "otto" engine); it is more descriptive to refer to such engines as spark ignition engines. The "spark" aspect of the term implies that combustion is initiated by introducing a high voltage electrical arc, typically through the use of a spark plug, into the reactant mixture at a point during the cycle where sustained combustion reaction will proceed. Most spark ignition engines (exceptions are described in section on "Future Directions of Internal Combustion Engines") induct a premixed and homogeneous mixture of fuel and air during the intake process. At a point near the end of compression, the spark plug discharges an arc into the mixture. This arc, in the near-by region of the spark plug, forms a high temperature plasma that evolves into a flame kernel. The flame kernel transitions from the plug region in a laminar sense establishing a flame front with initial velocity close to the laminar flame speed. This initial flame development is referred to as the flame-development period. Because of turbulent motion among particles composing the mixture, the flame then rapidly spreads throughout the mixture in a turbulent fashion, enveloping microscale eddies with flame. This rapid flame propagation is referred to as the turbulent entrainment period. Upon flame envelopment of the turbulent eddies within the mixture, the flame then laminarly burns the microscale eddies in the final stage of the process referred to as rapid burnup. Turbulent entrainment and rapid burn practically occur simultaneously, and together compose what is referred to as rapid-burning period. This described sequence of spark ignition combustion for a typical operating point (i.e., 1,400 rev/min, part-load condition) is shown in Fig. 14 [62], and, for the same typical operating point, requires about 10 ms from spark to finish.

One of the important features of Fig. 14, along with the shown in-cylinder pressure, is the corresponding mass fraction burned. Using the first law of thermodynamics, it possible to develop an expression that determines the chemical-to-thermal energy conversion rate (commonly called the "heat release rate") of the combustion process. Since the mass of fuel present in the cylinder, along with the fuel's heating value, are known, the mass fraction burned rate of fuel can be calculated from the heat release rate. Figure 15 illustrates a typical mass fraction burned curve, as a percent of the total fuel, and defines the above-described periods referred to as flame-development period (shown in Fig. 15 as the flame-development angle, $\Delta\theta_d$) and the rapid-burning period (shown in Fig. 15 as the rapid-burning angle, $\Delta\theta_b$). These periods are conventionally defined as the angle swept from spark release to 10% mass fraction burned and the angle swept from 10% mass fraction burned to 90% mass fraction burned, respectively.

The mass fraction burned profile of spark ignition combustion can be modified, and the parameters that



Internal Combustion Engines, Developments in. Figure 14
In-cylinder images capture the sequence of spark ignition combustion with the corresponding result on in-cylinder pressure and mass fraction burned. The various steps of spark ignition combustion shown include the spark release (indicated by "Spark"), flame kernel development (**a**–**c**), turbulent entrainment and burn up (**d**–**e**), and flame termination (**f**) (Used with permission from [62])

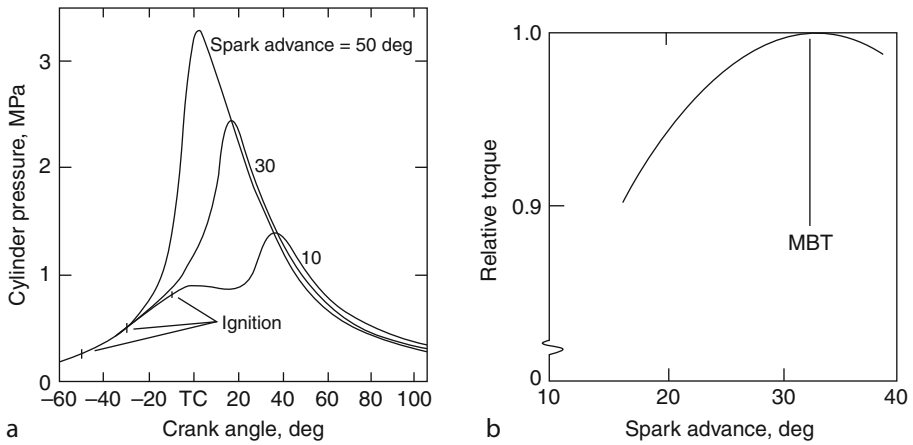


Internal Combustion Engines, Developments in. Figure 15
Mass fraction burn rate (as a percent of total fuel) of spark ignition combustion in an internal combustion engine, illustrating the definitions of the flame-development angle, $\Delta\theta_d$, and the rapid-burning angle, $\Delta\theta_b$ (Used with permission from [63])

cause alterations to the mass fraction burned profile are often coupled. Perhaps the most obvious parameter that can affect the mass fraction burned profile is the spark timing, or sometimes called the spark advance. By adjusting the time in the cycle when combustion is initiated (i.e., the spark advance), the times required for flame development and rapid-burning periods are altered. An advance in spark timing (i.e., spark timing is moved earlier into compression stroke) generally results in combustion occurring earlier in the cycle and in-cylinder pressures becoming higher in magnitude. Correspondingly, a retard in spark timing (i.e., spark timing is moved later into the compression stroke, or possibly even into the expansion stroke) generally results in combustion occurring later in the cycle and in-cylinder pressures becoming lower in magnitude. There are several other parameters, however, that can affect the burn profile of spark ignition combustion. Examples of such parameters include the mixture's fuel–air ratio (i.e., stoichiometry), the level of EGR, the level of turbulence in the mixture, the level of heat transfer, initial mixture pressure (i.e., load variation), initial mixture temperature, and engine speed. Analysis of each of these parameters is

outside the scope of this article; readers are referred to the "Books and Reviews" for additional reference text on the subject.

One item that is important to mention, however, is the effect of burn profile on engine-scale parameters such as performance, efficiency, and emissions. The rate of combustion, and its corresponding effect on in-cylinder pressure, flame temperature, and mixture gas temperature, alters the performance, efficiency, and emissions of the spark ignition engine. For example, Fig. 16 illustrates the effect of spark timing (spark advance) on in-cylinder pressure and relative torque to maximum brake torque (MBT). MBT is the maximum torque attained over a given parametric sweep, most commonly over a spark-timing sweep. Note that there is nearly a single spark timing that yields the maximum torque delivered by the engine. Too early of a timing (advanced) or too late of a timing (retarded) renders a lower torque than MBT. Three spark timings are shown in Fig. 16: 50°, 30°, and 10° before TDC (BTDC). The most advanced timing (50° BTDC) renders the earliest burn profile (of the studied timings), the highest magnitude of in-cylinder pressure, and the earliest location of peak pressure (occurring around 0° BTDC). Although it may appear from Fig. 16 that the spark advance of 50° BTDC yields the maximum area under the pressure-volume curve (and thus, the maximum work for the cycle), it should be noted that its pressure is lower through most of the expansion stroke than the other two shown spark timings (i.e., 30° and 10° BTDC). This lower pressure through the expansion stroke results from the increased level of heat transfer manifested by the higher gas mixture temperature caused by the faster burn rate-induced higher cylinder pressures. Thus, spark timings advanced of MBT timing result in lower torque due to higher levels of compression work, heat transfer, and friction. Conversely, the most retarded timing (10° BTDC) renders the latest burn profile (of the studied timings), the lowest magnitude of peak pressure, and the latest location of peak pressure (occurring around −40° BTDC). Because of the later-phased combustion, the location of the rise in pressure misses the full opportunity of the expansion stroke. Thus, spark timings retarded of MBT result in lower torque due to expansion losses. It is clear that MBT occurs in the balance of minimized compression work and maximized expansion work, both of which are affected by combustion phasing, heat transfer, and friction. In the example of Fig. 16, this balance occurs around 30° BTDC. Notice that peak pressure for this timing is around −20° BTDC; a general "rule of thumb" is that MBT occurs when peak pressure is positioned between −15° and −20° BTDC. As revealed in Fig. 16, this rule of thumb extends to the mass fraction burned profile, where maximum brake torque is timed when 50% of



Internal Combustion Engines, Developments in. Figure 16
(**a**) In-cylinder pressure as a function of engine crankangle and (**b**) relative torque to maximum brake torque (MBT) as a function of spark advance of a typical spark ignition engine (Used with permission from [64])

the fuel burns by −10° BTDC. In some cases, MBT timing may be "knock limited," meaning that a higher torque could be attained at an earlier timing if fuel knock were not present. Fuel knock is a combustion abnormality of spark ignition engines that, due to the combustion-generated compression of the reactive mixture in the end regions of the cylinder, results from autoignition of the mixture prior to its controlled burn by the propagating flame. Fuel knock can be very damaging because the uncontrolled combustion tends to cause dramatic rises in pressure near critical mechanical components (such as piston rings). The octane rating of a fuel indicates the fuel's resistance to knock in a spark ignition engine; a fuel with a higher octane has a higher resistance to autoignition. Additional discussion about this is provided in the sections. on "A Case Study: Diesel Engines Versus Gasoline Engines" and "Direct Injection, Spark Ignition Engines".

At MBT timing, since fuel flow rate is generally held constant during spark-timing sweeps, efficiency will also be correspondingly maximized. In the case of emissions, however, the trends are not straightforward and a brief discussion of emissions is reserved for the section on "Emissions Formation and Exhaust Pollution".

Finally, it is noted that several parameters affect the spark ignition burn profile. Thus, for each change to a given parameter (e.g., initial pressure, initial temperature, fuel–air equivalence ratio, engine speed, and level of mixture turbulence), a potentially different spark timing will correspond to maximum brake torque. Again, the general trend of spark timing will be such that 50% mass fraction burned occurs at −10° BTDC for maximum brake torque.

Compression Ignition Combustion

In contrast to spark ignition combustion – which uses an electrical arc, or spark, to initiate combustion of a reactive mixture – compression ignition combustion relies on compressive heating – or, the increase in temperature of a gas due to the increase in pressure resulting from the decrease in volume – where combustion initiation is kinetically driven by exceeding the ignition temperature of the fuel–air mixture. Compression ignition combustion is the typical mode of combustion used in the commonly called diesel engine, an engine which more descriptively is called a compression ignition engine.

It is immediately recognized that one challenge of compression ignition engines is the lack of a direct trigger of ignition; the spark acts as the direct trigger of combustion initiation in a spark ignition engine. Conventional applications of compression ignition engines (e.g., the conventional diesel engine) overcome this challenge by using the fuel injection event as the direct trigger. Thus, conventional compression ignition engines typically induct an air and residual mixture (i.e., no premixing of fuel) during the intake stroke. This same unreactive mixture is compressed during the compression stroke until near the point when combustion is desired to begin. At this point fuel is introduced into the mixture, which, due to compression, is a high temperature, high pressure environment. After several complex and coupled processes occur (described below), combustion initiates and chemical to thermal energy conversion takes place.

At this point, it is instructive to briefly describe the role of "glow plugs" often used in compression ignition engines. Glow plugs should not be confused with spark plugs. A glow plug is a resistive heating element inserted into the cylinder of a compression ignition engine to aid in the initial starting of the engine. It acts as a warming device during "cold start"; after engine warm up and stable operation, the glow plugs deactivate. They are not necessary on a cyclic-basis. Spark plugs, on the other hand, are integral components of spark ignition engines; they provide the source of ignition for every combustion cycle of a spark ignition engine.

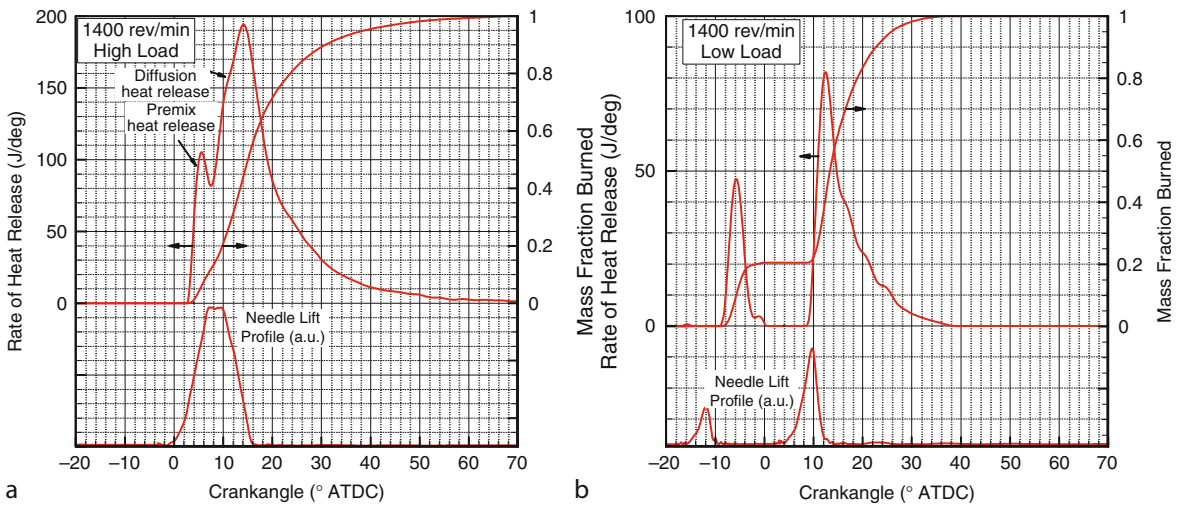
The method by which fuel is injected into the compressed mixture typically falls into one of two categories: (1) direct injection, or (2) indirect injection. Indirect injection involves injecting fuel into a "prechamber" which is connected to the main chamber. Fuel injected into the prechamber ignites and causes the mixture to issue into the main chamber where main heat release and work extraction occurs. Indirect injection engines are used primarily in applications where motion of the nonreactive mixture in the main chamber is too quiescent for ignition; this typically occurred in the early implementations of small, high speed automotive compression ignition engines. Use of a prechamber, where nonreactive mixture forced

through orifices or nozzles connecting the main chamber to the prechamber suitably increased mixture swirl and turbulence, generated sufficient fluid motion to allow ignition to occur. Technological developments in intake port design, combustion chamber design, and fuel injection systems have rendered indirect injection systems nearly obsolete and technically inferior to direct injection techniques.

Direct injection, as the name implies, indicates that the fuel is injected directly into the cylinder at the point near when combustion is desired to begin. Because of the compressed nature of the nonreactive mixture into which fuel is injected, fuel injection pressures must be very high; i.e., modern fuel injection systems inject fuel at pressures on the order of 100–2,000 bar (ca. 3,000–30,000 psi). The processes of fuel injection, combustion initiation, and subsequent combustion propagation (or burning) are very complex in compression ignition engines. The following discussion will briefly highlight these complexities.

Figure 17 illustrates the sequence of events that occur in a typical direct injection compression ignition engine. Two load conditions at 1,400 rev/min are shown, with a high-load condition (ca. 75% peak load which is about 11.3 bar BMEP) shown as Fig. 17a and a low-load condition (ca. 25% peak load which is about 1.9 bar BMEP) shown as Fig. 17b. Both plots illustrate the fuel injector needle lift profile (which roughly correlates to the fuel delivery rate), the rate of heat release profile, and the mass fraction burned profile. Focusing first on the high-load condition (Fig. 17a), notice that fuel injection occurs near 2° BTDC. Fuel in typical compression ignition engines is injected as a liquid. Thus, the fuel must undergo a series of physical processes (e.g., penetration, breakup, atomization, and vaporization) before it undergoes its chemical process of bond fragmentation and eventual ignition. This period of time – i.e., the time between start of fuel injection and ignition (commonly called start of combustion) – is often called the ignition delay period. As will be described below, the ignition delay period of a compression ignition engine is an important parameter that affects the remainder of the burn profile and is affected by several other parameters. The end of the ignition delay period after start of injection will witness start of combustion and noticeable heat release, as indicated in Fig. 17a as the positive rate of heat release. During heat release, the mass fraction burned profile steadily increases until end of combustion when all the fuel is most nearly completely burned.

An interesting and important feature of Fig. 17 is the rate of heat release profile between start of



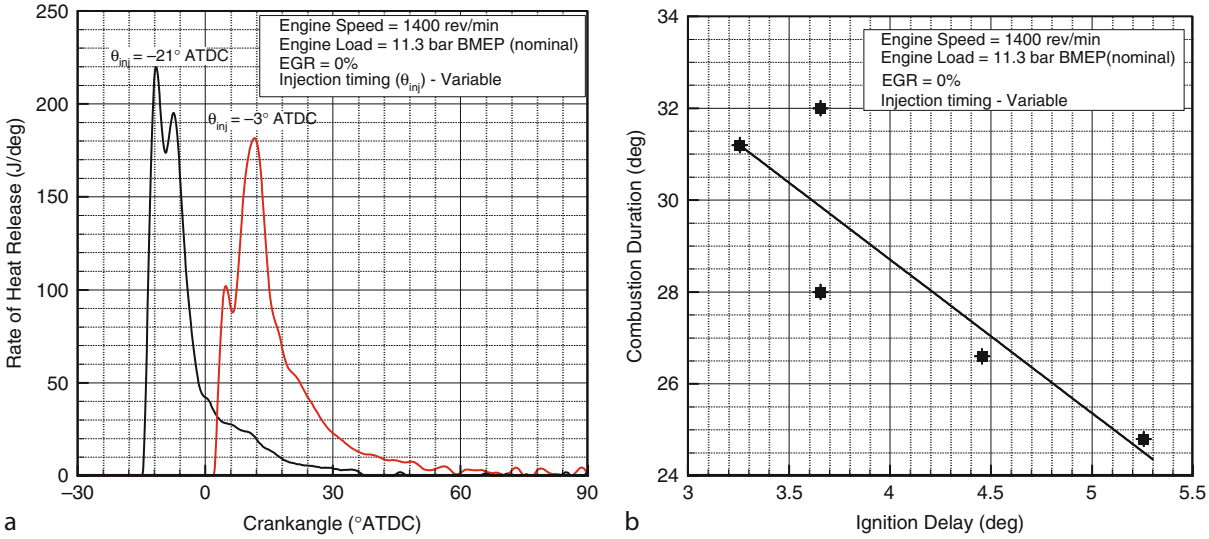
Internal Combustion Engines, Developments in. Figure 17
Fuel injector needle lift profile, rate of heat release, and mass fraction burned for a typical diesel engine operating at 1,400 rev/min (**a**) high-load condition (ca. 75% peak load = 11.3 bar BMEP) and (**b**) low-load condition (ca. 25% peak load = 1.9 bar BMEP) with the use of pilot injection (Data from author's laboratory, Texas A&M University)

combustion and end of combustion. Notice in Fig. 17a there are two distinct components of the rate of heat release: a premixed component and a diffusion component. These two respective components are nearly an exclusive feature of compression ignition combustion that relies on fuel injection as the ignition trigger. The premixed heat release component results from the complex processes involved in physically preparing the initial "packets" of fuel for chemical decomposition [65]; as described above, the injected liquid fuel must be prepared and sufficiently mixed with the cylinder gases before chemical decomposition can begin. Once ignition occurs, the parcels of fuel that prepared for burning during the ignition delay period react under kinetic control, with relatively high rates of burning for the given mixture temperature. Fuel continues to inject during the premixed portion, and the increase in temperature due to premixed combustion accelerates the physical preparation of the newly introduced parcels of fuel. In spite of this acceleration of fuel preparation, the burn rate becomes limited by the fuel's mixing rate with the air. Since the time scales for mixing are larger than the time scales for chemical kinetic decomposition [66], the burn rate becomes mixing-controlled, rather than kinetically controlled. The mixing-controlled, or diffusion-controlled, component of heat release is observed in Fig. 17a as the longer heat release following premixed heat release and is identified as "diffusion heat release." The burn rate of diffusion heat release is generally slower than premixed heat release at any given reaction temperature. Figure 17a shows diffusion heat release having a much higher burn rate than premixed heat release; this is manifested, however, by the associated higher mixture temperature effected by the combustion process.

The situation is slightly different for the low-load case, shown as Fig. 17b. Conventional compression ignition engines that use fuel injection as the ignition trigger typically also use the amount of fuel delivery as the means to control engine load. Thus, the need to produce less power at a given engine speed is effected by decreasing the amount of fuel delivered during the injection process. This is evident in Fig. 17b where the area under the fuel injector needle lift profile is less than that of the high-load case (for the moment, disregard the bimodal fuel injection profile, which will be explained below). Consequently, the distribution of premix versus diffusion heat release is not apparent on the low-load heat release profile. This does not imply that mixing-controlled combustion does not occur during low-load operation; it does imply, however, that combustion is predominantly kinetically controlled at low-load conditions whereas it is predominantly diffusion-controlled at high-load conditions.

The bimodal feature of the fuel injection profile in Fig. 17b is the result of the use of a double-injection strategy on this particular engine at this particular condition. Modern advanced engine fuel systems are capable of introducing fuel in sequences, or pulses, in what is commonly called "multi-injection strategies." In the case of the engine highlighted in Fig. 17b, the first injection is considered a "pilot" injection, while the second injection is considered a "main" injection. The pilot injection introduces only a small portion (e.g., ca. 10%) of the total fuel to be delivered in the cycle, creating an opportunity to prepare a smaller portion of fuel for burning during the ignition delay period. This small pilot increases cylinder temperature, so that the ignition delay period of the main injection is much shorter, and there correspondingly is a lower fraction of premixed heat release during the main combustion event. Such a strategy is used, for example, to reduce combustion-generated noise of a diesel engine. Interestingly, diesel engine noise is an "age-old" problem, serving as a primary focus of development for Sir Harry Ricardo (a pioneer in the development of high speed diesel engines) [67].

The above-described use of multiple injections hints at an ever-present relationship within compression ignition combustion systems using fuel injection as the direct trigger; i.e., a relationship among ignition delay, fraction of premix heat release, and fraction of diffusion heat release exists. This relationship is shown in Fig. 18a, which illustrates the rates of heat release at two different injection timings of a typical compression ignition engine operating at 1,400 rev/min, high-load condition (11.3 bar BMEP, nominal). Notice for the advanced injection timing there is a substantial level of premix heat release and a correspondingly lower fraction of diffusion heat release relative to the retarded injection timing. The advanced injection timing, due to the injection of fuel into a relatively cool environment when rates of physical and chemical processes are low, results in a relatively longer ignition delay. Because of



Internal Combustion Engines, Developments in. Figure 18
(**a**) Rate of heat release at an advanced injection timing (21° BTDC) and a retarded injection timing (3° BTDC) and (**b**) the relationship between ignition delay and combustion duration via an injection timing sweep for a typical compression ignition engine operating at 1,400 rev/min, high-load condition (11.3 bar BMEP, nominal) (Data from author's laboratory, Texas A&M University)

the longer ignition delay, there is more time for fuel preparation prior to ignition; once ignition occurs, there is a relatively higher concentration of fuel that is prepared for burning and correspondingly combusts in a premix and kinetically controlled fashion. As more fuel is prepared during the ignition delay period for premix burn, there consequently is less fuel available for diffusion heat release; thus, as premix burn fraction increases, diffusion burn fraction decreases. Likewise, at the retarded injection timing, fuel is injected in a relatively hotter environment which enables fuel preparation for burning at a faster rate and combustion commences relatively sooner (i.e., shorter ignition delay). Because of the shorter ignition delay, there is less fuel prepared for burning, thus there is lower fraction of premix heat release and correspondingly larger fraction of diffusion heat release.

Because premix heat release rate is relatively faster (at a given temperature) than diffusion heat release, the "combustion duration" – or, the time taken between start of combustion and end of combustion – for a mostly premix heat release combustion event will be shorter than that of a mostly diffusion heat release combustion event. Thus, there tends to be an inverse relationship between ignition delay and combustion duration, as shown in Fig. 18b, where changes to ignition delay are manifested by alterations to injection timing. The shorter ignition delay results in less premix heat release, more diffusion heat release, and correspondingly a longer combustion duration.

Injection timing is just one parameter that can affect ignition delay, and thus the burn profile of a compression ignition combustion event when fuel injection is used as the ignition trigger. Other engine parameters such as EGR level [68], initial temperature and initial pressure [69], injection pressure [70–72], and swirl and turbulence [69, 73] all have certain effects on ignition delay and the resulting relative fractions of premix and diffusion heat release. Like the effect of various parameters of spark ignition engine performance, efficiency, and emissions, the various effects of compression ignition engine parameters on ignition delay and burn profile also have an effect on engine efficiency, performance, and emissions. For example, overly advanced or retarded injection timings will cause decreases in engine torque for a given fuel delivery rate. Additionally, fuel quality is particularly important for conventional diesel combustion operation, where low

volatility fuels with high ignitability are generally used to ensure ignition occurs during the cycle. An important fuel parameter – i.e., its cetane number – is used to identify the quality of fuels appropriate for use in a diesel combustion system. A higher cetane corresponds to a fuel with a shorter ignition delay.

Finally, in closing, it is important to recognize the above-described phenomena are largely phenomenological observations that can be made using conventional diagnostics with relatively straightforward analysis. Considerable development (e.g., [49, 74, 75]) has been made to provide substantial insight into the complex fluid, heat transfer, and chemical processes that occur during compression ignition combustion where fuel injection is used as the direct ignition trigger. Description of these details is outside the scope of this work.

Emissions Formation and Exhaust Pollution

A discussion on the basics of internal combustion engines, actually any combustion-based device, is incomplete without a description of the associated harmful species that may exist in the products of combustion. Such harmful species are generally called "exhaust pollution" and many governing agencies around the world place restrictions on the emission of certain pollutants from combustion-based devices. Because the application of the internal combustion engine is so varied, emission regulations tend to be application-oriented. For example, in the USA, emission regulations are placed on passenger vehicles differently than emission regulations placed on heavy trucks or hand-held engine devices (such as lawn mowers). Because of such variability in regulation, the time-oriented nature of the regulations (i.e., regulations have, to this point, been in a state of flux), and variation in regulation among various governing agencies around the world, further description of specific emissions will not be provided. The basic issues at hand, however, can be briefly described.

As shown in Reaction (R1), there are several species formed during the combustion reaction of a typical hydrocarbon–air mixture. Some of these species are generally stable and nonreactive in the atmosphere, thus pose little to no harm to the five kingdoms of nature. Other species, however, are either harmful or reactive in ways that lead to harmful consequences. Specifically, there is little scientific debate about the harmful nature of certain combustion products such as CO, NO/NO_2, unburned hydrocarbons (HC), and particulate matter (PM, or the solid/liquid components of exhaust that can be collected on a filter. It is noted that historically sulfur oxides and sulfates have been considered either separately from [76] or in combination with [77, 78] particulate matter). There is, perhaps, continued debate about the potential consequences of other products of combustion; in particular, there is current debate on the consequence of CO_2 and the role it may play in the presently observed warming of the planet (i.e., so-called global warming or global climate change). Because of the certainty of the effects of the former species, attention will be given to them and basic information on their formation during combustion. Because of the uncertainty of the effects of the latter species, readers are referred to other literature to uncover the current state-of-debate of CO_2 and its potential impact on global trends currently believed to occur (see, e.g. [79, 80]).

The first such species to describe is CO, which due to its fatal effects on human/animal life is one of the first combustion products to be considered a pollutant [76]. It is well established [76] that all precursor hydrocarbon decomposition reactions firstly form CO. Thus, increased concentrations of CO in engine exhaust result from incomplete reaction of the principal CO oxidation step [26], given as Reaction (R2):

$$CO + OH \leftrightarrow CO_2 + H \qquad (R2)$$

In internal combustion engines, the incomplete oxidation of CO most typically occurs during rich engine operation [76], where available oxidants for final CO oxidation are lacking [81]. Fuel–air mixtures close to stoichiometric or even slightly lean, however, result in nonnegligible concentrations of CO.

The next species to consider is NO. NO emerged as combustion-generated pollutant due to its observed effect of reacting with hydrocarbons in the presence of sunlight to produce tropospheric ozone [82]. Its formation in a combustion system is rather complex, as there are several major pathways through which it can form. These major pathways include thermal (or commonly called Zeldovich), prompt, and fuel-based nitrogen [26]. For reciprocating-type internal

I

combustion engines, the primary NO formation mechanism is the thermal mechanism where atmospheric air serves as the principal source of nitrogen in the mechanism [27]. There are three reactions that compose the mechanism, given as Reactions (R3)–(R5):

$$O + N_2 \leftrightarrow NO + N \tag{R3}$$

$$N + O_2 \leftrightarrow NO + O \tag{R4}$$

$$N + OH \leftrightarrow NO + H \tag{R5}$$

The forward and reverse reaction rate constants of Reactions (R3) and (R4) are generally exponentially dependent on temperature. To demonstrate the substantial role temperature plays on NO formation, several simplifying assumptions (see below) are applied to Reactions (R3) and (R4) to yield Eq. 15 [26]:

$$\frac{d[NO]}{dt} = \frac{6 \times 10^{16}}{T^{\frac{1}{2}}} \exp\left(\frac{-69090}{T}\right) \left[O_{2,eq}\right]^{\frac{1}{2}} \left[N_{2,eq}\right] \cdot \left(\text{mol/cm}^3\text{-s}\right) \tag{15}$$

where [NO] is the concentration (mol/cm^3) of NO at time, t (s), and $[O_{2,eq}]$ and $[N_{2,eq}]$ represent the equilibrium concentrations (mol/cm^3) at temperature T (K) of oxygen and nitrogen, respectively. The several simplifying assumptions that go into Eq. 15 include the following. The first assumption is that the nitrogen chemistry is de-coupled from the combustion reactions. Although combustion reactions generally occur much faster than nitrogen chemistry [25], the presence of O and OH radicals in the thermal mechanism (which are also important species in combustion reactions) may require the chemistries to be coupled for accurate NO prediction [27]. By assuming the chemistries are de-coupled, O, O_2, OH, H, and N_2 can be approximated by their equilibrium concentrations at equilibrium temperature; assuming equilibrium temperature is the second assumption applied to Eq. 15. The third and last assumption applied to Eq. 15 is that nitrogen radical (N) is in steady-state concentration (i.e., $\frac{d[N]}{dt} = 0$). Finally, it is reinforced that the forward reaction rate constant of Reaction (R3) is used from [26] in Eq. 15; updated reaction rates are available in Dean and Bozzelli [29]. It is clear from Eq. 15 the strong dependency NO formation rate has on temperature.

Because of the dominance of the thermal mechanism on NO formation in internal combustion engines, combustion-based efforts to reduce NO center on reducing the reaction temperature and O_2 concentration. Such techniques include altering spark advance [83, 84] or injection timing [85–87] for spark ignition or compression ignition engines, respectively, and introducing EGR into the mixture [87–90].

As described above with NO, unburned hydrocarbons (HC) play a role in the formation of tropospheric ozone. Further, they represent unreacted fuel; if left to discharge into the atmosphere, there is lost opportunity to convert that chemical energy into useful work. Formation of HC species during combustion reaction are complex and varied depending on the type of fuel used [91]. A general cause for HC emissions, however, is insufficient mixing between fuel and air [76], where most of the HC emission species are formed during low temperature ($T < 1{,}000$ K) reactions [26]. Typically, in conventional reciprocating internal combustion engines, most of these species are oxidized as combustion enters high temperature mechanisms [26]. Reciprocating engines, however, contain "sources" for hydrocarbon storage that, if not oxidized upon release in the gas mixture in the later portion of the cycle, emit as HC emissions in the exhaust [92]. Although there are special considerations given to engines operating under cold-start conditions [93], the general storage locations of HC species (and thus, the major source of HC emissions [94]) include cylinder head gasket crevice, spark plug crevices, piston ring pack crevices, and valve seat crevices. In addition to crevice HC storage, other sources of HC emissions [94] include single-wall flame quenching, oil film layers, combustion-chamber deposits, exhaust valve leakage, and liquid fuel (i.e., HC species not vaporized during the process). Because of the relative importance of crevice storage on HC emissions, much of the combustion-based effort to decrease HC emissions has centered on reducing the volume and flow pattern of crevices in the piston/cylinder arrangement.

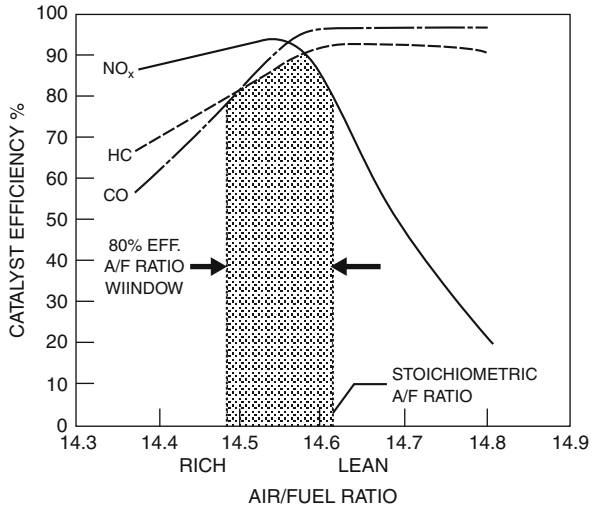
The final major pollutant to briefly describe is particulate matter, or PM. PM is essentially any exhaust specie that can be collected on a filter; it typically is structured on a solid organic component (which is mostly pyrolized carbon particles, or "soot") upon which organic (e.g., unburned HC) and nonorganic (sulfates) components build. Since soot serves as the building block for PM, much of the research efforts are

dedicated to understanding soot formation processes [95–99]. The general path for soot formation begins with the high temperature pyrolysis, or fragmentation to hydrocarbon radical, of the hydrocarbon fuel. Such pyrolysis typically leads to a certain group of hydrocarbon radicals called polyaromatic hydrocarbons. Such polyaromatic hydrocarbons that do not oxidize during the reaction serve as the nucleation site for soot growth. Nuclei then begin to coalesce and substantially increase surface area. Now particles, the high surface area soot particles agglomerate and allow other species (i.e., liquid HC, condensed gaseous HC, and sulfates) to absorb onto the surfaces [99]. Soot formation is strongly dependent on mixture fuel–air ratio and also temperature (both tend to govern the rate of pyrolysis). Formed soot can, however, undergo oxidation as well. Soot oxidation is likewise a function of temperature, but has a stronger dependency on temperature than soot formation [99]. The difference between soot that is formed and soot that oxidizes is ultimately released in the engine exhaust, and is commonly called "net soot release." Compression ignition engines (e.g., diesel engines) are often plagued with high PM emissions, largely due to the heterogeneity induced into the fuel/air mixture by use of fuel injection as a direct trigger for ignition. Spark ignition engines, however, are recently being considered as sources of nanoscale soot particles [100, 101].

It is clear from the above discussion that soot and PM emission processes are complex. From phenomenological considerations, however, Khan et al. [86, 102] and Ahmad [103] provide insight into the general behavior of PM with, in particular, diesel engine combustion. Specifically, it is observed that increases in diffusion heat release generally increase the emissions of PM. A portion of this may be due to the overall lower reaction temperatures, manifested by relatively lower premix heat release, decreasing the rate of soot oxidation through the diffusion flame. This observation is also supported, however, by Dec's [74] observation that precursor soot formation occurs in the standing premix reaction zone within a diffusion flame sheet that exists in diesel combustion. Thus, increased diffusion burn correspondingly results in increased soot emission. Either way, it becomes apparent when relating this discussion to the known effects of injection timing, for example, on diffusion combustion and NO emissions that an attempt to decrease soot by creating higher temperatures with, perhaps, more premix heat release will correspondingly increase NO emissions; hence, the establishment of the conventional "soot-NO_x" tradeoff of diesel engines.

As described briefly above with each species, there are in-cylinder and combustion-based methods to reduce the formation of various exhaust pollutants. Advanced development of combustion systems continues to focus on these in-cylinder methods, as described in section on "Future Directions of Internal Combustion Engines". Also in use to eliminate exhaust pollution are exhaust after treatment systems, which conventionally are catalyzed devices (hence, their common name of "catalyst" or "catalytic converter") [104–106]. The basic idea of a catalyst is to promote a reaction that otherwise would not proceed. In aftertreatment of engine exhaust, there are several catalysts in use depending on the species composition of the exhaust. For example, the conventionally named "three-way catalyst" is often used with conventional gasoline spark ignition engines. Such a catalyst is typically an exhaust flow-through device composed of a ceramic monolith (the substrate) with thru-hole passages to allow exhaust gases to flow and a metal-oxide "washcoat" that suspends catalytic particles on the surfaces of the monolithic passages. The monoliths are usually constructed of coerderite and the metal-oxide washcoat is often an alumina washcoat. Catalyzed particles are often from the precious metals group (e.g., platinum, palladium, and rhodium), with platinum being the most commonly used metal. The general operation of the three-way catalyst is to promote the reduction–oxidation reactions among CO and HC (the reductants) and NO/NO_2 (the oxidants). In other words, the catalyst promotes the reduction of NO or NO_2 to yield stabilized N_2 and the oxidation of CO and HC to yield stabilized CO_2 and H_2O. The efficiency of the three-way catalyst – or, the conversion effectiveness of converting a given species to its more stable species, e.g., CO to CO_2 – is strongly dependent on the constituent composition on the inlet mixtures, as shown in Fig. 19 [104]. Notice that maximum conversion efficiencies of NO_x, CO, and HC occur very near stoichiometric air–fuel ratios; a small departure from stoichiometric conditions – e.g., about 0.5% increase or decrease in A/F ratio – results in a nearly

I



Internal Combustion Engines, Developments in. Figure 19
Catalyst efficiency of a typical three-way catalyst interacting with exhaust from a typical gasoline spark ignition engine as a function of mixture air–fuel ratio (A/F ratio) (Used with permission from [104])

20% decrease in conversion efficiency. Because of such an intolerable response of the catalyst, effective control of gasoline spark ignition engines requires precise control of the mixture stoichiometry during operation.

Also clear from Fig. 19 is the challenge of outfitting a typical platinum-based catalyst with non-gasoline spark ignition engine technology. Diesel engines which typically operate fuel lean (i.e., oxygen rich), for example, create constituent exhaust species that are difficult to catalyze using conventional techniques. Advanced aftertreatment technologies for diesel engines, or other advanced combustion/engine systems, are under development and beginning to appear in production applications [105, 106].

A Case Study: Diesel Engines Versus Gasoline Engines

At this point, with the basics of internal combustion engines having been described, it is useful to do a "state-of-the-art" technology comparison between the two dominant conventional internal combustion engines – i.e., the gasoline spark ignition engine and the diesel compression ignition engine – to set the stage for discussions on future directions in internal combustion engines. The elementary comparison is provided in Table 2; note that this comparison is neither comprehensive nor general. It is intended to highlight the common state-of-the-art of the two technologies, and create a sense for the trends of future directions in internal combustion engine development.

It is clear from Table 2 that the two engine types differ in virtually every way, save for their common use of the kinematic elements of the crank-slider and piston/cylinder components. Consider first the spark ignition engine; its use of spark requires a relatively tight control on the fuel–air mixture equivalence ratio. The flame travel time is minimized when fuel–air equivalence ratio is near stoichiometric, and greater than about 20–40% departure from stoichiometric will typically cause misfire [107]. It is noted that this requirement [107] is only for ignition and flame propagation. The integration of a three-way catalyst with a conventional spark ignition engine requires even more precise control over fuel equivalence ratio, as described in section on "Emissions Formation and Exhaust Pollution" [104]. Thus, conventional spark ignition engines are not able to tolerate substantial departures from stoichiometric conditions. Further, because of this requirement, it is necessary to avoid mixture striations between fuel and air. Thus, conventional spark ignition engines make use of premixing devices (e.g., carburetors or throttle body and port fuel injectors) to prepare the fuel–air mixture prior to induction. The result is a homogeneous mixture of fuel, air, and other residuals such as residual fraction and purposefully introduced constituents such as EGR. In order to assist with the homogenization of the fuel–air mixture, light distillates and aromatics with high volatility are typically used as the fuel for conventional spark ignition engines. The most common, of course, is the mixture of light distillates and aromatics commonly called "gasoline" or "petrol." Interestingly, largely due to the reactive nature of the homogeneous and near-stoichiometric mixture [108], compression ratios of spark ignition engines are limited to around 11. Compression ratios higher than this promote the undesirable autoignition of certain regions of the fuel–air mixture during the combustion process; a phenomenon known as "fuel knock." Fuel knock can be very damaging to an engine, mostly due to its occurrence in the end regions of cylinder, where

Internal Combustion Engines, Developments in. Table 2 Elementary comparison between conventional spark ignition engines and conventional compression ignition engines. The comparison is not comprehensive nor general, rather it is intended to highlight the common state-of-the-art and create a sense for the trends of future directions in internal combustion engine development

Feature	Conventional spark ignition engine (homogeneous charge spark ignition)	Conventional compression ignition engine (heterogeneous charge compression ignition)
Also known as	Gasoline, petrol, Otto engine	Diesel engine
Method of ignition	Spark	Compression
Fuel equivalence ratio	Precisely controlled to stoichiometric	Varies depending on engine load; typically remains lean
Fuel/air mixture preparation	Carburetion, throttle-body injection, port injection	Direct injection or indirect injection
Degree of fuel/air mixing	Homogeneous	Heterogeneous
Fuels used	Gasoline, alcohols, ethanol, hydrogen, high volatility hydrocarbons	Diesel, oils (transesterified), hydrogen, low volatility hydrocarbons
Compression ratios	Ca. 8–11	Ca. 14–21
Method of load control	Throttle restriction of air or fuel/air mixture (depending on mixture preparation)	Fuel injection duration (i.e., fuel quantity)

uncontrolled autoignition of the mixture results in excessively high rises in pressure near susceptible components of the piston/cylinder arrangement (e.g., piston rings). As an aside, initially tetra-ethyl lead and eventually other lead alkyls were used in gasoline as anti-knock agents, allowing for compression ratios to increase in spark ignition engines [109]. Concerns over the effect of lead on human and animal life, as well as the implications of lead on catalytic devices, have mostly eliminated lead-based additives from fuels [110]. Finally, and similarly related to the spark's need to ignite a near-stoichiometric and homogeneous mixture, conventional spark ignition engines control their load, or power output, via the use of a throttle. In other words, the power of the engine at any given speed is minimized by decreasing the efficiency by which the engine can induct fresh mixture.

Consider now the conventional compression ignition engine, or commonly called diesel engine. It seems that from its first inception, the diesel engine was to be a compressively ignited machine, as Diesel intended to create an isothermal reaction where air and fuel were (separately) compressed to the combustion temperature and expansion occurred isothermally as chemical energy converted directly into mechanical energy [61]. Of course, Diesel never succeeded in attaining this isothermal process (nor has anyone since); but, the foundation had been laid for a compression ignition engine. By using compression ignition combustion, many of the constraints placed on the spark ignition engine (i.e., near-stoichiometric fuel equivalence ratio, homogeneous mixture, and throttle of intake mixture) no longer need apply to the diesel engine. It also seems that from its initial inception, Diesel intended the fuel to be directly delivered to the compressed air [61]; of course, it is also clear that direct (or indirect) injection of the fuel to compressed air is necessary so as to control start of combustion. Further, to aid in the gradual and isothermal conversion from chemical to mechanical energy, Diesel stipulated that there be a chemical abundance of air (i.e., fuel lean) in the mixture [61]; practical implementations of diesel engines make use of varying fuel-equivalence ratios to control the load of the engine. The direct (or indirect) injection of fuel into the compressed air, contrary to Diesel's conception, establishes a heterogeneous mixture that within itself has widely varying fuel–air ratios and reaction temperatures. Further, since ignition is

manifested through compression, fuels with high ignitability (e.g., heavy distillates and aromatics) are required (correspondingly, the fuels also have relatively lower volatility, further establishing the heterogeneous nature of diesel combustion). Also related to the use of compression to ignite the mixture is the need to have high compression ratios (i.e., typically varying between 14 and 21). Finally, as already described, engine load is controlled by the direct and exclusive control of fuel; thus, engine load varies as mixture fuel–air equivalence ratio varies. Due to the heterogeneous nature of diesel combustion, and the dependency of soot formation on mixture stoichiometry under conventional conditions, fuel equivalence ratios rarely exceed 90% of stoichiometric to avoid "smoke limitation."

In the context of the three common attributes of internal combustion engines (i.e., performance, efficiency, and emissions), and the comparison given in Table 2, brief comments will be made about the pros and cons of each conventional technology.

In terms of power, consider Eq. 7 and the qualitative relation of each parameter for a given technology at their "wide-open throttle" or full-load operation, as given in Table 3. Described in more detail below, diesel engines tend to have higher fuel conversion efficiencies. Because of not premixing the fuel with air and generally lower engine speeds (which allows the avoidance of flow choke), diesel engines tend to have higher volumetric efficiencies (refer to the discussion surrounding Fig. 6 and [111] for more detail on factors affecting volumetric efficiency of engines). Typical applications of diesel engines use turbocharging, which increase the inlet mixture density to above atmospheric conditions (typical spark ignition engines operate naturally aspirated). As described above, spark ignition engines generally operate stoichiometric, whereas compression ignition engines generally operate lean even at full power. The heating value of diesel fuel is marginally higher than that of gasoline. Displaced volumes of typical compression ignition engines tend to be larger than those of typical spark ignition engines. Correspondingly, peak power speeds for typical spark ignition engines tend to be higher than compression ignition engines. Mostly due to the larger displaced volumes, but also assisted by higher fuel conversion and volumetric efficiencies, higher inlet density, and higher heating value of the fuel, typical compression ignition engines tend to exhibit higher peak powers than typical spark ignition engines. If the power is normalized by displaced volume to render the specific power, however, spark ignition engines tend to have higher specific power than compression ignition engines (in spite of compression ignition engines having high efficiencies, density, and heating value, the stoichiometric and high speed operation of the spark ignition engine tends to yield higher specific power).

As described in section on "Thermodynamic Analysis of Internal Combustion Engines", engines operating with high compression ratios and lean fuel equivalence ratios will tend to have higher efficiencies than engines operating with relatively lower compression ratios and near-stoichiometric equivalence ratios. As a result, typical compression ignition engines tend to have higher efficiencies than typical spark ignition engines. In some instances, turbocharging that is

Internal Combustion Engines, Developments in.

Table 3 Comparison of the various parameters that control an engine's ability to make power (Eq. 7) between typical applications of conventional spark ignition and conventional compression ignition engines at wide-open throttle or full load condition. It should be noted that these are qualitative assessments of typical technology, and should not be viewed as absolute truths of the respective technologies. Since typical applications of internal combustion engines operate on the four-stroke principle, its effect on the comparison is neutral

Parameter	Parameter's effect on power[a]	Conventional spark ignition engine	Conventional compression ignition engine
η_f	↑		↑
η_v	↑		↑
$\rho_{a,i}$	↑		↑
(F/A)	↑	↑	
Q_{HV}	↑		↑
V_d	↑		↑
N	↑	↑	
n_R	↓	↔	

[a]In other words, an increase in the parameter will have the listed effect on power

typically found on compression ignition engines could increase efficiency if intake manifold pressure is boosted to higher than exhaust manifold pressure; the primary function of a turbocharger, however, is to increase inlet mixture density to increase the power capabilities of the engine (as shown in Eq. 7). The efficiency improvement of compression ignition engines at part-load conditions becomes amplified as (a) the mixture becomes leaner for the compression ignition engine and (b) the use of throttle to manifest part load in the spark ignition engine introduces a thermodynamic loss parameter in the cycle.

Finally, a brief comparison of emissions between the two engines at wide-open throttle or full load condition is made. Because of their near-stoichiometric operation, engine-out emissions of HC and CO for spark ignition engines tend to be relatively higher for those of compression ignition engines (where the latter uses fuel-lean mixtures, creating an oxygen rich exhaust and high level of oxidation of partially oxidized species such as CO and HC). Engine-out emissions of NO tend to be nearly the same between engine technologies. Engine-out emissions of PM are much higher for compression ignition engines, where mixture heterogeneity creates numerous opportunities for soot formation. Of course, it is important to note that conventional spark ignition engines are typically coupled with an effective catalyst, substantially lowering the catalyst-out emissions of the various species to levels well below the engine-out emissions of compression ignition engines. While the use of aftertreatment systems with conventional compression ignition engines are less straightforward, technology is becoming available to also allow substantial reduction of their engine-out emissions [106].

At this point, it becomes clear both engine technologies have features which are more favorable than the other for power, efficiency, and emissions considerations. For example, the high compression ratio and lean mixture required by compression ignition are attractive from efficiency perspectives. The homogeneous mixture of spark ignition is attractive from uniform combustion and emissions perspectives. The lack of a throttle to control load of a compression ignition engine is attractive, but the direct ignition trigger of a spark ignition engine is also attractive. Thus, it is clear that future engine developments could exploit the favorable features of engine technologies to create the next generation internal combustion engine. The next section will describe such efforts and offer insight into the likely future direction of internal combustion engines.

Future Directions of Internal Combustion Engines

There are several types of advanced technology that exist for internal combustion engines. Some of this technology is very prevalent on engines (e.g., turbocharging on diesel engines). Some technology is beginning to appear on production models (e.g., variable valve timing). Other technology is still in its development stages, awaiting its potential entry into full-scale production (e.g., homogeneous charge compression ignition combustion). This section will briefly describe such technology.

Engine Downsizing

A general idea that permeates much of the technology under development of internal combustion engines is that of engine downsizing – or, the effort to use advanced technology to enable the use of smaller-sized engines to produce the same power (i.e., increase power density) [112]. The application-oriented benefit is that a smaller engine likely weighs less, thus could improve application efficiency (e.g., better vehicle fuel economy with a lighter engine). The engine itself, however, will likely realize improved efficiency. For example, a smaller engine that uses turbocharging to maintain same power (as a larger-sized engine) will generally operate more often near the location of peak efficiency. Some benefit of such is realized in diesel engines, which typically have best efficiencies near mid-speed, and 75% peak load conditions. More benefit, however, is realized in gasoline engines where throttles are used; a smaller engine with higher power density will require less throttle and thus realize larger gains in efficiency improvement. Overall, the major purpose of engine downsizing is to improve parameters other than V_d in Eq. 7. This idea will become more apparent as specific technology is discussed below.

Turbocharging/Supercharging/Boosting

Boosting an engine – i.e., increasing the inlet mixture density to increase the trapped mass per cycle – is a very

common means to increase the power density of an engine. As evident from Eq. 7, an increase in the inlet air density will directly increase the power of the engine. There are two major types of boosting technology: turbocharging and supercharging. The major difference between the two technologies is the former uses a centrifugal device (i.e., a turbine) to exploit available exhaust energy for conversion to shaft work whereas the latter absorbs shaft work via a direct mechanical connection to the engine. In both cases, the shaft work of the device is coupled to a boosting component – typically, either a centrifugal-based compressor or a positive-displacement compressor – which acts as an "air pump" to increase density of the inlet air. In most applications, a turbocharger uses a centrifugal turbine/compressor configuration while a supercharger uses a positive-displacement compressor.

Diesel engines are typically favorable engines to outfit with boosting devices (further, usually with turbochargers as they assist over the entire engine operating map). Boosting compensates for the diesel engine's typical use of fuel-lean mixtures to improve its power density. Further, since diesel engines do not employ throttles for load control, boosting of a diesel engine provides benefit over the entire operating map. For this latter reason, and due to the general higher efficiency of a turbocharger over a supercharger, turbochargers are the most common boosting device on a diesel engine. Gasoline engines also can be outfitted with boosting devices, with additional complexity to consider. First, boosting devices on gasoline engines usually offer benefit only at wide-open throttle conditions; a point of operation for most applications of engines that is rarely used. Also, for this reason, directly coupled superchargers are often used where improved response to boost is provided. Second, increasing density of the inlet mixture tends to increase the propensity to knock. Thus, while boosting provides additional trapped mass to deliver increased power, a potential retard in timing to avoid fuel knock likely decreases fuel conversion efficiency and introduces a tradeoff in how much additional power can be expected from the boost. Third, because of the typical use of superchargers when boosting is applied to gasoline engines, the system efficiency tends to decrease as shaft work is transferred to provide the boosting action (whereas turbocharging uses available energy of the exhaust).

Often times, boosting devices – in particular, turbochargers – are viewed as devices to increase the overall system efficiency of the engine. In other words, it is thought that because a turbocharger uses exhaust energy that would otherwise be wasted, its conversion to useful work (i.e., boosting) should increase efficiency. This work transfer, however, usually does not leave the control system (i.e., the shaft work of the turbine is directly coupled to the pumping action of the compressor). In some instances, efficiency improvements can be realized if a "negative pumping loop" is created by boosting the intake manifold to a higher pressure than the exhaust manifold. It is also possible to improve the overall engine efficiency if boosting an engine enables the use of a smaller-sized engine for a given application; an effort, as described above, known as downsizing. In most instances, however, overall system efficiency may decrease even with the use of a turbocharger, as the major objective of increasing inlet density to increase power density requires additional energy transfer through the exhaust than out as shaft work. This latter aspect is often realized as an increase in exhaust manifold pressure to "drive" the turbocharger.

An area of technology development for turbocharging is the use of waste-gated and variable geometry turbochargers. In non-waste-gated or non-variable geometry turbochargers, the turbine has fixed geometry and thus fixed flow characteristics. As such, a fixed geometry turbocharger is restrictively designed to provide maximum benefit to the engine at a narrow operating range which usually centers on the peak power condition of the engine. As such, at low speed or low-load conditions, the turbocharger's boosting benefits are diminished. While decreasing the turbocharger's maximum benefit over a broader regime of the engine's operating map, such a constraint also typically creates a dynamic issue during engine accelerations known as "turbo lag." The large turbine designed for maximum engine flow rates requires substantial inertia to rotationally accelerate to maximum boosting benefit. Waste-gated and variable geometry turbochargers offer opportunities to overcome such issues. In the case of a waste-gated turbocharger, usually a smaller turbine with less inertia is used with an exhaust "waste-gate." The smaller turbine provides faster response and better boosting at low-loads and

speeds; upon approach of the engine's peak power condition (where either boosting becomes excessive for the intake system or turbocharger speeds exceed maximum limits) a waste-gate opens that allows exhaust energy to bypass the turbine. Thus, the turbocharger is supplied by a fraction of the available exhaust energy providing suitable boost at allowable rotational speeds of the turbocharger at the engine's peak power condition. A variable geometry turbocharger uses a similar concept, except that it is designed to change the flow momentum of the exhaust gases and create multiple pressure ratios for a given exhaust flow rate [113, 114].

Finally, it is noted that outfitting a boosting device to an engine is not a trivial task [115]. There is not an exclusive match between an engine and a given boosting device; instead, the match of a boosting device to its engine application is dictated by the objectives of adding the device, whether it is to exclusively increase power density, efficiency, and/or emissions of the engine system [116–118].

Advanced Engine Controls

Much of the advanced technology that appears on modern engines, or will appear on future engines, is enabled by advanced engine controls. Engine control has always been an integral component of the engine's success at delivering cost-effective and efficient power. Early engine control systems were purely mechanical and only concerned with controlling the level of power (e.g., throttle or rack position) or holding a constant speed with variable load (such as a generator system, using a speed governor). Modern-day engine control systems, however, are virtually all electrical-based, and at the very least sense several aspects of the engine's operation (e.g., cam position, throttle position, manifold temperatures and pressures, and air flow) and typically control most aspects of the engine (e.g., spark timing, injection timing, injection pressure, EGR level, and boost pressure) [119].

Like the engine itself, engine controls are becoming more sophisticated and advanced. Specifically, a general trend to use in-cylinder information as a feedback signal is an example of the type of complexity future engine control systems intend to resolve. Knowledge of in-cylinder pressure, for example, can provide immediate information to the engine controller about load produced by the engine. Additionally, in-cylinder pressure is the major property necessary for assessing the rate of energy release during the combustion process; having such information could allow engine control systems to change parameters based on a desired burn profile in the cylinder [120–122]. Further, along with the continuing advancements in engine model development (e.g., [46–50]), a trend toward model-based engine control [123, 124] intends to decrease engine development time and improve control over the several parameters now present on internal combustion engines.

Variable Geometry Engine Designs

Along with efforts to effect engine downsizing, and made possible with advancements in engine control systems, is the notion of variable geometry engine designs. In other words, conventional reciprocating internal combustion engines have mechanically fixed geometries, i.e., constant compression ratios and constant displaced volumes. A variable compression ratio is attractive, for example, since it might enable high compression ratio operation for a spark ignition engine at part-load conditions where there is decreased propensity for knock. Similarly, it could be used in diesel engines to avoid excessively high peak pressures at full load conditions. Variable displacement engines are attractive since they enable high peak powers, but use less throttle at part-load conditions (thus diminishing pumping losses).

An early concept of variable compression ratio was proposed by J. Atkinson, for whom the Atkinson Cycle is named. As described by [125], Atkinson's original conception involved mechanical linkages to displace the piston such that the engine's compression ratio is lower than its expansion ratio; the main idea being that more expansion work will yield higher efficiency. Similar in idea, but different in implementation, is the Miller concept [125] which uses either late intake valve closing or late exhaust valve opening to shorten or extend the compression or expansion strokes, respectively. Both concepts are attempted in modern-day applications, using both variable compression ratio techniques [112, 126, 127] and altered valve timing techniques [112, 127, 128]. The latter approach, of

using altered valve timing techniques, is fluidly made possible through the use of variable valve timing [112, 127], a concept discussed in more detail in the next section.

While variable compression ratio concepts attempt to increase engine efficiency by way of increasing expansion, variable displacement engines attempt to increase engine efficiency by decreasing the use of throttle (thereby, decreasing the pumping work associated with the gas exchange process of a conventional spark ignition engine). With variable displacement, an engine is able to deliver high power using full displacement; at part-load conditions, rather than use throttle, cylinders can be "deactivated" so that they produce no power and allow the engine to deliver part-load power [129]. The deactivated cylinders typically continue to stroke and exchange gases; the closed portion of the cycle (i.e., compression and expansion) realize some loss due to heat transfer and friction but this is intended to be less than the gain realized through decreased pumping work.

Variable Valve Timing

Briefly discussed in the above sections is the idea of variable valve timing, or, the ability to change the valve events (i.e., intake and exhaust valve opening and closing) at any given point during the engine's operation [130]. In conventional engine design, the valve events are "fixed" by mechanical positions of the lobes on the cam shaft. The effectiveness of an engine to induct and exhaust mixture depends on many things including, for example, engine speed, engine load, the use of EGR, spark timing, and injection timing. Thus, there are not valve events that will universally yield peak power, efficiency, and/or emission for any given engine design.

The idea of variable valve timing allows the engineer to decouple the valve events from the in-cylinder processes. In other words, flexibility of intake and exhaust is afforded with the use of variable valve timing. This not only allows the valve events to be uniquely tuned for each operating point of the engine for improved performance [131] and efficiency [132], but it also can enable other advanced technology. For example, having variable valve timing allows the implementation of the Miller approach to effecting variable compression ratio at part-load conditions, but enabling conventional operation at peak power conditions [127]. Another example is the use of variable valve timing to control the amount of residual fraction in the cylinder, which can affect emission of certain pollutants [133]; controlling residual fraction is a way to enable advanced modes of combustion, described in more detail below. Finally, variable valve timing can be used to replace a throttle, for example, for load control [134]; in doing so, trapped mass can be controlled without inducing pumping work in the engine. Although not widespread technology, variable valve timing is becoming more prevalent on modern-day engines.

Waste Heat Recovery

Waste heat recovery is the effort to take advantage of temperature gradients created between the engine and its environment. As described in section on "Thermodynamic Analysis of Internal Combustion Engines", thermal energy is transferred out of the system through heat transfer (e.g., through engine coolant) and exhaust flow. Because of the temperature gradient that exists between, for example, the high temperature exhaust and the low temperature surroundings, an orderly flow of thermal energy tends to cause the exhaust system to attain thermal equilibrium with the environment. Conventionally, this orderly flow of thermal energy is wasted (i.e., the heat transfer completely dissipates as generated entropy). It is practically possible to instead intercept the orderly flow of thermal energy and convert it to useful work. In the theoretical limit, this conversion of thermal energy to work energy is given by the completely reversible cycle, often called the Carnot cycle. The corresponding efficiency of useful work converted from thermal energy is thusly the Carnot efficiency, as is given by Eq. 16:

$$\eta_{\text{th,Carnot}} = 1 - \frac{T_{\text{L}}}{T_{\text{H}}} \tag{16}$$

where $\eta_{\text{th,Carnot}}$ is the Carnot efficiency (maximum possible conversion of thermal energy to work energy), T_{L} is the sink's temperature (e.g., the environment temperature), and T_{H} is the source's temperature (e.g., the exhaust temperature). Considering that a typical automotive exhaust temperature may be 900 K operating in a 300 K environment, ideal

efficiencies could be on the order of 67%. Of course, real process irreversibilities diminish actual device efficiencies from the ideal Carnot efficiency. In spite of the diminished efficiency from ideal, though, waste heat recovery in an automotive application, for example, is reported to decrease vehicle fuel consumption by as much as 7.4% [135].

There are several technologies available to make use of internal combustion engine exhaust waste heat energy; because of its spatial accessibility and concentrated high temperature, exhaust energy has been the focus of much of the waste heat recovery. Example major technologies include: (1) mechanical or electrical turbo-compounding/generating devices, (2) Rankine cycle-type devices, and (3) thermoelectric-type devices. Turbo-compounding or turbo-generating [136] converts exhaust thermal energy to mechanical energy (in the form of pressure and kinetic energy) and couples the mechanical energy either directly to the engine driveshaft (to deliver additional brake power) or to an electric generator. One disadvantage of turbo-compounding (like its turbocharging companion) is the method of converting thermal energy to mechanical energy; the centrifugal device increases an engine's exhaust pressure based on its operation. The increased exhaust pressure affects the engine's operation by altering the pumping work and initial mixture composition (increased exhaust species in the initial mixture); such factors can deteriorate the engine's cycle efficiency.

Rankine cycle-type devices make use of the thermodynamic Rankine cycle – typically with an organic fluid designed to undergo phase change within the temperature ranges of a typical engine exhaust system – to output shaft work for either direct-coupling to the engine driveshaft or electrical generation. In such a device, the engine exhaust stream provides the thermal energy to boil or vaporize the organic working fluid of the cycle. After becoming saturated vapor, the fluid expands through a turbine converting the thermal energy to mechanical energy; the cycle completes with the usual condenser and pump processes. Such a device is reported to increase the combined engine + waste heat recovery efficiency by up to 10% [137]; a potential downside is the added complexity of adding four processes (as opposed to one, for example, in the case of a turbomachine) to waste heat recovery system.

Another example major waste heat recovery device is the thermoelectric device. The thermoelectric effect, first observed by Seebeck in 1821, is the generation of a voltage due to a temperature difference between two junctions of two dissimilar materials [138]; principally, the Seebeck effect describes the operation of a thermocouple measurement of temperature. The practical use of the Seebeck effect to produce electricity as a thermoelectric device has recently emerged with semiconductor materials having favorable properties to transmit electricity with little resistance heating (Joule heating) and thermal conductivity (which would tend to "short-circuit" the thermoelectric device). In fact, the advent of semiconductor materials first made possible practical thermoelectric devices as refrigerators exploiting the Peltier Effect (i.e., the opposite of the Seebeck effect, where an applied voltage creates a temperature difference between two junctions of two dissimilar metals) [138]. Now, however, thermoelectric devices create promise to exploit the available thermal energy in an internal combustion engine's exhaust for conversion to electricity [135, 139].

Direct Injection, Spark Ignition Engines

At the end of the section on "A Case Study: Diesel Engines Versus Gasoline Engines", it is suggested that spark ignition and compression ignition engines each have favorable features for power, efficiency, and emissions, but that each has conventionally designed limitations. Thus, it becomes attractive to design each ignition system's limitations out of the engine, and potentially realize gains in the engine's attributes (i.e., power, efficiency, and emissions). Two now-common technologies exist to do so: (1) direct injection spark ignition combustion and (2) homogeneous charge compression ignition combustion. This and the next section will describe these two technologies.

Direct injection spark ignition combustion attempts to create a stratified fuel–air mixture "charge" around the spark plug so that, at the point of spark release, the spark ignites a near-stoichiometric mixture. It is intended that outside of the stratification zone there is little to no fuel, thus creating an overall lean fuel–air mixture. The use of the word "stratified" to describe the fuel–air mixture is used here, as opposed to heterogeneous, to reinforce the notion that under

ideal conditions the fuel–air mixture would be homogeneous (i.e., homogeneously stoichiometric) throughout the fuel–air mixture, but pure air outside the stratification zone (i.e., outside the fuel–air mixture). This is in contrast to the use of "heterogeneous" to describe a mixture (e.g., diesel engine mixture), where it is expected that substantial fuel–air ratio gradients exist throughout the fuel sprays.

In order to manifest the stratified mixture concept (see, e.g., [140–148]), fuel is injected directly into the cylinder. Typically, the combustion chamber of a direct injection spark ignition engine is specially designed to assist the stratification of the fuel–air mixture, and center it on the spark plug. The intake stroke draws air and residual mixture into the cylinder – notably, fuel is not inducted during intake as is done in conventional spark ignition operation. After intake, direct fuel injection into the cylinder usually occurs at some point during the piston's travel from BDC to TDC during the compression stroke. Spark advance is typically timed at around the same point as that in a conventional spark ignition engine (i.e., at a point near the piston reaching TDC-compression). The remaining processes of the direct injection spark ignition engine are basically the same as the conventional spark ignition engine.

There are several potential benefits of direct injection spark ignition combustion. Perhaps the clear benefit is the ability to use lean mixtures in a spark ignition engine. Because the combustion chamber is designed to stratify the mixture and create a stoichiometric mixture near the spark plug, the overall equivalence ratio of the mixture filling the entire chamber can be lean. As described in section on "Thermodynamic Analysis of Internal Combustion Engines", overall lean mixtures possess higher ratios of specific heats (γ), which translate to higher fuel conversion efficiencies. Another, less obvious, benefit of a stratified mixture is the decreased propensity to fuel knock (or, at least decreased propensity of fuel knock in the regions of the cylinder able to cause harm such as near cylinder walls and piston rings). The lack of a reactive mixture – manifested by charge stratification – in the regions furthest from the spark plug – which are the last to be controllably burned by the propagating flame – decreases the propensity that the mixture will autoignite and burn uncontrollably. Such a feature enables the direct injection spark ignition engine to have increased compression ratios relative to conventional spark ignition engines; again, this is an attribute that promotes an increase in efficiency of the novel engine concept. Finally, the use of direct fuel injection into the cylinder enables the elimination of a throttle to control engine load. In other words, engine load is controlled by the quantity of fuel injected into the cycle, similar to the load control of a diesel engine. Elimination of the throttle, as repeatedly described, will improve part-load efficiency by eliminating pumping losses effected by throttle.

Although the benefits are plentiful, the challenges are also present. From practical perspectives, perfect attainment of a stratified charge is difficult to accomplish. As such, heterogeneities within the stratified mixture emerge and can lead to products of incomplete combustion such as CO, HC, and PM. This, of course, is amplified at full load conditions; thus, peak power attainment through direct injection means alone would likely be limited by smoke limitations (similar to full load limitations of a diesel engine). Further, and like the challenges faced by diesel engines, outfitting direct injection spark ignition engines with after treatment devices is complicated by the use of overall lean mixtures (as described in section on "Emissions Formation and Exhaust Pollution"). In spite of such challenges, direct injection spark ignition engines are in production; continued development of in-cylinder flow modeling tools and engine controls contribute toward the concept's potential success.

Homogeneous Charge Compression Ignition Engines

An attractive feature of the conventional spark ignition engine is its use of a homogeneous mixture. Although this tends to promote knock (as described in section "Direct Injection Spark Ignition Engines"), it provides the benefit of being kinetically rate-limiting as opposed to mixing rate-limiting (as in the case of diesel engines). One issue with using a spark, or single point, to ignite a homogeneous mixture is the establishment of a flame that must propagate the mixture to convert the chemical energy. In order to more quickly react the mixture, in a volumetric sense, multi-point ignition is required; i.e., ignition that occurs in several locations throughout the mixture will result in a faster

burn rate. Such a voluminous ignition can be effected through compression of a homogenous mixture.

This is the basic idea of homogeneous charge compression ignition (HCCI); use compression to ignite a homogeneous mixture (similar to a diesel engine, except that diesel engines use compression to ignite an inherently heterogeneous mixture). The homogeneous mixture could, for example, be formed through premixing of fuel and air prior to mixture induction during the intake stroke. A faster burn rate effected by homogeneous charge compression ignition allows heat release to occur at near constant volume conditions; thus establishing the possibility to approach theoretical limits of maximum efficiencies of internal combustion engines. Further, because compression is used to ignite the mixture rather than spark, lean mixtures can be used at part-load conditions which further promotes higher efficiencies of the HCCI concept. Similar to the direct injection spark ignition concept, HCCI engines could eliminate throttles as load is controlled directly by the quantity of fuel mixed with the intake mixture. Finally, the common issue of particulate matter emissions faced by diesel engines (which also use compression ignition) are substantially decreased through the avoidance of locally rich fuel–air mixture zones (due to the use of a homogeneous mixture in an HCCI concept).

Of course, immediately the obvious problem becomes the method of ignition control. HCCI concepts do not have a direct ignition trigger, as do conventional gasoline (i.e., a spark) or diesel (i.e., direct fuel injection) engines. Controlling ignition in the HCCI concept depends on very precise control of the mixture's initial state at start of compression and the compression path followed up to the point of ignition. Several factors which are present – even in tightly controlled research environments – such as heat transfer, turbulence, and the history of preceding combustion events make the practical application of HCCI very challenging. The payoffs, of course, are correspondingly very high.

Practical implementation of HCCI is first reported by Onishi et al. [149] with theoretical developments experimentally provided by Najt and Foster [150]. Several control parameters could be adjusted such as compression ratio (e.g., with the use of variable valve timing), initial temperature, and quantity of residual fraction (e.g., effected either through exhaust gas recirculation or variable valve timing). Identifying the key control parameters, and the optimal way to adjust them during real-time operation of the engine, continue to be on-going research activities [151–156].

Advanced Compression Ignition Engines

A technique to control combustion of an HCCI engine is to use precisely metered amounts of residual fraction, which not only act to alter the kinetics of combustion but also result in substantially lower combustion temperatures. As such, much of HCCI combustion is characterized by low temperature mechanisms commonly referred to as *low temperature combustion* (LTC). LTC offers a few benefits. First, efficiency improvements in the engine can be realized (in spite of increased exergy destruction due to low reaction temperatures) due to more favorable thermodynamic properties (i.e., higher ratio of specific heats, see Fig. 11) of the burned mixture and lower rates of heat transfer. Second, and typically the driver for LTC technology development, lower nitric oxide formation per the discussion in section on "Emissions Formation and Exhaust Pollution".

With this in mind, and reconsidering the prevailing issue of HCCI implementation – i.e., control of start of combustion – it becomes plausible to consider developing an "HCCI-type" mode of combustion in a diesel engine. That is, rather than induct a homogeneous mixture of fuel and air and rely upon indirectly controlled parameters to control ignition, perhaps fuel can be injected directly into the cylinder allowing for better control of ignition. In order to manifest LTC and harvest its benefits (e.g., possibly higher efficiency and substantially emission), high levels of EGR and strategic injection timings are used to extend ignition delay and create a nearly all-premixed combustion event. The long ignition delay, coupled with low temperature mechanisms, establishes the phenomenological observation of two-stage ignition characterized by the presence of cool-flame reactions [150, 157]. Interestingly, because of attainment of LTC, soot precursor formation is substantially abated, and the engine is made to operate with very low emissions of nitric oxide and particulate matter [158–168]. The combustion concept has become known as premixed compression ignition or premixed charge compression ignition combustion.

The ability to attain LTC in compression ignition engines is attributed to the advancement of technology now in place on such machines, such as common-rail and electronic fuel pressure systems, variable geometry turbochargers, and exhaust gas recirculation systems.

Alternative Fuels

The term "alternative fuels" for an internal combustion engine is somewhat baseless, as an internal combustion engine has considerable flexibility in the type of fuel it uses. Of course, conventional fuels are the commonly called "gasoline" and "diesel" fuels, but generally engines have been shown to operate on virtually any gaseous, liquid, and solid dust particle specie that has heating value (i.e., will release thermal energy in a chemical oxidation process). Because of the wide variability of fuels available to internal combustion engines, this topic will not be expanded in this article. There are, however, certain considerations that should be given to the use of a fuel in an engine which was not intently designed for use with such fuel (e.g., use of ethanol in a gasoline engine or use of biodiesel in a diesel engine). First, ignition characteristics of the fuel may not be favorable for the particular engine design. For example, short-chain volatile hydrocarbons do not generally ignite well in conventional unmodified compression ignition engines of typical compression ratios. Likewise, long-chain nonvolatile hydrocarbons do not generally vaporize well in conventional unmodified spark ignition engines. Second, flame temperatures of the combustion of the fuel may exceed material limits of any given engine construction. Third, fuels may react with other support components of the engine system (e.g., solvency of fuels with rubber hoses). Finally, combustion process will likely be altered when using an unconventional fuel in a conventional unmodified engine yielding different emissions, efficiency, and peak power capabilities. Thus, although internal combustion engines have inherent fuel-flexibility, their use with unconventional fuels is not straightforward and requires careful design and engineering considerations.

Acknowledgments

The author wishes to thank several people who have helped to make this work possible. First, the State of Texas is acknowledged for their financial support of some of the research highlighted here; specifically, their support through the Texas Commission on Environmental Quality and the Houston Advanced Research Center is acknowledged. Second, Professor Jerald A. Caton of Texas A&M University is acknowledged for his contributions to this article and for providing a thorough proof of its contents. Third, Margaret Fisher is acknowledged for her assistance in preparing the copyrighted materials from other sources and securing permissions to use them. Fourth, Wiley is acknowledged for providing permission of copyrighted material *au gratis*. Fifth and lastly, but certainly not least, I acknowledge the contributions and assistance of my graduate and undergraduate students, some of whom have research highlighted in this work. Specifically, these individuals include: Mr. Josh Bittle, Mr. Jason Esquivel, Ms. Sarabeth Fronenberger, Mr. Blake Gettig, Mr. Mark Hammond, Mr. Bryan Knight, Mr. Jeffrey Kurthy, Mr. Jimmy McClean, Ms. Claire Mero, Mr. Yehia Omar, Ms. Gurlovleen Rathore, Mr. Kyle Richter, Mr. Sidharth Sambashivan, Mr. Chris Schneider, Ms. Amy Smith, Mr. Hoseok Song, Mr. Jiafeng Sun, Mr. Brandon Tompkins, Mr. Brad Williams, Mr. R. Kevin Wilson, Mr. Whit Wilson, and Mr. Jesse Younger.

Abbreviations

BDC	Bottom dead center
EGR	Exhaust gas recirculation
HCCI	Homogeneous charge compression ignition
IC	Internal combustion
LTC	Low temperature combustion
TDC	Top dead center

Bibliography

Primary Literature

1. Cummins C Jr (1976) Early IC and automotive engines. SAE Trans 85(SAE Paper No. 760604):1960–1971
2. Heywood J (1988) Internal combustion engine fundamentals. McGraw-Hill, New York, p 9
3. Heywood J (1988) Internal combustion engine fundamentals. McGraw-Hill, New York, p 10
4. Automot Eng Int (January 2010), 118(1):47. http://www.sae.org/automag/
5. Brown W (1967) Methods for evaluating requirements and errors in cylinder pressure measurement. SAE Trans 76(SAE Paper No. 670008):50–71

6. Lancaster D, Krieger R, Lienesch J (1975) Measurement and analysis of engine pressure data. SAE Trans 84(SAE Paper No. 750026):155–172
7. Randolph A (1990) Methods of processing cylinder-pressure transducer signals to maximize data accuracy. SAE Trans J Passenger Cars 99(SAE Paper No. 900170):191–200
8. Kuratle R, Marki B (1992) Influencing parameters and error sources during indication on internal combustion engines. SAE Trans – J Engines 101(SAE Paper No. 920233):295–303
9. Davis R, Patterson G (2006) Cylinder pressure data quality checks and procedures to maximize data accuracy. SAE Paper No. 2006-01-1346
10. Amann C (1983) A perspective of reciprocating-engine diagnostics without lasers. Prog Energy Combust Sci 9:239–267
11. Heywood J (1988) Internal combustion engine fundamentals. McGraw-Hill, New York, pp 56–57
12. Tompkins B, Esquivel J, Jacobs T (2009) Performance parameter analysis of a biodiesel-fuelled medium duty diesel engine. SAE Paper No. 2009-01-0481
13. Heywood J (1988) Internal combustion engine fundamentals. McGraw-Hill, New York, p 217
14. Heywood J (1988) Internal combustion engine fundamentals. McGraw-Hill, New York, p 154
15. Lauck F, Uyehara O, Myers P (1963) An engineering evaluation of energy conversion devices. SAE Trans 71(SAE Paper No. 630446):41–50
16. Foster D, Myers P (1984) Can paper engines stand the heat? SAE Trans 93(SAE Paper No. 840911):4.491–4.502
17. The K, Miller S, Edwards C (2008) Thermodynamic requirements for maximum internal combustion engine cycle efficiency, Part 1: optimal combustion strategy. Int J Engine Res 9:449–465
18. Carnot NLS (1824) Reflections on the motive power of heat (Trans and ed: Thurston RH), 2nd edn (1897). Wiley, New York
19. Borgnakke C, Sonntag R (2009) Fundamentals of thermodynamics. Wiley, New York, p 497
20. Heywood J (1988) Internal combustion engine fundamentals. McGraw-Hill, New York, p 177
21. Heywood J (1988) Internal combustion engine fundamentals. McGraw-Hill, New York, p 163
22. Edson M (1964) The influence of compression ratio and dissociation on ideal otto cycle engine thermal efficiency, Digital calculations of engine cycles. SAE, Warrendale, pp 49–64
23. Westbrook C, Dryer F (1984) Chemical kinetic modeling of hydrocarbon combustion. Prog Energy Combust Sci 10:1–57
24. Olikara C, Borman G (1975) A computer program for calculating properties of equilibrium combustion products with some applications IC engines. SAE Paper No. 750468
25. Lavoie G, Heywood J, Keck J (1970) Experimental and theoretical study of nitric oxide formation in internal combustion engines. Combust Sci Technol 1:313–326
26. Bowman C (1975) Kinetics of pollutant formation and destruction in combustion. Prog Energy Combust Sci 1:33–45
27. Miller J, Bowman C (1989) Mechanism and modeling of nitrogen chemistry in combustion. Prog Energy Combust Sci 15:287–338
28. Turns S (1995) Understanding NO_x formation in nonpremixed flames: experiments and modeling. Prog Energy Combust Sci 21:361–385
29. Dean A, Bozzelli J (2000) In: Gardiner WC Jr (ed) Combustion chemistry of nitrogen in gas-phase combustion chemistry. Springer, New York, pp 125–341
30. McBride B, Gordon S (1992) Computer program for calculating and fitting thermodynamic functions. NASA Report No. RP-1271
31. Svehla R (1995) Transport coefficients for the NASA Lewis chemical equilibrium program. NASA Report No. TM-4647
32. Gordon S, McBride B (1999) Thermodynamic data to 20000K for monatomic gases. NASA Report No. TP-1999-208523
33. McBride B, Gordon S, Reno M (2001) Thermodynamic data for fifty reference elements. NASA Report No. TP-3287/Rev 1
34. McBride B, Zehe M, Gordon S (2002) CAP: a computer code for generating tabular thermodynamic functions from NASA Lewis Coefficients. NASA Report No. TP-2001-210959-Rev1
35. Heywood J (1988) Internal combustion engine fundamentals. McGraw-Hill, New York, pp 136–137
36. Stull D, Prophet H (1971) JANAF thermochemical tables, NSRDS-NBS 37. http://www.nist.gov/data/nsrds/NSRDS-NBS37.pdf. Accessed July 5, 2010
37. Keenan J (1951) Availability and irreversibility in thermodynamics. Br J Appl Phys 2:183–192
38. Edson M, Taylor C (1964) The limits of engine performance – comparison of actual and theoretical cycles. In: SAE digital calculations of engine cycles, pp 65–81
39. Strange F (1964) An analysis of the ideal Otto cycle, including the effects of heat transfer, finite combustion rates, chemical dissociation, and mechanical losses. In: SAE digital calculations of engine cycles, pp 92–105
40. Patterson D, Van Wylen G (1964) A digital computer simulation for spark-ignited engine cycles. In: SAE digital calculations of engine cycles, pp 82–91
41. Woschni G (1967) Universally applicable equation for the instantaneous heat transfer coefficient in the internal combustion engine. SAE Trans 76(SAE Paper No. 670931): 3065–3083
42. Hohenberg G (1979) Advanced approaches for heat transfer calculations. SAE Trans 88(SAE Paper No. 790825): 2788–2806
43. Borman G, Nishiwaki K (1987) Internal-combustion engine heat transfer. Prog Energy Combust Sci 13:1–46
44. Heywood J, Higgins J, Watts P, Tabaczynski R (1979) Development and use of a cycle simulation to predict SI engine efficiency and NO_x emissions. SAE Paper No. 790291
45. Sandoval D, Heywood J (2003) An improved friction model for spark-ignition engines. SAE Trans J Engines 112(SAE Paper No. 2003-01-0725):1041–1052
46. Blumberg P, Lavoie G, Tabaczynski R (1979) Phenomenological models for reciprocating internal combustion engines. Prog Energy Combust Sci 5:123–167
47. Assanis D, Heywood J (1986) Development and use of a computer simulation of the turbocompounded diesel system for

engine performance and component heat transfer studies. SAE Trans 95(SAE Paper No. 860329):2.451–2.476
48. Filipi Z, Assanis D (1991) Quasi-dimensional computer simulation of the turbocharged spark ignition engine and its use for 2 and 4-valve engine matching studies. SAE Trans J Engines 100(SAE Paper No. 910075):52–68
49. Kamimoto T, Kobayashi H (1991) Combustion processes in diesel engines. Prog Energy Combust Sci 17:163–189
50. Reitz R, Rutland C (1995) Development and testing of diesel engine CFD models. Prog Energy Combust Sci 21: 173–196
51. Caton J (2003) Effects of burn rate parameters on nitric oxide emissions for a spark ignition engine: results from a three-zone, thermodynamic simulation. SAE Paper No. 2003-01-0720
52. Caton J (2000) A review of investigations using the second law of thermodynamics to study internal-combustion engines. SAE Trans J Engines 109(SAE Paper No. 2000-01-1081): 1252–1266
53. Rakopoulos C, Giakoumis E (2006) Second-law analyses applied to internal combustion engines operation. Prog Energy Combust Sci 32:2–47
54. Shyani R, Caton J (2009) A thermodynamic analysis of the use of exhaust gas recirculation in spark ignition engines including the second law of thermodynamics. Proc Inst Mech Eng Part D: J Automobile Eng 223:131–149
55. Dunbar W, Lior N (1994) Sources of combustion irreversibility. Combust Sci Technol 103:41–61
56. Som S, Datta A (2008) Thermodynamic irreversibilities and exergy balance in combustion processes. Prog Energy Combust Sci 34:351–376
57. Caton J (2000) On the destruction of availability (exergy) due to combustion processes – with specific application to internal-combustion engines. Energy 25:1097–1117
58. Keenan J (1941) Thermodynamics. Wiley, New York, p 269
59. Obert E (1970) Internal combustion engines, 3rd edn. International Textbook Company, Scranton, p 459
60. Patrawala K, Caton J (2008). Potential processes for "reversible" combustion with application to reciprocating internal combustion engines. In: Proceedings of the 2008 technical meeting of the central states section of the combustion institute, Tuscaloosa
61. Diesel R (1897) Diesel's rational heat motor. A lecture delivered at the general meeting of the society at Cassell, June 16, 1897. Original published in Zeitschrift des Vereines Deutscher Ingenieure (Trans: Leupold R). Progressive Age, New York (Reprinted)
62. Heywood J (1988) Internal combustion engine fundamentals. McGraw-Hill, New York, p 391
63. Heywood J (1988) Internal combustion engine fundamentals. McGraw-Hill, New York, p 390
64. Heywood J (1988) Internal combustion engine fundamentals. McGraw-Hill, New York, p 374
65. Lyn W (1963) Study of burning rate and nature of combustion in diesel engines. Proc Combust Inst 9:1069–1082
66. Plee S, Ahmad T (1983) Relative roles of premixed and diffusion burning in diesel combustion. SAE Trans 92(SAE Paper No. 831733):4.892–4.909
67. Ricardo H (1941) The high-speed internal combustion engine (Rev: Glyde HS), 3rd edn. Interscience Publishers, New York
68. Ladommatos N, Abdelhalim S, Zhao H, Hu Z (1998) Effects of EGR on heat release in diesel combustion. SAE Paper No. 980184
69. Meguerdichian M, Watson N (1978) Prediction of mixture formation and heat release in diesel engines. SAE Paper No. 780225
70. Lyn W, Valdmanis E (1968) Effects of physical factors on ignition delay. SAE Paper No. 680102
71. Kamimoto T, Aoyagi Y, Matsui Y, Matsuoka S (1981) The effects of some engine variables on measured rates of air entrainment and heat release in a DI diesel engine. SAE Trans 89(SAE Paper No. 800253):1163–1174
72. Dent J, Mehta P, Swan J (1982) A predictive model for automotive DI diesel engine performance and smoke emissions. Paper presented at the international conference on diesel engines for passenger cars and light duty vehicles. Institution of Mechanical Engineers, London. IMECE Paper No. C126/82
73. Binder K, Hilburger W (1981) Influence of the relative motions of air and fuel vapor on the mixture formation processes of the direct injection diesel engine. SAE Trans 90(SAE Paper No. 810831):2540–2555
74. Dec J (1997) A conceptual model of DI diesel combustion based on laser-sheet imaging. SAE Trans J Engines 106(SAE Paper No. 970873):1319–1348
75. Flynn P, Durrett R, Hunter G, zur Loye A, Akinyemi O, Dec J, Westbrook C (1999) Diesel combustion: an integrated view combining laser diagnostics, chemical kinetics, and empirical validation. SAE Trans J Engines 108(SAE Paper No. 1999-01-0509):587–600
76. Chigier N (1975) Pollution formation and destruction in flames – Introduction. Prog Energy Combust Sci 1:3–15
77. Beltzer M (1976) Non-sulfate particulate emissions from catalyst cars. SAE Trans 85(SAE Paper No. 760038):198–208
78. Khatri N, Johnson J, Leddy D (1978) The characterization of the hydrocarbon and sulfate fractions of diesel particulate matter. SAE Trans 87(SAE Paper No. 780111):469–492
79. Hoffert M, Caldeira K, Benford G, Criswell D, Green C, Herzog H, Jain A, Kheshgi H, Lackner K, Lewis J, Lightfoot H, Manheimer W, Mankins J, Mauel M, Perkins L, Schlesinger M, Volk T, Wigley T (2002) Advanced technology paths to global climate stability: energy for a greenhouse planet. Science 298:981–987
80. Ghoniem A (2011) Needs, resources and climate change: clean and efficient conversion technologies. Prog Energy Combustion Sci 37:15–51
81. Henein N (1976) Analysis of pollutant formation and control and fuel economy in diesel engines. Prog Energy Combust Sci 1:165–207
82. Haagen-Smit A, Fox M (1955) Automobile exhaust and ozone formation. SAE Trans 63(SAE Paper No. 550277):575–580

83. Huls T, Nickol H (1967) Influence of engine variables on exhaust oxides of nitrogen concentrations from a multicylinder engine. SAE Paper No. 670482
84. Starkman E, Stewart H, Zvonow V (1969) Investigation into formation and modification of exhaust emission precursors. SAE Paper No. 690020
85. Hames R, Merrion D, Ford H (1971) Some effects of fuel injection system parameters on diesel exhaust emissions. SAE Paper No. 710671
86. Khan I, Greeves G, Wang C (1973) Factors affecting smoke and gaseous emissions from direct injection engines and a method of calculation. SAE Trans 82(SAE Paper No. 730169):687–709
87. Yu R, Shahed S (1981) Effects of injection timing and exhaust gas recirculation on emissions from a D.I. diesel engine. SAE Trans 90(SAE Paper No. 811234):3873–3883
88. Newhall H (1967) Control of nitrogen oxides by exhaust recirculation, a preliminary theoretical study. SAE Trans 76(SAE Paper No. 670495):1820–1836
89. Benson J, Stebar R (1971) Effects of charge dilution on nitric oxide emission from a single-cylinder engine. SAE Trans 80(SAE Paper No. 710008):7–19
90. Komiyama K, Heywood J (1973) Predicting NO_x emissions and effects of exhaust gas recirculation in spark-ignition engines. SAE Trans 82(SAE Paper No. 730475):1458–1476
91. McEnally C, Pfefferle L, Atakan B, Kohse-Hoinghaus K (2006) Studies of aromatic hydrocarbon formation mechanisms in flames: progress toward closing the fuel gap. Prog Energy Combust Sci 32:247–294
92. Cheng W, Hamrin D, Heywood J, Hochgreb S, Min K, Norris M (1993) An overview of hydrocarbon emissions mechanisms in spark-ignition engines. SAE Trans J Fuels Lubricants 102(SAE Paper No. 932708):1207–1220
93. Henein N, Tagomori M (1999) Cold-start hydrocarbon emissions in port-injected gasoline engines. Prog Energy Combust Sci 25:563–593
94. Alkidas A (1999) Combustion-chamber crevices: the major source of engine-out hydrocarbon emissions under fully warmed conditions. Prog Energy Combust Sci 25:253–273
95. Haynes B, Wagner H (1981) Soot formation. Prog Energy Combust Sci 7:229–273
96. Smith O (1981) Fundamentals of soot formation in flames with application to diesel engine particulate emissions. Prog Energy Combust Sci 7:275–291
97. Kennedy I (1997) Models of soot formation and oxidation. Prog Energy Combust Sci 23:95–132
98. Richter H, Howard J (2000) Formation of polycyclic aromatic hydrocarbons and their growth to soot – a review of chemical reaction pathways. Prog Energy Combust Sci 26:565–608
99. Tree D, Svensson K (2007) Soot processes in compression ignition engines. Prog Energy Combust Sci 33:272–309
100. Hassaneen A, Samuel S, Morrey D, Gonzalez-Oropeza R (2009) Influence of physical and chemical parameters on characteristics of nanoscale particulate in spark ignition engine. SAE Paper No. 2009-01-2651
101. Ericsson P, Samson A (2009) Characterization of particulate emissions propagating in the exhaust line for spark-ignited engines. SAE Paper No. 2009-01-2654
102. Khan I (1969–1970) Formation and combustion of carbon in a diesel engine. Proc Inst Mech Eng 184(3J):36–43
103. Ahmad T, Plee S, Myers J (1982) Diffusion flame temperature – its influence on diesel particulate and hydrocarbon emissions. Paper presented at the international conference on diesel engines for passenger cars and light duty vehicles. Institution of Mechanical Engineers, London. IMECE Paper No. C101/82
104. Kummer J (1980) Catalysts for automobile emission control. Prog Energy Combust Sci 6:177–199
105. Koltsakis G, Stamatelos A (1997) Catalytic automotive exhaust aftertreatment. Prog Energy Combust Sci 23:1–39
106. Johnson T (2010) Diesel emission control in review. SAE Int J Fuels Lubricants 2(SAE Paper No. 2009-01-0121):1–12
107. Taylor C (1985) The internal combustion engine in theory and practice, Vol 2: Combustion, fuels, materials, design (rev. edition). The MIT Press, Cambridge, MA, pp 21–23
108. Taylor C (1985) The internal combustion engine in theory and practice, Vol 2: Combustion, fuels, materials, design (rev. edition). The MIT Press, Cambridge, MA, p 50
109. Heywood J (1988) Internal combustion engine fundamentals. McGraw-Hill, New York, pp 4–5
110. Heywood J (1988) Internal combustion engine fundamentals. McGraw-Hill, New York, p 475
111. Heywood J (1988) Internal combustion engine fundamentals. McGraw-Hill, New York, pp 217–220
112. Clenci A, Descombes G, Podevin P, Hara V (2007) Some aspects concerning the combination of downsizing with turbocharging, variable compression ratio, and variable intake valve lift. Proc Inst Mech Eng D J Automobile Eng 221:1287–1294
113. Van Nieuwstadt M, Kolmanovsky I, Morael P (2000) Coordinated EGR-VGT control for diesel engines: an experimental comparison. SAE Trans – J Engines 109(SAE Paper No. 2000-01-0266):238–249
114. Arnold S, Slupski K, Groskreutz M, Vrbas G, Cadle R, Shahed S (2011) Advanced turbocharging technologies for heavy-duty diesel engines. SAE Trans J Engines 110(SAE Paper No. 2001-01-3260):2048–2055
115. Kessel J, Schaffnit J, Schmidt M (1998) Modeling an real-time simulation of a turbocharger with variable turbine geometry (VGT). SAE Paper No. 980770
116. Hawley J, Wallace F, Pease A, Cox A, Horrocks R, Bird G (1997) Comparison of variable geometry turbocharging (VGT) over conventional wastegated machines to achieve lower emissions. In: IMechE autotech conference, Birmingham, UK, pp 245–259 (IMechE Seminar Publication: Automotive Engines and Powertrains, Paper No. C524/070/97)
117. Hawley J, Wallace F, Cox A, Horrocks R, Bird G (1999) Reduction of steady state NO_x levels from an automotive diesel engine using optimized VGT/EGR schedules. SAE Trans J Engines 108(SAE Paper No. 1999-01-0835):1172–1184

118. Tanin K, Wickman D, Montgomery D, Das S, Reitz R (1999) The influence of boost pressure on emissions and fuel consumption of a heavy-duty single-cylinder DI diesel engine. SAE Trans J Engines 108(SAE Paper No. 1999-01-0840):1198–1219
119. Cook J, Sun J, Buckland J, Kolmanovsky I, Peng H, Grizzle J (2006) Automotive powertrain control – A survey. Asian J Control 8(3):237–260
120. Leithgoeb R, Henzinger F, Fuerhapter A, Gschweitl K, Zrim A (2003) Optimization of new advanced combustion systems using real-time combustion control. SAE Paper No. 2003-01-1053
121. Corti E, Moro D, Solieri L (2007) Real-time evaluation of IMEP and ROHR-related parameters. SAE Paper No. 2007-24-0068
122. Leonhardt S, Muller N, Isermann R (1999) Methods for engine supervision and control based on cylinder pressure information. IEEE/ASME Trans Mechatron 4(3):235–245
123. Yoon M, Chung N, Lee M, Sunwoo M (2009) An engine-control-unit-in-the-loop simulator of a common-rail diesel engine for cylinder-pressure-based control. Proc Inst Mech Eng D J Automobile Eng 223:355–373
124. Turin R, Zhang R, Chang M (2008) Systematic model-based engine control design. SAE Int J Passenger Cars Electron Electr Syst 1(SAE Paper No. 2008-01-0994):413–424
125. Caton J (2008) Results from an engine cycle simulation of compression ratio and expansion ratio effects on engine performance. J Eng Gas Turbines Power 130(5):052809-1–052809-7
126. Wirbeleit F, Binder K, Gwinner D (1990) Development of pistons with variable compression height for increasing efficiency and specific power output of combustion engines. SAE Trans J Engines 99(SAE Paper No. 900229):543–557
127. Sugiyama T, Hiyoshi R, Takemura S, Aoyama S (2007) Technology for improving engine performance using variable mechanisms. SAE Trans J Engines 116(SAE Paper No. 2007-01-1290):803–812
128. Boggs D, Hilbert H, Schechter M (1995) The Otto-Atkinson cycle engine: fuel economy and emissions results and hardware design. SAE Trans J Engines 104(SAE Paper No. 950089):220–232
129. Leone T, Pozar M (2001) Fuel economy benefit of cylinder deactivation – Sensitivity to vehicle application and operating constraints. SAE Trans J Fuels Lubricants 110(SAE Paper No. 2001-01-3591):2039–2044
130. Gray C (1988) A review of variable engine valve timing. SAE Trans J Engines 97(SAE Paper No. 880386):6.631–6.641
131. Payri F, Desantes J, Corberaan J (1988) A study of the performance of an SI engine incorporating a hydraulically controlled variable valve timing system. SAE Trans J Engines 97(SAE Paper No. 880604):6.1133–6.1145
132. Ma T (1988) Effect of variable engine valve timing on fuel economy. SAE Trans J Engines 97(SAE Paper No. 880390):6.665–6.672
133. Meacham G (1970) Variable cam timing as an emission control tool. SAE Trans 79(SAE Paper No. 700673):2127–2144
134. Tuttle J (1980) Controlling engine load by means of late intake-valve closing. SAE Trans 89(SAE Paper No. 800794):2429–2441
135. Stobart R, Wijewardane A, Allen C (2010) The potential for thermo-electric devices in passenger vehicle applications. SAE Paper No. 2010-01-0833
136. Patterson A, Tett R, McGuire J (2009) Exhaust heat recovery using electro-turbogenerators. SAE Paper No. 2009-01-1604
137. Srinivasan K, Mago P, Zdaniuk G, Chamra L, Midkiff K (2008) Improving the efficiciency of the advanced injection low pilot ignited natural gas engine using organic Rankine cycles. ASME J Energy Resour Technol 130:022201-1–022201-7
138. Goldsmid H (1960) Principles of thermoelectric devices. Br J Appl Phys 11:209–217
139. Hussain Q, Brigham D, Maranville C (2010) Thermoelectric exhaust heat recovery for hybrid vehicles. SAE Int J Engines 2:1(SAE Paper No. 2009-01-1327):1132–1142
140. Barber E, Reynolds B, Tierney W (1951) Elimination of combustion knock ~ Texaco combustion process. SAE Trans 59(SAE Paper No. 510173):26–38
141. Davis C, Barber E, Mitchell E (1961) Fuel injection and positive ignition ~ A basis for improved efficiency and economy. SAE Trans 69(SAE Paper No. 610012):120–131
142. Mitchell E, Cobb J, Frost R (1968) Design and evaluation of a stratified charge multifuel military engine. SAE Trans 77(SAE Paper No. 680042):118–131
143. Alperstein M, Schafer G, Villforth F III (1974) Texaco's stratified charge engine – multifuel, efficient, clean, and practical. SAE Paper No. 740563
144. Pischinger F, Schmidt G (1978) Experimental and theoretical investigations of a stratified-charge engine with direct fuel injection. SAE Paper No. 785038
145. Hiraki H, Rife J (1980) Performance and NO_x model of a direct injection stratified charge engine. SAE Trans 89(SAE Paper No. 800050):336–356
146. Ullman T, Hare C, Baines T (1982) Emissions from direct-injected heavy-duty methanol-fueled engines (one dual injection and one spark-ignited) and a comparable diesel engine. SAE Trans 91(SAE Paper No. 820966):3154–3170
147. Giovanetti A, Ekchian J, Heywood J, Fort E (1983) Analysis of hydrocarbon emissions mechanisms in a direct injection spark-ignition engine. SAE Trans 92(SAE Paper No. 830587):2.925–2.947
148. Kato S, Onishi S (1988) New mixture formation technology of direct fuel injection stratified charge SI engine (OSKA) ~ Test result with gasoline fuel. SAE Trans J Engines 97(SAE Paper No. 881241):6.1497–6.1504
149. Onishi S, Jo S, Shoda K, Jo P, Kato S (1979) Active thermo-atmosphere combustion (ATAC) ~ A new combustion process for internal combustion engines. SAE Trans 88(SAE Paper No. 790501):1851–1860
150. Najt P, Foster D (1983) Compression-ignited homogeneous charge combustion. SAE Trans 92(SAE Paper No. 830264):1.964–1.979

151. Martinez-Frias J, Aceves S, Flowers D, Smith J, Dibble R (2000) HCCI engine control by thermal management. SAE Trans J Fuels Lubricants 109(SAE Paper No. 2000-01-2869):2646–2655
152. Law D, Kemp D, Allen J, Kirkpatrick G, Copland T (2001) Controlled combustion in an IC engine with a fully variable valve train. SAE Trans J Engines 110(SAE Paper No. 2001-01-0251):192–198
153. Rausen D, Stefanopoulou A, Kang J, Eng J, Kuo T (2005) A mean-value model for control of homogeneous charge compression ignition (HCCI) engines. J Dyn Syst Meas Contr 127:355–362
154. Shaver G, Gerdes J, Roelle M, Caton P, Edwards C (2005) Dynamic modeling of residual-affected homogeneous charge compression ignition engines with variable valve actuation. J Dyn Syst Meas Contr 127:374–381
155. Bengtsson J, Strandh P, Johansson R, Tunestal P, Johansson B (2006) Multi-output control of a heavy-duty HCCI engine using variable valve actuation and model predictive control. SAE Paper No. 2006-01-0873
156. Chiang C, Stefanopoulou A (2009) Sensitivity analysis of combustion timing of homogeneous charge compression ignition gasoline engines. J Dyn Syst Meas Contr 131:014506-1–014506-5
157. Fish A, Read I, Affleck W, Haskell W (1969) The controlling role of cool flames in two-stage ignition. Combust Flame 13:39–49
158. Takeda Y, Keiichi N, Keiichi N (1996) Emission characteristics of premixed lean diesel combustion with extremely early staged fuel injection. SAE Trans J Fuels Lubricants 105(SAE Paper No. 961163):938–947
159. Akagawa H, Miyamoto T, Harada A, Sasaki S, Shimazaki N, Hashizume T, Tsujimura K (1999) Approaches to solve problems of the premixed lean diesel combustion. SAE Trans J Engines 108(SAE Paper No. 1999-01-0183):120–132
160. Iwabuchi Y, Kawai K, Shoji T, Takeda Y (1999) Trial of new concept diesel combustion system – premixed compression-ignited combustion. SAE Trans J Engines 108(SAE Paper No. 1999-01-0185):142–151
161. Kimura S, Aoki O, Kitahara Y, Airoshizawa E (2001) Ultra-clean combustion technology combining a low-temperature and premixed combustion concept for meeting future emission standards. SAE Trans J Fuels Lubricants 110(SAE Paper No. 2001-01-0200):239–246
162. Kaneko N, Ando H, Ogawa H, Miyamoto N (2002) Expansion of the operating range with in-cylinder water injection in a premixed charge compression ignition engine. SAE Trans J Engines 111(SAE Paper No. 2002-01-1743):2309–2315
163. Shimazaki N, Tsurushima T, Nishimura T (2003) Dual mode combustion concept with premixed diesel combustion by direct injection near top dead center. SAE Trans J Engines 112(SAE Paper No. 2003-01-0742):1060–1069
164. Hasegawa R, Yanagihara H (2003) HCCI combustion in DI diesel engine. SAE Trans J Engines 112(SAE Paper 2003-01-0745):1070–1077
165. Okude K, Mori K, Shiino S, Moriya T (2004) Premixed compression ignition (PCI) combustion for simultaneous reduction of NOx and soot in diesel engines. SAE Trans J Fuels Lubricants 113(SAE Paper No. 2004-01-1907):1002–1013
166. Jacobs T, Bohac S, Assanis D, Szymkowicz P (2005) Lean and rich premixed compression ignition combustion in a light-duty diesel engine. SAE Trans J Engines 114(SAE Paper No. 2005-01-0166):382–393
167. Lechner G, Jacobs T, Chryssakis C, Assanis D, Siewert R (2005) Evaluation of a narrow spray cone angle, advanced injection timing strategy to achieve partially premixed compression ignition combustion in a diesel engine. SAE Trans J Engines 114(SAE Paper No. 2005-01-0167):394–404
168. Jacobs T, Assanis D (2007) The attainment of premixed compression ignition low-temperature combustion in a compression ignition direct injection engine. Proc Combust Inst 31:2913–2920

Books and Reviews

Ferguson CR, Kirkpatrick AT (2001) Internal combustion engines: applied thermosciences, 2nd edn. Wiley, New York
Heywood J (1988) Internal combustion engine fundamentals. McGraw-Hill, New York
Jennings BH, Obert EF (1944) Internal combustion engines: analysis and practice. International Textbook Company, Scranton
Pulkrabek WW (2004) Engineering fundamentals of the internal combustion engine, 2nd edn. Pearson Prentice-Hall, Upper Saddle River
Taylor CF (1985) The internal combustion engine in theory and practice – Vol 1: Thermodynamics, fluid flow, performance (rev), 2nd edn. The MIT Press, Cambridge, MA
Taylor CF (1985) The internal combustion engine in theory and practice – Vol 2: Combustion, fuels, materials, design (rev). The MIT Press, Cambridge, MA

Invasive Species

Anthony Ricciardi
Redpath Museum & McGill School of Environment, McGill University, Montreal, QC, Canada

Article Outline

Glossary

Biological invasion The process by which an organism is introduced to, and establishes a sustainable population in, a region beyond its native range.

Eradication The managed extirpation of an entire nonnative population.

Impact The effect of a nonnative species on its environment.

Invasibility The vulnerability of a habitat, community, or ecosystem to invasion.

Invasion ecology A multidisciplinary field that examines the causes and consequences of biological invasions.

Invasional meltdown The phenomenon in which multiple nonnative species facilitate one another's invasion success and impact.

Invasive species Nonnative species with conspicuously high colonization rates. Such species have the potential to spread over long distances. The term *invasive* is also used (often by policy makers) to describe colonizing species that cause undesirable ecological or economic impacts.

Nonnative species (synonyms: alien, exotic, foreign, nonindigenous) Species present in a region beyond their historic range.

Propagule pressure The quantity or rate of nonnative organisms released into an area.

Definition of the Subject

Biological invasion is the process by which a species is introduced, deliberately or inadvertently, into a new geographic region where it proliferates and persists. Outside their historic range (in which they evolved) such species are described as *nonnative* (or nonindigenous, exotic, alien). For a variety of reasons, the vast majority of introduced nonnative organisms fail to persist. Many of those that do establish self-sustaining populations do not spread very far or very fast beyond their point of introduction, and they often do not have conspicuous impacts on their environment. However, a small proportion (but a large and growing number) of nonnative species becomes *invasive* – that is, they may spread aggressively and/or have strong environmental effects. Invasive species are a global problem that threatens native biodiversity, the normal functioning of ecosystems, natural resources, regional economies, and human health. As such, they pose a major concern for conservation and management, and are the focus of a highly productive multidisciplinary field called *invasion ecology*.

Introduction

The potential impact of nonnative species has long been recognized by naturalists. In *The Origin of Species*, Darwin (1859) warned "Let it be remembered how powerful the influence of a single introduced tree or mammal has been shown to be [on native communities]." A century later, Charles Elton's groundbreaking monograph *The Ecology of Invasions by Animals and Plants* [1] helped inspire two generations of scientists to study what has become one of the world's most challenging environmental problems.

The major findings of this burgeoning research are summarized in recent texts by Lockwood et al. [2], Davis [3], Blackburn et al. [4], and Richardson [5].

This entry describes the causes and consequences of biological invasions, by synthesizing concepts from population biology, community ecology, evolution, biogeography, and conservation biology. First, the patterns and process of invasion are explored; then, some of its potential ecological and socioeconomic impacts are examined. Some major hypotheses and theoretical concepts explaining patterns of colonization and impact are presented. Next, management approaches to assessing, preventing, and mitigating this problem are considered. The entry ends with a brief glimpse at some of the emerging issues that will likely be the foci of future research.

Pattern and Process in Biological Invasion

The process of invasion comprises a sequence of events involving the transport, introduction, establishment, and spread of organisms into a new region. Organisms in various life stages may be moved by natural dispersal (e.g., passive transport by wind, water currents, or animals; active transport by the organism's own movements) or, far more frequently, by human activities

(e.g., transportation systems carrying people or material) across a geographic barrier that previously defined the limits of the historic range of the species. Most organisms will die soon after arrival, or reproduce for only a couple of generations; thus, the vast majority of introduction events fail to produce a sustainable population. If a sufficient number of healthy individuals arrive in a suitable habitat when conditions are favorable, then a self-sustaining population will develop and the species is said to be established. Although populations can sometimes establish from very small numbers, higher numbers of introduced individuals and more frequent introduction events (collectively termed *propagule pressure*) contribute to a higher probability of establishment [6].

In general, the more species introduced to an area, the more that become established in that area [7]. Lonsdale [8] presented an instructive model to describe the number of nonnative species in a region, E:

$$E = I \times S$$

where I is the number of species introduced (*colonization pressure* [7]) and S is the product of the survival rate of each species. S is a function of both the biological traits of the nonnative species and the environmental conditions of the target habitat; for example, all other things being equal, a higher survival rate would result from a closer match between the species' physiological requirements and the prevailing habitat conditions.

There is a variable time lag between initial introduction and establishment, followed by an exponential increase in abundance until the population reaches limits imposed by local abiotic and biotic conditions, at which point population growth diminishes. The range expansion of the species (increase in area occupied per unit time) is correlated with its population growth. The lag phase may range from being negligible (e.g., for a rapidly reproducing species) to extensive – during which the species may remain inconspicuous for years or decades prior to becoming abundant and widespread [9, 10]. For example, the first outbreak of the European gypsy moth (*Lymantria dispar*) in North America occurred two decades after it was initially released. A mussel introduced from the Red Sea remained rare for about 120 years prior to developing dense colonies on the Israeli Mediterranean coast [9]. Recognition of the lag phase phenomenon is critical to management; otherwise, it may lead to inaccurate assessments of benign invasion risk and low impact, as well as missed opportunities to control a nonnative species population while it was still small [10]. Nonmutually exclusive factors contributing to lag phases include: (1) density-dependent (Allee) effects, in which the organism's birth rate is correlated with its population density [11]; (2) adaptation and selection of new genotypes; (3) a change in the composition of the recipient community (e.g., the introduction of a pollinator or seed disperser [12], or the extinction of a dominant resident predator) that triggers the explosive growth of a previously subdued nonnative species; and (4) changing abiotic conditions (e.g., climate change [13]) that release the nonnative species from physiological constraints. Furthermore, the inability to detect an inconspicuous population in its early growth stages is often responsible for a substantial delay in the discovery of a nonnative species. Substantial lags in detection, caused by inadequacies in monitoring and taxonomic expertise, are a major hindrance to effective management [14].

The range expansion of an introduced species tends to fall into a few general patterns, each of which is characterized by an establishment lag phase, an expansion phase, and, when a geographic limit to suitable habitat is realized, a saturation phase [15]. In the simplest pattern, the species expands its range linearly through time; this pattern is the result of random short-distance dispersal outward in all directions through a homogeneous environment, and is often exhibited by rodents such as muskrats. The expanding range is modeled as a circle whose radius increases at a constant rate [16]. The probability of invasion at a given site is inversely proportional to the distance from the edge of the expanding colony and directly proportional to time.

A second pattern is defined by a slow initial rate of linear spread followed by an abrupt shift to a higher linear rate. This biphasic pattern, which has been observed in invasive birds such as the European starling (*Sturnus vulgaris*), occurs when long-distance migrants generate new satellite colonies not far from the primary colony; the coalescence of satellites into the expanding primary colony generates a higher linear rate of expansion. A third pattern occurs when long-distance

dispersers create numerous remote satellite colonies that begin to expand their range independent of each other; their continuous coalescence generates an exponential expansion phase, as exhibited by European cheatgrass (*Bromus tectorum*) in North America and tiger pear cactus (*Opuntia aurantiaca*) in South Africa [15, 17]. In this pattern, a prolonged lag phase often occurs prior to conspicuous exponential growth. Genetic adaptation is another mechanism that can produce the enhanced rate of expansion that characterizes the second and third patterns, but the occurrence of long-distance migrants is probably the more common cause. Via long-distance "jumps," migrants may establish satellite colonies that are remote from the expanding edge of the primary colony; the overall rate of range expansion is driven more by the number of these satellite colonies than by their individual size [16]. The pattern is more pronounced where human vectors dominate dispersal, such that there would be multiple introductions of satellite colonies within a region (e.g., the transport of zebra mussels and aquatic weeds between river basins by recreational boats, or introductions of a marine invertebrate along a coastline via ballast water release at various ports). In this case, the probability of dispersal to a given site is nearly independent of time and distance from the primary colony but instead is driven largely by human-mediated dispersal opportunity [18].

Factors Affecting Establishment Success

In addition to propagule pressure, other biotic and abiotic factors have been hypothesized to explain why some species are better invaders, and why some systems are more invaded, than others. Attributes associated with highly invasive species include an ability to rapidly reproduce from small numbers (a high intrinsic rate of population growth), broad environmental tolerance, and mechanisms of exploiting human transportation vectors and human-modified landscapes. A popular view is that generalist species are better invaders than specialists, because the former can thrive in a broader range of habitat conditions (*niche breadth-invasion success hypothesis* [19]). As such, traits that enable species to cope with new environments (e.g., diet breadth, physiological tolerance [20, 21]), or proxy variables that suggest broad tolerance (e.g., latitudinal range [22]), are generally good predictors of invasion success. Among vertebrates, brain size also generally predicts invasion success [23–25], perhaps because it facilitates behavioral flexibility in new environments (but see [26]). Similarly, invasive plants tend to be more phenotypically plastic than noninvasive plants [27]. Traits associated with reproduction are often correlated with the post-establishment success (abundance and range size) of plants [20, 28]. However, the most important factor limiting the large-scale distribution of a species is whether it is valued by humans for domestication [29–32] or, for a species that is not introduced deliberately, whether its life history allows it to be easily transported by human vectors operating on a global scale [33, 34].

Much research on the question of why some communities or systems are more invasible has addressed the concept of *biotic resistance*, which posits that biotic interactions between nonnative species and resident enemies can limit establishment and post-establishment success. The logical extension of this concept is that resident species diversity may act as a barrier to invasion – an idea promoted by Elton [1] to explain the seemingly disproportionate invasibility of species-poor systems such as oceanic islands and highly disturbed areas such as agricultural fields. Most support for Elton's hypothesis is derived from terrestrial plant communities and is equivocal. Over a range of scales, from small garden plots to regional landscapes, positive correlations between native and nonnative species richness have been observed, reflecting shared responses to external variables [35]. Where negative correlations exist, they are found only at local (m^2) scales in experimental manipulations [36]. Numerous studies suggest that competition, herbivory, and native species richness can strongly inhibit the performance (and impact) of nonnative plants following establishment [37, 38], but little evidence suggests that these interactions can prevent establishment when abiotic conditions are favorable and propagule pressure is high. The lesson for managers from these studies is that even highly diverse native communities are often readily invaded by nonnative species, but the reduction of local species richness may accelerate invasion [35].

Most recent studies of invasion mechanisms focus on two popular hypotheses: *fluctuating resource availability* and *enemy release*. The former hypothesis

proposes that a system's susceptibility to plant invasions varies with fluctuations in unused resources (e.g., light, water, space, nutrients). Where propagule pressure exists, invasion will be promoted by a sudden increase in resource supply (such as through nutrient pollution) or reduced uptake by resident species (following a disturbance such as clearcutting or fire) [39, 40]. Nutrient-rich habitats do experience more plant invasions, but native plants may not always outperform nonnatives in low-resource conditions [41]. Highly disturbed environments are also believed to be more invasible [1]. Nonnative species may dominate a habitat following a disturbance event that is outside the evolutionary experience of the natives; otherwise, natural disturbance may contribute to a system's resistance to invasion [42].

The enemy release hypothesis attributes the success of nonnative species to their escape from specialized natural enemies upon arrival to a new region, and their inherent advantage over resident competitors that are burdened by their own enemies [43]. One reason why plants that are subject to strong herbivory in their native range can thrive in novel regions is that, in the absence of specialized enemies, they may reallocate the energetic costs of defense toward reproduction and growth, and thus become more competitive [44]. It follows that fast-growing species adapted to resource-rich environments may benefit most from the absence of specialized enemies; thus, multiple mechanisms (enemy release, disturbance, resource addition) may act synergistically to drive such invasions [45].

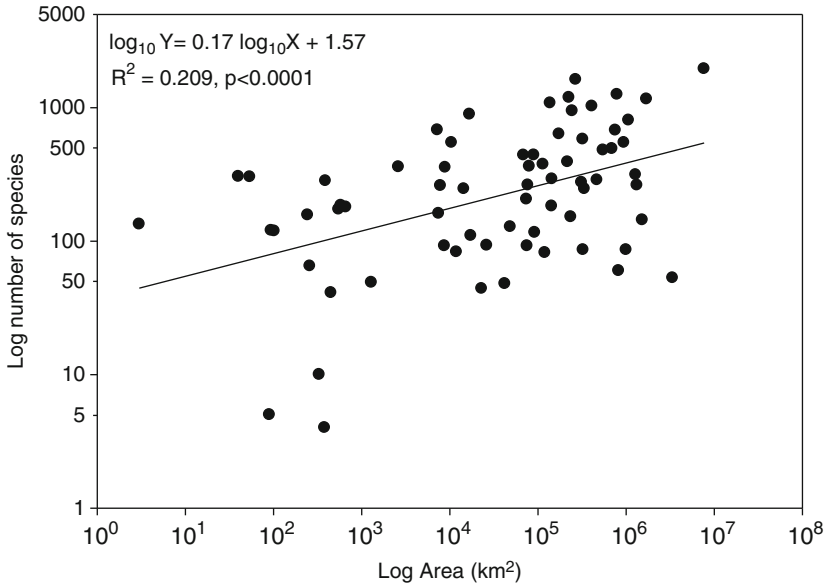
Modern Invasions as Unprecedented Global Change

The spread of species into regions beyond their native range has accelerated exponentially during the past millennium because of human activities such as agriculture, international travel, and global trade. There is a strong link between trade activity and the global distribution of nonnative species [46, 47]. International trade often involves cargo moved by transoceanic ships, which can carry an enormous number of organisms on their hulls and especially in their ballast tanks. Tens of thousands of ships are estimated to be collectively transporting several thousand species around the planet on any given day [48].

Most countries have recorded the establishment of several hundred nonnative species, including invertebrates, vertebrates, plants, bacteria, and fungi (Fig. 1). Human influence is reflected in the improbable composition of modern species assemblages worldwide: African grasses dominate large tracts of the Neotropical region [30], European mammals and birds are abundant in Australia and New Zealand [29, 32], Eurasian invertebrates and fishes dominate food webs in the North American Great Lakes [34], and over 25% of the nonnative species in the Baltic Sea originate from the Pacific and Indian Oceans [50]. Over a decade ago, it was estimated that nonnative plants covered at least 3% of the Earth's ice-free land mass, excluding the already immense area under agricultural cultivation [51]. Nonnative species comprise substantial fractions of flora and fauna on continental areas and, especially, on islands (Table 1). The majority of these invasions have occurred over the past few centuries, coinciding with steep increases in global trade, human travel, and land use. Invaders are presently colonizing new regions at rates that are several orders of magnitude faster than prior to human arrival (Fig. 2). Even the seemingly remote Antarctic continent and its surrounding islands have been colonized by nearly 200 nonnative species of terrestrial plants, invertebrates, and vertebrates within the past two centuries, owing to the effects of scientific exploration, increased accessibility by air and by sea, a burgeoning tourist industry (tens of thousands of visitors annually), and a changing climate [59]. The modern rate and geographic extent of invasion is without historical precedent [58].

Ecological Impacts

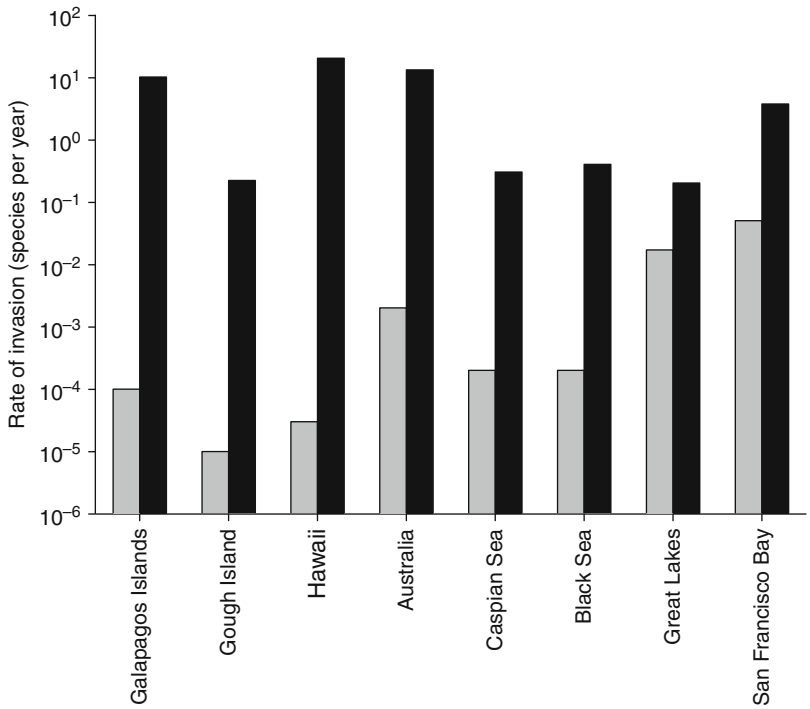
Most nonnative species appear to have only minor effects on their invaded systems, but this observation is tempered by two caveats: The impacts of the vast majority of invasions have not been studied [60], and even species that are generally benign can become disruptive at different times or different locations [61]. In many cases, nonnative species can profoundly affect ecosystems by altering community composition, resident species interactions, physical habitat structure, hydrology, nutrient cycling, contaminant cycling, primary production, and natural disturbance (fire, flood, erosion) regimes [17, 62–64]. They can disrupt food



Invasive Species. Figure 1
Number of nonnative vascular plant species versus area for regions worldwide (Data from [49]. Line is fitted by least-squares regression)

Invasive Species. Table 1 Proportion (%) of extant species comprised by established nonnative freshwater fishes, breeding birds, land mammals, and vascular plants in selected regions (Data from [32, 49, 52–57])

Region	Fishes	Birds	Mammals	Plants
Continental areas				
Europe	10	3	19	6
Russia	7	n/a	17	n/a
Southern Africa	11	1	12	4
North America (north of Mexico)	8	4	19	11
South America	<1	<1	4	n/a
Australia	13	6	14	1
Islands				
Puerto Rico	71	35	40	12
Bahamas	14	9	n/a	18
Bermuda	n/a	30	50	65
Hawaii	88	33	89	44
Madagascar	17	2	5	3
Japan	15	2	14	n/a
New Zealand	38	18	40	40



Invasive Species. Figure 2

Prehistoric versus modern rates of invasion (number of nonnative species established per year) for various regions. Prehistoric rates (grey bars) are before human settlement and were estimated from the fossil record or by calculating numbers of "native" species (excluding endemics) that have become established in the region over time. Modern rates (black bars) are inferred from discovery rates averaged over the past 40–100 years (Modified from [58])

webs [65, 66] and plant-animal mutualisms that are crucial for pollination and seed dispersal [67, 68]. Even where environmental stressors such as habitat degradation have already caused population declines of native species, invasions can accelerate these declines [69]. They are a major cause of animal extinctions [70, 71], particularly in insular habitats, such as lakes, river basins, and islands [72, 73]. The invasion-mediated loss of genetically distinct native populations in continental regions has likely been grossly underestimated. There are examples of once widely distributed species being reduced to near extinction as a result of introduced pathogens [17]. Some of the greatest impacts on biodiversity are caused by nonnative predators, and the most conspicuous examples involve introductions to oceanic islands [74, 75] and freshwater ecosystems [76]. Large mammalian herbivores have also had devastating effects on island biodiversity [77, 78]. Other factors contributing to species loss at local to global scales include hybridization [79, 80], competition [69], disease transfer [81], food web alteration [65, 66, 68], and physical habitat alteration [17].

Entire ecosystems may be transformed by invaders that alter resource availability, disturbance regimes, or habitat structure. Some invaders alter the disturbance regime of habitats through fire suppression (e.g., the shrub *Mimosa pigra* in Australian flood plains), fire enhancement (e.g., Eurasian cheatgrass *Bromus tectorum* in the Western United States), increased erosion (e.g., the Australian shrub *Acacia mearnsii* in South Africa), reduced erosion (e.g., exotic plants with extensive root systems that stabilize hills, stream banks, or sand dunes), and increased soil disturbance (e.g., the rooting activities of feral European pigs *Sus scrofa* can destroy the herbaceous understory of a forest, causing soil mineral depletion, rapid organic decomposition, and loss of habitat). Through its

filter-feeding activities, the zebra mussel (*Dreissena polymorpha*) has dramatically increased water transparency in North American and European lakes, thus stimulating the growth of benthic algae and macrophytes and altering physical habitat for invertebrates and fishes [82]. In Hawaii, a nitrogen-fixing tree, *Myrica faya*, significantly enriched nutrient-poor volcanic soils at a rate 90-times greater than native plants and thus has a dominant influence on ecosystem properties including soil chemistry and productivity [83]; *Myrica* has also added habitat structure, shading, and high-quality leaf litter that has promoted enhanced populations of nonnative earthworms [84].

Socioeconomic Impacts

The economic value of cultivated nonnative species (such as crop plants) is widely appreciated, but the same cannot be said for the enormous costs incurred by invasions in general. In several countries, nonnative species comprise more than 40% of all harmful weeds, 30% of arthropod pests, and 70% of plant pathogens, and cause substantial losses in total crop production each year [85]. A single invasive forest insect, the emerald ash borer beetle, is projected to cost the United States $10 billion over the next decade [86]. The 2001 outbreak of foot-and-mouth disease in the United Kingdom, linked to illegal meat imports, cost $25 million USD and required the slaughter of ~11 million animals [87]. The annual costs of 16 nonnative species to fisheries, agriculture, and forestry in Canada are projected to be as high as $34 billion CDN [88]. The combined annual costs of biological invasions in the United States, United Kingdom, Australia, India, South Africa, and Brazil are estimated to be $314 billion USD. Assuming similar costs worldwide, the global economic damage attributable to invasions amounts to US $1.4 trillion per year, which constitutes 5% of the global economy [85].

Whereas some nonnative species perform valuable roles, other nonnatives can degrade ecosystem services – including water purification, soil stabilization, agricultural yield, disease regulation, and climate regulation [89]. The conservation of water resources in African countries is threatened by introduced plants [90], whereas pollination services provided by European honeybees are threatened by Asian *Varroa* mites, whose parasitism has destroyed entire hives [91]. Animal (including human) health, in general, is threatened by invasions that spread parasites, diseases, and their vectors (e.g., mosquitoes [92]). Invasions can also alter the transmission of parasites to humans by introducing hosts to novel regions [93]. About 100 species (~6%) of nonnative invertebrates (e.g., spiders, mosquitoes, nematodes) in Europe adversely affect human or animal health, and these are a subset of ~1,300 nonnative species in the region that have documented socioeconomic impacts [94]. Climate change is expected to drive a new wave of such invasions, as suggested by the recent occurrence in Northern Europe of the tropical virus that causes "bluetongue disease" that resulted from the introduction of infected livestock from a Mediterranean country [95].

Management of Invasions

Risk Assessment

Managers have few tools for prioritizing invasion threats because reliable predictive methods are scarce (but see [96, 97]). Progress in developing a predictive understanding of impact has been hampered by the lack of standardized metrics. Parker et al. [60] proposed a metric for impact (I) that can be compared across species and invaded sites:

$$I = R \times A \times E$$

where R is the total area occupied by the nonnative species in its invaded range, A is its abundance (in numbers or biomass per square meters) in the invaded range, and E is its per-capita effect based on the functional ecology and behavior of individuals (e.g., filtration rate of mussels, functional response of predators, rate of habitat conversion for ecosystem engineers). Data on per-capita effects are often scarce, but inferences regarding the magnitude of impact may be drawn from abundance, which has been shown to be a useful predictor of impact [61]. Range size, in contrast, may not necessarily be a good predictor. Beyond the trivial expectation that the impacts of an invading species accumulate as it occupies more territory, there is no statistical correlation between the invasion success of a species (i.e., its rate of establishment success or spread) and the magnitude of its impact [98]. Even relatively poor invaders can have strong local

impacts on native populations (e.g., the Asian clam *Potamocorbula amurensis*; Atlantic salmon *Salmo salar*), whereas highly successful colonizers do not necessarily displace native species (e.g., freshwater jellyfish *Craspedacusta sowerbyi*). One generalization that has emerged from numerous case studies is that high-impact invaders often represent novel life forms in the invaded system. They acquire and use resources differently than resident species, possess defense mechanisms and "weapons" that are foreign to the invaded community [99], and may have predatory capabilities to which residents are poorly adapted. Such species tend to belong to taxonomic or functional groups that were not present in the ecosystem prior to invasion [100–102]. As such, the phylogenetic distinctiveness of the invader in its novel environment might be an indicator of its impact potential [101, 102].

A major challenge to prediction is context-dependent variation generated by site-specific environmental factors [60, 61]. The best predictor of the colonization success and impact of an introduced plant or animal is its invasion history [20, 61]. Although impacts vary across a heterogeneous environment, models may be developed to predict the impact (or abundance) of a species with a well-documented impact history [61], but the predictive power of such models is diminished at sites that have been highly invaded. Nonnative species can interact in multiple ways to produce unpredictable effects [12, 75], sometimes by facilitating each other's spread and impact (i.e., *invasional meltdown* [103]).

Prevention

Given the growing frequency of invasions, their profound impacts, and the substantive resources required to control rapidly spreading species after they become established, the most cost-effective management strategy is prevention [14]. Arguably, invasions warrant similar investments in preparedness and response planning as natural disasters; despite being slower in their onset, invasions have more persistent impacts and a greater scope of ecological and economic damage than natural disasters [104].

Prevention involves controlling either species entry or establishment. Preventing entry of nonnative species begins with the identification and control of dominant transportation vectors and pathways [14]. The effectiveness of vector-control policies requires rigorous inspection, enforcement, evaluation, and – where necessary – refinement, as has been demonstrated by the evolution of a management program to control ballast water–mediated invasions in the Great Lakes [105]. An additional preventative approach is to manage ecosystems so as to reduce their vulnerability to invasions – e.g., via restoration of intact native communities in degraded areas, managed disturbance (e.g., fire, river flow) regimes, and manipulation of resource supply (nutrients, water supply) [14, 106]. Cultivated systems can be designed with resistance in mind; for example, the use of polycultures (e.g., diversified crops, mixed forest stands) has been demonstrated to reduce harmful outbreaks of invasive pests [107]. The spatial modification of habitats (such as the use of small-scale dispersal barriers) may also be employed to limit colonization [11].

Eradication

The Convention on Biological Diversity [article 8(h)] directs signatory nations to "prevent the introduction of, control or eradicate those alien species which threaten ecosystems." Eradication, the removal of a nonnative population, can lead to the recovery of previously threatened native species [108, 109]. Several conditions must be met for an eradication program to be successful [110]: (1) The target species must be detected at low densities. (2) Its biology must make it susceptible to control measures. (3) Resources must be sufficient to complete the project. (4) Managers must have the authority and public support to take all necessary steps. (5) Re-invasion must be prevented. Also influencing the success of eradication are the reproductive and dispersal capabilities of the invader, both of which determine how fast it will spread. The probability of success is highest in the initial stages of invasion when spatial spread is still limited; hence, early detection and rapid response are crucial, particularly for species that can reproduce and disperse rapidly [14].

Owing to the indirect effects of nonnative species, eradication can have unanticipated negative consequences. Where multiple invaders exist, particularly in simple food webs (e.g., on islands), the removal of a nonnative predator or herbivore can cause the

proliferation of a second invader that was previously controlled by the target species through top-down regulation [111, 112]. For example, the eradication of feral cats from Macquarie Island led to a population explosion of an invasive herbivore – European rabbit [112]. The explosion of rabbits was accompanied by large-scale habitat alteration characterized by a shift in vegetation that favored fast-growing plants, some of which themselves were nonnative. Similarly, the removal of cats from Little Barrier Island, New Zealand, released the introduced Pacific rat (*Rattus exulans*) from top-down control and led to a reduction in the breeding success of an endangered endemic seabird (Cook's petrel, *Pterodroma cookii*), apparently due to nest predation by the rat; subsequent eradication of the rat was followed by a rapid rise in the seabird's breeding success [111]. Additional effects of eradication on multiply invaded systems might be to increase predation pressure on natives as a result of nonnative predators shifting their diets following the removal of nonnative prey, or to release one or more nonnative species from competition by removing a superior competitor.

Maintenance Control

When dealing with nonnative species with strong Allee effects, eradication may involve culling individuals to bring a population below sustainable levels [11]. If eradication fails, or is impossible, the next option is maintenance control of the invader at acceptable population levels, using mechanical, chemical, or biological control methods. Mechanical control, such as hunting, may be particularly effective on islands and other geographically restricted areas. Chemical control involves the application of pesticides to reduce the abundance of a target species, but high economic costs and human health risks constrain the application of chemicals over large areas. Moreover, pesticides often impact nontarget species (including native competitors), sometimes to the benefit of the target itself [113].

Biological control involves the introduction of a nonnative species (usually a predator, herbivore, or parasite) to reduce an established nonnative pest to less harmful densities. This technology is considered to be a more desirable alternative to pesticide use, despite its potential for unanticipated consequences. Because the introduced agents can disperse beyond the target area and evolve to exploit new hosts, nontarget species may be attacked and even driven to extinction [17, 114]. The assumption underlying biological control is that nonnative species proliferate to harmful levels because they have escaped their natural enemies. However, indirect (e.g., competitive) effects may sometimes be more important than top-down consumer regulation. Under these situations, the introduction of a biological control species may have a counterproductive effect [115]. Difficulties in predicting such complex community interactions can obviously compromise ecological risk assessments.

Future Directions

The questions underlying invasion ecology – that is, why some species are more successful and have greater impact than others, why some systems are more vulnerable to invasion, and how ecosystem functions and services are affected by invasion – are clearly of societal importance and will remain relevant in the future, as invasive species are increasingly viewed as a biosecurity issue [87]. The extent and impact of invasions will be further exacerbated by climate change, and synergies between nonnative species and other human-mediated stressors will become more frequent. Future research foci will include the consequences associated with cultivation of novel biofuels and bioenergy crops [116] and the expanded use of genetically modified organisms [117]. Moreover, there may be increasing interest among conservation biologists to relocate native species deemed to be threatened by climate change or other stressors, and some plants and animals could be moved well beyond their historical ranges [73]. Each of these practices will have potentially high ecological risks whose assessment will require more powerful forecasting methods than are currently available. Thus, we can anticipate a growing need for invasion ecology to develop a more predictive understanding of the impact of nonnative organisms.

Bibliography

Primary Literature

1. Elton CS (1958) The ecology of invasions by animals and plants. Methuen, London
2. Lockwood JL, Hoopes MF, Marchetti MP (2007) Invasion ecology. Blackwell, Oxford

3. Davis MA (2009) Invasion biology. Oxford University Press, Oxford
4. Blackburn TM, Lockwood JL, Cassey P (2009) Avian invasions. Oxford University Press, Oxford
5. Richardson DM (ed) (2011) Fifty years of invasion ecology – the legacy of Charles Elton. Wiley-Blackwell, Chichester
6. Simberloff D (2009) The role of propagule pressure in biological invasions. Annu Rev Ecol Evol System 40:81–102
7. Lockwood JL, Cassey P, Blackburn TM (2009) The more you introduce the more you get: the role of propagule and colonization pressure in invasion success. Divers Distrib 15: 904–910
8. Lonsdale WM (1999) Global patterns of plant invasions and the concept of invasibility. Ecology 80:1522–1536
9. Rilov G, Benayahu Y, Gasith A (2004) Prolonged lag in population outbreak of an invasive mussel: a shifting habitat model. Biol Invasions 6:347–364
10. Crooks JA (2005) Lag times and exotic species: the ecology and management of biological invasions in slow-motion. Ecoscience 12:316–329
11. Tobin PC, Berec L, Liebhold AM (2011) Exploiting allee effects for managing biological invasions. Ecol Lett 14:615–624
12. Richardson DM, Allsopp N, D'Antonio CM, Milton SJ, Rejmànek M (2000) Plant invasions – the role of mutualisms. Biol Rev 75:65–93
13. Witte S, Buschbaum C, van Beusekom EE, Reise K (2010) Does climatic warming explain why an introduced barnacle finally takes over after a lag of more than 50 years? Biol Invasions 12:3579–3589
14. Hulme PE (2006) Beyond control: wider implications for the management of biological invasions. J Appl Ecol 43:835–847
15. Shigesada N, Kawasaki K (1997) Biological invasions: theory and practice. Oxford University Press, Oxford
16. Hengeveld R (1989) Dynamics of biological invasions. Chapman and Hall, New York
17. Mack RN, Simberloff D, Lonsdale WM, Evans H, Clout M, Bazzazz FA (2000) Biotic invasions: causes, epidemiology, global consequences, and control. Ecol Appl 10:689–710
18. MacIsaac HJ, Grigorovich IA, Ricciardi A (2001) Reassessment of species invasions concepts: the Great Lakes basin as an example. Biol Invasions 3:405–416
19. Vasquez DP (2006) Exploring the relationship between niche breadth and invasion success. In: Cadotte M, McMahon SM, Fukami T (eds) Conceptual ecology and invasion ecology. Springer, Dordrecht, pp 307–322
20. Hayes KR, Barry SC (2008) Are there any consistent predictors of invasion success? Biol Invasions 10:483–506
21. Blackburn TM, Cassey P, Lockwood JL (2009) The role of species traits in the establishment success of exotic birds. Glob Chang Biol 15:2852–2860
22. Goodwin BJ, McAllister AJ, Fahrig L (1999) Predicting invasiveness of plant species based on biological information. Conserv Biol 13:422–426
23. Sol D, Timmermans S, Lefebvre L (2002) Behavioural flexibility and invasion success in birds. Anim Behav 63:495–502
24. Sol D, Bacher S, Reader SM, Lefebvre L (2008) Brain size predicts the success of mammal species introduced to novel environments. Am Nat 172:S63–S71
25. Amiel JJ, Ringley R, Shine R (2011) Smart moves: effects of relative brain size on establishment success of invasive amphibians and reptiles. PLoS One 6(4):e18277
26. Drake JM (2007) Parental investment and fecundity, but not brain size, are associated with establishment success in introduced fishes. Funct Ecol 21:963–968
27. Davidson AM, Jennions M, Nicotra AB (2011) Do invasive species show higher phenotypic plasticity than native species and, if so, is it adaptive? A meta-analysis. Ecol Lett 14:419–431
28. Ordonez A, Wright IJ, Olff H (2010) Functional differences between native and alien species: a global-scale comparison. Funct Ecol 24:1353–1361
29. Long JL (1981) Introduced birds of the world. Universe Books, New York
30. Mack RN, Lonsdale WM (2001) Humans as global plant dispersers: getting more than we bargained for. BioScience 51:95–102
31. Reichard SH, White P (2001) Horticulture as a pathway of invasive plant introductions in the United States. BioScience 51:103–113
32. Long JL (2003) Introduced mammals of the world. CSIRO Publishers, Collingwood
33. Suarez AV, Holway DA, Case TJ (2001) Patterns of spread in biological invasions dominated by long-distance jump dispersal: insights from Argentine ants. Proc Natl Acad Sci USA 98:1095–1100
34. Ricciardi A (2006) Patterns of invasion in the Laurentian Great Lakes in relation to changes in vector activity. Divers Distrib 12:425–433
35. Fridley JD, Stachowicz JJ, Naeem S, Sax DF, Seabloom EW, Smith MD, Stohlgren TJ, Tilman D, Von Holle B (2007) The invasion paradox: reconciling pattern and process in species invasions. Ecology 88:3–17
36. Herben T, Mandák B, Bímova K, Münzbergova Z (2004) Invasibility and species richness of a community: a neutral model and a survey of published data. Ecology 85:3223–3233
37. Levine JM, Adler PB, Yelenik SG (2004) A meta-analysis of biotic resistance to exotic plant invasions. Ecol Lett 7:975–989
38. Olofsson J, Oksanen L, Callaghan T, Hulme PE, Oksanen T, Suominen O (2009) Herbivores inhibit climate-driven shrub expansion on the tundra. Glob Chang Biol 15:2681–2693
39. Davis MA, Grime JP, Thompson K (2000) Fluctuating resources in plant communities: a general theory of invasibility. J Ecol 88:528–534
40. Davis MA, Pelsor M (2001) Experimental support for a resource-based mechanistic model of invasibility. Ecol Lett 4:421–428
41. Funk JL, Vitousek PM (2007) Resource-use efficiency and plant invasion in low-resource systems. Nature 446:1079–1081
42. Leprieur F, Hickey MA, Arbuckle CJ, Closs GP, Brosse A, Townsend CR (2006) Hydrological disturbance benefits a native fish at the expense of an exotic fish. J Appl Ecol 43:930–939

I

43. Keane RM, Crawley MJ (2002) Exotic plant invasions and the enemy release hypothesis. Trends Ecol Evol 17:164–170
44. Blossey B, Nötzold R (1995) Evolution of increased competitive ability in invasive nonindigenous plants: a hypothesis. J Ecol 83:887–889
45. Blumenthal D, Mitchell CE, Pyšek P, Jarošik V (2009) Synergy between pathogen release and resource availability in plant invasion. Proc Natl Acad Sci USA 106:7899–7904
46. Westphal MI, Browne M, MacKinnon K, Noble I (2008) The link between international trade and the global distribution of invasive alien species. Biol Invasions 10:391–398
47. Lin W, Cheng X, Xu R (2011) Impact of different economic factors on biological invasions on the global scale. PLoS One 6(4):e18797
48. Carlton JT (1999) The scale and ecological consequences of biological invasion in the World's oceans. In: Sandlund OT, Schei PJ, Viken Å (eds) Invasive species and biodiversity management. Kluwer, Dordrecht, pp 195–212
49. Vitousek PM, D'Antonio CM, Loope LL, Rejmanek M, Westbrooks R (1997) Introduced species: a significant component of human-caused global change. N Z J Ecol 21:1–16
50. Leppäkoski E, Olenin S (2000) Non-native species and rates of spread: lessons from the brackish Baltic Sea. Biol Invasions 2:151–163
51. Mack RN (1997) Plant invasions: early and continuing expressions of global change. In: Huntley B, Cramer W, Morgan AV, Prentice HC, Allen JRM (eds) Past and future rapid environmental changes: the spatial and evolutionary responses of terrestrial biota. Springer, Berlin, pp 205–216
52. Bogustkaya NG, Naseka AM (2002) Freshwater fishes of Russia database. http://www.zin.ru/animalia/pisces/. Accessed 22 July 2011
53. Callmander MW, Phillipson PB, Schatz GE, Andriambololonera S, Rabarimanarivo M, Rakotonirina N, Raharimampionona J, Chatelain C, Gautier L, Lowry PP (2011) The endemic and non-endemic vascular flora of Madagascar updated. Plant Ecol Evol 144:121–125
54. Froese R, Pauly D (2011) FishBase. www.fishbase.org, version (06/2011). Accessed 11 July 2011
55. Hockey PAR, Dean WRJ, Ryan PG (2005) Roberts – birds of southern Africa, 7th edn. The Trustees of the John Voelcker Bird book Fund, Cape Town
56. Koehn JD, MacKenzie RF (2004) Priority management actions for alien freshwater fish species in Australia. N Z J Mar Freshw Res 38:457–472
57. Lehtonen H (2002) Alien freshwater fishes of Europe. In: Leppakoski E, Gollasch S, Olenin S (eds) Invasive aquatic species of Europe: distribution, impacts and management. Kluwer, Dordrecht
58. Ricciardi A (2007) Are modern biological invasions an unprecedented form of global change? Conserv Biol 21:329–336
59. Frenot Y, Chown SL, Whinam J, Selkirk PM, Convey P, Skotnicki M, Bergstrom DM (2005) Biological invasions in the Antarctic: extent, impacts and implications. Biol Rev 80:45–72
60. Parker IM, Simberloff D, Lonsdale WM, Goodell K, Wonham M, Kareiva PM, Williamson MH, Von Holle B, Moyle PB, Byers JE, Goldwasser L (1999) Impact: toward a framework for understanding the ecological effects of invaders. Biol Invasions 1:3–19
61. Ricciardi A (2003) Predicting the impacts of an introduced species from its invasion history: an empirical approach applied to zebra mussel invasions. Freshw Biol 48:972–981
62. Brooks ML, D'Antonio CM, Richardson DM, Grace JB, Keeley JE, DiTomaso JM, Hobbs RJ, Pellant M, Pyke D (2004) Effects of invasive alien plants on fire regimes. BioScience 54:677–688
63. Croll DA, Maron JL, Estes JA, Danner EM, Byrd GV (2005) Introduced predators transform subarctic islands from grassland to tundra. Science 307:1959–1961
64. Simberloff D (2011) How common are invasion-induced ecosystem impacts? Biol Invasions 13:1255–1268
65. Spencer CN, McClelland BR, Stanford JA (1991) Shrimp stocking, salmon collapse, and eagle displacement. BioScience 41:14–21
66. Burghardt KT, Tallamy DW, Philips C, Shropshire KJ (2010) Non-native plants reduce abundance, richness, and host specialization in lepidopteran communities. Ecosphere 1(5):1–22
67. Traveset A, Richardson DM (2006) Biological invasions as disruptors of plant reproductive mutualisms. Trends Ecol Evol 21:208–216
68. Sekercioglu CH (2011) Functional extinctions of bird pollinators cause plant declines. Science 331:1019–1020
69. Ricciardi A (2004) Assessing species invasions as a cause of extinction. Trends Ecol Evol 19:619
70. Clavero M, García-Berthou E (2005) Invasive species are a leading cause of animal extinctions. Trends Ecol Evol 20:110
71. Clavero M, Brotons L, Pons P, Sol D (2009) Prominent role of invasive species in avian biodiversity loss. Biol Conserv 142:2043–2049
72. Ebenhard T (1988) Introduced birds and mammals. Swed Wildl Res (Viltrevy) 13(4):1–107
73. Ricciardi A, Simberloff D (2009) Assisted colonization is not a viable conservation strategy. Trends Ecol Evol 24:248–253
74. Fritts TH, Rodda GH (1998) The role of introduced species in the degradation of island ecosystems: a case history of Guam. Annu Rev Ecol Syst 29:113–140
75. Blackburn TM, Petchey OL, Cassey P, Gaston KJ (2005) Functional diversity of mammalian predators and extinction in island birds. Ecology 86:2916–2923
76. Witte F, Goldschmidt T, Wanink J, Vanoijen M, Goudswaard K, Wittemaas E, Bouton N (1992) The destruction of an endemic species flock – quantitative data on the decline of the haplochromine cichlids of Lake Victoria. Environ Biol Fish 34:1–28
77. Spear D, Chown SL (2009) Non-indigenous ungulates as a threat to biodiversity. J Zool 279:1–17
78. Carrete M, Serrano D, Illera JC, López G, Vögeli M, Delgado A, Tella JL (2009) Goats, birds, and emergent diseases: apparent and hidden effects of exotic species in an island environment. Ecol Appl 19:840–853

79. Simberloff D (2006) Hybridization between native and introduced wildlife species: importance for conservation. Wildl Biol 2:143–150
80. Ayres DR, Zaremba K, Strong DR (2004) Extinction of a common native species by hybridization with an invasive congener. Weed Technol 18S:1288–1291
81. Wyatt KB, Campos PF, Gilbert MTP, Kolokotronis S-O, Hynes WH, DeSalle R, Ball SJ, Daszak P, MacPhee RDE, Greenwood AD (2008) Historical mammal extinction on Christmas Island (Indian Ocean) correlates with introduced infectious disease. PLoS One 3(11):e3602
82. Higgins SN, Vander Zanden MJ (2010) What a difference a species makes: a meta–analysis of dreissenid mussel impacts on freshwater ecosystems. Ecol Monogr 80:179–196
83. Vitousek PM, Whiteaker LR, Mueller-Dombois D, Matson PA (1987) Biological invasion by *Myrica faya* alters ecosystem development in Hawai'i. Science 238:802–804
84. Aplet GH (1990) Alteration of earthworm community biomass by the alien *Myrica faya* in Hawaii. Oecologia 82:414–416
85. Pimentel D, McNair S, Janecka J, Wightman J, Simmonds C, O'Connell C, Wong E, Russel L, Zern J, Aquino T, Tsomondo T (2001) Economic and environmental threats of alien plant, animal, and microbe invasions. Agric Ecosyst Environ 84:1–20
86. Kovacs KF, Haight RG, McCullough DG, Mercader RJ, Siegert NW, Liebhold AM (2010) Cost of potential emerald ash borer damage in U.S. communities, 2009–2019. Ecol Econ 69:569–578
87. Chomel BB, Sun B (2010) Bioterrorism and invasive species. Rev Sci Tech 29:193–199
88. Colautti RI, Bailey SA, van Overdijk CDA, Amundsen K, MacIsaac HJ (2006) Characterised and projected costs of nonindigenous species in Canada. Biol Invasions 8:45–59
89. Pejchar L, Mooney HA (2009) Invasive species, ecosystem services and human well-being. Trends Ecol Evol 24:497–504
90. Le Maitre DC, Versfeld DB, Chapman RA (2000) The impact of invading alien plants on surface water resources in South Africa: a preliminary assessment. Water SA 26:397–408
91. Cook DC, Thomas MB, Cunningham SA, Anderson DL, De Barro PJ (2007) Predicting the economic impact of an invasive species on an ecosystem service. Ecol Appl 17:1832–1840
92. Lounibos LP (2002) Invasions by insect vectors of human disease. Annu Rev Entomol 47:233–266
93. Lv S, Zhang Y, Steinmann P, Yang G-J, Yang K, Zhou X-N, Utzinger J (2011) The emergence of angiostrongyliasis in the People's Republic of China: the interplay between invasive snails, climate change and transmission dynamics. Freshw Biol 56:717–734
94. Vilà M, Basnou C, Pyšek P, Josefsson M, Genovesi P, Gollasch S, Nentwig W, Olenin S, Roques A, Roy D, Hulme PE (2010) How well do we understand the impacts of alien species on ecosystem services? A pan-European, cross-taxa assessment. Front Ecol Environ 8:135–144
95. Purse BV, Mellor PS, Rogers DJ, Samuel AR, Mertens PPC, Baylis M (2005) Climate change and the recent emergence of bluetongue in Europe. Nat Rev Microbiol 3:171–181
96. Reichard S, Hamilton CW (1997) Predicting invasions of woody plants introduced into North America. Conserv Biol 11:193–203
97. Kolar CS, Lodge DM (2002) Ecological predictions and risk assessments for alien species. Science 298:1233–1236
98. Ricciardi A, Cohen J (2007) The invasiveness of an introduced species does not predict its impact. Biol Invasions 9:309–315
99. Callaway RM, Ridenour WM (2004) Novel weapons: invasive success and the evolution of increased competitive ability. Front Ecol Environ 2:436–443
100. Short J, Kinnear JE, Robley A (2002) Surplus killing by introduced predators in Australia – evidence for ineffective antipredator adaptations in native prey species? Biol Conserv 103:283–301
101. Ricciardi A, Atkinson SK (2004) Distinctiveness magnifies the impact of biological invaders in aquatic ecosystems. Ecol Lett 7:781–784
102. Strauss SY, Webb CO, Salamin N (2007) Exotic taxa less related to native species are more invasive. Proc Natl Acad Sci U S A 103:5841–5845
103. Simberloff D, Von Holle B (1999) Positive interactions of nonindigenous species: invasional meltdown? Biol Invasions 1:21–32
104. Ricciardi A, Palmer ME, Yan ND (2011) Should biological invasions be managed as natural disasters? BioScience 61:312–317
105. Bailey SA, Deneau MG, Jean L, Wiley CJ, Leung B, MacIsaac HJ (2001) Evaluating efficacy of an environmental policy to prevent biological invasions. Environ Sci Technol 45: 2554–2561
106. Huston MA (2004) Management strategies for plant invasions: manipulating productivity, disturbance, and competition. Divers Distrib 10:167–178
107. Risch SJ, Andow D, Altieri MA (1983) Agroecosystem diversity and pest control: data, tentative conclusions, and new directions. Environ Entomol 12:625–629
108. Pascal M, Siorat F, Lorvelec O, Yésou P, Simberloff D (2005) A pleasing consequence of Norway rat eradication: two shrew species recover. Divers Distrib 11:193–198
109. Donlan CJ, Campbell K, Cabrera W, Lavoie C, Carrion V, Cruz F (2007) Recovery of the Galapagos rail (*Laterallus spilonotus*) following the removal of invasive mammals. Biol Conserv 138:520–524
110. Myers JH, Simberloff D, Kuris AM, Carey JR (2000) Eradication revisited: dealing with exotic species. Trends Ecol Evol 15:316–320
111. Rayner MJ, Hauber ME, Imber MJ, Stamp RK, Clout MN (2007) Spatial heterogeneity of mesopredator release within an oceanic island system. Proc Natl Acad Sci U S A 104:20862–20865
112. Bergstrom DM, Lucieer A, Kiefer K, Wasley J, Belbin L, Pedersen TK, Chown SL (2009) Indirect effects of invasive species removal devastate World Heritage Island. J Appl Ecol 46:73–81

113. Rinella MJ, Maxwell BD, Fay PK, Weaver T, Sheley RL (2009) Control effort exacerbates invasive-species problem. Ecol Appl 19:155–162
114. Louda SM, Arnett AE, Rand TA, Russell FL (2003) Invasiveness of some biological control insects and adequacy of their ecological risk assessment and regulation. Conserv Biol 17:73–82
115. Callaway RM, DeLuca TH, Belliveau WM (1999) Biological-control herbivores may increase competitive ability if the noxious weed *Centaura maculosa*. Ecology 80:1196–1201
116. Lonsdale WM, FitzGibbon F (2011) The known unknowns – managing the invasion risk from biofuels. Curr Opin Environ Sustain 3:31–35
117. Andow DA, Zwahlen C (2006) Assessing environmental risks of transgenic plants. Ecol Lett 9:196–214

Books and Reviews

Catford JA, Jansson R, Nilsson C (2009) Reducing redundancy in invasion ecology by integrating hypotheses into a single theoretical framework. Divers Distrib 15:22–40

Ionizing Radiation Detectors

Wm. David Kulp, III
Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Article Outline

Glossary

Alpha particle A particle emitted during radioactive decay that is comprised of two protons and two neutrons, equivalent to the nucleus of a ^{4}He atom.

Beta particle An electron or a positron (the positively charged antimatter twin of an electron) emitted during radioactive decay.

Electron volt (eV) A unit of energy measurement defined by the kinetic energy gained by a free electron when accelerated through a potential difference of 1 V; approximately equivalent to 1.602×10^{-19} joule.

Gamma radiation Highly energetic electromagnetic radiation (energy greater than approximately 100 keV) emitted from the nucleus during radioactive decay.

Ionizing radiation Particles or light with sufficient energy to remove an electron from an atom or molecule.

Nuclide A species of atomic nuclei, defined by the number of protons and neutrons present in the nucleus; nuclides are represented by the chemical symbol and atomic mass number. Two examples are ^{14}C (carbon-14, six protons and eight neutrons) and ^{235}U (uranium-235, 92 protons and 143 neutrons).

Radioactive Describes an unstable atomic nucleus that releases energy through ionizing radiation.

Scintillator A type of detector that uses fluorescence to detect radiation.

Spectroscopy The measurement of radiation intensity as a function of radiation energy; a device or system of detectors capable of spectroscopy is referred to as a spectrometer.

Definition of the Subject

Equipment to detect, identify, and measure radioactivity is a key component in the safe and responsible development of nuclear science and technology. Whether designed to monitor radioactive processes, provide an alert, or characterize the radiation measured, these systems "see" what is undetectable to human senses. Used in nuclear power, industry, medical imaging, nuclear medicine, scientific exploration, and nuclear security, radiation detectors provide information about the radiation present and can be used to interpret what the source of the radioactivity is.

Experimental data exist for about 2,900 nuclides, or species of atomic nuclei, characterized in the laboratory. Yet less than 300 nuclides are found in measurable abundance in the environment. Most of these naturally occurring nuclides are stable nuclei, meaning that they do not decay to other nuclei over time. However,

some unstable, or radioactive, nuclei are found in everyday objects. Examples of naturally occurring radioactive material (NORM) are ^{40}K in bananas and the nuclides in the uranium and thorium decay series that are found in cat litter. The identification of radioactive nuclides is accomplished through detection and measurement of the radiation emitted during the decay of the unstable nucleus.

The decay of a radioactive atomic nucleus results in energy being released in the form of particles or electromagnetic radiation. Particles emitted during radioactive decay include alpha particles, beta particles, neutrons, and photons. Alpha and beta radiation are electrically charged particles ejected from the decaying nucleus. Alpha particles are positively charged helium nuclei. Beta particles are electrons or positrons (the antiparticle of an electron), which carry negative and positive charges, respectively. Neutrons have no electric charge. Photons, X and gamma rays, are electromagnetic radiation; treated as particles, they have zero charge, zero rest mass, and travel at the speed of light.

Introduction

The detection and characterization of radiation originated with Wilhelm Conrad Röntgen, who was awarded the first Nobel Prize in Physics in 1901 [1]. This work was continued and led to the discovery of spontaneous radioactivity, for which Antoine Henri Becquerel, Pierre Curie, and Marie Curie shared the 1903 Nobel Prize in Physics [1]. The mature nature of the field is demonstrated by the many textbooks available on nuclear physics and radiation detection. Suggested reading for in-depth study with focused discussion on radiation detection are Glasstone [2], Kantele [3], Knoll [4], Krane [5], Leo [6], and Tsoulfinidis [7].

Nuclear radiation ranges in energy from a few thousand electron volts (kilo-electron volts, or keV) to millions of electron volts (mega-electron volts, or MeV). Particles in the keV or MeV energy range are energetic enough that as they pass through matter they can cause the ejection of one or more electrons from a neutral atom in the material, ionizing the atom. Because of this interaction, nuclear radiation is also referred to as ionizing radiation. The physical processes that lead to ionization as radiation passes through matter depend upon the kind of radiation. Charged particles, photons, and neutral particles all interact with matter in different ways.

Alpha particles and other heavy, charged particles interact with matter through a variety of mechanisms, but the primary reaction is simply Coulomb scattering, an interaction between charged particles that is kinematic in nature. When energy is imparted to a target atom in the material, an inelastic collision has occurred with atomic electrons. Where no energy is transferred to the target material, the incident particle has elastically scattered from a target nucleus. These interactions have two basic results for the incident particle: (1) the particle loses energy, and (2) the particle is deflected from its initial trajectory.

Electrons and positrons also lose energy through Coulomb scattering in matter. However, they are more easily deflected in the electric field near an atomic nucleus due to the small mass of these particles or in collisions with atomic electrons (same mass). When electrons collide, energy is directly transferred to the atomic electrons. When electrons are accelerated or decelerated, electromagnetic energy is emitted in a process known as bremsstrahlung. Above a few MeV in energy, this mechanism is the predominant interaction for high-energy electrons and positrons.

Gamma rays and X rays are very different from the charged particles discussed above. They are electromagnetic radiation, called photons; a photon has zero electric charge and zero rest mass. Photons have three main interactions with matter:

1. The photoelectric effect, where an atomic electron is ejected from an atom after the absorption of the photon
2. Compton scattering, the scattering of photons by free electrons
3. Pair production, where a photon is transformed into an electron-positron pair

Neutrons are similar to photons in that they lack electric charge and will not interact with matter through Coulomb scattering. Instead, a neutron interacts with nuclei through the strong force. This is a relatively rare occurrence due to the short range of the strong force (effective only within 10^{-15} m). The result of interaction may be:

1. Elastic scattering from nuclei, so that no nuclear reaction takes place.
2. Inelastic scattering, where the target nucleus is left in an excited state.
3. Neutron capture, where the target nucleus is transformed through absorption of the neutron; most of the time the new nucleus is radioactive and decays by emitting beta particles and/or gamma rays (and neutrons, too, in a few cases).
4. Nuclear reactions with the emission of a charged particle.
5. Fission, the splitting of a heavy atomic nucleus.

Radiation detectors make use of these interactions with matter to produce a measurable effect that signals the presence of radioactivity. In general, a radiation detector can be characterized through three traits: (1) the radiation absorber (the materials of which the detector is made), (2) an observable that signals the interaction with radiation, and (3) a way to measure the signal.

The radiation absorber may be gas, liquid, or solid and can be made from a range of materials. The choice of detection medium phase depends on the type of radiation to be measured. Heavy charged particles have a range of less than about 100 μm (0.01 cm or 0.004 in.) in a solid absorber, but the resulting signals may be hard to distinguish from electronic noise. Neutrons and gamma rays, on the other hand, may penetrate centimeters of solid matter without producing any observable response. For detecting neutrons, the use of enriched isotopes may be used in order to take advantage of specific nuclear reactions that have higher probabilities of occurring.

The choice of observable effect produced by the detector is usually more dependent upon the application and the material used as the radiation absorber, rather than the type of radiation. Early researchers Henri Becquerel and Marie and Pierre Curie recorded data on photographic plates. While this method of observation provided long-lasting visible evidence of radiation, other detection methods such as electronic signals, scintillation light emissions, and changes in temperature are more advantageous for modern radiation detection. For example, light emissions produce the fastest detector response, thus a scintillation detector is the best choice for a measurement that requires precision timing. On the other hand, semiconductor detectors provide excellent energy resolution with good timing resolution, and are used for detailed nuclear spectroscopy.

If the radiation detection application requires only a qualitative measure of the presence of radiation, then an effective method of measurement would be an audible alarm that sounds when a threshold radiation level is reached, measured as a current generated within the detector volume. Nuclear science, however, requires quantitative analysis of the number and energy of individual particles emitted by atomic and nuclear transitions. For nuclear spectroscopy, it is therefore necessary to measure each electronic pulse registered in the detector, amplify the pulses and perhaps shape the signals as necessary, and record these signals for later analysis.

Given the different types of radiation and the range of energies, no single detector will be sensitive to all nuclear radiations at all applicable energies. Further, the diverse applications for radiation detection preclude a general list of radiation detectors that could be considered comprehensive. In the sections that follow, the most common detector types will be discussed and some recent advancements in the field will be introduced.

Gas-Filled Detectors

Ionizing radiation produces pairs of positively charged ions and negatively charged electrons as it passes through matter. It follows that a simple way to measure radioactivity is to apply an electric field across the radiation-absorbing material and count the ion-electron pairs produced in the detector. Such a detector can be envisioned as a parallel-plate capacitor filled with a gas. An electric potential applied across the capacitor creates an electric field that separates the electrons and ions. The electrons drift toward the positively charged anode plate, while the ions drift in the opposite direction toward the negatively charged cathode. This separation prevents the electrons and ions from recombining and enables measurement of the electronic signal produced by the ionizing radiation.

Ionization Chambers

The applied voltage across the capacitor influences how quickly charged particles move in the ionization

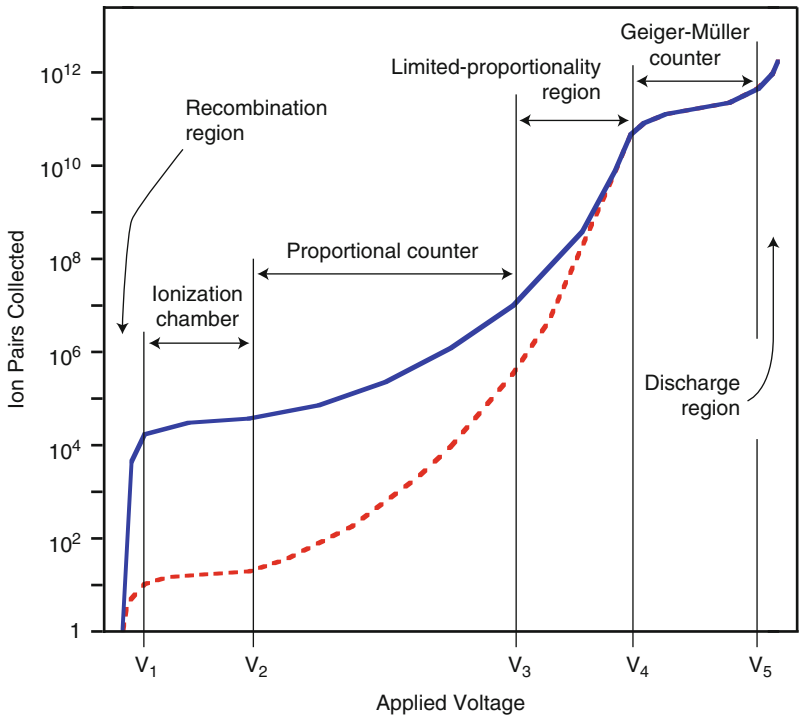
chamber. Electrons and ions tend to recombine to form neutral atoms at low voltages, with the result that only a weak signal is collected. This recombination region is indicated in the range where $V < V_1$ as in Fig. 1. Above some threshold potential, the electric field prevents recombination. This is indicated in the region where $V_1 < V < V_2$, where the total charge detected is insensitive to the applied voltage, as all of the electron-ion pairs that are created by the initial ionizing event are collected. A detector operating in this region collects only the charge produced directly by the incident radiation and is thus called an ionization chamber.

How big is the output electronic signal from an ionization chamber? The average energy required to produce an ion in dry air is about 30 electron volts (eV). An ionization chamber consisting of two square plates, each 10 cm long on a side, separated by a 1-cm air gap has a capacitance of 9×10^{-12} farads. Based on the energy to produce one ion in air, a 1-MeV gamma ray that deposits all of its energy in this detector would produce a maximum of about 3×10^4 electron-ion pairs, and a 2-MeV gamma ray would produce twice as many pairs. The voltage pulse resulting from these events would be about 0.5 or 1 mV, respectively.

To analyze individual pulses, the small signals produced by the direct radiation interaction require amplification. The two curves illustrated in Fig. 1 correspond to radiations that deposit different energies in the detector, e.g., an alpha particle and a beta particle or two gamma rays of different energies. The more energetic radiation produces more electron-ion pairs, resulting in a larger output signal.

Proportional Counters

A larger output signal can also be generated by increasing the applied voltage across the capacitor. The increased electric field accelerates the ions and electrons in the chamber to higher kinetic energy. Above a second threshold voltage, indicated as V_2 in



Ionizing Radiation Detectors. Figure 1
The number of ion-electron pairs collected in a gas-filled detector depends on the applied voltage and on the energy deposited in the active volume of the detector

Fig. 1, free electrons, produced by the incident radiation, are accelerated to sufficient energy such that they ionize additional gas atoms during collisions and produce more free electrons. This process is known as *gas multiplication* and results in a larger output signal.

The electrons produced in the knock-on reactions are called *secondary electrons*. The secondary electrons accelerate and produce additional ionization, resulting in a *Townsend avalanche*, where 10^3–10^5 secondary events occur for each original ion produced. As shown in region $V_2 < V < V_3$, the number of electron-ion pairs is proportional to the number of pairs produced in the primary event. Detectors operating in this range are called *proportional counters*. Using such a detector, the measurement of the incident particle energy is possible because the final signal is proportional to the energy deposited in the detector.

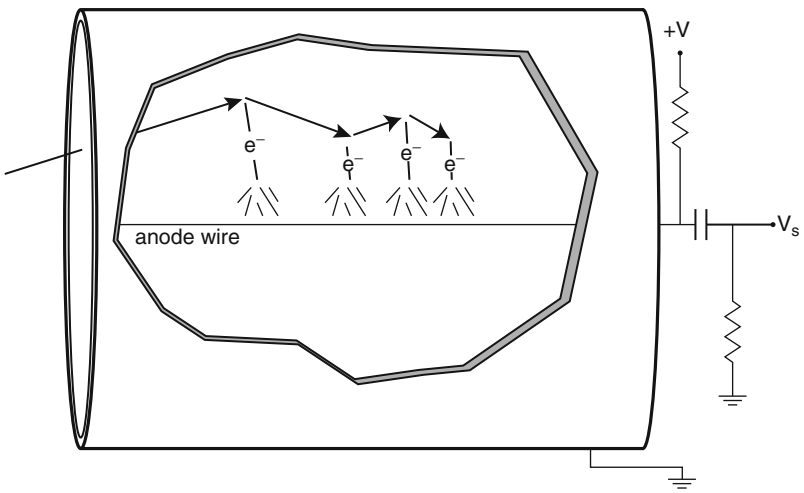
Proportional counters are typically cylindrical in shape, as shown in Fig. 2. This geometry results in an electric field that has an inversely proportional 1/r dependence, where r is the distance from the distance from the center of the detector. The site of the original interaction is not critical in such a detector. However, as an electron accelerates closer to the central anode wire, the field becomes very intense, resulting in a Townsend avalanche, indicated as a shower of electrons in Fig. 2. Because this occurs near the anode, the secondary electrons created are highly localized and no additional cascades form.

Increasing the applied voltage beyond $V = V_3$, the total ionization produced through gas multiplication continues to increase, but with reduced proportionality. This is the result of the creation of clouds of ions near the anode wire that have significantly lower drift speeds than the electrons. As a result, as the voltage increased in the region $V_3 < V < V_4$, the ions build up a space charge that shields the anode and changes the effective electric field.

Further increasing the voltage beyond V_4 results in a discharge occurring in the gas. Instead of a single, localized avalanche for each original electron-ion pair, *secondary avalanches* occur all along the anode wire. The secondary avalanches are the result of photons emitted by de-exciting gas molecules in the detector causing further ionization and avalanches elsewhere in the detector. A saturation effect thus takes place in the region indicated by $V_3 < V < V_4$: the discharge always has the same output, independent of the energy of the initial event.

Geiger-Müller Counters

Detectors operating in the $V_4 < V < V_5$ region are called *Geiger-Müller counters*. As shown in Fig. 1, there is no difference in count rate due to the energy initially

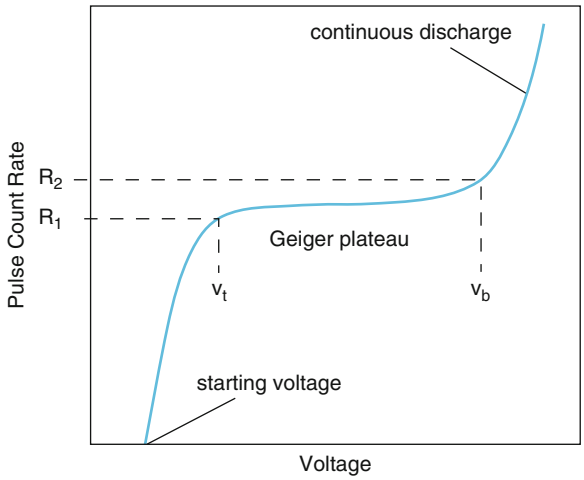


Ionizing Radiation Detectors. Figure 2
Radiation enters a proportional counter through a thin window and interacts with the gas within the cylinder, creating electron-ion pairs. The electrons accelerate toward the anode wire and produce avalanches of secondary electrons

deposited in the detector. Moreover, while the measured pulse size changes because the charge collected increases with increasing voltage, the pulse count rate does not change significantly, as shown in Fig. 3.

What is happening here is that the potential difference is so large in the active region of the detector that secondary avalanches cause a chain reaction of avalanches and total breakdown occurs. The discharge ends only when a large number of slow-moving secondary ions are formed near the anode wire. This localized concentration of ions represents a *space charge* that reduces the magnitude of the electric field, diminishing the attractive force accelerating the secondary electrons, and quenching the breakdown so that all radiation interacting with the detector produces the same current, regardless of particle type or initial energy.

Increasing the voltage to $V > V_5$ results in continuous breakdown in the gas, producing a steady current, whether radiation is present or not. This discharge region should be avoided in order to prevent damage to the detector. For this reason, Geiger-Müller tubes are typically operated at a voltage in the middle of the Geiger plateau.



Ionizing Radiation Detectors. Figure 3
The operating point of a Geiger-Müller tube is the middle of a region between a threshold voltage, V_t, and a breakdown voltage, V_b. In this region, called the *Geiger plateau*, the count rate changes very little as a function of the applied voltage. Beyond the breakdown voltage, the tube discharges continuously

In general, gas-filled detectors are the simplest detectors to operate, but have relatively low radiation detection efficiency. For electrons, ions, and low-energy X rays and gamma rays, the low density of matter in the active volume of the detector is sufficient. However, for high-energy photons, gas-filled detectors lack sufficiently high density to stop the radiation effectively. Practically, this is demonstrated using a rule of thumb: the thickness of material required to attenuate by half the intensity of a beam of 1 MeV photons is ~10 g/cm^2. To halve the intensity of a 1 MeV source of gamma rays using air (density = 0.00129 g/cm^3 at standard temperature and pressure) would require a detector 78 m thick. In comparison, only 2.7 cm thickness of sodium iodide (density = 3.667 g/cm^3) is needed to reduce the beam intensity by half.

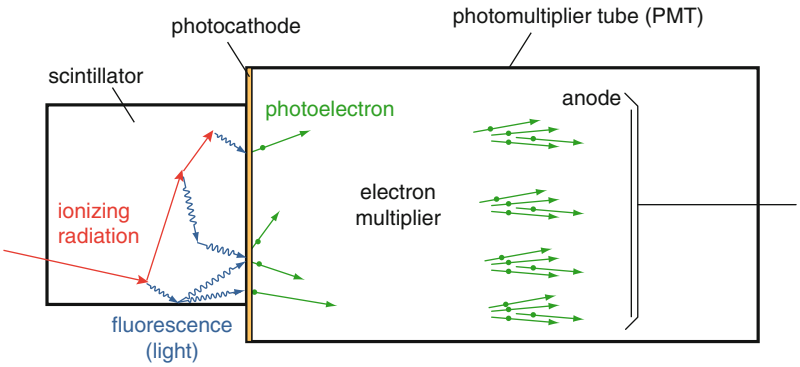
Scintillation Detectors

The principle of operation for a scintillation detector is very different from that of a gas-filled detector. Rather than collecting the electrons produced directly in the ionizing event in the radiation absorber, scintillation detectors use light as the observable that signals radiation detection.

The basic principle of operation of a scintillation detector, illustrated in Fig. 4, is as follows:

1. Incident radiation ionizes an atom in the scintillator material.
2. The excited atom *fluoresces*, i.e., produces light, as it relaxes to its initial state.
3. The light strikes the front surface of a *photomultiplier tube* (*PMT*) called a *photocathode* that yields a *photoelectron* through the photoelectric effect.
4. The photoelectrons are accelerated and multiplied through a series of electrodes (called *dynodes*) to produce a shower of secondary electrons.
5. The secondary electrons are collected at the anode as an output signal pulse.

The scintillator medium may be a solid, liquid, or gas. Scintillator material may be selected from organic crystals, organic liquids, plastics, glasses, inorganic



Ionizing Radiation Detectors. Figure 4
Ionizing radiation produces flashes of light in a scintillation detector. This light is focused to produce photoelectrons, which are multiplied to produce a measurable signal

crystals, glasses, and gases. With the exception of the crystalline detectors, organic scintillators are referred to using manufacturer designations. For example, a common liquid scintillator used in fast neutron detection is known as NE-213 (Nuclear Enterprises), BC-501A (Bicron/St. Gobain), and EJ-301 (Eljen). Plastic scintillation detectors, such as NE-102A (alternatively marketed as BC-400 or EJ-212), are manufactured by dissolving organic scintillators in a solvent such as styrene or polyvinyltoluene (PVT) that can be polymerized. These detectors have a notable advantage in that they may be cut or shaped as needed and are fairly durable, but the choice of material ultimately depends on the detection application.

The characteristics of a good scintillator are:

1. Efficient *luminescence*, i.e., it converts most of the energy deposited in the material into light.
2. Transparent to its own light output to enable light transmission through the absorbing material.
3. Has an index of refraction approximately that of glass ($n = 1.5$) to allow coupling to a light sensor.
4. Emits light within a wavelength range that matches existing light sensors.
5. Emits light pulses with a short decay time constant (τ).

Organic detectors are characterized by the shortest decay time constants; however, these lighter compounds lack the efficiency of detectors made using materials of higher atomic number. Inorganic crystals can be made from materials as heavy as bismuth (atomic number, Z = 83), and typically have better energy resolution than organic detectors. Scintillators in common use are:

- Anthracene ($C_{14}H_{10}$), an organic crystal with a short decay constant ($\tau = 30$ ns) used in general radiation detection
- Stilbene ($C_{14}H_{12}$), a very fast ($\tau < 5$ ns) organic crystal used in neutron detection
- NE-102A, a general-purpose plastic scintillator ($\tau = 2.4$ ns)
- NE-213, an organic liquid used in neutron detection ($\tau = 3.7$ ns)
- NaI(Tl) (sodium iodide activated with a thallium dopant), an inorganic crystal in wide use in radiation detection ($\tau = 230$ ns)
- LiI(Eu) (lithium iodide doped with europium), an inorganic crystal used in neutron detection ($\tau = 1{,}200$ ns)
- $Bi_4Ge_3O_{12}$ (bismuth germanate, or BGO), used in PET scanners and in nuclear spectroscopy ($\tau = 300$ ns)
- BaF_2 (barium fluoride), used for fast timing in nuclear spectroscopy (two components to the pulse $\tau = 0.6$ and 630 ns, respectively)

Semiconductor Detectors

Semiconductor detectors are essentially solid-state ionization chambers. With higher mass density, and requiring less energy per charge generated (on the

order of 3 eV, compared with about 30 eV in gaseous detectors), however, semiconductor detectors provide both increased detection efficiency and superior energy resolution compared with gaseous detectors. Ionizing radiation excites electrons into the conduction band of the semiconductor crystal. These electrons and the positively charged *holes* left behind in the valence band of the semiconductor migrate under the influence of the applied electric field, which is on the order of a few thousand volts.

The band gap in a semiconductor is on the order of about 1 eV. Such a small energy difference results in a measurable output signal from semiconductor detectors at room temperature. To reduce this thermal noise, some semiconductor detectors are operated at cryogenic temperatures.

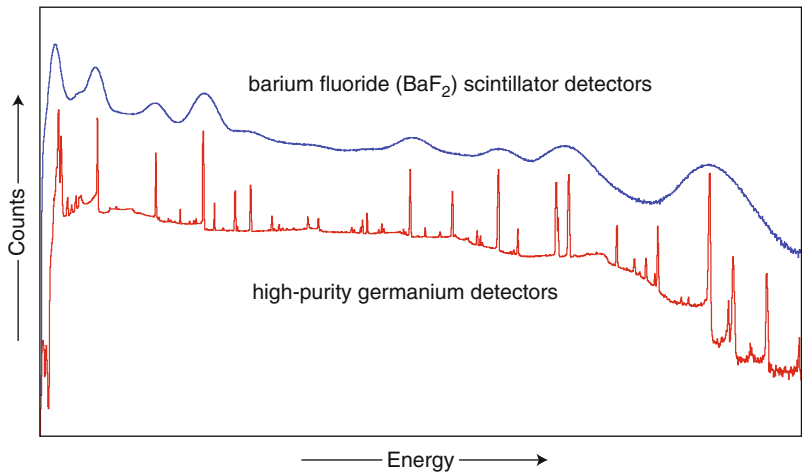
The two operational constraints on semiconductor detectors, high voltage and cryogenic temperatures, limit widespread use of these devices. However, semiconductor detectors typically have significantly better energy resolution than scintillator detectors. This is illustrated in the gamma-ray spectrum measured during the decay of ^{152}Eu, shown in Fig. 5.

The upper pulse-height spectrum in Fig. 5 shows the response of an array of barium fluoride (BaF_2) scintillation detectors, while the lower spectrum is that collected using an array of high-purity germanium (HPGe) semiconductor detectors. Individual lines resolved by the HPGe detector appear as a continuum with mound-like features in the BaF_2 spectrum. This difference makes semiconductor detectors the tool of choice in nuclear spectroscopy. With a higher atomic number (Z = 32), germanium detectors are typically used for gamma-ray spectroscopy, while silicon detectors (Z = 14) are used in X-ray spectroscopy and in charged particle spectroscopy.

Neutron Detectors

The absence of an electric charge complicates neutron detection. Neutrons do not directly produce ionization in matter. However, neutrons interact with atomic nuclei through absorption or scattering, and are thus detected through reaction products that do produce ionization.

Absorption through neutron-induced nuclear reactions is most probable at very low energies (eV); the probability of a reaction occurring increases with decreasing neutron energy with a 1/v relationship, i.e., inversely proportional to the neutron velocity. Most nuclear power reactors are designed to work at thermal energies (on the order of 0.025 eV), where a neutron has a speed of 2,200 m/s, and the probability is high for neutron-induced fission in uranium. The 1/v



Ionizing Radiation Detectors. Figure 5
Gamma-ray spectra from the radioactive decay of ^{152}Eu demonstrate the difference in energy resolution between that of a barium fluoride (BaF_2) scintillator detector array (*top*) and that of a high-purity germanium semiconductor detector array (Spectra courtesy of D. Cross)

relationship implies that nuclear reactions will be most useful for detecting *slow* neutrons, categorized by a neutron energy < 0.5 eV. Neutrons with an energy above this threshold are more effectively detected through scattering in the detector material.

Slow Neutron Detectors

Nuclear reactions used to detect slow neutrons (neutron energy <0.5 eV) typically produce heavy charged particles such as protons and alpha particles. These reactions are referred to as activation reactions, because they usually leave the product nucleus in an excited state, which subsequently decays through gamma-ray emission. Two common reactions used in detectors are the (n,p) and (n,α) activation reactions. In the (n,p) reaction a neutron, n, is absorbed and a proton, p, is emitted. Similarly, in the (n,α) reaction an alpha particle, α, is emitted. These reactions release considerable energy (approximately 1 MeV or more), so that the incident neutron energy (< 0.5 eV) cannot be determined from the reaction. Subsequently, detectors designed for slow neutrons are used only for indicating the presence of neutrons, and not for neutron spectroscopy.

The primary reactions used to detect slow neutrons are:

- $^{10}B(n,\alpha)^{7}Li$, where the detector requires enriched boron that is >90% ^{10}B (boron is naturally found in ratios of 19.8% ^{10}B and 80.2% ^{11}B)
- $^{6}Li(n,\alpha)^{3}H$, which uses ^{6}Li enriched to over 90% (the natural abundance of lithium is 7.59% ^{6}Li and 92.41% ^{7}Li)
- $^{3}He(n,p)^{3}H$, where the detector relies on rare ^{3}He gas that has a natural abundance of 0.00137% and is very expensive to produce
- $^{157}Gd\ (n,\gamma)^{158}Gd$, used in liquid scintillator detectors

Because the nuclear reactions require the use of specific isotopes, the availability of enriched isotopes contributes to the cost of fabrication for these detectors. In the case of ^{3}He, this cost is significant, if sufficient quantities of the material can be acquired at all. Helium-3 is not only for neutron detection, but also for cryogenics and is in high demand in many fields of research. Manufactured through the decay of tritium produced in a nuclear reactor, ^{3}He was available in greater quantities during the Cold War, because tritium is a critical component in thermonuclear weapons.

Fast Neutron Detectors

Neutron-induced nuclear transformations such as the $^{6}Li(n,\alpha)$ and $^{3}He(n,p)$ reactions may be used to detect fast neutrons. Unlike the case for slow neutrons, where the incident neutron energy is negligible compared with the reaction energy, it is possible to measure the neutron energy. However, the efficiency of these detectors is limited because the reaction probability decreases rapidly with increasing neutron energy. More commonly, scattering reactions are used to detect and measure the energy of fast neutrons.

Kinematics limit the energy that may be transferred in the neutron-nucleus collision. Because the mass of the neutron and the mass of the proton are nearly the same, it is only possible to transfer all of the neutron energy in a single collision in the (n,p) reaction. As the mass of the recoil nucleus increases, the fraction of energy transferred decreases. For the case of a deuterium recoil nucleus (atomic mass A = 2), a maximum of 88.9% of the energy can be transferred. In the case of ^{3}He, this maximum value falls to 75%. It is evident that a radiation absorber made from light nuclei is preferred, as it is possible to transfer more energy to the detector nuclei in fewer collisions. To provide higher efficiency, a solid-state detector is preferable, and materials with relatively high concentrations of hydrogen are desired.

The *proton recoil scintillation detector* takes advantage of kinematics and the high availability of scintillators that contain hydrogen. The kinematic advantage of these detectors is that the energy distribution of the recoil protons does not depend on the collision angle, resulting in a rectangular-shaped distribution in an ideal case. The shape of the detector output pulse may be used to separate gamma rays from neutrons, and the energy of the incident neutrons may be determined by comparing the response of the detector with calibration spectra obtained using a monoenergetic neutron source.

The availability of organic scintillators in many forms, including plastic and liquid detectors, provides a broad range of available materials for detector

construction. The hydrogen in the aromatic compounds provides an efficient mechanism for energy transfer in the absorbing medium, but the response of the detector is complicated by the presence of other elements such as carbon and oxygen. If the source of neutrons is pulsed, such as at an accelerator facility, then the energy may be extracted using a *time-of-flight* method. This method relates the detection time to the pulse structure of the beam used to create the neutrons and extracts the neutron energy from the amount of time it takes for the neutron to travel to the detector.

Moderating Detectors

A third type of neutron detector uses a layer of hydrogen-containing material to slow down, or *moderate*, neutrons in order to use neutron-induced nuclear reactions as the detection method. Called *moderating detectors*, these systems are useful if there is a broad range of neutron energies to be detected or if the neutron energy distribution is unknown. This type of detector has a relatively slow response time due to the moderating process. This is unsuitable for situations where the neutron energy distribution changes with time or when it is desirable to relate the neutron with another event, such as the detection of a gamma ray.

Examples of moderating detectors include *Bonner sphere spectrometers* and *spherical neutron dosimeters*. Used to detect fast neutrons, the Bonner sphere spectrometer consists of a set of different-diameter solid polyethylene moderating spheres that slow incident neutrons and a thermal neutron detector such as a lithium iodide scintillator. The spheres are placed over the detector in turn, and the count rate is recorded for each sphere. The neutron energy spectrum is then interpreted from this data using calibration data.

A spherical neutron dosimeter is essentially the same construction as a Bonner sphere spectrometer, but only uses a single sphere. The sphere is modified to provide a response that coincidentally resembles the neutron dose equivalent delivered curve as a function of energy. This detector is often used for neutron monitoring to provide neutron dose estimates.

Future Directions

Advances in radiation detection may result from the development of new radiation absorber materials, the refinement of methods to signal radiation interactions, and innovations in signal measurement. Some recent advances are discussed in the following section. The impact of these new developments and future directions may have an effect on the ability to detect small quantities of nuclear materials for safety and security, provide better tools for medicine and medical imagining, characterize the rarest nuclei in the cosmos, provide key data for understanding astronomical interests such as supernovas and neutron stars, and investigate dark matter and the nature of neutrinos.

Advancements in Detector Materials (Radiation Absorbers)

It is the interaction of radiation with matter which provides a signal to be measured. Perhaps because of this, and because of the limitations of current radiation detectors, it may be presumed that the primary need in radiation detection is in the characterization of new materials that are sensitive to radiation. Indeed, the physical characteristics of some of these materials in use are not optimal. Some examples are found in gamma-ray detectors: HPGe crystals require cryogenics temperatures for operation and the hygroscopic nature of NaI(Tl) leads to performance degradation as water is absorbed in the crystal. In other cases, the optimal materials are simply difficult to acquire: the scarcity of ^{3}He for cryogenics and for neutron detectors since the reduction of tritium production is a case in point [8].

There are many open areas of research, including applications in nanotechnology and crystal growth. Suspension of nanoparticles in liquid scintillators may lead to improved scintillation detectors or detectors based on novel new materials. Development of crystal growth techniques for newer CdZnTe-based detectors [9] may reduce the defects found in these detectors and enable the growth of larger crystals. It is important to note, however, that the common materials already in use have been selected through years of research, and that a breakthrough in detector materials may take decades to come to fruition.

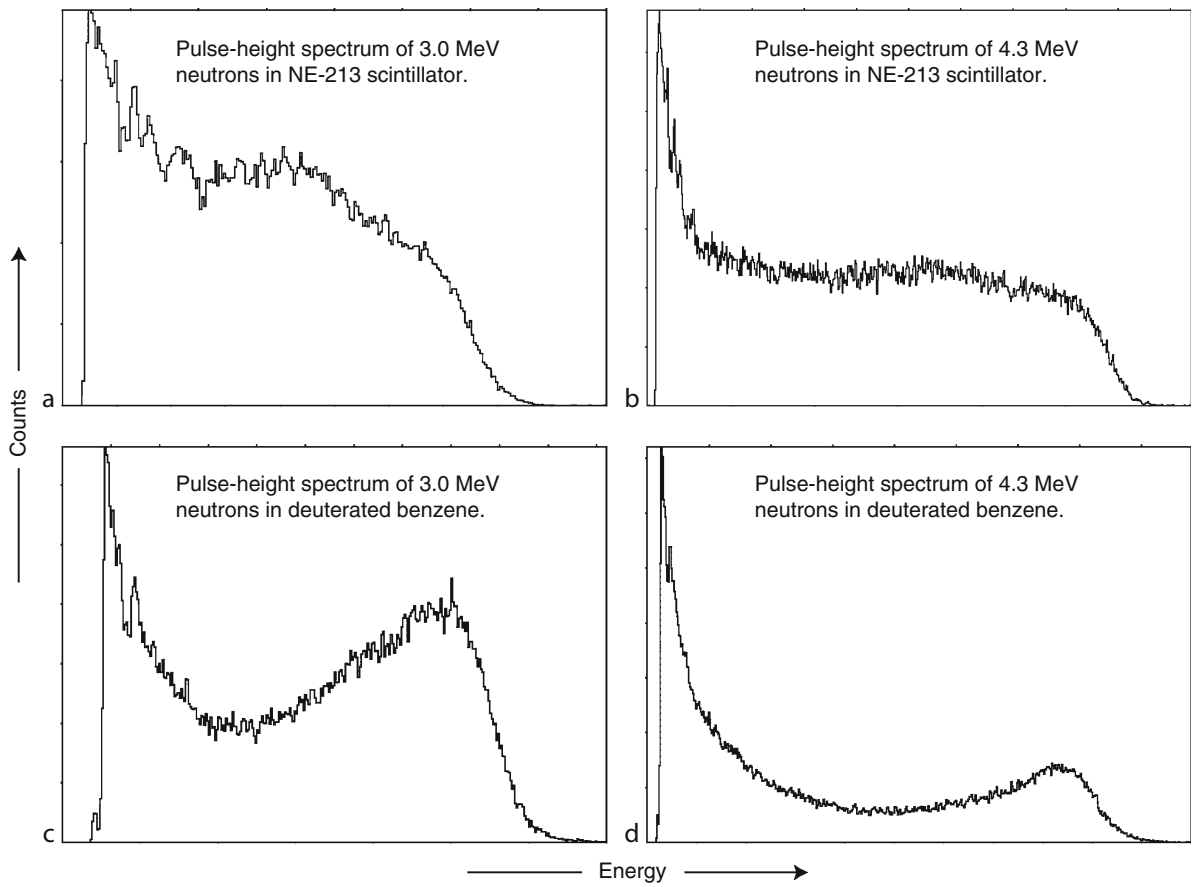
Near-term improvements may come from refining methods to make detectors, as in the case of the high-purity germanium crystal. Originally, germanium semiconductors required a lithium dopant in the

crystal matrix. This dopant would degrade as the lithium migrated out of the crystal, unless the detector was constantly kept at liquid nitrogen temperature. By refining the technique for growing germanium crystals, the lithium dopant is no longer used in these detectors.

Simply modifying compounds already in use can advance the field as well. A case in point is in the use of deuterated benzene scintillators for neutron detection. Benzene (C_6H_6) is commonly used in neutron scintillation detectors. In neutron spectroscopy, neutron energy is generally extracted using the time-of-flight method. This is because the detector response is essentially featureless, as shown in Fig. 6a, b taken with NE-213 scintillators. On the other hand, Fig. 6c, d are from deuterated benzene scintillators measured at the same energies. The peaks in these spectra appear due to the kinematics of scattering from deuterium. By combining these detectors with advanced pulse-shape analysis electronics, the energy information may be extracted in addition to the time-of-flight method, allowing for fast neutron spectroscopy in laboratory experiments.

Advancements in Detection Methods (Observables)

The types of detectors discussed here, gas-filled ionization detectors, scintillation detectors, semiconductor detectors, and neutron detectors, represent the majority of radiation detectors currently in use.



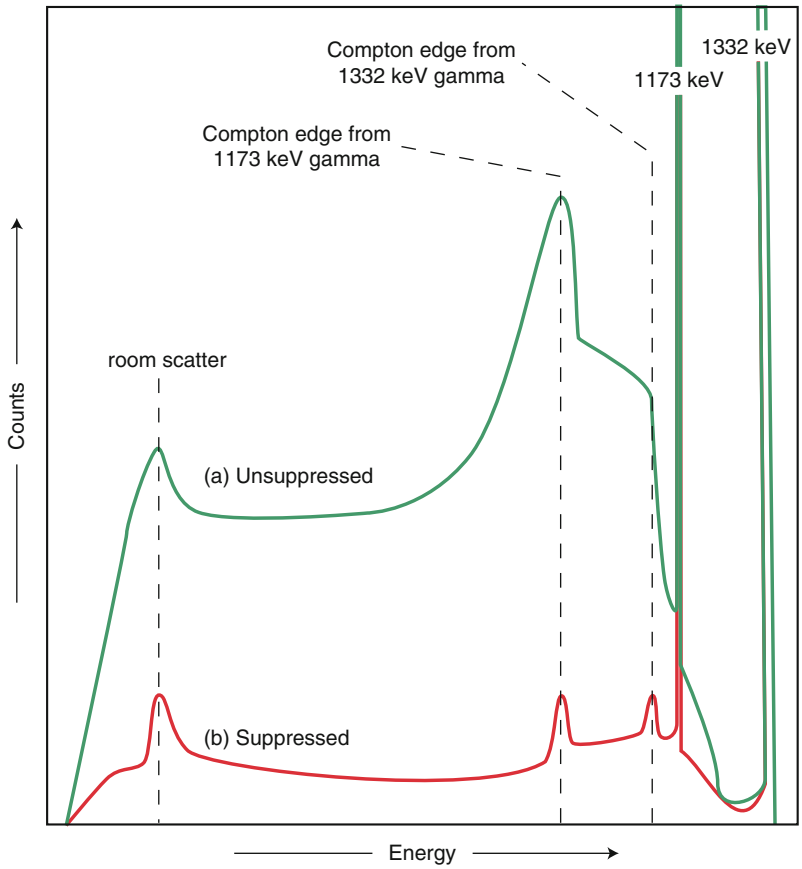
Ionizing Radiation Detectors. Figure 6

Pulse-height spectra of monoenergetic neutrons collected using NE-213 scintillators and deuterated benzene scintillators. Neutron energy may be extracted from the deuterated benzene detectors using pulse-shape discrimination electronics (Spectra courtesy of P. E. Garrett)

The primary observables have been the electronic signals or light output based on direct or indirect ionization to signal the interaction of radiation in the detector. These represent the basic interactions which can be measured in common detectors.

Another possible observable is to measure temperature changes in a material to indicate radiation detection. Such a detector is called a *bolometer*, and the radiation absorber in this kind of detector is a material that has electric resistance that is highly dependent upon the material temperature. These detectors are at the forefront of dark matter and neutrinoless double-beta decay experiments [10] and are typically small metal, semiconductor, or even superconductor devices. The size is limited in order to maximize the temperature rise and the measured change in resistance. A drawback to this kind of detector is the need to maintain a consistent temperature.

One simple advancement in this area is to combine multiple types of detectors to filter the signal that is recorded for later use. An example of this is found in the *Compton-suppressed* germanium detector [11]. A germanium semiconductor detector is surrounded by high-efficiency scintillation detectors that act as an anticoincidence shield. If a gamma ray is detected only in the HPGe detector, but not in the surrounding detectors, it is presumed to have deposited the full energy in the germanium crystal. However, when signals are detected in coincidence with the surrounding scintillators, this indicates that the gamma ray has



Ionizing Radiation Detectors. Figure 7
The spectrum of a ^{60}Co source with characteristic gamma-ray lines at 1,173 and 1,332 keV collected using (**a**) a high-purity Ge semiconductor (HPGe) detector, and (**b**) the same HPGe detector operated in anticoincidence with a Compton-suppression shield of scintillators

scattered out of the HPGe detector, and a full-energy pulse will not be detected. By suppressing these Compton scattering events using nanosecond coincidence timing circuits, the background spectrum of the detector can be greatly reduced, as shown in Fig. 7. This enables the detection of much lower intensity peaks than would normally be visible with an unsuppressed detector.

Advancements in Signal Measurement

The basic method for radiation spectroscopy is to measure signal outputs and record the data to build up statistics for interpretation. Typically, each signal is passed through electronic circuits comprised of amplifiers, discriminators, and analog-to-digital convertors in order to electronically record each event as it is detected. Computer analysis is later used to scan the recorded data, sort out events that fit requisite criteria, and fit the data for interpretation.

In some cases, such as in neutron detection, pulse-shape discrimination is used to distinguish between the types of radiation detected. This is typically done off-line during the computer analysis in order to separate the neutrons from gamma rays. An advancement to this technique is to use a computer in the data acquisition system in order to fit and digitize pulses during the data collection, and record only the signals of interest.

Very advanced gamma-ray spectrometers GRETA (Gamma-Ray Energy Tracking Array) [12] and AGATA (Advanced GAmma Tracking Array) [13] are being constructed in the United States and in Europe, respectively, to follow gamma-ray interactions as they scatter in germanium detectors. These *gamma-ray tracking spectrometers* use segmented germanium crystals connected by electronic contacts to determine where gamma rays interact in the detector. Off-line computer analysis is used to reconstruct the history of interaction. The full energy of the gamma ray is determined by summing up the energy of the individual interactions, and the first point of interaction in the detector may be determined for use in analysis based on angular distributions.

The GRETA and AGATA spectrometers are the most advanced research-class systems in gamma-ray spectroscopy and come with multimillion dollar (euro) price tags. Such systems are in development to support large groups of scientists at national laboratories, and as such are of specialized interest, rather than directly applicable to the general field of radiation detection. However, the future in this area will be closely tied to computational power and to the development of specialized electronics and programs for signal processing and data analysis. Adoption of better in-line and off-line computational power may overcome some of the inherent barriers in radiation detection and open opportunities for the development of new materials and new observables for detectors.

Bibliography

1. Nobel Lectures, Physics 1901–1921. Elsevier, Amsterdam, 1967
2. Glasstone S (1958) Sourcebook on atomic energy. D. Van Nostrand, Princeton
3. Kantele J (1995) Handbook of nuclear spectroscopy. Academic, San Diego
4. Knoll GF (2010) Radiation detection and measurement. Wiley, New York
5. Krane KS (1988) Introductory nuclear physics. Wiley, New York
6. Leo WR (1994) Techniques for nuclear and particle physics experiments. Springer, New York
7. Tsoulfanidis N (1995) Measurement and detection of radiation. Taylor & Francis, Washington, DC
8. Kouzes RT et al (2010) Nucl Instr Meth A 623:1035
9. Erickson JC, Yao HW, James RB, Hermon H, Greaves M (2000) J Electron Mater 29:699
10. Arnaboldi C et al (2008) Phys Rev C 78:035502
11. Nolan PJ, Gifford DW, Twin PJ (1985) Nucl Instr Meth A 236:95
12. Deleplanque MA et al (1999) Nucl Instr Meth Phys Res A 430:292
13. Eberth J, Simpson J (2008) Prog Part Nucl Phys 60:283

Ionizing Radiation on Nonhuman Biota, Effects of Low Levels of

Glen A. Bird
EcoMetrix Incorporated, Mississauga, ON, Canada

Article Outline

Glossary

Acute dose A single dose generally greater than 500 mGy of ionizing radiation to most or all of the body in a short time, usually a matter of minutes.

Adaptive response Adaptation to the presence of a low concentration of a contaminant, rendering subsequent treatment with high doses of the same agent less effective.

Alpha radiation The emission of alpha particle (the nucleus of a helium atom consisting of two protons and two neutron) from the nucleus of an unstable atom (radionuclide). Since alpha particles transfer their energy in a very short distance and cannot penetrate the outer layer of skin, alpha radiation is only an internal radiation hazard if it is inhaled or absorbed following ingestion. Most members of the U and Th decay chains are alpha emitters, e.g., ^{210}Po, ^{226}Ra, ^{232}Th, ^{238}U, ^{239}Pu.

Becquerel (Bq) The SI unit of radioactivity for measuring the rate of decay of a radioactive substance. It is equivalent to the disintegration of one radioactive nucleus per second.

Beta radiation The emission of electrons or positrons from the nucleus of an unstable atom (radionuclide). Beta particles can penetrate biological tissue to a depth of 1–2 cm. They may pose both an internal and an external hazard, e.g., ^{3}H, ^{14}C, ^{90}Sr.

Cancer An abnormal growth of cells which tend to proliferate in an uncontrolled way and in some cases to metastasize or spread. Cancer can involve any tissue of the body and have many different forms in each body area.

Chronic dose A dose of ionizing radiation received either continuously or intermittently over a prolonged period of time comprising a major portion of an organisms life cycle.

Deoxyribonucleic acid (DNA) Nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms with the exception of some viruses.

Deterministic effect Both the incidence and severity increase as a function of dose after a threshold dose is reached and usually involves cell killing.

Gamma radiation The emission of photons (gamma rays), which carry energy but no charge, by nuclear transition or interaction (radionuclide). Gamma radiation is the most penetrating radiation.

Gray (Gy) The SI unit of absorbed dose for ionizing radiation, equal to 1 J of radiation energy absorbed in 1 kg of the material of interest.

LD_{50} The dose of a substance that causes mortality in 50% of the organisms exposed.

Linear energy transfer (LET) The rate of energy loss per unit path length.

Radiation dose The energy absorbed per unit mass of any material exposed to ionizing radiation, measured in grays (Gy).

Relative biological effectiveness (RBE) The ratio of the absorbed doses of reference to the test radiation types that produce the same biological effect.

Stochastic effect The severity of the effect is independent of the absorbed dose, there is no threshold, and the probability of the effect occurring is proportional to the dose absorbed.

Track The path of a particle of ionizing radiation.

Definition of the Subject

Radiation is present in the environment from both natural and anthropogenic sources. Civilian use of nuclear materials results in chronic releases of low levels of radiation to the environment, particularly from the nuclear fuel cycle, i.e., the mining and milling of uranium ore to the operation of nuclear power plants, the reprocessing of spent nuclear fuel and nuclear waste disposal. Genetic effects such as chromosomal aberrations are observed in biota living in areas of elevated natural background levels of radiation. High exposure levels affect rates of morbidity, reproduction, and mortality. In introducing the topic of radiation effects on biota, a brief discussion of biological response is presented covering stochastic and deterministic effects, tissue sensitivity, dose fractionation, relative biological effectiveness, adaptive response, and radiation as just another environmental contaminant.

The focus of the chapter is on the effects of radiation on biota at low dose levels similar to those

associated with chronic releases from normal operations of nuclear facilities. However, a scarcity of information is available on the effects of exposure to low doses possibly because the effects are small, are difficult to quantify, and are only temporary, i.e., are reversible. Acute doses of radiation, like those of most other contaminants, produce more severe effects. Acute exposures of radiation are never observed under natural conditions and are seldom observed in the environment except for severe accident situations, i.e., Kyshtym and Chernobyl accidents or in tailing ponds. Information from laboratory studies, field irradiator studies, industrial contaminated sites, and accidents is reviewed to ascertain the effects of radiation on the environment. Information provided on the effects of low levels of radiation on biota may be useful to practitioners in establishing ecotoxicological benchmarks to meet their regulatory requirements to protect the environment.

Introduction

A growing global human population has fuelled increased energy demand, which together with concern over climate change and dependence on foreign supplies of fossil fuels have created renewed interest in nuclear power [1]. Most of this interest is in countries that already use nuclear power. There were 438 reactors in operation globally at the end of 2008 with a total nuclear capacity of 372 GW(e) [2]. The IAEA [2] projected that nuclear capacity would increase to between 473 and 748 GW(e) by 2030. The World Nuclear Association [1] projection is for at least 1,100 GW(e) of nuclear capacity by 2060, and possibly up to 3,500 GW(e). In association with this expansion, there will also be increased activity in the remainder of the nuclear fuel cycle: mining and milling of uranium (and thorium) ore, refining, conversion, enrichment, fabrication of fuels, and reprocessing and disposal of spent fuel. All components of the nuclear fuel cycle release some contaminants to the environment. Radionuclide releases from the fuel cycle are greatest from the mining and milling of ore and from reprocessing of spent fuel. In recent years, strict environmental regulations, advances in pollution prevention technology, and more modern designs have greatly curtailed releases to the environment compared with earlier operations. Future releases will result in large areas being influenced by low levels of radiation. The present chapter focuses on the potential effects that low levels of radiation may have on the environment.

Characteristics of Radiation

Radiation is energy that is transmitted in the form of rays, waves, or particles as either ionizing or nonionizing radiation. Ionizing radiation has energy of at least 12.4 eV, and the ability to remove an orbital electron from an atom, thereby forming an ion–electron pair from a neutral atom. Nonionizing radiation is also electromagnetic radiation but does not have enough energy to ionize atoms or molecules. Nonionizing radiation interacts with biological tissue primarily by generating heat. Examples of nonionizing radiation include radio waves, microwaves, and visible light. This chapter focuses on ionizing radiation.

Radiation is measured using a gamma-, beta-, or alpha spectroscopy, and other mass spectroscopy techniques. Advances in the level of detection of radiation have improved with the development of low-level and ultralow-level gamma ray spectrometry, and the ability to directly count atoms using mass spectrometry (i.e., accelerator mass spectrometry (AMS), inductive coupled plasma mass spectrometry (ICPMS), thermal ionization mass spectrometry (TIMS), and resonance ionization mass spectrometry (RIMS) (see "► Radionuclides as Tracers of Ocean Currents"). Improvements in detection limits have increased our ability to measure the movement of radionuclides in the environment but do not affect dose calculations in any significant manner or the interpretation of older literature with respect to radiation dose measurements and effects, i.e., the older literature is still valid today.

Major types of ionizing radiation are: X-rays, gamma (e.g., ^{60}Co, ^{134}Cs, ^{65}Zn), beta (e.g., ^{3}H, ^{12}C, ^{90}Sr) and alpha (e.g., ^{210}Po, ^{226}Ra, ^{232}Th, ^{238}U, ^{239}Pu), and neutrons. X-rays and gamma rays are ionizing electromagnetic radiations produced from atomic and nuclear transitions respectively. When absorbed in matter, energy is deposited unevenly in discrete packets with enough energy to break chemical bonds, hence, the termed ionizing.

Gamma radiation is the emission of photons from the nucleus. Gamma radiation is much more

penetrating than alpha and beta radiation, and has no precise range. Gamma radiation is both an external and internal hazard. The intensity of gamma radiation attenuates exponentially with distance in dense media. Beta and alpha are particulate ionizing radiation. Beta particles are highly energetic electrons that originate in the nucleus and may carry either a −1 or +1 charge. Alpha particles are highly energetic helium nuclei lacking orbital electrons with 2^+ charges when ejected from the nucleus during decay. Alpha and beta particles are less penetrating, so their energy is more likely to be transferred within the organism and cause more damage locally. Beta particles may be both an external and an internal hazard, whereas alpha particles are only an internal hazard.

Alpha radiation has high linear energy transfer (LET) and is more biologically damaging for the same absorbed dose than gamma and beta radiation with low LET [3]. Radiation damage can be divided into two general types: direct effects due to molecular damage occurring, where the energy has been absorbed in the molecule (target has been hit); and indirect effects where the molecular damage is brought about by the chemical reactions of free radicals produced by the radiation (e.g., in cell water) [4]. The free radicals can then diffuse and damage the critical target, an indirect action verses a direct effect. Direct and indirect ionization result in essentially the same effect, an ionized atom. DNA is the most important target, although a large diversity of molecules may be affected.

In the following, radiation doses presented in rads in referenced material have been converted to grays (Gy), whereas doses given in roentgens (R or r) have been converted to rads then grays using the following conversion factors, 1 R = 0.9 rad [5] = 10 mGy. Most dose rates are presented in milliGray per day to allow easier comparison of chronic exposures. Where possible, the period of exposure is given along with the radiation exposure rate. Where exposure rates to biota are given for field studies or nuclear accidents, it is assumed that the exposure period is over most of the organism's life cycle or for perhaps several generations. Likewise, implicit for international guidance on radiation dose rates protective of nonhuman biota is that exposure is over a large portion of the organism's life cycle.

Biological Response to Radiation

Radiation effects may be either stochastic or deterministic. A stochastic effect is "a radiation-induced health effect, the probability of occurrence of which is greater for a higher radiation dose and the severity of which (if it occurs) is independent of dose." A deterministic effect is "a radiation effect for which generally a threshold level of dose exists above which the severity of the effect is greater for a higher dose" [6]. Cancer is a stochastic effect. Cell death, reddening of the skin, opacity of the eye lens, and permanent sterility are examples of deterministic effects.

For short-lived animal species, cancer is unlikely to be important at the population level. At the population level, cancer may be more important in more long-lived species such as marine mammals, and terrestrial animals with long reproductive life spans, slow recruitment, and low numbers of individuals, although this needs to be demonstrated.

Much of the information on the effects of radiation on humans comes from laboratory studies primarily with small mammals such as mice and rats. The findings of these studies are equally applicable to other animals.

For all endpoints from cell death to tumor induction, cancer induction and life shortening, reduction in dose rate in general results in reduced biological effects. The risk per unit dose of low-LET radiation has been observed in experimental systems to depend upon both the magnitude of the dose and its temporal distribution. Dose–response curves for low-LET radiation for late effects and genetic effects generally increase in slope with increasing dose and dose rate. The response for tumorigenesis may pass through a maximum and turn downward after a single high-dose-rate exposure at doses above 2.4–4 Gy; the dose often attributed to cell killing. Although dose–response relationships differ from one biological effect to another, qualitatively, the relationships are similar. Linear interpolation of effects from high doses and dose rates to effects of either low doses or dose rates overestimate effects by a factor of 2–10. Hence, the assumption of a linear, no-threshold dose–response relationship is a conservative approach to estimate risk for low dose and dose rate exposure. This is the dose rate effectiveness factor [7].

Radiation doses are defined as:

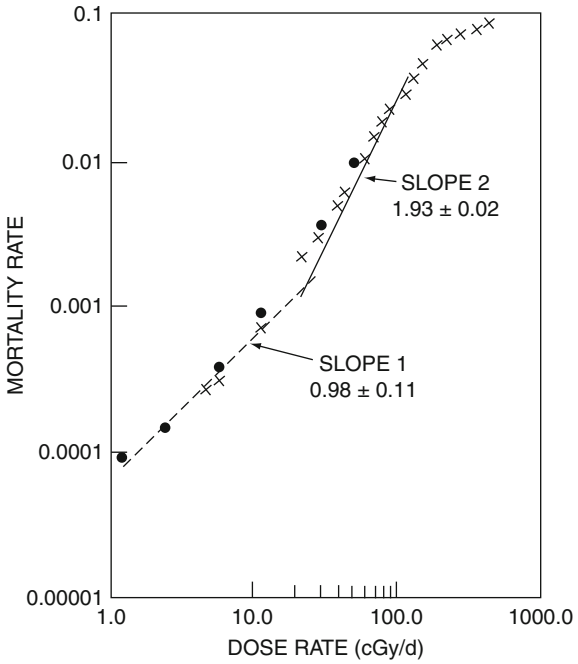
- Low – between 0 and 200 mGy;
- Intermediate – between 200 and 1,500 mGy;
- High – between 1.5 and 3.5 Gy; and
- Ultrahigh – greater than 3.5 Gy [8].

The biological effect of a given dose of low-LET radiation is dependent on the duration of exposure, time frame for biological repair, and biological target, which may change with age [8]. Lesions at low doses are formed almost entirely by single-hit kinetics, i.e., radiation events singly capable of inducing the complete lesion. The resultant effect is expected to be proportional to dose over the low-dose range. However, it is extremely difficult to detect radiation-induced effects in animals in the low-dose range at any dose rate, and thus the shape of the dose–response curve is better defined for high doses and dose rates [8].

Linear-Quadratic Model The linear-quadratic model adequately describes the dose response for solid cancers and describes chromosome aberrations, incidence of myeloid leukemia, and breast and lung cancer in mice [9] and plants such as the spiderwort (*Tradescantia*) [8]. Radiation effects on the spider plant can be used to illustrate the model. The spiderwort has flower buds that contain stamen hairs consisting of 25 blue colored cells. Mutation events cause the cells to turn pink. Pink mutations can be quantified down to less than 3 mGy of X-ray radiation and lower doses for high-LET radiation. The region below a radiation dose of 100 or 150 mGy shows an almost linear response, whereas the region of the curve beyond about 1 Gy shows a flattening and then declines, consistent with the effect of cell killing. The mortality of mice exposed to gamma radiation (Fig. 1) shows a similar respond [7]. The intermediate portion of the curve between 100 and 1,000 mGy is essentially linear on a log plot with a slope of 2, a value which indicates a quadratic relationship [8]. The dose–response relationship can be defined by the equation

$$I_\gamma = aD + bD^2$$

where I_γ is the induced incidence, D is absorbed dose in mGy, and the a and b terms are coefficients for the specific radiobiological conditions.



Ionizing Radiation on Nonhuman Biota, Effects of Low Levels of. Figure 1
Log-log plot of mortality rate of mice exposed to ^{60}Co at various dose rates for about 10 h per day (From Fry [7])

Radiation acts as if it is made of two separate components, an aD component and a bD^2 component. The aD component is considered the single hit response, which is all or none. The bD^2 component is a two-hit response, i.e., two hits are necessary at a sensitive site to produce the effect. The second hit can only produce the effect if it occurs soon enough after the first. The probability of an effective second hit increases with dose and dose rate, and is negligible at low doses and dose rates, leaving only the aD component to describe the effect. Most of the effect below about 1 Gy is considered to be the sum of the linear (aD) and squared (bD^2) components. The liner component contributes the same degree of effect per unit of absorbed dose at all doses, while the quadratic term dominates at higher doses. Therefore, the linear term coefficient for low-LET radiation is important only at low absorbed doses.

Radiation induced chromosome damage is detectable in cultured human lymphocytes for an

adsorbed dose of 100–250 mGy. The relation between chromosomal aberrations and dose is

$$I = C + aD + bD^2$$

where C is the spontaneous aberration frequency. Of the two-event lesions induced in irradiated cells, the dicentric chromosome is considered a reliable indicator of radiation exposure in lymphocytes. Most studies of dicentric chromosomes use a dose of 0.5–4 Gy since few dicentrics occur at lower doses. Most dicentric lesions (94%) are attributed to the linear component at 50 mGy and follow the two component theory at doses above 300 mGy [8].

Tissue Sensitivity to Radiation Rapidly dividing cells are generally the most sensitive to acute radiation, suggesting that chromosomal structure is most vulnerable during division. Thus, cell absorption of ionizing radiation leads to abnormal mitosis, growth, and metabolism. Undifferentiated cells are also usually more sensitive to radiation than differentiated cells. In male fish, sensitive germ cells are type B spermatogonia and spermatocytes [10]. In female fish, the most sensitive germ cells are the oogonia in the process of mitosis and oocytes at the start of the prophase changes of meiosis. Oocytes and lymphocytes are very sensitive despite being differentiated (resting) cells. In the small intestinal crypts, stem cells are more sensitive to ionizing radiation than their differentiated progeny and are more sensitive than the stem cells of the large intestinal crypts. Although rapidly dividing cells are sensitive to radiation and the small intestine is characterized by high cell proliferation, the small intestine is more resistant to cancer formation than the large intestine [11].

Effects of prenatal irradiation on growth and development depend on the gestational age when irradiated, the total dose, the dose rate, the linear energy transfer (LET) of the radiation, and on the particular endpoint or type of response expected. The period of organogenesis is a susceptible period for the induction of growth retardation, microcephaly, and mental retardation. The sensitivity of the stage of development and total dose delivered during a sensitive period are critical factors in determining the radiation effect.

The primitive germ cell (gonocyte) is the most radiosensitive cell type, and the level of response depends upon the prenatal time period during which the gonocyte persists. The rate of gonocyte killing in mammals with long gestation periods (>50 days) and, therefore, with long gonocyte life spans, appears to depend only upon the total dose and not upon dose rate down to rates of 10 $\mu Gy \cdot min^{-1}$ or less. Biological injury to the fetus occurs following exposure to either low single doses (~10 mGy) at a sensitive stage or low-dose-rate exposures (~14.4 mGy day^{-1}) over most of the prenatal period. Effects include germ cell depletion, growth retardation, central nervous system damage, or depressed central nervous system growth and cytogenetic abnormalities [8]. Immature mouse oocytes are very sensitive to radiation with a LD_{50} of 81 mGy. At 450 mGy, 99% of the cells are killed. Nevertheless, sufficient oocytes survive that mice are able to produce litters and maintain their population in both the laboratory and the environment. This may be because germ cells in the adult mouse remain in an arrested, nondividing oocyte stage, which in the late follicles is not easily killed by radiation [8].

Mutations are proportional to the amount of radiation absorbed and depend upon the amount of DNA in the cell, the functional state of the cell, and the effectiveness of the repair mechanisms. Generally, a higher dose rate will cause a greater impact than a lower dose rate, given the same total dose. This is assumed to be because repair mechanisms are better able to keep pace with damage caused by a lower dose rate. Note that cells that are irradiated during dormancy can store the damage without being able to repair it. When conditions permit cell division, accumulated damage may be manifested to a greater degree than in cells that are continuously active. Most examples of this occur in plants or in animals that greatly reduce their metabolic rates during cold or dry weather conditions. For example, fish, amphibians, reptiles, and certain mammals avoid cold weather by going into diapauses or hibernation. Likewise, certain insect species go into diapause to avoid periods of drought or warm weather.

The biological effect on cells is dependent on the number of cells struck by ionizing radiation, and the radiation energy and type – alpha, beta, or gamma. The lowest possible dose to a cell is the dose deposited by a single photon. For ^{60}Co gamma radiation, 1 mGy corresponds to an average of one track (path of a particle of ionizing radiation) per cell. This is

the lowest possible dose to a cell [12]. Below 1 mGy, not all cells receive a track of damage, but those that do still receive 1 mGy. Exposures at low doses are not uniform and, at 1 mGy, some cells receive two tracks, whereas about a third of the cells receive no track. However, at 1 mGy, all cells respond as if they received a track because of communication or the bystander effect [13]. There are about 100 tracks per nucleus at a dose of 100 mGy of low LET [14]. At moderate to high doses (>100 mGy) of sparsely ionizing radiation, all cells and tissues receive a nearly uniform exposure.

Dose Fractionation Dose fractionation is the repeated exposure to fractions of the total dose, rather than the same total dose being delivered at an essentially constant average dose rate. The effect of dose fractionation is highly dependent on the size of the fractions, the dose rate within each fraction, and age-dependent changes that may occur over the exposure period, which affect the radiosensitivity of the target. The magnitude of dose, dose rate, and protraction of dose interact to influence the dose effect [8]. Protraction allows for more repair and more effective repair because of the time between depositions of energy [7].

When the fractions of dose are administered at a time longer than the time required to rejoin the breaks (4 h), the yields of translocations are generally equal to the sum of those resulting from the individual fractionations. For example, 2.7 Gy delivered as 30 equal exposures reduces the yield of translocations to about 25% of that after an acute dose, whereas five exposures of 540 mGy reduced the yield to about 40%. Repair of chromosomal damage in oocytes takes between 1½ to 3 h. In general, closely spaced, large fractions are about as effective as a single dose. As the interval between fractions is prolonged and/or the dose per fraction is reduced, the effect on life shortening lessens [8]. Once-weekly exposure was less than half as effective in shortening the life of mice as the single exposure, and continuous exposure was only 20% as effective as the single exposure [15]. UNSCEAR [16] estimated that a radiation dose of 7 Gy to the mouse over its lifetime is equivalent to 5% life shortening due to cancer induction.

For immobile organisms such as barnacles and tube-dwelling benthic invertebrates and organisms with small home ranges, such as mice and muskrats, it is continuous (chronic) exposure over the life cycle of the organism that is of interest. For more mobile organisms with a large home range such as deer, moose, and wolves, exposure may be intermittent or fractured.

Adaptive Response The response by organisms to preexposure to a stressor is termed an adaptive response and is attributed to the stimulation of repair mechanisms by low dose exposure [17]. An adaptive response is a general response to stress induced by many different contaminants. For example, exposure of lymphocytes and other cells to low doses of ionizing radiation and subsequently to a high dose lessens the genetic damage. Fewer chromosomal aberrations are found in cells that had been preexposed before exposure. Various stress conditions induce an adaptive response to subsequent radiation-induced chromosome damage, including exposure to low-dose radiation, low concentrations of chemical mutagens, anticancer drugs (bleomycin, mitomycin C, and actinomycin D), free radical–generating chemicals such as hydrogen peroxide (H_2O_2), mild hypothermia, heavy metals (zinc), and low levels of double-strand DNA breaks [17, 18]. For example, pretreatment of rabbit peripheral lymphocytes with Zn results in resistance to gamma radiation (2 Gy) induced chromosome aberrations such as dicentrics, centric rings, and cells with chromosome aberrations. An increase in apoptosis is a cytogenetic adaptive response because cell death occurs at a lower dose than more serious effects [18]. For example, exposures of mouse spermatogonial cells to about 4.5–5.4 Gy results in cell killing, selectively eliminating germ cells with chromosomal damage.

Stimulated effects at low doses are often considered a positive effect but should be considered a negative effect as they can entail disorder of homeostasis, change of dominating species, and reduction of biological diversity [19]. Stimulation is an early response to stress and is a warning of potential harm should stress levels increase.

Relative Biological Effectiveness The extent and type of damage from radiation exposure depends on the type of radiation and the amount and rate of energy absorbed at the site of impact. Gamma rays and X-rays penetrate through biological tissue, deposit less energy (i.e., low-LET) and, at low dose rates, tend to produce

single-strand breaks in chromosomes. At low dose rates, alpha radiation is of much greater concern than gamma or beta. Alpha particles penetrate only about 70 μm into tissues (i.e., does not penetrate the outer layer of dead skin and hence is an internal hazard), deposit much more energy at the site of impact (i.e., high LET), and are more likely to cause double-strand breaks and more rapid cell death. The repair of single-strand breaks is typically efficient and rapid, whereas double-strand breaks are more difficult to repair and may be fatal to the cell but not the organism. The vast majority of genetic effects are removed by selection at the tissue, organism, or population level. At the population level, genetic damage to individuals may be offset by immigration of unaffected individuals from surrounding areas that contribute to the local gene pool as observed following the Chernobyl accident [20].

The higher energy deposition associated with high-LET radiation has been shown to produce unique types of damage that are not observed with low-LET radiations [21–24]. The greater clustering of ionization damage in high-LET tracks is apparently more favorable to induce mutation than cell death. The increased damage to cells from high-LET radiation is caused by the cluster of ionizations in the region of an alpha track. The biological effect of low-LET radiations are predominantly due to track-end clustered ionizations rather than from the large number of sparse ionizations, i.e., sparse ionizations are biologically unimportant, and clusters of different sizes are likely the cause of biological damage of different repairability.

The clustering of ionizations on the scale of nanometers is a prime determinant of radiation effectiveness. High-LET lethal damage is predominantly from clusters of about 10 or more ionizations in a 3.4 nm target thickness, the diameter of a DNA molecule, whereas low-LET is predominantly from clusters of $\geq$3 ionizations in a similar target thickness. For mutagenic damage, the increase in RBE with high-LET is due to clusters of about 15–20 ionizations or more in 3–5 nm. High-LET radiation is also qualitatively different than low-LET due to the greater complexity of damage in larger volumes, comparable to nucleosomes, which include numerous breaks, base damages, and cross-links that are much less repairable [23]. Despite the greater damage caused by alpha radiation, cells have a high probability of surviving the passage of a high-LET particle through the nucleus of the cell. Studies also indicate that cells of irradiated tissue that have not been hit by an alpha particle contribute significantly to the response because of the "bystander effect," which is mediated through gap junction cell–cell communication [25].

As discussed, radiations differ in their relative biological effectiveness (RBE) per unit of absorbed dose. The RBE is the ratio of the absorbed doses of two different radiation types (e.g., alpha radiation to a reference radiation usually either X-rays or gamma) that produce the same biological effect. At low dose rates (e.g., <100 mGy), high-LET radiations such as alpha particles and neutrons have a higher effectiveness for producing biological effects than low LET at the same absorbed dose. Use of a weighting factor (w_R) to account for RBE allows the summation of the absorbed dose for each type of ionizing radiation to give the total dose for biological effects such as cancer induction and genetic defects at low doses [3].

In human radiological protection, it is the absorbed dose averaged over a tissue or organ and weighted for the radiation quality that is of interest for setting dose limits [26]. The w_R is used for this purpose. The w_R is selected for the type and energy of the radiation incident on the body or within the body. The values of w_R are based on review of the biological information for a variety of exposure circumstances that include inducing stochastic effects (e.g., cancer) at low doses as well as the quality factor (Q), which is related to LET, a measure of the density of ionization along the track of an ionizing particle [26, 27]. W_Rs of 20 for alpha emitters and 1 for beta particles and gamma are recommended by the ICRP [26] for radiation protection of workers and the public. These w_Rs are set against X-rays as the reference radiation. More recently, a RBE of 2 has been recommended for beta radiation for tritium ([28] and see "► Tritium, Health Effects and Dosimetry"). The RBE values for high-LET radiation are greatest at low dose rates and for stochastic effects such as carcinogenesis and lowest for high dose rates and deterministic effects such as impairment of fertility. At high radiation doses, all cells are exposed to radiation and, at very high doses, there is in essence no difference between the effect of alpha and gamma radiation (damage is extreme) and RBE is about one.

The increase in the RBE at low dose rates is not due to the effect of the alpha radiation increasing. The effect of alpha remains the same, it is the effect of the gamma radiation that becomes smaller and harder to measure in comparison to the effect of alpha which causes the RBE to increase at low doses. RBE does not increase to infinity. Double-strand breaks induced by alpha particles are repaired more slowly and a higher fraction remains unrepaired, whereas repair is more effective at low gamma doses, so their effect is harder to quantify. This accounts for higher RBE values at low doses [29, 30].

There is currently no consensus on the value of a radiation w_R that should be used in calculating doses to nonhuman biota from exposures to alpha particles. Weighting factors of 1–40 have been used to estimate the radiation dose from alpha radiation to nonhuman biota. Some investigators have not modified the calculated absorbed dose (in Gy) due to alpha particles (e.g., Amiro [31]), whereas others have used a value of 20 [32, 33]. UNSCEAR [34] suggested a lower w_R of 5 for alpha radiation and a w_R of unity for beta and gamma radiation. Kocher and Trabalka [35] expressed the view that an appropriate w_R for alpha particles for use in protection of biota probably lies in the range of about 5–10 as the endpoints of concern in protection of biota are deterministic effects (e.g., cell killing). A value of 20 is used for stochastic effects for humans and 5–10 for deterministic effects [36] such as impairment of fertility [3]. Pentreath and Woodhead [37, 38] recommended a value of about 40 for provisional application until sufficient data become available, whereas EC and HC [39] used a value of 40 to represent the low dose rates and endpoints relevant to their assessment of releases from Canadian nuclear facilities.

The subject of RBE is reviewed by NCRP [40], EC, and HC [39] and, subsequently, by Chambers et al. [41]. Chambers et al. [41], in establishing an RBE of 5 for alpha, rejected studies giving high RBE values as being of poor quality, whereas EC and HC [39] considered these studies of acceptable quality and their RBE included these studies in deriving a RBE value of 40.

Studies conducted using plant systems have also shown that low doses of radiation yield large RBE values, while high doses yield lower values. For example, exposure of plant systems to high doses and high dose rates resulted in an average RBE value of 12±3 (average of 16 studies), while exposure to a low dose administered at any dose rate resulted in an average RBE value of 65±5 (average of 23 studies) [40]. NCRP [40] also concluded that for many plant systems with either greater DNA content per cell or larger nuclear volumes than mammalian cells, the RBE values tend to be larger than those observed in mammalian test systems by a factor of 2 or more.

The choice of a radiation w_R to account for the difference in the RBE per unit of absorbed dose is currently not defined but is dependent on the endpoint and the radiation dose rate of interest. A lower value for w_R seems appropriate when protection is at the population level for deterministic effects at higher doses. Where protection is at the individual level from low chronic doses typical of modern nuclear facilities, a larger value for w_R seems more appropriate, e.g., EC and HC [39].

Radiation: Just Another Environmental Contaminant

Radiation differs from most other contaminants in that it can be quantified precisely at very low concentrations at the atom level and that external exposure may lead to health effects. Otherwise, exposure to low levels of radiation is much like that for many other contaminants. For certain radionuclides, it is possible to measure their presence down to a small fraction of a Becquerel (Bq), where a Becquerel is one disintegration per second. For example, 1 $Bq{\cdot}L^{-1}$ of tritium is one tritium molecule in 100,000 million molecules (“► Tritium, Health Effects and Dosimetry”). In the case of tritium, the drinking water quality guideline in Canada is 7,000 $Bq{\cdot}L^{-1}$, which sounds like a huge number, but this represents only a tenth of the public dose limit of 1 $mSv{\cdot}year^{-1}$ [42], which in itself has a large safety factor associated with it and is much less than the 2.4 billion Becquerels of ^{99m}Tc routinely injected into human patients for nuclear medicine diagnostic purposes.

As with other contaminants, radiation effects increase with dose and there is a minimum dose-rate below which further reductions in the dose rate do not result in further diminution of response per unit of

dose [13]. High-radiation doses and high concentrations of other contaminants are harmful to the environment; however, their effects are much less than natural ecological events that regulate animal abundance [43].

Expression of radiation-induced biological effects and responses may be at either the cell, organ, or organism level. Induction of cancer and genetic effects can have their origin in the interaction of a single charged particle within the cell nucleus. Changes in the cell can initiate organ effects and responses and, because the cell nucleus contains essentially all of the genetic material, it is the target where this change occurs [12].

Deoxyribonucleic acid (DNA) strands in our cells are continuously being exposed to and broken by ionizing radiation such as ultraviolet light, X-rays, and gamma rays, and by other natural and man-made mutagenic chemicals. The latter include heavy metals, polycyclic aromatic hydrocarbons, polychlorinated biphenyls, pesticides, oxyradicals generated by ionizing radiation or processes such as redox cycling by heavy metal ions, radio-mimetic drugs which create oxyradicals, asbestos fibers, certain plant toxins, hydrolysis, thermal disruption, and viruses [44–47]. Therefore, many contaminants produce single- and double-strand breaks (DSBs) and chromatid aberrations. DSB can also be caused by mechanical stress on chromosomes, or when a replicative DNA polymerase encounters a DNA single-strand break or other type of DNA lesion [44]. DSBs occur when DNA polymerase runs into an unrepaired nick in the DNA. Topoisomerase inhibitors also cause breaks: topoisomerase breaks and rejoins DNA in the course of its function, and inhibitors can block the rejoining step. Cells may break their DNA on purpose for special functions, most notably during the gene shuffling that occurs as lymphocytes mature, which generates diversity in antibodies, T-cell receptors, and other highly variable immune system proteins [45].

Cancer is not caused by just one unique cause, e.g., radiation, therefore one must be very cautious in attributing any excess cancers to a sole cause, be it radiation or any other singe contaminant [11]. The ability of ionizing radiation to induce a unique fingerprint in DNA, which leads to a molecular marker mutation in a tumor suppressor gene of tumors in experimental animals, has not been demonstrated. The non-specificity of final effects of ionizing irradiation and chemical toxicants on the environment makes it difficult to recognize their independent contribution to harmful effects to organisms [48]. It has been suggested that the analysis of the distribution of chromosome aberrations in cells and of the frequency of the different types of aberrations may be a means of differentiating between the effects of radiation and that of other chemicals [49], although this needs to be confirmed.

Repair of DNA

Two major methods to repair double-stranded DNA breaks are (1) homologous recombination – the break is repaired using the duplicate set as a template which is very precise since the cell can use the undamaged DNA strand to ensure that the repair is correct; and (2) nonhomologous end joining – repairs the break directly without any outside information. The latter is less accurate and may result in the addition or removal of a few nucleotides at the repair site [45]. DSBs are the most dangerous form of DNA damage and are formed when the two complementary strands of the DNA double helix are broken simultaneously at sites that are sufficiently close to one another that base-pairing and chromatin structure are insufficient to keep the two DNA ends juxtaposed. As a consequence, the two DNA ends generated by a DSB are liable to become physically dissociated from one another, making repair difficult. This may lead to inappropriate recombination with other sites in the genome. Error also occurs because DNA termini have sustained base damage, which means DSB ligation cannot occur until processing by DNA polymerases and/or nucleases has taken place. DSBs are potent inducers of mutations and of cell death [44].

Inaccurate repair or lack of repair of a DSB can lead to mutations or to larger-scale genomic instability through the generation of dicentric or acentric chromosomal fragments. Such genome changes may have tumorogenic potential [44]. Breaks can cause serious problems as a single break in a key gene can kill a cell or cause it to kill itself by apoptosis. For this reason, cells have powerful methods to repair damage as soon as it happens. In the human lifetime, each of the body's cells

will have repaired more or less successfully several thousand double-stranded DNA breaks. In cancer treatment, radiation therapy is used to overwhelm the natural repair mechanism using high doses of radiation to fragment the DNA in cancer cells [45].

Protection of the Environment

Biota residing downstream of nuclear facilities will be exposed over their entire life cycle, and perhaps for several generations, to low levels of radionuclides released in effluents. Therefore, protection of the environment is concerned with chronic exposures from routine releases. The exposures also are mainly internal from alpha, beta, and gamma emitters incorporated into the tissues of organisms. In the case of biota exposed to releases from uranium mines and mills, over 90% of the dose is from alpha emitters deposited internally from ingestion of food and water. For organisms exposed to routine discharges from nuclear power plants and some waste management facilities, tritium, "a beta emitter," may comprise a major component of the radiation dose. Other radionuclides such as ^{14}C, ^{60}Co, ^{90}Sr, and ^{137}Cs may also be important.

The goal of environmental protection is to prevent or minimize the effects of ionizing radiation on organisms, including the induction of cytogenetic effects, morbidity, decreased reproduction, and early mortality [39, 50, 51]. Cytogenetic effects such as mutations include genetic changes in a somatic cell that may alter its function or in a germ cell that may be inherited. Morbidity includes effects on growth, immune system, and behavioral modifications which may lead to a loss of individual fitness. Decreased reproduction results from reduction in fertility and fecundity rates, whereas early mortality occurs due to damage to tissues and organs [50, 51].

The effects of exposure to radiation have been reviewed by several international agencies for the purpose of identifying a radiation dose below which effects on populations of organisms would not likely occur (Table 1) [34, 52–54]. These reports have suggested that doses of approximately 10 mGy day^{-1} to maximally exposed individuals for aquatic organisms and 10 mGy day^{-1} for terrestrial plants, and 1 mGy day^{-1} to maximally exposed individuals of terrestrial animals would not put populations at risk. This approach assumes that in a heterogeneously contaminated environment, only a few individuals in a population would be exposed to elevated doses of radiation, and that if doses to only a limited number of individuals of a large population are greater than 1 or 10 mGy day^{-1}, then the population should not be adversely affected. In contrast, the United States Department of Energy adopted a radiation dose standard of 10 mGy day^{-1} for the protection of aquatic animals and

Ionizing Radiation on Nonhuman Biota, Effects of Low Levels of. Table 1 Radiation levels defined as protective of the population by international agencies and lower effect levels to individuals recorded in the present paper

	Terrestrial animals	Terrestrial plants	Aquatic organisms	References
Population	1 mGy day^{-1}	10 mGy day^{-1}	10 mGy day^{-1}	IAEA [52, 53], NCRP [54], UNSCEAR [34]
Individual	0.13 mGy day^{-1}[a] to a mouse	0.5 – 2.4 mGy day^{-1} [c,d,e]	0.4 mGy day^{-1} to fish[f]	Present paper
	0.84 mGy day^{-1}[b] to the earthworm		0.12 mGy day^{-1} to invertebrate[g]	

[a]Fluctuating asymmetry in yellow-necked mouse [55]
[b]Integument epithelium and midgut epithelium changes [56]
[c]Delayed growth of wheat, barley, and beans [57]
[d]Reduced growth and morphoses in pine trees [58]
[e]Reduced growth of Scotch pine seedlings [59]
[f]Reduced spermatogenesis and survival in *Tilapia mossambica* [60]
[g]Abnormalities and mortality of embryo-larvae of *Mytilus edulis* [61]

proposed dose standards of 10 mGy day^{-1} for terrestrial plants and 1 mGy day^{-1} for terrestrial animals [62] for the population as opposed to the maximally exposed individual.

The above radiological protection guidelines based on the maximally exposed individuals assumes that populations of organisms possess compensatory capabilities such that impacts on a few individuals should not affect the integrity of a population or community [63]. These radiological protection guidelines have shortcomings for the protection of populations composed of a small number of organisms or organisms with a small home range. Another shortcoming of guidance based on the maximally exposed individual is demonstrating that the radiation dose is to the maximum exposed individual as opposed to the larger population. This is because large sample sizes are required to demonstrate the maximally exposed individual, i.e., sampling such large numbers may have a greater impact on the local population than the contaminant. In support of this guidance, the FASSET Radiation Effects Database [64] contained few indications of readily observable effects on biota at dose rates less than 2.4 mGy day^{-1} and the LC_{50} for radiation exposures over a period of 30 days is more than 150 mGy day^{-1} for even the most radiosensitive organisms [65].

Protection of the environment from radiation has been governed by the anthropocentric approach that "radiological control of the environment to the standard necessary to protect humans will ensure that other species are not put at risk" [26, 53, 54]. In the past, this has been considered reasonable except when human access is restricted without restricting access by biota; unique exposure pathways exist; rare or endangered species are present; or other stresses are significant [66]. However, this approach is not defensible for several reasons: the importance of exposure pathways to biota may differ from that of humans; many organisms are just as sensitive to radiation as humans; and in most environments, humans are likely to be the least exposed to radiation. For example, from 1949 to 1952, large quantities of liquid radioactive waste were discharged into the Techa River, Russia, from the Mayak processing facility [67]. Evaluation of radiation doses from 1950 to 1951 data and again in 1991–1994 showed doses to aquatic biota were much higher than those to humans. Doses to humans decreased from maximum individual rates of 2–4 Sv and average of 0.1 Sv to an average of 0.2 mSv·year^{-1} (0.1–2 mSv·year^{-1}) from consumption of local foods. In comparison, doses to fish decreased from 10 Gy·year^{-1} to 4 mGy·year^{-1}, mollusks from 20 Gy·year^{-1} to 20 mGy·year^{-1}, and algae from 30 Gy·year^{-1} to 3 mGy·year^{-1}; values 100–300 times higher than for humans. In the upper reach of the river, radiation doses were estimated to be as high as 200 Gy·year^{-1} to fish, 500 Gy·year^{-1} to mollusks, and 700 Gy·year^{-1} to algae [67]. Unfortunately, radiation effects were not reported. Likewise, following the Chernobyl accident, terrestrial biota were exposed to higher levels of radiation than humans [58]. The same may also apply to the uranium mining and milling industry, where human access to the receiving water is usually restricted.

The anthropocentric approach for radiation also differs from that taken for the protection of the environment from other contaminants and does not demonstrate that the environment is adequately protected [68]. There is no reason to treat ionizing radiation differently to other environmental stressors [68–71]. Nonhuman species should be protected in their own right [71, 72]. The level of protection depends on the component of the environment to be protected: individual, population, community, or ecosystem. Most jurisdictions focus on protection of populations rather than individuals. As noted by Oughton [72], individuals can suffer, individual humans matter, but individual animals tend not to.

Protection of individuals (within reason) will ensure that the population is protected. When population changes are observed in response to a stressor, it may already be too late to protect the population as the effect has occurred. The philosophy that the loss of the local population (maximally exposed individuals) is acceptable as migration of healthy individuals from surrounding populations will repopulate the area once radiation levels decline is not acceptable for other contaminants and is in conflict with the philosophy of sustainable development. With conventional contaminants, emphasis is placed on protection of exposed individuals, whereas with radiation protection, emphasis is on protection of populations. Individuals may be harmed as long as the larger population is protected.

Once the radiation contamination has decayed or been diluted to a low enough level that the quality of the environment has improved enough, organisms from the surrounding areas can recolonize the area.

When assessing the effects of radiation on biota, the life history of the organism should be taken into consideration along with the dose rate and total radiation dose. For example, chromosomal aberrations were seen in the benthic invertebrate *Chironomus tentans* exposed to 6 mGy day^{-1} in White Oak Lake, Tennessee, USA [73]. In the warm waters of White Oak Lake, presumably about 22°C, *C. tentans* would complete its life cycle in about 30 days and be exposed to a radiation dose of about 180 mGy. However, at low temperatures found in northern climates, over-wintering *C. tentans* may take 8–10 months to complete its life cycle. This would be an exposure of about 1.8 Gy at a dose rate of 6 mGy day^{-1}. Would this same exposure rate over a longer time frame produce a more severe effect than chromosomal aberrations? Longevity should be considered in assessing dose effects. For example, in a breeding facility, the American alligator (*Alligator mississippiensis*) had complete reproductive failure by 21 years of age due to the accumulation of high levels of lead from their diet [74]. The accumulation of contaminants over time in long-lived, late-reproducing species such as the alligator should be of environmental concern.

Populations generally exhibit compensatory mechanisms to increase reproduction in response to population decline. For example, following control of rodents and insects, their populations are often seen to bounce back to original levels or increase to greater numbers. Pollutants that disturb the reproductive function of animals have the greatest potential for consequences [43]. Inhibition of reproduction in response to population decline may lead to the destruction of the population. In assessing potential effects, the presence of high abundance of a species in certain habitats is not always a reliable indicator of population well-being. This is because high abundance may be supported by the inflow of migrants after reproduction has ceased. For example, elevated water temperature in industrial reservoirs in the Dnieper region, Russia, created favorable conditions for marsh frogs (*Rana arvalis*) which became abundant. However, discharge of wastewater to the reservoirs resulted in marsh frogs ceasing reproduction and their high abundance was maintained only due to in-migration [43]. Similar observations have been observed with tundra voles at contaminated sites (section "Technically Enhanced Naturally Occurring Radioactive Material (TENORM)").

Effects of Radiation on Animals

Information on the effects of radiation on biota comes from laboratory studies, field irradiation studies, and observations at sites of industrial releases, nuclear tests, and accidents. Radiosensitivity generally increases with increasing organism complexity [34, 53, 75]. Polikarpov [48] provides a conceptual model for the response of organisms, populations, and ecosystems to radiation dose rates (Table 2).

The following sections briefly review the relevant radiation effects data with emphasis on more sensitive studies for both aquatic and terrestrial organisms. Other information on the effects of radiation exposure can be found in EC and HC [39], Harrison [76], Anderson and Harrison [77], Real et al. [78], and the reviews cited above.

In the following sections, dose rates reported are for gamma radiation, unless otherwise noted.

Ionizing Radiation on Nonhuman Biota, Effects of Low Levels of. Table 2 Response of biota to ionizing radiation as described by Polikarpov [48]

Zone	Radiation dose (Gy·year^{-1})	
Obvious action	4–>3,000	Above lower boundary of evident effects
Ecological masking	0.05–4	Above natural background where see effects, but effects are not significant
Physiological masking	0.005–0.05	Area of overlap between minor radiation effects and natural variability
Radiation well-being	0.00004–0.005	Natural background
Uncertainty	<0.00001–0.00004	Below natural background

In comparison to the effects of acute radiation on organisms, relatively few studies have investigated the effect of chronic exposure on biota. Within the literature, it is common to find studies presenting information on radiological effects but without providing information on absorbed doses, although concentrations may be provided. Information on absorbed dose is essential to interpret radiation effects [79]. For example, in the case of the Chernobyl literature, Polikarpov and Trytsugina [79] noted that 10 out of 16 studies reporting radiation effects did not provide radiometric or dosimetric data, but only the location of the study, i.e., zone of radiation contamination.

Interpretation of the literature on radiation effects is difficult because of the variety of dose rates, total dose exposure, time frame, and endpoints studied, and uncertainty in the relevance of certain endpoints with respect to protection of species. It is also sometimes difficult to estimate the radiation dose that organisms have been exposed to. The presence of other contaminants or stressors in field studies also complicates matters, i.e., it is difficult to determine whether the effect is due to radiation or other contaminants. The response of organisms to contaminants is essentially the same, whether it is exposure to radiation or another contaminant. Most studies have employed high doses of gamma radiation over short-time periods from an external source [3, 39]. Few studies have looked at the effects of internal exposure to alpha radiation or effects on more long-lived, slow-reproducing organisms.

Acute Doses

At high doses, effects are expressed by the death of cells, which may ultimately result in loss of tissue and organ function and, if the dose is high enough, the death of the organism. The mechanisms of radiation-induced mortality in invertebrates and lower vertebrates are similar to those observed in mammals. Major systems affected by high acute exposures are the hematopoietic, gastrointestinal, and immune systems. Lower doses and dose rates affect reproduction and can cause chromosomal aberrations and mutations. Genetic effects, such as point mutations, single-strand breaks, double-strand breaks, and sister chromatid exchanges occur at lower doses in both somatic and germinal cells and may be expressed at the population level.

Most research has been done using acute exposures with high doses generally given as a single dose but, in some cases, continuously for a specific period or until fatal. In the case of earlier studies, this is because the focus was on the potential effects of a nuclear war and accidental acute exposures to humans. Although information has been gained on potential effects of radiation to the environment from acute exposure studies, it should be recognized that most of the exposures used in these studies would be lethal to larger mammals, including humans, and are not representative of levels of radiation seen in the environment as a result of releases from nuclear facilities under normal operations. For our purposes, a radiation dose of greater than 100 mGy in less than a day may be considered an acute exposure, although mortality is not expected following such an exposure.

Chronic Doses

The data on the effects of chronic radiation exposure on mammals are extremely variable. The IAEA [53] concluded that a dose rate of 10 mGy day^{-1} is a threshold at which reproductive capacity may be affected in some mammals. Acute doses of 100 mGy are very unlikely to produce persistent, measurable deleterious changes in populations or communities of terrestrial plants or animals; and irradiation at a rate of 1 mGy day^{-1} is unlikely to cause observable changes in terrestrial animal populations. UNSCEAR [34] concluded that dose rates below 10 mGy day^{-1} to the most exposed members of the population would not seriously increase the death rate of mammal populations, and reproductive effects are unlikely at 10% of this dose rate, or 1 mGy day^{-1}. UNSCEAR [34] also cited a study in which the developing oocytes of the squirrel monkey (*Saimiri sciureus*) had an LD_{50} of 1 mGy day^{-1}, giving total doses in the range of 40–200 mGy over the study. As a general rule, chronic effects begin at 10% of the LD_{50} value, and effects on oocytes occur at 1% of the LD_{50}. Harrison and Knezovich [80] concluded that in mammals, adverse effects on fertility are first observed at a critical range of 0.48–4.8 mGy day^{-1}. Rose [81] concluded that the lowest dose rate producing a spectrum of effects is around 2.7 mGy day^{-1}, although effects have been observed at lower dose rates. For example, a dose rate of 0.2 mGy day^{-1}

(42 mGy) increased the mortality of offspring of laboratory mice provided with tritiated water [82], and a dose rate of 1 mGy day^{-1} (210 mGy) reduced testicular mass and epididymal sperm counts of mice after 30 weeks exposure to alpha radiation from plutonium [83].

Genetic effects in nonhuman species are the major consequence from radiation exposure at low and moderate dose rates. Radiation causes many types of damage to genetic material. Small errors in DNA repair and changes in the gene expression regulatory mechanisms can lead to cancer. Cancer is the primary concern in human radiation protection strategies. Nonhuman vertebrate species also develop cancers in response to radiation [84]. The development of cancer and mortality later in life, after the reproductive life cycle is complete, is considered to have little effect on the population. However, this may not be the case for long-lived, slowly reproducing species, such as elephants and whales or perhaps slow-growing fish in northern lakes. Can cancer be a limiting factor for local populations of animals like the bear that frequently live for more than 20 years in the wild and reproduce every 2 or 3 years, or for local populations like the alligator, which does not start to breed until 10–12 years of age in the wild [74], in areas where their population numbers are low? Certainly the loss of older members of the community in long-lived animals like the elephant that depend on the experience and leadership of elders for guidance under adverse conditions such as drought can be detrimental to the local population.

Survival, reproduction, and growth are the endpoints most relevant to protection of the population. With respect to reproduction, the processes from gametogenesis to embryonic development are limiting endpoints in terms of survival of the population [76]. Endpoints that measure changes in the ability to reproduce are the factors that affect fertility and sterility, such as reduction in number of gametes produced, gamete death, and increased abnormalities and mortality of early-life-stages [76]. Changes in fertility (an indicator of reproductive success) have been attributed to damage to cytogenetic material [54, 78, 85–87]. Although there is much interest in genetic effects because they may be transmitted to subsequent generations, genetic damage per se is generally not considered a limiting endpoint because of the difficulty in interpreting the significance at the population level (i.e., population fitness and survival). However, genetic damage provides an early warning that damage is occurring and may be important at the individual level.

Minor effects may be seen at low dose rates in more sensitive species and systems, the threshold for statistically significant effects in most studies is about 2.4 mGy day^{-1}. The significance of effects on the individual or population of minor responses at dose rates $<$2.4 mGy day^{-1} has not been determined, especially for morbidity and cytogenetic effects. The response increases progressively with increasing dose rate and is usually very clear at $>$24 mGy day^{-1} sustained for a large fraction of the lifespan [78]. In the present chapter, exposure to chronic radiation doses of $<$10 mGy day^{-1} to biota are considered most relevant to exposures biota may receive from continuous routine releases from modern nuclear facilities.

Effects at Elevated Background Concentrations

When assessing the effect of radiation on biota, it is the incremental radiation dose above background that is of concern. Normal background radiation is about 2.4 mGy·$year^{-1}$. Background radiation is generally in the range of 0.6–7 mGy·$year^{-1}$ for leaves and needles of terrestrial plants, 1–5 mGy·$year^{-1}$ for terrestrial mammals, and 0.7–1.7 mGy·$year^{-1}$ for freshwater organisms [34]. However, some species of wildlife may receive much higher doses of background radiation. For example, the accumulation of ^{210}Pb and ^{210}Po in the liver and kidney of some herds of Canadian caribou (*Rangifer tarandus*) [88] may result in doses 10–100 times higher than background (see "► Radiation Effects on Caribou and Reindeer"). Similarly, burrowing animals living in radon-rich soils may receive exposures resulting from short-term peaks in radon concentrations, equivalent to over 100 mGy·$year^{-1}$ [89].

Naturally occurring alpha-emitting radionuclides appear to be the most significant sources of background radiation exposure to organisms. In the terrestrial environment, background radiation is mainly from ^{222}Rn and its short-lived decay products, whereas in aquatic environments, ^{210}Po is the major contributor [34], although it is important in the terrestrial environment as seen in the example for caribou above. Po-210 is

about 8,000–50,000 times less abundant than ^{40}K, but the dose rate from ^{210}Po is sometimes 10–100 times higher than for ^{40}K [90]. K-40 is a major source of internal and external gamma radiation, and together with members of the ^{238}U and ^{232}Th decay chain, are part of the natural background dose (see "► Radiation in the Environment, Sources of").

In certain areas, normal background radiation may be greatly elevated with radiation levels of up to about 100 mGy·year^{-1} (see "► Radiation in the Environment, Sources of"). In the state of Kerala in South India, natural radiation levels are as high as 1 mGy day^{-1}, or about 250 times the normal background (4.3 μGy·day^{-1}), due to ^{232}Th and its daughters in the monazite soil. At these elevated radiation levels, a higher incidence of chromosomal aberrations is reported. For example, higher frequency of meiotic abnormalities and pollen sterility is reported in plants in the presence of monazite soil in comparison with plants from control areas [91]. Bats (*Chiroptera*) living in an abandoned monazite mine showed a dose-dependent increase in DNA damage at 0.5 mSv·day^{-1} and 2.4 mSv·day^{-1} compared to bats from a control area [92]. Rabbits exposed to naturally high radioactivity in areas in southwest France of about 1.9 mGy·day^{-1} and up to 700 mGy·day^{-1} gamma rays and >6 Gy of alpha radiation to the bronchial region from inhalation of radon showed a small but significant increase of chromosomal aberrations (dicentrics) in their lymphocytes. Chromosomal aberrations are also reported in spermatocytes of scorpions (*Tityus bahiensis*), inhabiting areas with radiation levels of 0.7 mGy·day^{-1} and 252 mGy·year^{-1} in the state of Minas, Brazil [93]. Increased levels of chromosomal aberrations in peripheral blood lymphocytes are also observed in humans living in areas with elevated natural background radiation levels in India, China, and Iran [94]. Thus, naturally elevated background levels of radiation may be having subtle effects on biota.

Effects on Mammals

There is an apparent inverse relationship between the $LD_{50/30}$ and the weight of the animal [34]. Lethal radiation doses for small mammals are in the range of 6–10 Gy for several species of rodents (Table 3) with the lowest doses to cause death in the range of 3–6 Gy·year^{-1} [96]. Lethal radiation doses (LD_{50}s) for large domestic animals such as cattle, sheep, goats, pigs, burros, and horses are in the range of 1.2–3.9 Gy [34] and that for large wild animals is in the same range [96]. Dose rates < 4 Gy·year^{-1} do not seriously affect the population of small mammals. However, individual specimens are affected at these doses.

Effects of radiation on larger agricultural animals were observed following the Kyshtym (Mayak) and

Ionizing Radiation on Nonhuman Biota, Effects of Low Levels of. Table 3 The LD_{50} (Gy) of several types of rodents and larger mammals exposed to acute doses based primarily on Dunaway et al. [95]

Species/strain	LD_{50} (Gy)
Larger mammals, including humans	1.5–6 Gy
Four strains of mice and two reciprocal crosses	4.9–6 Gy
Marsh rice rat (*Oryzomys palustris*)	5.3 Gy $LD_{50/30}$
Feral house mouse (*Mus musculus*)	5.5 Gy
Wild house mouse (*M. musculus*)	6.3 Gy
Thirteen strains of laboratory mice	5.1–7.2 Gy
Short-tailed shrew (*Biarina brevicauda*)	7.8 Gy
Wild house mouse (*M. musculus*)	8.2 Gy
Least shrew (*Cryptotis parva*)	8.4 Gy
Norway rat (*Rattus norvegicusl*)	8.7 Gy
RF strain (*M. musculus*)	8.8 Gy
Pine vole (*Microtus pinetorum*)	9.4 Gy $LD_{50/30}$
Eastern harvest mouse (*Reithrodontomys humulis*)	9.5 Gy $LD_{50/30}$
Hisoid cotton rat (*Sigmodon hispidus*)	9.6 Gy $LD_{50/30}$
Golden mouse (*Peromyscus nuttalli*)	10 Gy $LD_{50/30}$
Old-field mice (*Peromyscus polionotus*)	10.1 Gy
Cotton mice (*Peromyscus gossypinus*)	10.1 Gy
Cotton rat (*Sigmodon hispidus*)	10.4 Gy
White-footed mouse (*Peromyscus leucopus*)	10.7 Gy $LD_{50/30}$
Eastern harvest mouse (*Reithrodontomys humulis*)	10.8 Gy
Little pocket mouse (*Perognathus longimembris*)	13.7 Gy

Chernobyl accidents [55]. Following the Kyshtym accident in the Urals, cows feeding in pastures received a radiation dose of 1.4–3 Gy from external gamma exposure within 12 days following the accident, whereas their large intestine received a radiation dose of about 4–23 Gy. Cows began dying on days 9–12 from acute radiation exposure. Similarly, sheep received about 1.4–3 Gy from external gamma exposure and from 8 to 15 Gy up to 30–54 Gy to the large intestine through ingestion of forage. The sheep died within 9–12 day due to acute radiation sickness. Where fallout was less (170 $MBq{\cdot}m^{-2}$), cows and sheep received 0.13 Gy from external radiation and 1–4 Gy to the large intestine during the first 12 days after the accident. These animals showed radiation induced changes to their blood but, when evacuated from the area, they recovered. Based on a decline in the population of elk and roe deer in areas with 3.7–37 $MBq{\cdot}m^{-2}$ of ^{90}Sr, a die off of elk and roe deer occurred. The estimated radiation dose to the intestines was 0.1–1 Gy for elk and ~10–30 Gy for roe deer [55]. Similar findings were reported for the Chernobyl accident [20, 97].

The lowest acute dose in the literature that caused sublethal effects is 10 mGy to pregnant rats, which impaired the reflexes of their offspring (Semagin 1975 cited in [81]). Mice are among the most sensitive species to the reproductive effects of radiation with impairment of reproduction being observed at an acute dose as low as 0.2 Gy for females, while males are less sensitive (3.2 Gy) [53]. An increase in the number of mice embryos resorbed and, in the frequency of abnormalities, is seen in specimens exposed to a radiation dose of 5 mGy or more [98]. However, increases in fertility are also observed at low dose rates, and normal breeding mice are reported at high doses. For example, no effect was observed on fertility for 10 generations of rats exposed to 7 $Gy{\cdot}year^{-1}$ or for six successive litters produced by female rats exposed to 36 $Gy{\cdot}year^{-1}$ [99]. Exposure of male mice to high doses (9 Gy for each of eight generations, 3 Gy for 15 generations, or 2 Gy for 35 generations) resulted in no change in health or fitness of offspring. Slight changes in bull semen occur at an acute dose of 0.5 Gy, but recovery is complete in about 30 weeks post-irradiation.

The dog (beagle) is another sensitive species with a dose rate of 4.3 $mGy{\cdot}day^{-1}$, producing cellular regression and sterility in a few months. However, no effects are observed at a dose rate of 0.9 $mGy{\cdot}day^{-1}$ throughout the dog's lifespan of 12 years [96].

Natality is generally a more radiosensitive parameter than mortality. In general, within the same species, the ovary is more sensitive to radiation than the testis. The most sensitive endpoint for mammals appears to be the killing of 50% of immature oocytes as a result of exposure to 1 $mGy{\cdot}day^{-1}$ (total of 210 mGy) of tritium (^3H) during the last trimester of fetal development in monkeys [82]. A dose of 3.1 $mGy{\cdot}day^{-1}$ to neonatal mice from ^3H produced a 50% reduction in number of immature oocytes. However, more immature oocytes are produced than can be utilized for reproduction, so oocyte killing does not necessarily affect reproduction and population size. For example, chronic exposure to gamma radiation at a rate of 13–26 $mGy{\cdot}day^{-1}$ over 10 generations did not produce changes in the litter size of mice or sex ratios of progeny. Likewise, survival of mammals in an irradiated hardwood forest was not affected at a dose rate of 20 $mGy{\cdot}day^{-1}$ [100]. Wild rodents living on the dry bed of White Oak Lake exposed to lifetime doses of 2–3 Gy showed no effects that could be ascribed to radiation [101].

In their review on the effects of chronic irradiation on animals, Real et al. [78] found no effects on morbidity and mortality at a radiation dose rate of <2.4 $mGy{\cdot}day^{-1}$, but evidence of life shortening at dose rates of 2.4–24 $mGy{\cdot}day^{-1}$ to the dog and mouse, and reproduction effects to the pig, rat, and mouse.

Chromosomal damage may occur at low doses. An increase in micronuclei in the bank vole (*Clethrionomys glareolus*) was observed at dose rates up to 14.4 $mGy{\cdot}year^{-1}$ [102]. Similar chromosomal effects were observed in caribou in Norway after Chernobyl (see "► Radiation in the Environment, Sources of," "► Radionuclide Fate and Transport in Terrestrial Environments") at doses of 50–60 mGy [103]. However, no difference was seen in the frequency of micronuclei in bank voles exposed to a maximum dose rate of 86 $mGy{\cdot}day^{-1}$ within the Chernobyl exclusion zone and unexposed voles [104]. This observation is in conflict with reports of increased micronuclei in rodents exposed at low dose rates. Evidence is also accumulating that damage occurs to genetic material that is not expressed in the irradiated generation but may be expressed several generations later [105].

Effects on Life Span: Life Shortening Life-shortening studies have been mainly focused on rodents using gamma radiation. Mortality at daily doses above 216 mGy·day^{-1} primarily reflects radiation effects to the blood system such as depression of platelets, red cell count, white cell count, etc. Radiation induced life-span shortening of mammals at low and moderate doses (<216 mGy·day^{-1}) are essentially due to carcinogenesis, i.e., premature death due to the induction of tumors and to specific diseases such as neoplasias, amyloidosis, kidney degeneration, etc., which have opportunity to more fully develop with radiation survival times beyond 200 days. There is about a 6 month mean latent time for neoplastic and degenerative diseases in the mouse [5].

Carnes et al. [106] reported that the lowest doses in studies performed at Argonne National Laboratory at which radiation induced mortality caused by primary tumors could be detected was about 1–2 Gy for gamma rays and 100–150 mGy for neutrons. Increased pathological burdens were detected in irradiated mice at doses lower than those that increased mortality, i.e., at 220 mGy for gamma rays and 20 mGy for neutrons. Adverse health effects at low doses, other than cancer, involve multiple organ systems (cardiovascular, kidney, lungs and pleura, and reproductive organs). These pathologies were speculated to have as great an effect on the animal's health as those reported for tumors [106].

Life shortening of mice averaged 35 days·Gy^{-1} at high doses of >0.5 Gy·d^{-1} but only about 4 days·Gy^{-1} at daily doses of 3–560 mGy over their life span (Table 4) [8]. At lower daily doses of a few mGy·day^{-1}, a consistent life shortening affect is not seen. Sensitivity decreases with age and is strongly dependent on age at exposure during the first 2 months of life [5]. Because the response to radiation exposure decreases with age, the lifetime accumulated dose is not truly representative of the biological response [5].

At dose rates <2.4 mGy·day^{-1} of gamma or beta, no detriment effects on morbidity or reproductive capacity are seen. However, neutrons (1 or 5 MeV) at dose rates lower than 2.4 mGy·day^{-1} for 475 days result in increased mortality (50%) of mice. Dose rates of 2.4–24 mGy·day^{-1} of gamma shorten the life of mice and dogs, and reduces the number of primitive stem germ cells by 41% and the weight of ovary and testes by 44% in prenatal pigs compared to controls. The

Ionizing Radiation on Nonhuman Biota, Effects of Low Levels of. Table 4 Estimated life shortening for the mouse at low radiation dose rates and total dose measured in the laboratory at 400–450 mGy·min^{-1} of gamma

Dose rate (mGy day^{-1})	Total dose (mGy)	Life shortening (days)
1	650	3
2	1,290	5
3	1,930	8
5	3,190	13
8	5,030	21

Source: Modified from NCRP [8]

mortality rate of dogs exposed to dose rates ranging from 3–540 mGy·day^{-1} is dependent on the accumulated total dose rather than the dose rate with a slope of 1 for tumor deaths and slope of 2 for non-tumor deaths [7]. Thompson and Grahn [15] concluded that at dose rates lower than 2.0 mGy·day^{-1}, life shortening depends only on total dose and is independent of dose rate. These findings are in contrast to the statement that for life shortening, daily or monthly exposure rate throughout life is more important than the total dose [8].

Searle [107] irradiated C3H mice with ^{60}Co for many generations at a rate of 9 mGy over a 11–17 h night starting at the age of 4–8 weeks through an average generation time of 80 days to first reproduction. Three sublines persisted successfully for 25–30 generations. However, after the second litter, the successive litter sizes became smaller than those of controls. Most mice became sterile before producing a fifth litter. For specimens receiving 18 mGy per night, only one mouse out of 66 produced more than three litters. The effect on litter size was attributed to the killing of oocytes in the ovary, which reduced the number of oocytes ovulated resulting in the early onset of sterility. Early oocyte stages in 10-days old mice were particularly sensitive to irradiation with a LD_{50} of 75.6 mGy for stage I oocytes after gamma irradiation of 26 mGy·min^{-1}. Searle [107] concluded that significant depletion of oocyte stocks occurred at low dose rates of 9 mGy per night and 720 mGy per generation. Vivarium studies have also demonstrated life shortening of mice at a few mGy·day^{-1} [55].

Continuous exposure to gamma radiation at 8 mGy·day^{-1} and 10 mGy·day^{-1} in two studies resulted in about a 50% reduction in oocytes on day 14 of life in mice irradiated from conception [108]. Effects on female fertility occurred over 33 days at a total dose of only 0.26 Gy. Chronic irradiation of sibling matings of four strains of mice at 12–24 mGy·day^{-1} for ten generations did not alter reproductive success, and average litter sizes remained the same because there is a huge over production of oocytes in mice [107]. This is why population effects are usually not seen in small mammals exposed to low levels of radiation in the field.

Hibernation In poikilotherms and hibernating animals at low temperatures radiation damage may be latent, so radiation effects may not be apparent in winter, but at warmer temperatures the lesions may manifest themselves [109].

The pocket mouse (*Perognathus* subfamily Perognathinae Heteromyidae) exhibits a high degree of radiation resistance compared to the other rodents. This is attributed to the capability of heteromyids to become hypothermic under certain adverse environmental conditions [110]. Hibernation after exposure to X-rays prevents damage leading to death in the ground squirrel (*Citellus tridecemlineatus*) and door mouse [111]. However, this was not the case for the brown bat (*Myotis lucifugus*). Therefore, there are species differences with respect to the development of cellular damage in hibernation [111].

Field Irradiator Studies Several field irradiation studies were carried out to assess the effect of radiation on various natural ecosytems (section "Field Irradiator Studies" – Table 10). These studies in general were not very successful in assessing effects to small mammals because of their mobility, the small areas irradiated, the nonuniform delivery of exposure and the short duration of some studies. The results of several of these irradiation studies have been reviewed by Turner [112]. Only a brief synopsis is presented here.

French et al. [113] irradiated Mojave Desert shrub type habitat in the Nevada test site to assess the effect of radiation on plant and animal life, particularly on heteromyid rodents. Nine hectare enclosures were continuously exposed to radiation for 5 years on average 310 days·year^{-1} with about 3.1 Gy in 1965, 3.2 Gy in 1966, and 1.9 Gy in 1967. A sixfold variation in dose rate was observed along the 550-m radius of the plot with most (77%) of the area receiving between 2.7 and 54 mGy day^{-1}. Radiation dose to individual animals was measured using a thermoluminescent dosimeter attached to the skin of their back.

Desert rodents are relatively long lived (e.g., 2–5 years) and maintain reproduction throughout life. Food supply is an important limiting factor for rodents. In good years, the population increased about fivefold, but in poor years there was no reproduction. Sampling to assess the population was nondestructive using live traps. Trapping was successful during summer months when the rodents were active, but less successful during the winter, when they largely remain underground and inactive. This seasonal behavior resulted in animals receiving a radiation dose of about 13.5 mGy day^{-1} in summer and $<$0.9 mGy day^{-1} in mid-winter [113].

The pocket mouse *Perognathus formosus* was exposed to 5–9 mGy day^{-1} (1.9–3.2 Gy·year^{-1}), which reduced survival particularly before the age of 6 months. The instantaneous rate of death for the irradiated population was 0.219 during this period and 0.075 and 0.104 for the two control populations. Life expectancy at age 1 month for the irradiated population was 9.2 months, and 11.4 and 14.4 months for the control populations. Laboratory studies confirmed that these levels of chronic radiation were sufficient to double mortality in preweaning age animals [113]. Thus, the pocket mouse suffered increased early mortality due to irradiation. The intrinsic rate of increase for the irradiated population was 0.314 compared to 0.493 and 0.498 per year for controls. This is a rate that would result in a 40% reduction in the multiplication rate per generation for the irradiated animals. However, Turner [112] considered this an over interpretation of the data for survival since plot-specific estimates of fertility were not available. Two long-lived specimens of *P. formosus*, a female 4 years 9 months old and a male 4 years 10 months of age (sacrificed about 8 months after radiation ceased) received an average dose of about 4.5 mGy day^{-1} for a lifetime exposure of about 16.9 Gy.

French et al. [113] speculated that a population of 3-year-old *P. formosus* subjected to such radiation likely would not be able to reproduce. Hence, when reproduction is curtained by unfavorable conditions,

radiation effects would be detrimental to the population. The rapid turnover of generations prevents the population from accumulating damaging effects at low levels of exposure.

Large kangaroo rats *Dipodomys merriami* and *Dipodomys microps* did not perform well in the enclosures and became extinct in the irradiated plot in <1 year and in one of two control plots after 5 years. The enclosures were too small for kangaroo rats and, with immigration curtailed, these rats became extinct. French et al. [113] concluded that chronic exposure to 8.1 mGy day^{-1} of gamma radiation was clearly detrimental to both the large kangaroo rat and pocket mouse.

Mihok [114] reported that exposure of meadow voles (*Microtus pennsylvanicus*) to radiation from a ^{137}Cs irradiator on six 1-ha meadows in Manitoba over 1–1.5 years resulted in no effects to the population or individuals at measured dose rates as high as 81 mGy day^{-1}. Third generation voles received up to about 5.7 Gy at a dose rate of 44 mGy day^{-1}. Mihok [114] concluded that voles that received a chronic dose of 5–6 Gy survived as well as controls and small numbers of overwintering animals survived and reproduced at doses up to 10 Gy.

Technically Enhanced Naturally Occurring Radioactive Material (TENORM) Tundra voles (*Microtus oeconomus*) inhabit an area of technically enhanced natural radiation near old plants where radium was processed from radium rich groundwater. Dose rates were 800 times reference values at Radium Site No 1 and Uranium-Radium Site No 1 and 400 times at Radium Site No 2 [55, 57, 115]. Tundra voles spend considerable time in their holes in the top 20 cm of soil, which results in radiation doses of 0.84–2.5 mGy day^{-1} in contaminated areas [57] and 2–7 mGy day^{-1} exposures from radon in burrows. Of 3,590 tundra voles evaluated from the exposed sites, 61% showed a decreased fat level compared to 21% for 2,135 control animals. All exposed voles were parasitized compared to 41% of those for control sites. Other effects included: changes in blood composition indicating chronic radiation sickness; low liver weight in young mice and degenerative changes in the liver, spleen, kidneys, testicles and ovaries; cytogenic changes in chromosomal aberrations in bone marrow; individuals with changed karyotype; significant alterations in the lipid peroxidation regulation in tissues; and serious reproductive problems.

Sexual maturity of male voles was inhibited up to 9 months of age compared to controls, which matured at 1–3 months of age. A lower number of females were involved in reproduction and these had half the number of embryos per female than the control population. The reproduction period and survival of young voles under 3 months of age were lower than reference levels (Table 5). Young males voles showed decreased fertility, reduction of testicular weight, and vacuolization of the seminiferous tubules parenchyma from Radium Site No 2, which suggests inhibition of spermatogenesis to complete termination of spermatogenesis [57]. This was more severe in young 1–2 months old specimens. The results reported for tundra voles are in agreement with other studies [107, 116, 117] that have reported increased mutation rate and reduction in life span, litter size, and reproductive capacity at low doses and dose rates of about 0.8 Gy per generation.

Despite a reduced reproduction period and decreased life span, exposed voles showed compensatory effects in fertility which together with immigration helped maintain the population size. Immigrants formed 30% of the population, or 1.5 to 2 times more of the population at contaminated sites than the

Ionizing Radiation on Nonhuman Biota, Effects of Low Levels of. Table 5 Radiation effects on tundra voles exposed to elevated radium concentrations near Vodnyi, Komi Republic, Russia

Parameter	Reference site	Ra No. 1 Site	U-Ra No. 2 Site
^{226}Ra in voles ($mBq{\cdot}g^{-1}$ ash)	26±12	201±99	265±198
Life span (months)	16.2±2.3	12.6±2.8	16.4±2.9
Reproductive period (months)	9.6±1.8	7.9±1.8	4.8±2.2
No of litters	6.7±0.8	8.6±1.6	6.3±1.1
Young per female	25.7±3.1	31.2±5.8	28.6±4.0
% loss of young in 3 months	13.9±2.32	31.8±3.75	51.4±4.73

Source: Based on Geras'kin et al. [57]

reference site. However, within a month, the immigrants were as contaminated as resident specimens [55].

Effects were observed in otters (*Lutra lutra*) and game birds living in an area of elevated natural thorium in the northern taiga, Komi region. The area has up to 1 mg Th·g^{-1} ashed soil and a gamma radiation background of 0.2–0.25 mGy day^{-1}. Otters living along a river in the elevated thorium area weigh less than controls, and spent less time in the water, i.e., hunt less, possibly due to disruption of thermoregulation. They were also 33% less numerous along a 127-km stretch of river than the reference area. In the same area, populations of great grouse (*Tetrao urogallis*) and black grouse (*Lyrurus tetrix*), two large game birds, had higher radionuclide concentrations than smaller grouse, hazel grouse (*Tetrastes bonsai*) and ptarmigan (*Lagopus lagopus*). They also have smaller populations per km^2 than in a reference area; 13–25% less for great grouse and 14–35% less for black grouse. Large grouse weighed less than reference specimens and were more heavily infested with feather parasites and endoparasites [55].

Nuclear Accidents

Kyshtym Accident In September 1957, the Kyshtym accident at the Mayak nuclear materials production complex (Mayak Production Association) east of the town of Kyshtym in the southern Urals, Russia, resulted in the release of ~7.4 × 10^{17} Bq of fission products from a concrete tank of liquid radioactive waste contaminating a stretch of forest called the East Urals Radioactive Track. Initially, about 90% of the radiation was from short-lived radionuclides (^{89}Sr, ^{144}Ce, ^{95}Zr, and ^{95}Nb) with ^{90}Sr accounting for only about 2.7% of the radioactivity. Initial maximum contamination was up to 1.5 × 10^8 Bq·m^{-2} for ^{90}Sr near the explosion. An area of 23,000 km^2 was contaminated with ^{90}Sr at a density of 3.7 kBq·m^{-2} and an area of 1,000 km^2, about 105 km in length by 8–9 km wide, with 74 kBq·m^{-2} of ^{90}Sr [55].

In the first fifteen years following the Kyshtym accident, during which radiation doses decreased from 100 mGy day^{-1} to 1 mGy day^{-1}, an increase in mortality of mice and a reduction in their fertility and life span was observed (Table 6)[55, 118]. Mortality of mice increased by a factor of two to ten, life span decreased by a factor of 1.5 to 2, and their fertility decreased by a factor of 1.25 at a radiation dose of 10 mGy day^{-1}. These changes had no effect on the population, but population structure and condition changed with bloodsucking ectoparasites increasing by a factor of five on small mammals. More than fifteen years after the accident, deviations were still evident in the morphology of blood and marrow of mice at total doses of 0.18 to 1.8 Gy·year^{-1}.

Chernobyl Accident The Chernobyl reactor accident in April 1986 released about 3–6 × 10^5 TBq of ^{137}Cs and 2–4 × 10^5 TBq of ^{90}Sr to the atmosphere of which about half was deposited within 20 km [119]. This resulted in the death of trees and other biota, and the establishment of a human exclusion zone. In the first 10–20 days, radiation doses in the exclusion zone may have reached up to 880 Gy mainly due to beta radiation [119]. From April 26 to June 1, 1986, external radiation was one to two orders of magnitude higher than that of the internal radiation dose. Initial exposure was mainly due to aerosols precipitated on the trees. Two months after the accident, most radionuclides moved to the leaf-litter layer with 95% or more of the radionuclide inventory residing in the litter layer 1–2 years after the cessation of radioactive fallout [120] and remained in the upper 3–5 cm soil layer for ~7 year. After the initial exposure to aerosols, exposure to vegetation was by root uptake [121].

Maximum absorbed dose following the Chernobyl accident occurred in the first year after deposition. External exposure to beta radiation, because of the path length of beta particles, was more an issue with smaller animals than large animals since their skin and fur absorbed most of the external beta dose [120]. Radiation dose declined monotonic mainly as a result of the decline in external gamma radiation. Exceptions were due to delayed maximum internal exposure [122]. Maximum levels of radiocesium occurred at 1–2 years after deposition, followed by an exponential decrease. Incorporation of ^{90}Sr increased up to the tenth year after deposition. Uptake of transuranic elements (^{238}Pu, ^{239,240}Pu, ^{241}Pu, and ^{241}Am) were much lower. Considerable ^{241}Am was first detected in animals, the bank vole (*C. glareolus*) and yellow-necked mouse (*Apodemus flavicollus*), in areas of high contamination five years after deposition and concentrations continued to increase over the next 5 or more years.

Ionizing Radiation on Nonhuman Biota, Effects of Low Levels of. Table 6 Radiation dose and dose rates to biota and observed effects following the Kyshtym nuclear accident at Mayak Production Association in the Urals, Russia, in 1957

Dose rate and accumulated dose	Effect
0.6–5 mGy day^{-1} and total dose of 0.1–1 Gy	No effect on average number of embryos in female northern red-backed mouse (*Clethrionomys rutilus*)
20 mGy day^{-1} and external dose of 0.13 Gy, 1–2 Gy large intestine	12 days after the accident, cows and sheep had abnormal blood but recovered when moved to an uncontaminated area
0.28 Gy day^{-1} and 10 Gy	Slight life shortening of northern red-backed mouse (*C. rutilus*)
60 mGy day^{-1} and 12–20 Gy	Life shortening of short-tailed vole (*Microtus agrestis*), 16% of contaminated adults had anomalous growth of upper teeth, which interfered with feeding – not seen in controls
11 mGy day^{-1} and total dose of 4 Gy	Shortening of the reproductive period of the European wood mouse (*Apodemus sylvaticus*) because of early death of adults
15 mGy day^{-1} and total dose of 5 Gy per lifetime, 0.57 mGy day^{-1} to bones with total dose of 0.2 Gy	Increased leukocyte concentration in short-tailed vole (*M. agrestis*) and weakened resistance to blood parasites (*Leucocytgrenarinae mieroti*) by a factor of 6 from control specimens; life shortening, i.e., 344±53 days for Kyshtym mice in vivarium versus 433±134 days for controls; poor health of the European wood mouse (*A. sylvaticus*), i.e., heavily infested with mites, eye diseases, loss of hair, and a reduced immune response compared to controls
11 mGy day^{-1} and 4.3 Gy	Altered blood composition with signs of leucopoenia, anemia and inversion in lymphocyte/neutrophile proportion, percent composition relative to controls was: 60%±5 for erythrocytes; 96%±8 for hemoglobin; 127%±20 for thrombocytes; 46%±7 for leucocytes; 295%±60 for neutrophiles; and 37%±4 for lymphocytes. Life shortening decreased the reproductive period of adults with 5–10% of the control females being pregnant verses 0.8–1% in exposed females. In autumn, the proportion of older animals in the population was 5–10% less than in the control population; number of progeny/female was the same as the control
3 mGy day^{-1} and total dose of 1 Gy	*A. sylvaticus* had a higher average rectal temperature and average breathing rate than controls. At a 3.7–370 kBq ^{90}Sr·kg^{-1} mice were more vulnerable to predation by the bird buzzard (*Buteo buteo*) than less contaminated mice, i.e., about 80% mice eaten had ~35 kBq ^{90}Sr·kg^{-1} in their bones, while only 20% trapped had this level, most (60%) had about 3.7 kBq·kg^{-1}

Source: Based on Sazykina and Kryshev [55]

Within 2–3 years after the Chernobyl accident, mice populations practically recovered in the Chernobyl zone. However, radiation effects typical of animals living in contaminated areas were present in subsequent years, e.g., negative changes in blood, infestation with parasites, hypooxygenia, and cytogenetic effects [55]. Ecological differences in mice species affected their response to the accident. Species that either fed on less contaminated food items, root voles (*M. oeconomus*), and/or lived in sheltered places, the house mouse (*M. musculus*), substantially increased in numbers after the Chernobyl accident due to evacuation of people from the area and abandonment of agricultural plants in fields. Fluctuating asymmetry in yellow-necked mouse (*Apodemus flavicollis*) was higher in close proximity to the Chernobyl reactor for both size and shape. Detectable effects of radiation on developmental stability probably start to occur between 3 and 7 μGy day^{-1}. Fluctuating asymmetry was highest (3.6 times reference) in specimens from the 10-km exclusion zone at a dose rate of 100 μGy day^{-1}, intermediate (2.3 times reference) in the 30-km exclusion zone at a dose rate of 7 μGy day^{-1}, and lowest in the reference area with a dose rate of 1.3 μGy day^{-1} [123].

Chesser et al. [124] reported that the bank vole (*C. glareolus*) living in the highly contaminated Red Forest area near Chernobyl Reactor 4 experienced an average dose of 18 mGy day^{-1} from ^{137}Cs and 25 mGy day^{-1} from ^{90}Sr, most of which is taken up from the diet. Total accumulated dose (internal and external) for voles was about 1.1 Gy over 30 day and 1.5 Gy for the house mouse (*M. musculus*). No evidence was found for reproductive failure in mammals from the Red Forest, i.e., the rodents were able to maintain their populations. In another study, Baker et al. [125] trapped small mammals within the 10-km exclusion zone of the Chernobyl Reactor 4, within the 30-km exclusion zone that received minimal radioactive pollution, and five sites outside of the 30-km exclusion zone. In total, 355 specimens representing 11 species were collected within the exclusion zone, and 224 specimens representing 12 species outside the exclusion zone. They concluded that diversity and abundance of small mammals were not reduced in the exclusion zone, and there were no aberrant gross morphological features other than enlargement of the spleen within the contaminated zone. Gross chromosomal rearrangements were not recorded, and 8–9 years after the Chernobyl accident, the small mammal community appeared normal [125]. These latter findings differ from those of Oleksyk et al. [123] and Sazykina and Kryshev [55].

Within the 10-km exclusion zone red fox (*Vulpes vulpes)* grey wolf *(Canus lupus*), moose (*Alces alces*), river otter (*L. lutra*), roe deer (*Capreolus capreolus*), Russian wild boar (*Sus scrofa*), and brown hare (*Lepus europaeus*) were observed, while only a single brown hare was observed beyond the 30-km exclusion zone. Exodus of the human population from the exclusion zone allowed wildlife to flourish [125], although the data for small mammals indicates that they are heavily contaminated. Actual radiation exposure to large animals and birds immigrating to the exclusion zone is less than that for small mammals because of their mobility, large feeding areas, and the nonuniform distribution of the contamination.

Animal populations were able to maintain themselves in the Chernobyl exclusion zone several years following the accident and even flourished in the absence of humans. However, effects typical of radiation exposure were observed in small mammal populations, and these populations should not be considered normal. In a review of the effects of the Chernobyl accident on biota, Polikarpov and Trytsugina [78] noted that 10 out of 18 studies reported deleterious effects over a wide range of dose rates (3 μGy day^{-1} up to 100–200 mGy day^{-1}), whereas the other 8 studies reported no damage below dose rates of 4–0.8 mGy day^{-1}. In three studies where the exposed dose was controlled, individual effects were observed at dose rates below 0.07–0.7 mGy day^{-1}.

Turner [112] concluded that chronic radiation of $>$ 1 mGy day^{-1} reduced fertility in rodents. Real et al. [78] concluded that dose rates lower than 24 mGy day^{-1} do not produce clear irreversible effects on morbidity, mortality, or reproductive capacity in nonhuman mammals. Impairment of reproductive capacity occurs at a threshold of ~2.4 mGy day^{-1}, although detrimental effects are reversible. The data support the finding that 1 mGy day^{-1} for mammals is protected at the population level, although there are effects on individuals [78]. However, it must be recognized that there is a lack of information on alpha emitters on mammals and for the effect of radiation on other species than rodents. Based on Chernobyl data, Fesenko et al. [58] derived a threshold dose for radiation effects of 1.6 mGy day^{-1} for cattle and 1 mGy day^{-1} for small mammals. Bird et al. [126] and EC and HC [39] also concluded 1 mGy day^{-1} is protective for small mammals.

Birds

Wild birds appear to exhibit LD_{50} values (5–12 Gy) in the same general range as small mammals [34]. The effect of radiation on birds has been reviewed by Brisbin [127] and Mellinger and Schultz [128]. Effects are seen at dose rates that may be considered acute exposures. Three eggs in a swallow's nest exposed to 1.5 Gy day^{-1} did not hatch despite prolonged incubation by the adults [129]. In four nests exposed to dose rates of 1 Gy day^{-1} (total dose of about 16–20 Gy), hatching was suppressed (56.3±14.7% compared with 82% and 91% for controls), but fledging success was not affected. Increased embryonic mortality was observed in birds irradiated at about 300 mGy day^{-1} in a Wisconsin forest [130], and impaired gametogenesis of irradiated chicken embryos was observed at 240 mGy day^{-1} in the laboratory [131].

Breeding tree swallows (*Tachycineta bicolor*) were exposed to a radiation dose of up to 45 times background levels (up to 0.14 mGy day^{-1}) [129, 132] without affecting breeding success [132]. In a field irradiation study, breeding tree swallows and house wrens (*Troglodytes aedon*) were exposed to radiation dose rates up to 5 mGy day^{-1} without any apparent effects on hatching, fledging, or breeding.

In the East Urals Radioactive Track, a reduction in hatching and a growth delay of chicks by 20% in nest volume was observed after the accident at radiation doses of 10 mGy day^{-1} [133]. Reproductive failure of flycatchers in man-made nests was reported in the Kyshtym area [56], but reproductive failure was not associated with specific radiation doses or dose rates. A loss of fitness was observed in the barn swallow *Hirundo rustic* breeding close to Chernobyl with an increase in partial albinism observed in the population. The number of breeding pairs also decreased 74% from 292 pairs in 1991 to 76 pairs in 1996 in nine villages near Chernobyl compared to a reduction of 19.8% from 202 pairs in 1991 to 162 pairs in 1996 in six villages in control areas. Unfortunately, radiation dose estimates were not available [134]. UNSCEAR [34] reported that swallows and sparrows produced young that appeared normal 4 months after the Chernobyl accident. From the above, it appears that birds do not seem to be particularly sensitive to radiation.

Amphibians and Reptiles

Data on the effects of radiation on survival of amphibians and reptiles are generally available only from studies with acute exposures where the post-irradiation observation period is often 30 days (30-d LD_{50}). Extending the observation period usually lowers the dose causing 50% mortality. In the case of poikilothermic (cold-blooded) animals, temperature can control the time of expression of radiation effects. Therefore, for fish, amphibians and reptiles, which are poikilotherms, a 60- or 90-day study period is more appropriate than the 30-day period normally employed for mammals.

LD_{50} values for frogs, salamanders, turtles, snakes, and lizards tend to be in the range of 2–22 Gy [34]. Although both reptiles and amphibians appear to be less sensitive to radiation than mammals [34, 135], work by Sparrow et al. [136] indicates that the LD_{50} for the mud puppy (*Necturus maculosus*) is less than 1 Gy when a longer study period is used, putting it in the same range of sensitivity as mammals. For the mud puppy, the 30-d LD_{50} of 35.5 Gy drops to 0.8 Gy at 200 days following a single exposure (i.e., 200-d LD_{50}). Other examples can be found in Woodhead and Pond [137]. It is hypothesized that *Necturus* is more sensitive than other amphibians because of a very large interchromosomal volume [138] and lack of a suitable system to repair the radiation damage [136]. LD_{50} values for adult anurans (frogs and toads) range from about 6 to 20 Gy [135]. Panter [139] showed that the most sensitive stage for acute exposure for the spotted grass frog (*Limnodynastes tasmaniensis*) is the fertilized egg, with a 40-d LD_{50} of 0.6 Gy. Urodeles (e.g., newts, mud puppy) are also sensitive, with LD_{50} values between 1 and 5 Gy, assuming up to 300-days postexposure for the time of lethality [136]. Juvenile life stages have lower LD_{50} values, ranging as low as 0.9 Gy (4.5 mGy day^{-1}) for Fowler's toad (*Bufo woodhousei fowleri*). The LD_{50} for acute radiation changes markedly through the developmental stages of a frog, increasing from <1 Gy in the early stages of development to over 25 Gy in the adult [140]. LD_{50} values for reptiles are in the same range as those for adult amphibians (>8 Gy).

In the Kyshtym area, the brown frog (*R. arvalis*) had smaller eggs, lower reproductive success, and 17% more morphological deformities than control specimens. Young brown frogs grew faster, but adults were smaller than control specimens [55]. Also in the Kyshtym area, the viviparous lizard *Lacerta viviparous* had 26.6% morphological deformities verses 2.1% in the nonexposed population [55]. Similarly, at the highly contaminated Izumrudnoye site 3 km from the Chernobyl NPP, in the first year after the accident, over 33% of the frog eggs were infertile, 17% in 1988, and 3% in 1989 compared to 0.01% in the control area [55]. Unfortunately, in the above studies, effects were not associated with specific radiation doses or dose rates.

For the moor frog (*R. arvalis*), the dose rate was highest when the frog was buried in sediment and lowest in the egg stage when suspended in water. This is because of the low ^{137}Cs concentration in water. A 1-cm layer of water between the egg mass and the sediment reduces the dose by 45% [141]. No effects were specified at external doses of 21–160 mGy·y^{-1} and

an internal dose rate between 1 and 14 mGy·y^{-1} measured in their study [141].

Continuous irradiation of lizards at about 20 mGy day^{-1} for 4 years from a ^{137}Cs source resulted in sterility at a cumulative dose of 15 Gy [112]. The annual radiation dose to the iguanid lizard (*Uta stansburianna*) was about 3.6–18 Gy, depending on location in the enclosure with most individuals receiving about 7.5 Gy. Dose rates to leopard lizard (*Crotaphytus wislizenii*) are estimated at 6.8–13.5 mGy day^{-1}, to the horned lizard (*Cnemidophorus tigris*) at about 3.5–7 mGy day^{-1}, and as high as 20–50 mGy day^{-1} to the iguanid lizard (*U. stansburianna*). Population survival depended on the individual species' life span, time to sexual maturity, and population age structure. *U. stansburiana* was protected by its rapid turnover (early sexual maturation and high fecundity), whereas longer-lived lizards, the horned lizard (*C. tigris*) and the leopard lizard (*C. wislizenii*), that matured later in life and produced fewer offspring suffered population losses, becoming nearly extinct [112]. All mature females became sterile, and there was no reproduction. Some of the irradiated *U. stansburianna* also became sterile by the age of 19–20 months. In the fourth year of the study, no female leopard lizards went into breeding colors, and three specimens sacrificed in the fifth year all lacked ovaries.

In summary, there is a lack of information on the effects of prolonged exposure to low levels of ionizing radiation for sensitive species such as frogs and the mud puppy, and most amphibian and lizard species in general. Based on the available information, amphibians and reptiles are as sensitive of chronic radiation exposure as mammals and a protective dose would be $\leq$1 mGy day^{-1}. However, no information is available on long-term exposures to more long-lived, slower-reproducing species such as alligators, crocodiles, and turtles. In the wild, the American alligator does not reach sexual maturity until about 10–12 years of age [74] and hence could accrue a considerable radiation dose even at low chronic dose rates before their first reproduction.

Fish

Several reviews on the effects of radiation on aquatic organisms have been published [39, 142–144]. Fish are considered the most radiosensitive of the nonmammalian aquatic organisms, with reproductive capacity being the most sensitive endpoint. Behavior and feeding habits affect the exposure of fishes to contaminants. Predatory fish tend to accumulate greater concentrations of contaminants than nonpredatory fish. For example, Koulikov [145] reported that the average ^{137}Cs levels were about 6.3 times higher in perch (*Perca fluviatilis*) and 4.4 higher in pike (*Esox lucius*) than nonpredatory species bream (*Abramis brama*), silver bream (*Blicca bjoerkna*), and rudd (*Scardinius erythrothalmus*). For tench (*Tinca tinca*) and goldfish (*Carassius* sp.), the factor was ~2. Differences were due to ecological and physiological factors. Bream fed mainly on *Chironomus* larvae while trench show a preference for mollusks and insect larvae. Other examples are given in "► Nuclear Accidents, Chernobyl Fallout in Scandinavian Watersheds."

Fish is the only non-mammal vertebrate studied in relatively more detail for dose rate or total dose effects. Most studies have involved acute exposures. Only 34 chronic studies were found at low-dose exposure to fish in the FASSET Radiation Effects Database [78]. The IAEA [87] concluded that reduced reproductive success would likely occur at dose rates in the range of 24–240 mGy day^{-1}, well above chronic exposures normally seen at modern nuclear facilities. Real et al. [78] reported a reduction in testis mass and sperm production, lower fertility and delayed spawning at a radiation dose of 2.4–24 mGy day^{-1} in plaice, medaka and roach. Anderson and Harrison's [146] synthesis of the data on effect levels for a number of endpoints indicated that a dose rate of 5–100 mGy day^{-1} would encompass the level at which a variety of low-level effects on reproduction, development, and genetic integrity are detectable in sensitive tissues and organisms. Increased mortality is also expected at a sustained dose rate of 240 mGy day^{-1} [146]. Dose rates (5–100 mGy day^{-1}) at which detrimental effects on fertility are first observed in fish are similar to those observed in mammals [146, 137].

The 30-d LD_{50} values, which are generally between 8.8 and 44 Gy, suggest that fish are less sensitive to radiation than mammals. These conclusions are based primarily on short-term acute toxicity studies. As noted above (section "Amphibians and Reptiles"), the

30-day period to assess radiation effects is too short for poikilotherm vertebrates. However, extending the study period reduces the LD_{50} considerably and into the lethal range for mammals [137]. For example, for mummichog (*Fundulus heteroclitus*), the 30-day LD_{50} is 12 Gy, whereas at 60 days, the LD_{50} is 3–3.5 Gy [147].

The LD_{50} for adult fishes ranges from 3.8 to 30 Gy, whereas the LD_{50} for fish embryos ranges from 0.16 to 25 Gy [39]. The lowest LD_{50} reported for fish was 0.16 Gy in coho salmon (*Oncorhynchus kisutch*) exposed at the one-cell stage and observed for 150 days [148]. The next lowest LD_{50} was 0.58 Gy for the one-cell stage of rainbow trout (*Oncorhynchus mykiss*) [149], then 0.9 Gy from exposure to X-rays for plaice (*Pleuronectes platessa*) larvae irradiated at the blastula stage [150]. The lowest acute exposure causing effects on reproductive tissue of fish appears to be 0.25 Gy [146]. A significant reduction in growth rate was observed when rainbow trout embryos were acutely exposed to 0.38 Gy [149]. In the same study, an increased frequency of abnormalities was observed in embryos irradiated at 2 Gy. Major malformations were observed when developing eggs of the mummichog were exposed to 3–4 Gy [151]. An acute dose of 5 Gy caused a 50% reduction in hatching of carp (*Cyprinus carpio*) eggs [152].

Few studies have been conducted to determine radiation doses that would cause mortality in fish as a result of chronic exposures. Significant reductions in fecundity have been observed at chronic doses ranging from $<$14.4 to 312 mGy day^{-1}. Chinook (*Oncorhynchus tshawytscha*) and coho salmon embryos exposed for 16–20 days and for periods up to 80 days to an external ^{60}Co source at a dose rate of 5 mGy day^{-1} (total dose of 330–400 mGy) showed a significant increase in deformities [153, 154] but no difference in survival at time of release, or in the number of adult salmon returning. The latter is possibly because at best only 1.2% of the salmon released returned to spawn. Opercular defects and fusion of vertebrae were observed in first-generation (F1) irradiated coho alevins [153]. Deformities were also more severe in F1-generation chinook salmon. When controls from the 1960 brood that returned in 1963 were mated with control and experimental (irradiated brood) fish, the resultant F1 generation, which had not been irradiated but were the offspring of irradiated parents, showed an increase in deformities. Twinning of the head was the most common deformity observed in chinook alevins, then curvature of the body and shortening of the body (fusion of vertebrae, omissions and incomplete tails) and opercular defects [153, 154]. Deformities were not observed in adults that returned to spawn, which may suggest that deformed fish were removed from the population. Exposure to 100 mGy day^{-1} or more was required to produce a significant change in the number of adult salmon returns [155]. These high radiation doses resulted in decreased growth and increased mortality of young fish in freshwater, decreased return of adult salmon, increased age at return and sterility in adult males [155].

In mosquitofish (*Gambusia affinis*) living in White Oak Lake (Oak Ridge National Laboratory, Tennessee, USA), a pond contaminated with radionuclides, exposure to 0.6 mGy day^{-1} and other contaminants over a lifetime produced a small but significant increase in embryo mortality over controls [156]. This small increase in mortality, although statistically significant, is not considered biologically significant, as the rapidly reproducing (spawning one or more times a year) mosquitofish population was thriving. Chronic exposure of the guppy (*Poecilia reticulata*) to 41 mGy day^{-1} caused reduced fecundity [157]. However, the guppy is a radio-resistant fish species with a 30-day LD_{50} of 23.5 Gy when irradiated with an acute dose of X-rays and no significant increase in mortality is seen in young guppies exposed to a total dose of 3.4–47 Gy from ^{3}H [158]. In a short-lived, rapidly reproducing fish species, the zebra fish (*Danio rerio*), exposure to a chronic radiation dose of 178 mGy day^{-1} from ^{137}Cs, caused complete sterility, with an EC_{25} for reproduction of approximately 24 mGy day^{-1} from ^{137}Cs [159]. In the same study, the highest dose rate from the alpha emitter ^{210}Po that caused no effects was reported to be 5 mGy day^{-1}.

Tilapia mossambica raised in ^{90}Sr contaminated aquaria for 800 days showed essentially normal reproduction at 4–5 μGy day^{-1}. At 0.4–0.5 mGy day^{-1}, male gonads were smaller in mass than those of control specimens and spermatogenesis was reduced, whereas no effects were observed on female gonads. At 30–40 mGy day^{-1}, 100% of males were sterile, 30% of females had underdeveloped ovaries, and 80% of females had anomalies of ovaries. Complete

suppression of reproduction occurred. Impregnation of exposed females with control males resulted in larvae that died by day 160 [60]. Overall survival decreased with radiation dose: 71% for controls, 54% for the 0.4 mGy day^{-1} fish, and 33% for 30 mGy day^{-1}. Weakening of the immune system also occurred with radiation exposure, i.e., experimental infection of fish exposed to 30 mGy day^{-1} with parasites increased mortality by 2–4 times compared to controls [60]. Reduced fecundity of the roach (*Rutilus rutilis*) that lived their entire lives in aquaria exposed to ^{90}Sr was also observed at dose rates > 5 mGy day^{-1} and all males became sterile [60].

Exposure of the carp, *C. carpio*, to concentrations of 1,850 Bq·L^{-1} and 37,000 Bq·L^{-1} of ^{90}Sr in aquaria for 15, 30, 90, 180, and 270 days resulted in radiation doses to the kidney of 0.2 mGy, 1.2 mGy, 42 mGy, 200 mGy, and 340 mGy. This affected the lymph system. Leucocyte concentrations and their phagocytic activity decreased with dose, and the immune response was delayed and weakened. Early signs of effects were detected at dose rates of 1 mGy day^{-1} and accumulated doses above 0.05–0.2 Gy. At lower doses, recovery was seen in 180 days. At a dose rate of 7–12 mGy day^{-1} and an accumulated dose of 2.5–3 Gy, distinct effects on fish immunity were seen. At 90 days and a radiation dose of >0.04 Gy, leucocyte concentrations were lower and were depressed at doses greater than 0.7 Gy by 40–50%. Changes in proportions of leucocytes were also seen along with negative biochemical changes in male gonads. At 0.4 Gy, the glycogen concentration in gonads was not detectable and testicles had up to 95 mg of fat per gram gonad. This increased to 127 mg fat per gram gonad at 0.7 Gy [60]. In nature, inhibition of the immune system results in increased parasite infection, development of morbid effects such as loss of competitive capacity and early mortality. Early signs of effects on the lymph system are detected at chronic dose rates of 1 mGy day^{-1} and accumulated doses above 0.05–0.2 Gy. At dose rates of 7–12 mGy day^{-1} and accumulated doses of 2.5–3 Gy, the effects on fish immunity are clearly distinct [60].

Studies on the effect of chronic exposure to radiation in the contaminated Chernobyl cooling pond demonstrated radiation effects to fish at low dose rates [10, 160–162]. Both caged silver carp (*Hypophthalmochthys molitrix*) and free-swimming specimens from the Chernobyl cooling pond were monitored from 1989 to 1992 to assess the effect radiation exposure had on their reproductive system.

Silver carp are planktivores; therefore, they accumulate lower concentrations of radionuclides than benthic and predatory fishes. Silver carp reaches sexual maturity at three years of age. From 1989 to 1992, first-generation silver carp received a radiation dose of about 0.2 Gy annually (0.55 mGy day^{-1}), i.e., about 0.6 Gy for the 1989 generation, 0.4 Gy for the 1990 generation, and 0.2 Gy for the 1991 generation [160]. Over the study period, the internal dose to silver carp was about 150 mGy·$year^{-1}$. This was primarily from consumption of plankton in summer and periphyton (filamentous algae) that colonized the cages in winter. In comparison, the parent stock, which was present in the cooling pond in 1986 as 1–2 year olds, received a radiation dose of 7–10 Gy, primarily from external radiation [160]. The external radiation dose was measured using dosimeters suspended at different depths, whereas the internal radiation dose was based on radionuclide concentrations in tissues. Internal radionuclide contamination was from ^{137}Cs and ^{134}Cs in flesh, ^{90}Sr in bone and ^{144}Ce, ^{141}Ce, ^{106}Ru, ^{95}Zr, and ^{95}Nb absorbed to the gut wall.

An increase in the number of silver carp (*H. molitrix*) with anomalies of their reproductive system occurred following the Chernobyl accident. From 47% to 90% of the fish exhibited anomalies in their sex cells. Anomalies included sterile individuals, morphological changes in the gonads such as gonad asymmetry, and degeneration of sexual cells. Fish with gonad asymmetry had one normal ovary or testis, and the other either degenerated or completely absent. Other anomalies such as those associated with the eye, e.g., absence of eye pupil and lens, and blindness in one eye may have been due to radiation or parasites.

The number and severity of reproductive anomalies in caged silver carp increased with time (Table 7). Despite abnormalities in their reproductive system, the high fecundity of silver carp allowed them to maintain their population in the cooling pond [60, 160].

Radiation effects were also investigated in the bighead carp (*Aristichthys nobilis*) [161]. Bighead carp was not as radiosensitive as silver carp. Only one female specimen exhibited gonadal abnormalities, large vacuoles in the oocyte cytoplasm. Male specimens showed

Ionizing Radiation on Nonhuman Biota, Effects of Low Levels of. Table 7 Effect of radiation on caged silver carp in the Chernobyl cooling pond following the Chernobyl accident

Year	Sterility (%)	Male gonadal malfunction (%)	Anomalies of male sex cells (%)	Anomalies of female oocytes (%)	Gonad asymmetry (%)
1989	5.7	25	25		8.6
1990	12.5	33	47.1	55	16.7
1991	0	57	68.8	78	23.1
1992	0 15.4[a]	100	100 89.5[a]	33 89[a]	9.1[a]
Reference[b]	0.25				4.6

[a]Free – swimming silver carp in Chernobyl cooling pond

[b]Experimental fish farm near Uzbekistan, which had water temperatures close to that in the cooling pond; 10–14°C in winter and 25–30°C in summer

Source: Based on Makeyeva et al. [160]

some effects, i.e., enhanced growth of connective tissue in gonads, lower sperm concentrations, and anomalous spermatozoa, but not to the same extent as for silver carp [161]. Catfish (*Ictalurus punctatus*) in the Chernobyl cooling pond were also found to exhibit greater genetic damage in relation to the radiocesium concentration in individual fish [163]. Similar finding were reported for the frequency of micronucleated erythrocytes in northern pike and the concentration of ^{137}Cs in pike in the Tom River downstream of the Siberian Chemical Complex [164].

In the Mayak Southern Urals field studies, the effect of radiation on several species of fish was investigated. The Siberian roach (*R. rutilus*) exposed to an external radiation dose of 4.8 mGy day^{-1}, 7 mGy day^{-1} to gonad, and 15 mGy day^{-1} to bone had lower fertility by a factor of ~2 and more morphological deformities (6–9% vs 0.5–1) than reference fish. A small percent (0.5–2%) had underdeveloped gonads and signs of hermaphroditism. In the roach (*R. rutilus*), asymmetry in the number of soft rays in pectoral fins was observed in fish living in the canal for liquid wastes from the Leningrad nuclear power plant. These fish were exposed to a dose rate of about 2 mGy day^{-1} [162].

A sample of 358 goldfish (*Carassius auratus gibelio*) was collected over a 3-year period from a triploid population of female fish from 1972 to 1975 in two lakes in the Mayak Southern Urals area. In Berdenish, the goldfish were exposed to a radiation dose of 30–40 mGy day^{-1} in 1957, which had decreased to 0.5 mGy day^{-1} in 1972–1975. Absorbed doses were 350–600 mGy for 2-year-old specimens, 730–1,300 mGy for 3-year-old specimens and 1.1–2 Gy for 5-year-old specimens. Morphological deformities, mainly deformities of gonads, were observed in about 24% of the specimens. The frequency of sterility was 17% for fish in the age group 3+ fish, and up to 25% for 4+ and older fish; 13.4% of the fish had only one ovary, 18.2% were without gonads, and 18% had curvature of the fins and tail, and irregular scale structure. Chlorine was also noted as present, but no information was given on concentrations. For the Lake Uruskul population, the radiation dose was about 30–40 mGy day^{-1} in 1957 and 3–5 mGy day^{-1} in 1972–1975. Two-year-old fish received 2.2–3.3 Gy, 3-year-old fish 3.3–4.45 Gy, and 5-year-old fish 5.5–7.75 Gy. About 15% of the goldfish in Lake Uruskul had morphological deformities and gonad anomalies (unpaired gonads and sterility). Despite increased numbers of deformities in gonads, growth and nutritional status of fish were normal in the contaminated lakes. In the two Ural Lakes, Uruskal and Berdenish, no goldfish specimens older than 8 years were present; 4–6-year-old specimens dominated spawning shoals. This suggests that radiation had a life shortening effect since older specimens are usually present [60, 90].

In the Southern Urals, anomalies are seen in the development of northern pike (*Esox lucius*) living in

Reservoir No. 10, which receives radioactive wastes (^{90}Sr, ^{137}Cs) and chemicals from radiochemical plants at the Mayak complex, i.e., the water was sulfuric with a pH of 4.5–5 [162]. The pike were tested for their ability to reproduce by mixing roe from 100 fish and examining the development of 1,000 fore larvae in comparison to control specimens from Lake Alabuga. An order of magnitude increase in anomalies above the reference population was observed at dose rates of 5–10 mGy day^{-1} (about 6.5 mGy day^{-1} from beta radiation and 1 mGy day^{-1} of gamma). About 8.3% of the exposed pike had serious deformities – no eyes, no yolk sac compared to 1.1% for controls. About 19–27% of the fish had curvature of the spinal cord in both exposed and control specimens. The latter was attributed to transportation from the distant lake to the fish farm. Nine types of deformities were present in exposed fish verses two types in controls. Immature fish exhibiting anomalies were removed from the population in the first few months of their life by natural selection. Despite the anomalies and loss in recruitment, the pike population in the reservoir remained viable over a period of 30 years [162].

Based on the observation of radiation effects on fish in the Chernobyl cooling pond and on fish at other contaminated sites in the former Soviet Union, Kryshev and Sazykina [162] concluded that dose rates below 10 mGy day^{-1}, in combination with other anthropogenic factors, affect reproduction and cause anomalies in fish species. Sazykina and Kryshev [60] concluded that obvious morphological and functional effects occur in the reproduction system of fish at chronic exposures above 2–5 mGy day^{-1}, and minor chemical changes occur around 0.5 mGy day^{-1}. Nevertheless, fish are able to maintain their population in highly contaminated water bodies.

For fish, the lowest acute LD_{50} is 0.16 Gy in coho salmon exposed in the one-cell stage [148]. At 5 mGy day^{-1}, anomalies are seen in the development of northern pike [162] and deformities are seen in coho salmon (total dose of 0.4 Gy) [153, 154]. The lowest reported chronic exposures to produce reproductive effects were observed in carp subjected to chronic exposures of 0.6 mGy day^{-1} in the Chernobyl cooling pond. The value of 0.6 mGy day^{-1} for reproductive effects in carp was used as an estimated-no-effect-value by EC and HC [39]. The carp, a more long-lived species that is slower to mature and reproduce, is affected by 0.6 mGy day^{-1}, whereas the mosquito fish, a short-lived, early to mature, fast reproducing fish, is not affected at this dose rate.

Considering the large number of eggs produced and natural losses, radiation may not be a controlling factor for most fish populations. For fish with long development times, such as the fish eggs of Arctic fish or viviparous fish with a relatively small number of eggs, radiation effects on embryos may limit the population [60]. This needs more detailed investigation. Effects on fish are in general similar to the effects reported for warm-blooded vertebrate animals. At the organism level, the following radiation effects are important for the survival of populations: weakening of the immune system, decreased reproduction, increased number of abnormalities, and life shortening [60].

At low doses, an increase in mortality is not observed directly, but mortality usually occurs in the form of a reduction in age-dependent survival. The effect of life shortening may be a cumulative result of effects on morbidity as well as abnormalities in reproduction and cytogenetic damages [60]. Slight life shortening may occur at relatively low dose rates of chronic exposure and may have a greater effect on long-lived species than on short-lived organisms.

Sazykina and Kryshev [60] concluded approximate threshold levels of chronic exposure above which specific types of effects can be detected are:

- 0.5–1 mGy day^{-1} with total doses above 0.05–0.2 Gy for the first appearance of negative changes in fish blood composition and early signs of a decrease in immune system function.
- 2–5 mGy day^{-1} with accumulated doses above 1.5 Gy for the appearance of negative effects on the reproductive system.
- 5–10 mGy day^{-1} of chronic lifetime exposure leads to life shortening of adult fish.

These effect levels are lower than the 10 mGy day^{-1} recommended or adapted for protection of aquatic populations. Sazykina and Kryshev [165] provided a threshold dose level of 0.3 mGy day^{-1} for fish, which is similar to the 0.6 mGy day^{-1} of Bird et al. [126] and EC and HC [39], whereas Fesenko et al. [58] used a value of 1.6 mGy day^{-1}.

Radiation is a pollutant that is frequently found in combination with other stressors. In the field, it is difficult to separate out the effect of radiation from that of other contaminants. In cooling ponds, water temperature is elevated. Elevated temperatures accelerate sexual maturation resulting in an increase in duration of gonad maturation of stages II-III in females, shorter duration of vitellogenesis, and reduced size and weight of mature eggs. In species that deposit eggs in batches, the number of eggs per batch decreases, and the spawning period is longer. Individual fecundity decreases in most species. For example, in the roach, fecundity increases as a result of higher temperature, but the percentage of oocyte resorption is high, and fecundity is actually lower than without elevated temperatures. In bream inhabiting cooling ponds, mass asynchronism of oocyte maturation is observed. Oocytes in the second generation usually undergo resorption with mass resorption occurring at all development stages in cooling reservoirs [43].

In water bodies heavily polluted with mixtures of contaminants (heavy metals, pesticides, organics), the reproductive system of fishes may be affected. Disturbances include asymmetrical development of ovaries and testicles, constriction of gonads, the absence of gonads in some individuals, testicular tissue degeneration, substitution of generative tissues by connective and adipose tissue in testicles, hermaphroditism in various forms (one gonad is an ovary and another is a testicle, or both types of generative tissue are found in the same gonad), mass reabsorption of oocytes, spermatocytes at different developmental stages, and mass abortive ovulation long before spawning. Many of these responses are associated with radiation effects. When mixtures of contaminants are present, it is difficult to assess what is causing the effect, i.e., a particular contaminant acting alone or in combination to produce the effect.

Invertebrates

Terrestrial Invertebrates Adult insects are relatively hardy when it comes to radiation exposure because very little cell division and differentiation occur in adults. Juvenile stages of insects are much more sensitive to radiation because of higher rates of cell turnover and differentiation. About 0.1 Gy kills the eggs of the braconid wasp (*Bracon hebetor*) [166], and <1.3 Gy kills the eggs of the housefly (*Musca domestica*) [167].

An impact on soil invertebrates as a result of elevated radium concentrations in soil was reported in an area where radium was commercially extracted from groundwater at Vodnyi in the Komi Republic in northern Russia (see section "Technically Enhanced Naturally Occurring Radioactive Material (TENORM)"). In these areas, soil invertebrates populations, particularly earthworms, are reduced [56, 57, 133, 168]. Immature earthworms were absent in plots with elevated radium concentrations but common in the reference plot. Earthworm numbers were reduced by a factor of 7 at 0.021 mGy day^{-1} recorded at the soil surface and changes in the earthworm integument epithelium, and midgut epithelium were observed at a radiation dose of 0.84 mGy day^{-1} [56]. In contrast to these findings, Hertel-Aas et al. [169] reported that the lowest dose rate to produce an effect on the earthworm *Eisenia fetida* was 96 mGy day^{-1}. Adult earthworms are radioresistant, but their juvenile stages are much more sensitive [170]. Slow-developing organisms and tardigrade species were also identified as being sensitive to elevated levels of radiation at the Vodnyi site [57].

Following the Kyshtym accident, earthworms and myriapods populations in the East Urals Radioactive Track were reduced by more than a factor of 10–100, where the radiation dose exceeded 6 Gy [119]. An estimated radiation dose of 30 Gy 2 months after the Chernobyl accident did not impact the number of adult invertebrates, but juvenile numbers were reduced due to mortality of eggs and early-life stages and the reproductive failure of adults [170]. Invertebrate numbers, including those for earthworms, were reduced only in the first two years or so after the accident. Earthworm numbers were more abundant than at the reference site in 1988 with only one species *Dendrobaena octaedra* present at contaminated sites and two species (*D. octaedra* and *Apporectodea caliginosa*) at the reference site [171]. Sampling of four sites within the Chernobyl exclusion area in 2002 revealed only one earthworm from the Parshev area (low contamination), three earthworms from the Forestry area (medium contamination), and no earthworms from the Pine Trees (medium contamination) and the Red Forest (high contamination) areas [119]. It was suggested

that the low pH 4.5–5.5 of the soils may not have been ideal habitat for the earthworms [119]; however, worms tend to be pH tolerant.

Jackson et al. [119] noted a general loss of invertebrate diversity in the Chernobyl exclusion zone with increasing radiation levels, but no change in biomass [119]. A noticeable reduction in number of soil invertebrates (mites, earthworms, myriapods, etc.) was observed at a radiation dose of 5–10 Gy [58, 120]. Asynchronous development was also reported between the hatching of leafworm eggs and their food source, the blossoming of leaves, following the Chernobyl accident [120, 172].

In field irradiation studies, invertebrates are more affected by indirect effects of chronic exposure from an external radiation source, such as loss of litter, than by the irradiation itself [112, 173]. Some examples are as follows:

- Increased diversity of microarthropods with increased diversity of lower vegetation
- Increase density in leaf miners due to scarcity of leaves
- Increase in ant populations as a result of removal of standing dead plant material which increased the area's attractiveness to foraging
- An outbreak of aphids on white oak leaves as irradiated leaves contained a higher sugar concentration [112]

A decrease in numbers only occurred at high doses. The effects of field irradiation studies on animals, including invertebrates, have been reviewed by Turner [112].

Fesenko et al. [58] suggested a radiation dose of 2.5 mGy day^{-1} as a threshold effects level based on Chernobyl data and because inhibition of reproduction usually occurs at doses an order of magnitude lower than lethal doses. Effects on earthworms are reported at a dose of <1 mGy day^{-1}.

Aquatic Invertebrates Most invertebrates are relatively tolerant of radiation. For example, Marshall [174, 175] showed that *Daphnia* are very resistant to radiation with sterilization occurring at about 105 Gy. Populations tolerated 13.3 Gy day^{-1} [174]. A significant decrease in percent hatch occurs in the calanoid copepod *Diaptomus clavipes* at a dose of 10 Gy [176]. The data on marine invertebrates suggest that acute LD_{50} values range from 2.1 Gy (*Palaemonetes pugio*) to 560 Gy (*Callinectes sapidus*, adults) [146]. Significant reductions in fecundity occur at doses ranging from 1.7 to 13,200 mGy day^{-1}. More recent studies exposing *Daphnia magna* to alpha radiation revealed that low doses of $\geq$ 2.6 mGy day^{-1} produced a significant (15%) reduction in body mass, smaller egg masses, and smaller neonates [177]. At a radiation dose of $\geq$7.2 mGy day^{-1}, the proportion of breeding females in the second generation was reduced, body mass was smaller, and oxygen consumption was increased [178].

A radiation dose of only 0.24 mGy day^{-1} is reported to accelerate cell division in the protozoa *Colpoa* sp [81, 90]. Embryos of the goose barnacle (*Pollicipes polymerus*) were most sensitive to radiation, showing a reduction in molting relative to controls when exposed to tritiated water for 32 days at a dose rate of 1.7 mGy day^{-1} [179] or dose rate of 3.4 mGy day^{-1} when the RBE of 2 for tritium is accounted for. However, the cultures were under considerable stress. Only about 55% of the control larvae molted from naupluis stage I to stage II, and antibiotics were required for successful culture. In a subsequent study with the marine mollusc *Mytilus edulis* exposed to tritium, significant effects on abnormalities and mortality of embryo-larvae accrued at a dose of about 0.12 mGy (corrected for a RBE of 2 and a 72 h exposure period) [61]. Tritium was genotoxic to *M. edulis* as measured by micronucleated hemocytes and induction of DNA strand breaks (Comet test) in a dose dependent manner at dose rates of 0.3, 2.2 and 11.6 mGy day^{-1} [180].

Chronic exposure of polychaete worms, *Neanthes arenaceodentata*, to a dose rate of 4.6 mGy day^{-1} for a total radiation dose of 0.55 Gy from a ^{60}Co source over their life cycle resulted in a significant reduction in live embryos per brood and hatchlings per brood and a significant increase in abnormal embryos per brood [77]. In contrast to the findings with *N. arenaceodentata* [77], the lowest dose rate to produce a significant effect on the polychaete worm *Ophryotrocha diadema* was 77 mGy day^{-1} [181]. Sensitivity of *N. arenaceodentata* is attributed in part to its gametogenic and spawning strategy; the gametes develop synchronously, and the female spawns only once [80]. In addition, little or no repair appears to

occur during egg development. In comparison, *O. diadema* lays eggs over several weeks [181]. A large variety of life-history strategies are observed in invertebrates, from synchronous development of gametes following a single spawning to laying several egg sacs to parental care of offspring. Considerable variation in radiosensitivity is observed and should be expected in invertebrates, much like that seen in other taxonomic groups.

In aquatic ecosystems, benthic organisms are likely to be the most highly exposed organisms due to the partitioning of radionuclides to sediment. Chronic irradiation of *C. tentans* larvae in White Oak Lake, Oak Ridge, Tennessee, at about 6 mGy day^{-1} increased the frequency of chromosomal aberrations but had no apparent additional effects on the population [73]. Reduction of the dose rate to about 0.3 mGy day^{-1} due to radiological decay resulted in the frequency of chromosomal aberrations decreasing to that found in reference populations. In the warm waters of White Oak Lake, *C. tentans* would complete its life cycle in about 30 days for a total dose exposure of about 180 mGy at a dose rate of 6 mGy day^{-1}. In a colder climate, it would take *C. tentans* much longer to complete its life cycle, especially for the over-wintering generation exposing the larvae to a much larger total dose and possibly greater effects.

In White Oak Lake, the snail *Physa heterostropha* experienced a reduction in egg capsule production at a radiation dose of about 6 mGy day^{-1}, but there was no difference in reproduction compared with the nonirradiated population because the number of eggs per capsule increased in the irradiated population [182–184]. In the laboratory, a dose rate of 240 mGy day^{-1} from ^{60}Co had no significant effect on reproduction, mortality, or size of the snail [183]. The pond snail (*Physa acuta*) exposed as a four-celled embryo had an acute LD_{50} of 10.8 Gy [185]. In contrast to the findings for the pond snail, the freshwater gastropod *Pila luzonica* appears to be more sensitive to irradiation [186]. Five velagis stages (embryos) were exposed to ^{3}H at concentrations of 3.7×10^{3}, 3.7×10^{5}, and 3.7×10^{7} Bq·L^{-1}. Concentrations of 3.7×10^{5} Bq·L^{-1} and greater resulted in histological abnormalities in the digestive tract. Abnormalities included infolding of the stomach wall, smaller stomach, variation in the thickness of the stomach wall, and hypertrophied cells. This was a dose-dependent response. At an estimated dose of about 0.06 mGy day^{-1}, 10–20% of the embryos exhibited abnormalities, whereas 45–65% of the exposed embryos were deformed at a dose rate of 6 mGy day^{-1} corrected for a RBE of 2. These abnormalities of the digestive tract occurred in nonfeeding embryo stages and were interpreted by the authors as effects on growth. It is not known whether the nature of these abnormalities could adversely influence growth and survival of affected individuals in later stages of life.

Tsytsugina and Polikarpov [187] studied the genotoxic effect of radiation on three species of oligochaetes *Dero obtuse, Nais pseudobtusa*, and *Nais pardalis* in a Chernobyl contaminated lake located in the 5-km zone of the Red Forest near Yanov. The lake was contaminated with metals, chloro-organic compounds, and high concentrations of ^{90}Sr, 5,872 Bq·kg^{-1} dw sediment with a dose rate of 0.2 mGy day^{-1}. The reference lake was located 20-km south of the Chernobyl NPP and had a ^{90}Sr sediment concentration of 105 Bq·kg^{-1} dw and a dose rate of 16 μGy day^{-1}. In the exposure lake, asexual reproduction and paratomous division was stimulated in *D. obtuse*, whereas activation of sexual reproduction occurred in the other two species. Asexual reproduction is the norm in these worms, with sexual reproduction occurring under stressed conditions. The change from asexual to a sexual mode of reproduction occurred at about ~0.3 mGy day^{-1}.

Increased cytogenetic damage in somatic cells of the three worms (6–8.3% in the contaminated lake verses 1.7–4.9% in the reference lake) was possibly due to low levels of radiation (0.2 mGy day^{-1}), although the other contaminants may also be responsible [187]. Where radiation effects could be excluded and the damage attributed to chemicals, chromosomal aberrations more closely followed a geometric distribution and the aberration spectrum included mostly bridges. With beta and gamma exposure, the distribution of chromosome aberrations between cells is described by the Poisson distribution. In the presence of both chemicals and radiation, the distribution is closer to a Poisson distribution if the effect is mainly from radiation and to a geometric distribution if the effect is mainly from the chemicals [49, 188]. Tsytsugina and Polikarpov [187] concluded more work is required on irradiation effects in the ~0.005 to ~4 Gy·$year^{-1}$ range

since these doses affect a larger area and greater population than higher doses. Further information is required on the toxicity of mixtures that include radiation at low doses and whether there is indeed a difference in the type and distribution of chromosomal aberrations in the presence of different contaminants.

Circumstantial information suggests that tardigrade may be sensitive to radiation. Tardigrada were reported to be abundant in a control pond but were absent from an adjacent pond contaminated with ^{137}Cs [189]. In the experimental pond, the radiation dose to benthic invertebrates was estimated to vary from about 0.2 mGy day^{-1} in the centre to 9 mGy day^{-1} at the perimeter in the pond [39]. The effect of radiation on Tardigrada needs further study.

Fesenko et al. [58] derived threshold effect values of 7 mGy day^{-1} for zooplankton and 2.5 mGy day^{-1} for benthic invertebrates. Bird et al. [126] and EC and HC [39] in their assessment of radiation effects on nonhuman biota used the value of 4.6 mGy day^{-1} as a no effect level based on the response of the polychaete worm to radiation [77] because the endpoint (reproductive success) is most relevant to population health. Chronic exposure of *C. tentans* larvae to about 6 mGy day^{-1} resulted in measurable effects on individuals but did not result in population-level effects [73]. It can be concluded that radiation doses of $<$10 mGy day^{-1} may not affect populations but may affect individuals.

Effects of Radiation on Plants

Terrestrial Plants

Larger plants are generally more radiosensitive than small ones [96] and young plants are much more sensitive than mature plants [190]. Plant parts vary in their radiosensitivity. The apical meristem, the growing tip of the main root or stem is most sensitive, whereas the dry seed is considered most resistant. In general, plants that have many small chromosomes are more resistant to ionizing radiation than plants with a few large chromosomes. Woody species tend to be about twice as sensitive as herbaceous species, for a given interchromosomal volume [75]. Low-stature plants and dormant seeds are more resistant to radiation effects [75]. The order of radiosensitivity is as follows: coniferous trees $>$ deciduous trees $>$ shrubs $>$ herbaceous species $>$ lichen and fungi. In general, strong growth inhibition occurs at 40–50% of the lethal dose and inhibition of seed production occurs at 25 - 35% of the lethal dose [96].

The LD_{50} values for 60 woody plants range from 4.1 Gy for sugar pine (*Pinus lambertiana*) to 77 Gy for bitternut hickory (*Carya cordiformis*) and mockernut hickory (*C. tomentosa*) [39]. Gymnosperms (conifers trees) are more sensitive, by almost an order of magnitude, than angiosperms (deciduous trees) and are among the most sensitive of all plants. Plant morphology affects the radiation dose, the size, shape, and density of plant stands alters exposure and consequently radiation dose, e.g., shades sensitive parts [78].

There are relatively few studies involving chronic exposure of plants to radiation because of the logistic difficulties in having plants grow for an extended time in elevated radiation fields. Studies are primarily from field irradiator studies and from areas contaminated by nuclear releases and older contaminated waste disposal sites. Long-term irradiation experiments with plants showed that the most sensitive endpoint is the loss of viability or mortality. A few data indicate that the production of viable seed is at least as sensitive an indicator of radiation damage as mortality [191].

Some more sensitive studies in which effects were observed for plants exposed to low acute doses include: a 50% reduction in seed yield at 0.51 Gy to a flowering plant; a small portion of young Douglas fir trees (*Pseudotsuga douglasii*) were killed by a dose of 0.8 Gy [192]; a radiation effect at 0.1 Gy on an enzyme related to auxin, a growth regulator hormone in plants [193]; and an increase in root growth of seeds at 0.2 Gy. Lethal radiations are usually between 10 and 1,000 Gy for plants [96]. Some plant species predicted to be more sensitive to radiation at acute doses of $<$10 Gy are given in Table 8. UNSCEAR [194] noted that spruce trees are more radiosensitive than pine trees with absorbed doses of 0.7–1 Gy, resulting in malformed needles, buds, and shoot growth.

For chronic exposures, a more sensitive finding was reduced growth (13%) of Scotch pine seedlings grown in small pots above a waste management facility at a dose rate of 2.4 mGy day^{-1} [195]. Amiro [196] and Amiro and Sheppard [197] reported a no-observed effect level for sensitive species of 2.4 mGy day^{-1} in

Ionizing Radiation on Nonhuman Biota, Effects of Low Levels of. Table 8 Predicted sensitivity to gamma radiation of major woody ecosystems and the dominant plant species that are affected by an acute radiation dose of less than 10 Gy

Major ecosystem and vegetation type	Dominant species	Common name	Predicted H 16-h acute gamma LD_{50} (Gy)
Coniferous forests			
Boreal	*Abies balsamea*	Balsam fir	8.9 I
	Picea glauca	White spruce	8.5 I
	Picea abies	Black spruce	5 I[a]
Subalpine and montane	*Abies lasiocarpa*	Alpine fir	6.2
	Picea engelmanni	Englemann spruce	7.3
	Pinus ponderosa	Ponderosa pine	5.8
	Pseudotsuga douglasii	Douglas fir	9.9
	Pseudotsuga douglasii	Douglas fir	4 I[a]
Sierra-Cascades	*Abies concolor*	White fir	8.1
	Pinus jeffreyi	Jeffrey pine	6.7
	Pinus lambertiana	Sugar pine	4.1
	Pinus ponderosa	Ponderosa pine	5.8
	Pseudotsuga douglasii	Douglas fir	4.6
Pacific conifer	*Abies grandis*	Grand fir	6.2
	Tsuga heterophylla	Western hemlock	8.0
Deciduous forests			
Beech-maple-basswood	*Tsuga canadensis*	Canada hemlock	7.2
Hemlock-hardwood	*Pinus resinosa*	Red pine	7.8 I
	Pinus strobus	Eastern white pine	4.7 I
	Tsuga canadensis	Canada hemlock	7.0 I
Oak-hickory	*Pinus taeda*	Loblolly pine	6.3

H – Based on calculations of interphase chromosome volumes from active meristems
I – Observed mortality in actual experiments
I[a] – from UNSCEAR [34]
Source: Modified from Sparrow et al. [192]

a field irradiation study. These studies and those of Gunckel and Sparrow [193], Sparrow et al. [198] and Whicker and Fraley [199] all reported effects at dose rates between 2.4 and 20 mGy day^{-1}.

Pine Trees

Kyshtym Accident The Kyshtym accident at the Mayak nuclear materials production complex in the southern Urals Russia contaminated an area of 23,000 km^2 with ^{90}Sr at a density of 3.7 $kBq{\cdot}m^{-2}$, and a smaller area of 1,000 km^2, about 105 km in length by 8–9 km wide, with 74 $kBq{\cdot}m^{-2}$ of ^{90}Sr [55]. About half the contaminated area was covered with birch trees and to a lesser extent birch-pine forests. An average radiation dose of 40 Gy to the crown of pine trees resulted in their death. The LD_{100} was 30 Gy to up to 50 Gy and represented an area of 20 km^2. At doses of 15–20 Gy, a wide spectrum of radiation effects were observed such as a 50% and greater reduction in sprout growth, under development of needles, up to 70%

die-off of needles in the crown and inhibition of radial growth of wood tissue [118]. More sensitive herbaceous plants showed a 10–25% reduction in growth at an absorbed dose of 5 Gy. Stimulation of plant growth such as an increase in length of sprouts and needles by up to 20% occurred at doses below 1 Gy.

Chernobyl Accident The Chernobyl accident had a marked effect on the surrounding pine forest. Pine trees close to the reactor in an area of 500–600 ha were estimated to be exposed to doses greater than 80–100 Gy, doses above the acute LD_{50} level which killed the trees. In a second zone, with an area of approximately 3,000 ha, pine trees received doses above 8–10 Gy. Die back of new vegetative shoots of pine trees and damaged needles and buds were observed. In a third zone of about 12,000 ha, pine trees received doses of about 3.5–4 Gy. In this zone, moderate effects to pine trees were observed, including suppression of growth and needle loss, reduced capacity, and genetic damage [34].

Pine Trees as a Sentinel Species As already noted, pine trees are highly sensitive to radiation (and other contaminants). This, together with their widespread distribution, makes pine trees an ideal organism to biomonitor for ecological and genetic effects of radiation. The reproductive organs of conifers have a complex organization with a long generative cycle, so sensitive tissues are exposed to contaminants for prolonged periods and contaminants are highly retained. Most angiosperms species have a reproductive cycle lasting several months, but Scotch pine (*P. sylvestris*) seeds require at least 18–20 months to mature from micro- and megaspore formation. This allows for significant DNA damage to accumulate in undifferentiated stem cells at low dose rates and low contaminant concentrations. The cytogenetic anomalies in the intercalary meristem of young pine needles are sensitive biomarkers of contaminant effects. Damage to the DNA mainly appears as chromosome aberrations at the first mitosis [200].

The Chernobyl accident occurred in spring when the reproductive organs of plants were highly radiosensitive. Short-lived radionuclides made up most the dose in the first 10–20 days. By late summer–early autumn, the dose rate on the soil had dropped to 20–25% of the initial value. Thus, the initial acute radiation dose was replaced by low-dose chronic exposure. Variation in the distribution of radionuclides resulted in greater than an order of magnitude variation in the radiation dose even in small localized areas. For example, at four sites within the exclusion zone, the air kerma levels were: 2.4–12 $\mu Gy\ day^{-1}$ at Paryshev, 74–142 $\mu Gy\ day^{-1}$ at Pine Trees, 86–398 $\mu Gy\ day^{-1}$ at Forestry, and 1.2–3.3 $mGy\ day^{-1}$ at the Red Forest [119].

In the first acute period of the Chernobyl accident, the absorbed dose was primarily due to external β and γ-irradiation, which resulted in cytogenic damage similar to acute gamma irradiation at comparable doses [201]. Mass mortality of pine trees occurred at a radiation dose of more than 60 Gy, the dying of the trees being accelerated by an evasion of pathogenic insects. Injury to pine stands was high at 10–60 Gy, medium at 1–10 Gy, and low at 0.1–1 Gy [172]. By May 1986, the pine forest had became known as the "red forest" because of the orange colored needles on the dead trees. These trees were cut down and buried. Since 1988, the forest has been replaced with grasses, shrubs, and young trees of deciduous species. By autumn 2000, young pine trees had started to reappear at the periphery of the red forest [121].

Geras'kin et al. [200] studied the frequency and spectrum of cytogenetic anomalies for reproductive (seeds) and vegetative (needles) tissues to assess the effect of different levels of radiation on Scotch pine (*P. sylvestris*) in the Chernobyl NPP 30-km zone and areas used for the processing and storage of radioactive wastes, the Leningrad regional waste processing enterprise radon site at Sosnovy Bor, Leningrad Region, Russia. In the Chernobyl, 30-km exclusion zone ionizing radiation predominated, whereas chemical toxins were the main contributors to contamination in the Sosnovy Bor area. Pine seeds in the Chernobyl area were collected from two areas in the exclusion zone, the asphalt–concrete plant at Chernobyl which received 10–20 Gy in 1986 and the village of Cherevach a relatively clean area within the 30-km zone, and the town of Obninsk in Kaluga Region, a remote control area [200]. Pine seeds in the Sosnovy Bor area were obtained from cones collected in 1995 from three areas: the Radon Leningrad regional waste processing enterprise site, the town of Sosnovy Bor, and a control area.

Cell aberrations in the root apical meristem are the result of aberrations induced during the period from gametogenesis to the maturation and harvesting of the seeds. Cytogenetic damage found in seed (roots from seedlings) and needle samples from the Chernobyl NPP 30-km zone increased with radiation exposure. Likewise, cytogenetic damage at Radon Leningrad regional waste processing enterprise site was higher than controls but cannot be solely attributed to radiation. Tripolar mitoses, which are rare anomalies possibly linked to spindle damage, were found in preparations from seeds sampled at both the Leningrad regional waste processing enterprise radon site and Sosnovy Bor, but not in samples from the Chernobyl NPP 30-km zone [200].

Cytogenetic effects of radiation on chromosome aberrations in anaphase cells in seedlings in the first mitosis of embryonic wood were higher in seeds of the first two reproductions than seeds produced subsequently [121]. The effect of radiation on *P. sylvestris* in the 30-km exclusion zone following the Chernobyl accident is given in Table 9.

Exposure to 0.5 mGy day^{-1} over long time periods increased radiosensitivity of pine trees indicating the accumulation of unrepaired damage. A decrease in pine shoot growth rate was observed at 0.43 Gy and cessation in growth occurred at 3.45 Gy in the Chernobyl area. Morphological alterations in pine needles occurred at a dose rate of 24 mGy day^{-1} and an accumulated dose of 13 Gy [78]. In the Chernobyl 0.2–20.4 TBq·km^{-2} area, a linear dose-dependent relationship was observed for cytogenic and genetic effects induced in *P. sylvestris* [121]. Acute radiation induced cytogenic and genetic effects that were significantly higher at 0.5 Gy than controls. Acute radiation at doses >1 Gy induced formation of morphoses and depressed growth and, at doses >2 Gy, the reproductive ability of the trees declined. The minimum dose at which

Ionizing Radiation on Nonhuman Biota, Effects of Low Levels of. Table 9 Effect of radiation dose rate on the Scotch pine (*Pinus sylvestris*) in the 30-km exclusion zone following the Chernobyl accident in 1986

Radiation dose (Gy)	Effect
0.5–1.0	Stimulation – an increase in secondary shoot growth
1–5	Slight damage, decrease in annual growth, morphological alteration of vegetative organs (variable needle length, intense budding at tips of annual shoots, secondary growth) in the first two years, then disappeared, frequency of chromosomal aberrations approached control levels by 7 years
5–15	Frequency of chromosomal aberrations was seven times control over the first two years then decreased to 2–3 times control by the eleventh year; moderate damage, depression of shoot, needle and wood growth, partly damaged crown and mortality of low tree classes, disturbed morphogenesis of vegetative organs and needle ultrastructure, shortened shoots with intense needle growth (broomlike shoots), distorted spatial orientation of needles and shoots (crushed needles and shoots), gigantic and dwarfed needles and shoots, significantly altered reproductive organs (small or absent male flowers, low number of seeds in cones), seeds had low germinating capacity and power, normal growth and development restored in 3 years, maintained its initial structure and function, i.e., the population did not decline in number, and more radioresistant offspring were produced
5–10	Threefold decrease in seed number per cone, a 12-fold increase in yield of unfilled seeds and 23% decrease in weight of 1,000 seeds. Effects were far less pronounced after 11 years
10–20	Most pine trees perished, surviving trees showed delayed reproduction in the sublethal zone for 5–7 years
0.36–0.96 mGy day^{-1}	After 14 years, restoration of trees was in progress, with mortality absent, increase in shoot growth, amount of needles, number of male strobiles, and pines reappearing

Source: Based on Kal'chenko and Fedotov [121]

morphologic effects occurred to pine trees following the Chernobyl accident was 1.2 mGy day^{-1} and involved reduction of shoot growth and morphoses [58].

Gene expression was assessed in *P. sylvestris* needles in the year 2000 at about 10 km from the Chernobyl NPP, where the absorbed dose in the first year was about 3–5 Gy but had subsequently decreased to 0.24–0.48 mGy day^{-1}. Changes in gene expression were small and in agreement with other studies of the flora and fauna in that considerable recovery had occurred in the Chernobyl area over the 10 years or more after the accident [202].

Other Plants Morphological effects were observed in plants at doses of 4.8–7.2 mGy day^{-1} with enhanced vegetative reproduction observed in heather and gigantism in other plants at 18–36 mGy day^{-1} following the Chernobyl accident [194]. The effect of chronic radiation up to 17.3 mGy day^{-1} on herbaceous plants was assessed at fifteen sites within the Chernobyl 30-km exclusion zone based on mass of 1,000 seeds, germination of seeds, and frequency of aberrant cells in roots of germinated seeds. No significant radiation effect was observed possibly because herbaceous plants are less sensitive to radiation than conifer trees [203]. Fesenko et al. [58] identified 8 mGy day^{-1} as a threshold effect dose for herbaceous plants at Chernobyl as cytogenic disturbances and point mutations were observed at this dose rate. Sterility, decrease in germination of seeds and morphological anomalies occurred at radiation doses greater than 27 mGy day^{-1}.

Radiation-induced effects were evident in agricultural plants. Increased cytogenetic damage was seen in the root apical meristem of seedlings in rye and wheat in the10-km zone of the Chernobyl NPP in 1987–1989. At an absorbed dose of 3.1 Gy, plants showed a significant increase in the yield of aberrant cells. Microsatellite mutations increased from 1.03×10^{-3} to 6.63×10^{-3} per locus over one generation at a total dose of about 300 mGy in wheat grown on contaminated soil (27 $MBq \cdot m^{-2}$) near Chernobyl. Percent germination decreased by about 50% at 12 Gy [201].

Mutation induction in cornflower *Arabidopis* and barley plants near Chernobyl indicated that doses of >2.5 mGy day^{-1} are less effective than low doses of 0.1–1.1 mGy day^{-1} in inducing mutations when exposure was mainly from internal irradiation [78]. At the East Ural RadiationTrack, an elevated frequency of chlorophyll mutations in cornflower *Arabidopis* was still observed 38 years after the Kyshtym accident. Thus, decline in cytogenic damage induced by radiation lags the decrease in radiation dose as a result of radioactive decay. Mutation rates significantly greater than the spontaneous rate occurred at 0.3 mGy day^{-1} at a total dose of between 1 and 10 mGy.

External gamma radiation at 0.52 mGy day^{-1} delayed the growth and development and yield of wheat, barley, and beans in uncontaminated soil at Uranium-Radium Site No. 2 [57]. A dose rate of 0.3 mGy day^{-1} and total dose between 1 and 10 mGy gave mutation rates significantly higher than the spontaneous rate in barley plants exposed to internal irradiation from ^{90}Sr [78]. Apparently, exposure to both internal and external gamma radiation results in increased frequencies of both fragments and vagrant chromosomes, in suppressed DNA synthesis, and in homologous recombination, whereas exposure to only external gamma radiation resulted in an increased frequency of bridges [57].

Seeds of wild vetch collected at Uranium–Radium Site 2 had low weight per 1,000 seeds and when germinated in the control field, more than half the plants did not survive. Those that survived had lower productivity than reference plants [57]. In wild vetch seedlings, a significant increase in cytogenetic disturbances was observed in meristem root tip cells, sterile buds (4–6%), and partial sterility (11–23%), including polyploidy in exposed plants compared to <2 and 3–12% in control plants. However, interpretation of these findings is complicated by the fact that the Uranium–Radium site had nutrient deficiency (sand-gravel soil), high concentrations of natural radionuclides, toxic metals, chlorides, and sulfates [57].

Field Irradiator Studies Studies on the effects of external radiation from a radiation source on the surrounding environment are not directly analogous to the potential effects that may be caused by chronic releases to the environment. This is because radionuclides are not incorporated into the organisms and alpha radiation may be more harmful in an organism than gamma radiation. Also, for larger plants, such as mature trees, radiation exposure may be much greater on one side of the tree than the other, which is largely shielded. Field irradiation

studies allow one to relate absorbed doses in vegetation to observed effects. However, it is hard to draw conclusions from these studies because of differences in the quality, intensity, and duration of radiation in each of the field irradiator studies (Table 10).

Irradiation of a slash pine (*Pinus ellottii*) and longleaf pine (*Pinus palustris*) forest sharply decreased the growth of long-leaf pine at 3.6 Gy for small trees, 4.5 Gy for medium sized trees, and 5.4 Gy for the largest individuals. Species diversity increased in the portion of the forest, where all trees were killed (>27 Gy) and numerous old field species (weeds) such as *Heterotheca subaxillaris* colonized the area. At high doses (>27 Gy), there was a complete change in species composition. Only minor changes in species composition occurred at 7.7 Gy [215]. Irradiation of a white oak and evergreen oak forest in southern France for 18 years resulted in little change at radiation doses of 360 mGy day^{-1} or less. The original woody plants remained in place, though visibly deformed [216].

Irradiation of the Canadian Boreal forest for 14 years from a point source resulted in the die-back of sensitive coniferous trees such as *Abies balsamea* and *Picea mariana* at dose rates greater than 48 mGy day^{-1} [69]. These conifers were replaced by more resistant species, such as *Populus tremuloides* and *Salix bebbiana*, in some locations. Effects on canopy cover could not be detected at dose rates of <2.4 mGy day^{-1}. Dose rates greater than 24 mGy day^{-1} suppressed growth of needles and branches and decreased the number of lateral and terminal buds. Buds were often killed at dose rates greater than 48 mGy day^{-1}. Germination of jack pine seeds was sensitive to radiation with deleterious effects observed at 26 mGy day^{-1} [217]. Dugle and Mayoh [218] reported on the response of 56 species of shrubs to radiation. LC_{50}s at the end of the sixth year ranged from 48 to >1488 mGy day^{-1}. The estimated no effect value of 2.4 mGy day^{-1} for mortality in this study was used by EC and HC [39] in their assessment of the potential effects of radiation released by Canadian nuclear facilities on plants.

Aquatic Plants

There are few studies on the effects of radiation on aquatic plants (macrophytes and algae), particularly macrophytes. The lowest dose causing sublethal effects on aquatic plants is 0.07–0.12 mGy day^{-1}, which caused a loss of synchrony in growth of *Chlorella pyrenoidosa* cultures [195]. Circumstantial evidence that low radiation doses may impede algal development also comes from analysis of algae species invading

Ionizing Radiation on Nonhuman Biota, Effects of Low Levels of. Table 10 Field irradiator studies

Project	Location	References
Enterprise radiation forest	Enterprise, Wisconsin, USA	Rudoph [204], Zavikoewski [205]
Mediterranean forest	Cadarache, France	Fabries et al. [206]
Brookhaven oak-pine forest	Upton, New York, USA	Woodwell [207]
Puerto Rico radiation forest (Montane tropical rain forest)	El Verde, Puerto Rico	Odum and Pigeon [208]
Mojave desert	Rock Valley, Nevada, USA	French et al. [113]
Short-grass Prairie	Nunn, Colorado, USA	Fraley and Whicker [209]
Savannah river irradiations (old field)	Aiken, South Carolina, USA	Miller [210], McCormick and Golley [211]
Field Irradiator Gamma (FIG) (boreal forest)	Pinawa, Manitoba, Canada	Amiro [69], Guthrie and Dugle [212]
Zoological Environment Under Stress (ZEUS) (meadows surround by boreal forest)	Pinawa, Manitoba, Canada	Turner and Iverson [213], Mihok [114]
USSR pine and birch forest	South Urals	Alexakhin et al. [214]

a ^{137}Cs-contaminated pond and a control pond [189]. The control pond developed a thick bloom of algae, but surface algal growth was sparse in the contaminated pond. Green algae were dominant in the contaminated pond, whereas blue-green algae were dominant in the control pond. Whether this was a radiation effect or a chance event is not clear. However, radiation appeared to have affected the succession of algal species in the artificial pond. Properly designed studies employing replication should be performed to confirm these observations.

Algae tend to be more resistant to radiation than higher plants when exposed to acute doses. This may be because of the minute size of their chromosomes and polyploid condition in many algal species. *Prorocentrum* and *Oedogonium* have relatively large chromosomes of about the same size as common in higher plants and are more sensitive algal genera; a lethal dose for *Prorocentrum* is 20 Gy [219] and, for *Oedogonium*, about 6.7 Gy [4].

In general, chronic exposure to radiation from routine releases from nuclear facilities is unlikely to have an effect on algae because of the low dose rates involved and the rapid turnover time for algae. A radiation dose rate of 20.9 mGy·y^{-1} increased the growth rate of *Synechococcuslividus* in the laboratory [220]. Whether low levels of radiation exposure stimulate algal production leading to more eutrophic conditions in the environment has not been documented. A threshold effect value of 8 mGy day^{-1} was suggested by Fesenko et al. [58].

Radiation Effects Synthesis

International guidance on radiation levels protective of the population of 1 mGy day^{-1} for the maximally exposed individuals of terrestrial animals, and 10 mGy day^{-1} for the maximally exposed individuals of aquatic organism is inadequate. The concept that it is okay to impact the local population as long as a nearby healthy population remains intact to repopulate the area through immigration is not in the spirit of sustainable development and is in noncompliance with national environmental regulations supporting environmental protection and pollution prevention. Although, the guidance is in all likelihood protective of populations over a relatively large area, since most individuals would be exposed to a much lower radiation dose, the guidance is difficult to administer. Large sample sizes are required to demonstrate that the maximally exposed individuals are within the criteria, such that sampling can have a major impact on populations. Further, when impacts are seen at the population level, it is too late, the impact is there. It is preferable to monitor effects at the individual level. If the individual is protected, then the population will also be protected. Radiation effects are observed at levels substantially lower than the guidance levels (section "Effects of Radiation on Animals" and Table 1), especially for aquatic organisms, and there are many taxa for which no radiation effects information is available, particularly long-lived, slow reproducing species that are theoretically more sensitive to radiation. It is difficult to rationalize the use of guidance levels that represent more severely polluted environments, i.e., those found following a major nuclear accident or in/near uranium mine/mill waste management areas and effluent discharge areas.

Under normal operating conditions and effluent releases, radionuclide members of the uranium and thorium decay chains potentially represent the most risk to the environment. This is because most members are alpha emitters and give the greatest radiation dose to organisms when ingested, several radionuclides are generally present (the decay chain) and the sum of these individual radionuclide doses gives the total dose (the total dose is additive). Finally, the effect of an internal radiation source appears to be greater than the effect from an external source. These radionuclides are most closely associated with the first stage of the nuclear fuel cycle; the mining and milling of uranium ore. Other radionuclides such as ^{131}I, ^{90}Sr, and ^{137}Cs are of most concern following a nuclear reactor accident.

Research is required into the chronic effects of internally ingested radionuclides, particularly alpha emitters, on biota. Several studies showing effects at low exposure levels warrant follow-up. For example, the effects of tritium on the goose barnacle [179] and on the gastropod *Pila luzonica* [186], the radiosensitivity of Tardigrada [58, 189], and the effect of radiation on the growth of the green alga *Chlorella pyrenoidosa* [195]. Further study on the effect of radiation on long-lived, slow-growing organisms is essential as is the influence of radiation in contaminant mixtures. The results of radiation exposure studies should always

report the exposure levels (radiation dose), duration of exposure, effect level, and dose–response relationship so that data may be standardized to provide estimated-no-effect-levels or low-effect levels.

Future Directions

Radiation like most contaminants shows a spectrum of effects from being harmless at very low doses to lethal at high doses. Radiation differs from other contaminants in that external exposure to elevated concentrations of radioactivity can be harmful in the case of gamma radiation. Most studies have focused on the effect of radiation at high doses where major effects are documented and expected. Relatively few studies have investigated the effects of low doses of radiation representative of chronic releases. Of these studies, most used gamma radiation as an external source. Studies are required on the effects of low doses of internal radiation, particularly alpha, on a variety of organisms. Most studies have investigated the effects of radiation on small, rapidly reproducing organisms such as mice and rats, or in the aquatic environment, small-bodied fish and zooplankton. This is because these organisms are easy to study in the laboratory. However, they are also much more resistant to radiation at the population level than long-lived, slow-reproducing species.

Since radiation is seldom the only contaminant present in the environment, it is important to assess the role of radiation in inducing effects when present in contaminant mixtures.

Finally, the issue of voluntary and regulatory levels protective of the environment needs to be addressed, although it is recognized that it is up to individual governments to set criteria. Presently, protection is at the population level, but there are suggestions that protection should be at the individual level similar to the trend for other contaminants, although even here some jurisdictions protect the individual and others the population. Environmental thresholds are usually based on "no effect" to populations typically defined as the EC_{10} or less, or low effects (EC_{15-25}). It is difficult to justify an international guidance level of 10 mGy day^{-1} for protection of the aquatic environment when a radiation dose level of this magnitude is essentially restricted to major accidents (Kyshtym and Chernobyl) and waste management areas, e.g., uranium tailing ponds. Plants are clearly more tolerant than many animals to radiation. Therefore, international guidance on radiation dose rates protective of plants (10 mGy day^{-1}) is not protective of many inhabitants of plant communities. Garnier-Laplace et al. [221] have suggested a generic screening value of 0.24 mGy day^{-1}, and others [39, 58, 126, 222] have suggested various values for different taxonomic groups. Clearly, better direction is required as to what radiation dose rate is truly protective of the environment.

Bibliography

Primary Literature

1. World Nuclear Association (2009) The nuclear renaissance. http://www.world-nuclear.org/info/inf104.html. Accessed 10 Nov 2010
2. IAEA (International Atomic Energy Agency) (2009) Nuclear technology review 2009. http://WWW.iaea-org/cgl-bin/db.page.pllpris.opris.oprconst.htm. Accessed 10 Nov 2010
3. Jones S, Copplestone D, Zinger-Gize I (2003) A method of impact assessment for ionising radiation on wildlife. In: Third international symposium on protection of the environment from ionizing radiation (SPIER 3). International Atomic Energy Agency (IAEA), Vienna, Australia, pp 248–256
4. Godward MBE (1962) Invisible radiations. In: Lewin AR (ed) Physiology and biochemistry of algae. Academic, New York, pp 551–566
5. Sacher GA, Grha D (1964) Survival of mice under duration-of-life exposure to gamma rays. 1. The dosage-survival relation and the lethality function. J Nat Cancer Inst 32:277–321
6. UNSCEAR (United Nations Scientific Committee on the Effects of Atomic Radiation) (1977) Sources and effects of ionizing radiation. Report to the general assembly with annexes UN sales publication NOE.77.IX.I New York
7. Fry RJM (1996) Effects of low doses of radiation. Health Phys 70:823–827
8. NRCP (National Council on Radiation Protection and Measurements) (1980) Influence of dose and its distribution in time on dose-response relationships for low-LET radiations. Recommendations of the National Council on Radiation Protection and Measurements, Washington, DC, RA121.R2/U58/No.64/1980
9. Gong JK, Glomski CA, Bruce AK (1983) The effects of low dose (less than 1 rad) X-rays on the eythropoietic marrow. Cell Biophysics 5:143–162
10. Belova NV, Verigin BV, Yemel'yanova NG, Makeyeva AP, Ryabov IN (1994) Radiobiological analysis of silver carp (*Hypophthalmichthys molitrix*) from the cooling pond of the Chernobyl Nuclear Power Station in the post-disaster period. Reproductive system of fish exposed to radioactive contamination J Ichthyol 34:16–38

11. Trosko JE (1996) Role of low-level ionizing radiation in multi-step carcinogenic process. Health Phys 70:812–822
12. Bond VP, Feinendegen LE, Booz J (1988) What is a 'low dose' of radiation? Int J Radiat Bio 53:1–12
13. Ulsh BA, Miller SM, Mallory FF, Mitchel REJ, Morrison DP, Boreham DR (2004) Cytogenetic dose-response and adaptive response in cells of ungulate species exposed to ionizing radiation. J Environ Radioact 74:73–81
14. Goodhead DT (1988) Spatial and temporal distribution of energy. Health Phy 55:231–240
15. Thompson JF, Grahn D (1989) Life shortening in mice exposed to fission neutrons and γ rays VIII. Exposures to continuous γ radiation. Rad Res 118:151–160
16. UNSCEAR (United Nations Scientific Committee on the Effects of Atomic Radiation) (1982) Ionizing radiation: sources and biological effects. Report to the General Assembly with annexes. United Nations, New York
17. Wolff S (1996) Aspects of the adoptive response to very low doses of radiation and other agents. Mutat Res 358:135–142
18. Cai L, Cherian MG (1996) Adaptive response to ionizing radiation-induced chromosome aberrations in rabbit lymphocytes: effect of pre-exposure to zinc, and copper salts. Mutation Res 369:233–241
19. Polikarpov GG (1998) Biological aspect of radioecology: objective and perspective. In: Imaba J, Nakamura Y (eds) International workshop on comparative evaluation of health effects of environmental toxicants derived from advanced technologies, Chiba, Japan, 28–31 Jan 1998. Kodansha Scientific, Tokyo
20. Kryshev II (1992) Radioecological consequences of the Chernobyl accident. Nuclear Society International, Moscow, p 142
21. Blöcher D (1988) DNA double strand break repair determines the RBE of α-particles. Int J Radiat Biol 54:761–771
22. Goodhead DT, Nikjoo H (1989) Track structure analysis of ultrasoft X-rays compared to high- and low-LET radiations. Int J Radiat Biol 55:513–529
23. Goodhead DT, Thacker J, Cox R (1993) Effects of radiation of different qualities on cells: molecular mechanisms of damage and repair. Int J Radiat Biol 63:543–556
24. Prise KM (1994) Use of radiation quality as a probe for DNA lesion complexity. Int J Radiat Biol 65:43–48
25. Zhou H, Suzuki M, Randers-Pehrson G, Vannais D, Chen G, Trosko JE, Waldren CA, Hei TK (2001) Radiation risk to low fluences of α particles may be greater than we thought. Pro Natl Acad Sci USA 98:14410–14415
26. ICRP (International Commission on Radiological Protection) (1991) 1990 recommendations of the International Commission on Radiological Protection. Ann ICRP 21(1–3):1–201 (ICRP Publication 60)
27. Simmons JA, Watt DE (1999) Radiation protection dosimetry: a radical reappraisal. Medical Physics Publishing, Madison
28. Health Protection Agency (2007) Review of risks from tritium. Report of the independent Advisory Group on ionizing radiation. Chilton, Doc HPA, RCE-4, pp 1–90
29. Barendsen GW (1993) Sublethal damage and DNA double strand breaks have similar RBE–LET relationships: evidence and implications. Int J Radiat Biol 63:325–330
30. Barendsen GW (1994) RBE–LET relationships for different types of lethal radiation damage in mammalian cells: comparison with DNA double strand breaks and an interpretation of differences in radiosensitivity. Int J Radiat Biol 66: 433–436
31. Amiro BD (1997) Radiological dose conversion factors for generic non-human biota used for screening potential ecological impacts. J Environ Radioact 35:37–51
32. Blaylock BG, Frank ML, O'Neal BR (1993) Methodology for estimating radiation dose rates to freshwater biota exposed to radionuclides in the environment. Oak Ridge National Laboratory, Oak Ridge, (ES/ER/TM-78)
33. Copplestone D, Bielby S, Jones SR, Patton D, Daniel P, Gize I (2001) Impact assessment of ionizing radiation on wildlife. Environment Agency, UK, p 222 (R&D Publication 128)
34. UNSCEAR (United Nations Scientific Committee on the Effects of Atomic Radiation) (1996) Effects of radiation on the environment, vol V92–53957. United Nations, New York
35. Kocher DC, Trabalka JR (2000) On the application of a radiation weighting factor for alpha particles in protection of non-human biota. Health Phys 79:407–411
36. ICRP (International Commission on Radiological Protection) (1989) RBE for deterministic effects. ICRP Publication 58. Ann ICRP 20:1–57
37. Pentreath RJ, Woodhead DS (2000) A system for environmental protection: reference dose models for fauna and flora. 10th international congress of the international radiation protection association (IRPA-10), Hiroshima, Japan, 14–19 May 2000, p 6. (P-4b-228). http://www.irpa.net/irpa10/cdrom/00662.pdf. Accessed 10 July 2010
38. Pentreath RJ, Woodhead DS (2001) A system for protecting the environment from ionizing radiation: selecting reference fauna and flora, and the possible dose models and environmental geometries that could be applied to them. Sci Total Environ 277:33–43
39. Environmental Canada, Health Canada (EC and HC 2003) Canadian Environmental Protection Act 1999. Priority substances list assessment report releases of radionuclides from nuclear facilities (Impact on non-human biota). Final Report May 2003
40. NCRP (National Council on Radiation Protection and Measurement) (1990) The relative biological effectiveness of radiations of different quality. Recommendations of the National Council on Radiation Protection and Measurement. Washington, DC, p 218. (NCRP Report No. 104)
41. Chambers DB, Osborne RV, Garva AL (2006) Choosing an alpha radiation weighting factor for doses to non-human biota. J Environ Radioact 87:1–14
42. IAEA (Intenational Atomic Energy Agency) (1996) International basic safety standards for protection against ionizing radiation and for the safety of radiation sources. Safety Series No. 115, Schedule II, Public Exposure, Dose Limits pp 92–93

43. Shilova SA, Shatunovskii MI (2005) Ecophysiological indicators of the state of animal populations exposed to damaging factors. Russian J Ecology 36:27–32
44. Jackson SP (2002) Sensing and repairing DNA double strand breaks. Carcinogenesis 23:687–696
45. Goodsell DS (2005) The molecular perspective: double-stranded DNA breaks. Oncologist 10:361–362
46. Ulsh B, Hinton TG, Congdon JD, Dugan LC, Whicker FW, Bedford JS (2003) Environmental biodosimetry: a biologically relevant tool for ecological risk assessment and biomonitoring. J Environ Radioact 66:121–139
47. Xu A, Smilenov LB, He P, Masumura K-I, Nohmi T, Yu Z, Hei TK (2007) New insights into intrachromosomal deletions induced by chrysotile in the gpt delta transgenic mutation assay. Environ Health Perspect 115:87–92
48. Polikarpov GG (1998) Conceptual model of responses of organisms, populations and ecosystems to all possible dose rates of ionizing radiation in the environment. Rad Prot Dosimetry 75:181–185
49. Tsytsugina VG (1998) An indicator of radiation effects in natural populations of aquatic organisms. Rad Prot Dosim 75:171–173
50. Strand P, Larsson C-M (2001) Delivering a framework for the protection of the environment from ionizing radiation. In: Bréchignac F, Howard BJ (eds) Radioactive pollutants, impact on the environment. EDP Sciences, les Veix France
51. Copplestone DC, Howard BJ, Bréchignac F (2004) The ecological relevance of current approaches for environmental protection from exposure to ionizing radiation. J Environ Radioact 74:31–41
52. IAEA (International Atomic Energy Agency) (1976) Effects of ionizing radiation on aquatic organisms and ecosystems, Vienna, p 131 (Technical Reports Series No. 172)
53. IAEA (International Atomic Energy Agency) (1992) Effects of ionizing radiation on plants and animals at levels implied by current radiation protection standards, Vienna (Technical Reports Series No. 332)
54. NCRP (National Council on Radiation Protection and Measurement) (1991) Effects of ionizing radiation on aquatic organisms. Washington, DC (NCRP Report No. 109)
55. Sazykina T, Kryshev II (2006) Radiation effects in wild terrestrial vertebrates – the EPIC collection. J Environ Radioact 88:11–48
56. Krivolutsky D, Turcaninova V, Mikhaltsova Z (1982) Earthworms as bioindicators of radioactive soil pollution. Pedobiologia 23:263–265
57. Geras'kin SA, Evseena TI, Belykh ES, Majstrenko TA, Michalik B, Taskaev AI (2007) Effects on non-human species inhabiting areas with enhanced level of natural radioactivity in the north of Russia: a review. J Environ Radioact 94:151–182
58. Fesenko SV, Alexakhin RM, Geras'kin SA, Sanzharova NI, Spirin YV, Spiridonov SI, Gontarenko IA, Strand P (2005) Comparative radiation impact on biota and man in the area affected by the accident at the Chernobyl nuclear power plant. J Environ Radioact 80:1–25
59. Chandorkar KR, Dengler NG (1987) Effect of low level continuous gamma irradiation on vascular cambium activity in Scotch pine *Pinus sylvestris* L. Environ Exp Bot 27:165–175
60. Sazykina TG, Kryshev AI (2003) EPIC database on the effects of chronic radiation in fish: Russian/FSU data. J Environ Radioact 68:65–87
61. Hagger JA, Atienzar FA, Jha AN (2005) Genotoxic, cytotoxic, developmental and survival effects of tritiated water in the early stages of the marine mollusc, *Mytilus edulis*. Aquat Toxicol 74:205–217
62. Higley KA, Domotor SL, Antonio EJ (2003) A probabilistic approach to obtaining limiting estimates of radionuclide concentration in biota. J Environ Radioact 66:75–87
63. Barnthouse LW (1997) Extrapolating ecological risks of ionizing radiation from individuals to populations to ecosystems. In: Symposium on radiological impacts from nuclear facilities on non-human species, Ottawa, 1–2 Dec 1996. Canadian Nuclear Society, Toronto, pp 53–60
64. Copplestone D, Zinger-Gize I, Woodhead DS (2002) The FASSET Radiation Effects Database: a demonstration SPIER: 3rd international symposium on the protection of the environment from ionising radiation, Darwin, 2002
65. IAEA (International Atomic Energy Agency) (1991) The international Chernobyl project technical report. International Atomic Energy Agency, Vienna, p 640
66. Barnthouse LW (1995) Effects of ionizing radiation on terrestrial palants and animals: a workshop report, ORNL/TM-1341. Oak Ridge National Laboratory, Oak Ridge
67. Kryshev II, Romanov GN, Chumichevs VB, Sazykina TG, Isaeva LN, Ivanitskaya V (1998) Radioecological consequences of radioactive discharges into the Techa River on the southern Urals. J Environ Radioact 38:195–209
68. Smith J (2005) Effects of ionizing radiation on biota: do we need more regulation? J Environ Radioact 82:105–122
69. Amiro BD (1995) Effects of ionizing radiation on the boreal forest. Prepared for the Atomic Energy Control Board, Ottawa, Ontario by AECL Research, Whiteshell Laboratories, Manitoba, AECB INFO-0581
70. Delistraty D (2008) Radioprotection of non-human biota. J Environ Radioact 99:1863–1869
71. Stone R (2002) Radioecology's coming of age – or its last gasp? Science 297:1800–1801
72. Oughton D (2003) Protection of the environment form ionizing radiation: ethical issues. J Environ Radioact 66:3–18
73. Blaylock BG (1965) Chromosomal aberrations in a natural population of *Chironomus tentans* exposed to chronic low-level radiation. Evolution 19:421–429
74. Lance VA, Horn TR, Elsey RM, de Peyster A (2006) Chronic incidental lead ingestion in a group of captive-reared alligators (*A. mississippiensis*): possible contribution to reproductive failure. Comparative Biochem Phys Part C 142:30–35
75. Whicker FW, Schultz V (1982) Radioecology: Nuclear energy and the environment, vols I and II. CRC Press, Boca Raton
76. Harrison F (1997) Radiobiological endpoints relevant to ecological risk assessment. In: Symposium on radiological

impacts from nuclear facilities on non-human species, Ottawa, 1–2 Dec 1996. Canadian Nuclear Society, Toronto, pp 39–52
77. Harrison FL, Anderson SL (1994) Effects of acute irradiation on reproductive success of the polychaete worm, *Neanthes arenaceodentata*. Radiat Res 137:59–66
78. Real A, Sundell-Bergman S, Knowles JF, Woodhead DS, Zinger I (2004) Effects of ionizing radiation exposure on plants, fish and mammals: relevant data for environmental radiation protection. J Radiol Prot 24:A123–A137
79. Polikarpov GG, Trytsugina VG (1996) Radiation effects in the Chernobyl and Kyshtym aquatic ecosystems. In: Luykx FF, Frissel MJ (eds) Radioecology and the restoration of radioactive-contaminated sites. Kluwer Academic Publishers, Netherlands, pp 269–277
80. Harrison FL, Knezovich JP (2001) Effects of radiation on aquatic and terrestrial organisms. In: Van der Stricht E, Kirchmann R (eds) Radioecology: radioactivity and ecosystems. International Union of Radioecology, Liege, pp 317–375
81. Rose KSB (1992) Lower limits of radiosensitivity in organisms, excluding man. J Environ Radioact 15:113–133
82. Dobson RL (1982) Low-exposure tritium radiotoxicity in mammals. In: Matsudaira H, Yamaguchi T, Nakazawa T, Saito C (eds) Proceedings of a workshop on tritium radiobiology and health physics, Chiba City, 27–28 Oct 1981. Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, pp 120–134, NIRS-M-41
83. Searle AG, Beechey CV, Green D, Humphreys ER (1976) Cytogenetic effects of protracted exposures to alpha-particles from plutonium-239 and to gamma rays from cobalt-60 compared in male mice. Mutat Res 41:297–310
84. Gilbert ES, Cross FT, Dagle GE (1996) Analysis of lung tumor risks in rats exposed to radon. Radiat Res 145:350–360
85. IAEA (International Atomic Energy Agency) (1979) Methodology for assessing impacts of radioactivity on aquatic ecosystems, Vienna, p 416. (Technical Reports Series No. 190)
86. Kligerman AD (1979) Cytogenetic methods for the detection of radiation-induced chromosome damage in aquatic organisms. In: Methodology for assessing impacts on radioactivity on aquatic ecosystems. International Atomic Energy Agency, Vienna. pp. 39–367 (Technical Reports Series No. 190)
87. IAEA (International Atomic Energy Agency) (1988) Assessing the impact of deep sea disposal of low level radioactive waste on living marine resources, Vienna, p 127. (Technical Reports Series No. 288)
88. Thomas P, Sheard JW, Swanson S (1992) Transfer of ^{210}Po and ^{210}Pb through the lichen–caribou–wolf food chain of northern Canada. Health Phys 66:666–677
89. Macdonald CR, Laverock MJ (1998) Radiation exposure and dose to small mammals in radon-rich soils. Arch Environ Contam Toxicol 35:109–120
90. Polikarpov GG (1979) Effects of ionizing radiation upon aquatic organisms. In: Methodology for assessing impacts of radioactivity on aquatic ecosystems. Report of an advisory group meeting IAEA, Vienna, 21–22 Nov 1979, STI/DOC/190.1, pp 173–194
91. Gopal-Ayengar AR, Nayar GG, George KP, Mistry KB (1970) Biological effects of high background radioactivity: studies on plants growing in the monazite-bearing areas of Kerala Coast and adjoining regions. Indian J Exp Biol 8:313–318
92. Meehan KA, Truter EJ, Slannert JP, Parker MI (2004) Evaluation of DNA damage in a population of bats (*Chiroptera*) residing in an abandoned maonazite mine. Mutat Res 557:183–190
93. Léonard A, Delpoux M, Decat G, Léonard ED (1979) Natural radioactivity in southwest France and its possible genetic consequences for mammals. Rad Res 77:170–181
94. Ghiassi-Nejad M, Zakeri F, Assaei RGh, Kariminia A (2004) Long-term immune and cytogenetic effects of high level natural radiation on Ramsar inhabitants in Iran. J Environ Radioact 73:107–116
95. Dunaway PB, Lewis LL, Story JD, Payne JA, Inglis JM (1969) Radiation effects in the Soricidae, Cricetidae, and Muridae. In: Nelson DJ, Evans FC (eds) Symposium on radioecology proceedings the second national symposium, United States Atomic Energy Commission Report CONF-670503. Michigan, Ann Arbor, pp 173–184
96. Bréchignac F (2001) Impact of radioactivity on the environment: problems, states of current knowledge and approaches for identification of radioprotection criteria. Radioprotection 36:51–535
97. IAEA (International Atomic Energy Agency (2001) Present and future environmental impact on the Chernobyl accident. IAEA-TECDOC-1240
98. Ohzu E (1965) Effects of low-dose X-irradiation on early mouse embryos. Rad Res 26:107–113
99. Brown SO, Krise GM, Pace HB, de Boer J (1964) Effects of continuous radiation on the reproductive capacity and fertility of albino rat and mouse. In: Carlson W (ed) Effects of ionizing radiation on the reproductive system. Pergamon Press, New York, pp 103–110 [cited in Rose, 1992]
100. Buech RR (1977) Small mammals in a gamma-irradiated northern forest community. In: Zavitkovski J (ed) The enterprise, Wisconsin, radiation forest: radioecological studies. US Energy Research and Development Administration, Washington, DC, pp 167–180, (ERDA Report TID-26113-P2
101. Dunaway PB, Kaye SV (1963) Effects of ionizing radiation on mammal populations on the White Oak Lake bed. In: Schultz V, Klement AW Jr (eds) Radioecology. Division of Technical Information Extension, U.S. Atomic Energy Commission, Oak Ridge, p 333
102. Cristaldi M, Ieradi LA, Mascanzoni D, Mattei T (1991) Environmental impact of the Chernobyl accident: mutagenesis in bank voles from Sweden. Int J Radiat Biol 59:31–40
103. Røed KH, Eikelmann IMH, Jacobsen M, Pedersen Ø (1991) Chromosome aberrations in Norwegian reindeer calves exposed to fallout from the Chernobyl accident. Hereditas 115:201–206
104. Rodgers BE, Baker RJ (2000) Frequencies of micronuclei in bank voles from zones of high radiation at Chernobyl, Ukraine. Environ Toxicol Chem 19:1644–1648

105. Mothersill C, Seymour C (1997) Lethal mutations and genomic instability. Int J Radiat Biol 71:751–758
106. Carnes BA, Gavrilova N, Grahn D (2002) Pathology effects of radiation doses below those causing increased mortality. Radiation Res 158:187–194
107. Searle AG (1964) Effects of low-level irradiation on fitness and skeletal variation in an inbred mouse strain. Genetics 50:1159–1178
108. Dobson RL, Kwan TC (1976) The relative biological effectiveness of tritium radiation measured in mouse oocytes increases at low exposure levels. Radiat Res 66:615–625
109. Sazykina TG (2005) A system of dose-effects relationships for the northern wildlife: radiation protection criteria. Radioprotection 40(Suppl 1):S889–S892
110. Gambino JJ, Lindberg RG (1964) Response of the pocket mouse to ionizing radiation. Rad Res 22:586–597
111. Smith DE (1960) The effects of ionizing radiation in hibernation. Bull Museum of Comp Zool 124:493–506
112. Turner FB (1975) Effects of continuous irradiation on animal populations. Adv Radiat Biol 5:83–144
113. French NR, Maza BG, Hill HO, Aschwanden AP, Kaaz HW (1974) A population study of irradiated desert rodents. Ecol Monogr 44:45–72
114. Mihok S (2004) Chronic exposure to gamma radiation of wild populations of meadow voles (*Microtus pennsylvanicus*). J Environ Radioact 7:233–266
115. Kudyasheva AG, Shishkina LN, Shevchenko OG, Bashlykova LA, Zagorskaya NG (2007) Biological consequences of increased natural radiation background for *Microtus oeconomus* Pall. populations. J Environ Radioact 97:30–41
116. Muramatsu S, Sugahara T, Okazawa Y (1963) Genetic effects of chronic low-dose irradiation on mice. Int J Radiat Biol 6:49–59
117. Russell WL (1963) The effect of radiation dose rate and fractionation on mutation in mice. In: Sobels FH (ed) Repair for genetic radiation damage and differential radiosensitivity in germ cells. Pergamon Press, London, pp 205–217
118. Spirin DA (1996) Effects of ionizing radiation on organisms of terrestrial ecosystems in the East Urals radioactive track territory. In: Luykx FF, Frissel MJ (eds) Radioecology and the restoration of radioactive-contaminated sites. Kluwer, Dordrecht, pp 235–244
119. Jackson D, Copplestone D, Stone DM, Smith GM (2005) Terrestrial invertebrate population studies in the Chernobyl exclusion zone Ukraine. Radioprotection 40(Suppl 1):S857–S863
120. Tikhomirov FA, Shcheglov AI (1994) Main investigation results on the forest radioecology in the Kyshtym and Chernobyl accident zones. Sci Tot Environ 157:45–57
121. Kal'chenko VA, Fedotov IS (2001) Genetic effects of acute and chronic ionizing irradiation on *Pinus sylvestris* L. inhabiting the Chernobyl meltdown area. Russian J Genetics 37:341–350
122. Ryabokon NI, Smolich II, Kudryashov VP, Goncharova RI (2005) Long-term development of the radionuclide exposure of murine rodent populations in Belarus after the Chernobyl accident. Radiat Environ Biophys 44:169–181
123. Oleksyk TK, Novak JM, Purdue JR, Gashchak SP, Smith MH (2004) High levels of fluctuating asymmetry in populations of *Apodemus flavicollis* from the most contaminated areas in Chornobyl. J Environ Radioact 73:1–20
124. Chesser RK, Rodgers BE, Wickliffe JK, Gaschak S, Chizhevsky I, Carleton JP, Baker RJ (2001) Accumulation of ^{137}Cesium and ^{90}Strontoium from abiotic and biotic sources in rodents at Chornobyl, Ukraine. Environ Toxicol Chem 20:1927–1935
125. Baker RJ, Hamilton MJ, Van Den Bussche RA, Wiggins LE, Sugg DW, Smith MH, Lomakin MD, Gaschak SP, Bundova EG, Rudenskaya GA, Chesser RK (1996) Small mammals from the most radioactive sites near the Chornobyl Nuclear Power Plant. J Mammalogy 77:155–170
126. Bird GA, Thompson PA, Macdonald CR, Sheppard SC (2002) Ecological risk assessment approach for the regulatory assessment on the effects of radionuclides released from nuclear facilities. In: Third international symposium on the protection of the environment from ionising radiation. Darwin, Australia, 22–26 July, 2002
127. Brisbin IL Jr (1991) Avian radioecology. Curr Ornithol 8:60–140
128. Mellinger PJ, Schultz V (1975) Ionizing radiation and wild birds: a review. Crit Rev Environ Contam 5:397–421
129. Zach R, Mayoh KR (1982) Breeding biology of tree swallows and house wrens in a gradient of gamma radiation. Ecology 63:1720–1728
130. Buech RR (1977) Observations of nesting avifauna under gamma radiation exposure. In: Zavitkovski J (ed) The enterprise, Wisconsin, radiation forest: radioecological studies. U.S. Energy Research and Development Administration, Washington, DC, pp 181–184 (ERDA Report TID-26113-P2)
131. Mraz FR, Woody MC (1972) Effect of continuous gamma irradiation of chick embryos upon their gonadal development. Radiat Res 50:418–425
132. Zach R, Hawkins JL, Sheppard SC (1993) Effects of ionizing radiation on breeding swallows at current radiation protection standards. Environ Toxicol Chem 12:779–786
133. Krivolutsky DA (1987) Radiation ecology of soil animals. Biol Fertil Soils 3:51–55
134. Ellegren H, Lindgren G, Primmer CR, Møller AP (1997) Fitness loss and germline mutations in barn swallows breeding in Chernobyl. Nature 389:593–596
135. Ewing LL, Macdonald CR, Amiro BD (1996) Radionuclide transfer and radiosensitivity in amphibians and reptiles. Atomic Energy of Canada Limited, Pinawa, Manitoba, p 34. (AECL Technical Report TR-731, COG-96-06)
136. Sparrow AH, Nauman CH, Donnelly GM, Willis DL, Baker DG (1970) Radiosensitivities of selected amphibians in relation to their nuclear volume and chromosome volumes. Radiat Res 42:353–371
137. Woodhead AD, Pond V (1987) Effects of radiation exposure on fishes. In: Capuzzo JM, Kester DR (eds) Oceanic processes in marine pollution, vol 1, Biological processes and wastes in the ocean. Robert E Krieger Publishing, Malabar, pp 157–180

138. Conger AD, Clinton JH (1973) Nuclear volumes, DNA contents, and radiosensitivity in whole-body irradiated amphibians. Radiat Res 54:69–101
139. Panter HC (1986) Variations in radiosensitivity during the development of the frog *Limnodynastes tasmaniensis*. J Exp Zool 238:193–199
140. Panter HC, Chapman JE, Williams AR (1987) Effect of radiation and trophic state on oxygen consumption of tadpoles of the frog *Limnodynastes tasmaniensis*. Comp Biochem Physiol 88A:373–375
141. Stark K, Avila R, Wallberg P (2004) Estimation of radiation dose from ^{137}Cs to frogs in a wetland ecosystem. J Environ Radioact 75:1–14
142. Blaylock BG, Trabalka JR (1978) Evaluating the effects of ionizing radiation on aquatic organisms. Adv Radiat Biol 7:103–152
143. Egami N, Ijiri K-I (1979) Effects of irradiation on germ cells and embryonic development in teleosts. Int Rev Cytol 59:195–248
144. Woodhead DS (1984) Contamination due to radioactive materials. In: Kinne O (ed) Marine ecology: a comprehensive, integrated treatise on life in oceans and coastal waters, vol 5, Ocean management. Pt. 3. Pollution and protection of the seas — radioactive materials, heavy metals and oil. John Wiley, Chichester, pp 1111–1287
145. Koulikov AO (1996) Physiological and ecological factors influencing the radiocaesium of fish species from Kiev reservoir. Sci Tot Environ 177:125–135
146. Anderson SL, Harrison FL (1986) Effects of radiation on aquatic organisms and radiobiological methodologies for effects assessment. U.S. Environmental Protection Agency, Washington, DC (Report No. 520/1-85-016)
147. Angelovic JW, White JC Jr, Davis EM (1969) Interactions of ionizing radiation and temperature on the estuarine fish, *Fundulus heteroclitus*. In: Nelson DJ, Evans FC (eds) Proceedings of the 2nd national symposium on radioecology. U.S. Atomic Energy Commission, Washington, DC, pp 131–141, CONF-670503
148. Bonham K, Welander AD (1963) Increase in radioresistance of fish to lethal doses with advancing embryonic development. In: Schultz V, Klement AW (eds) Proceedings of the 1st National Symposium on Radioecology. Reinhold Publishing Corp, New York, pp 353–358
149. Welander AD (1954) Some effects of X-irradiation of different embryonic stages of the trout (*Salmo gairdnerii*). Growth 18:227–255
150. Ward E, Beach SA, Dyson ED (1971) The effect of acute X-irradiation on the development of the plaice *Pleuronectes platessa* L. J Fish Biol 3:252–259
151. Rugh R, Clugston H (1955) Effects of various levels of X-irradiation on the gametes and early embryos of *Fundulus heteroclitus*. Biol Bull 108:318
152. Blaylock BG, Griffith NA (1971) Effect of acute beta and gamma radiation on developing embryos of carp (*Cyprinus carpio*). Radiat Res 46:99–104
153. Donaldson LR, Bonham K (1964) Effects of low-level chronic irradiation of chinook and coho salmon eggs and alevins. Trans Am Fish Soc 93:333–341
154. Bonham K, Donaldson LR (1966) Low-level chronic irradiation of salmon eggs and alevins. In: Proceedings of the symposium on disposal of radioactive wastes into sea, oceans, and surface waters. International Atomic Energy Agency, Vienna, pp 869–883
155. Hershberger WK, Bonham K, Donaldson LR (1978) Chronic exposure of chinook salmon eggs and alevins to gamma irradiation: effects on their return to freshwater as adults. Trans Am Fish Soc 107:622–631
156. Trabalka JR, Allen CP (1977) Aspects of fitness of a mosquitofish *Gambusia affinis* population exposed to chronic low-level environmental radiation. Radiat Res 70:198–211
157. Woodhead DS (1977) The effects of chronic irradiation on the breeding performance of the guppy, *Poecilia reticulata* (Osteichthyes: Teleostei). Int J Radiat Biol 32:1–22
158. Erickson RC (1973) Effects of chronic irradiation by tritiated water on *Poecilia reticulata*, the guppy. In: Nelson DJ (ed) Proceedings of the 3rd national symposium on radioecology. U.S. Atomic Energy Commission, Washington, DC, pp 1091–1099 (CONF-710501)
159. SNIFFER (Scotland & Northern Ireland Forum for Environmental Research) (2002) An investigation into the effects of chronic radiation in fish. Environment Agency, Bristol, UK. (SNIFFER R&D Technical Report P3-053/TR)
160. Makeyeva AP, Yemel'yanova NG, Velova NB, Ryabou IN (1995) Radiobiological analysis of silver carp, *Hypophthalmichthys molitrix*, from the cooling pond of the Chernobyl Nuclear Power Plant since the time of the accident. 2. Development of the reproductive system in the first generation of offspring. J Ichthyol 35:40–64
161. Makeyeva AP, Belova NV, Emer'yanova NG, Verigin BV, Ryabou IN (1996) Materials on the state of reproductive system of bighead *Aristichthys nobilis* from the cooling pond of Chernobyl Nuclear Power Station in the post-disaster period. J Ichthyol 36:181–189
162. Kryshev II, Sazykina TG (1998) Radioecological effects on aquatic organisms in the areas with high levels of radioactive contamination: environmental protection criteria. Radiat Prot Dosim 75:187–191
163. Sugg DW, Bickman JW, Brooks JA, Lomakin MD, Jagoe CH, Dallas CE, Smith MH, Baker RJ, Chesser RK (1996) DNA damage and radiocesium in channel catfish from Chernobyl. Environ Toxicol Chem 15:1057–1063
164. Ilyinskikh NN, Ilyinskikh EN, Ilyinskikh IN (1998) Micronucleated erythrocytes frequency and radiocesium bioconcentration in pikes (*Esox lucius*) caught in the Tom River near the nuclear facilities of the Siberian Chemical Complex (Tomsk-7). Mutat Res 421:197–203
165. Sazykina TG, Kryshev II (1999) Radiation protection of natural ecosystems: primary and secondary dose limits to biota. In: Proceedings of the international symposium on radioactive

waste disposal –health and the environment criteria and standards. Stockholm Environment Institute, Stockholm, pp 115–118
166. O'Brien RE, Wolfe LS (1964) Non genetic effects of radiation. In: Radiation, radioactivity and insects. Academic, New York, pp 23–54
167. Cole MM, LaBrecque GC, Burden GS (1959) Effect of gamma radiation on some insects affecting man. J Econ Entomol 52:448–450 [cited in Rose, 1992]
168. Krivolutsky DA (1980) The effect of an increased Ra content in the soil of soil animals. In: Proceedings of the VII international college of soil zoologists, Syracuse, pp 391–396
169. Hertel-Aas T, Oughton DH, Jaworska A, Bjerke H, Salbu B, Brunborg G (2007) Effects of chronic gamma irradiation on reproduction in the earthworm *Eisenia feida* (Oligochaeta). Radiat Res 168:515–526
170. Krivolutzskii DA, Pokarzhevskii AD (1992) Effects of radioactive fallout on soil animal populations in the 30 km zone of the Chernobyl atomic power-station. Sci Tot Eviron 112:69–77
171. Krivolutzskii DA, Pokarzhevskii AD, Viktorov AG (1992) Earthworm populations in soils contaminated by the Chernobyl atomic power-station accident. Soil Biol Biochem 24:1729–1731
172. Arkhipov NP, Kuchma ND, Askbrant S, Pasternak PS, Musica V V (1994) Acute and long-term effects of irradiation on pine (*Pinus sylvestris*) stands post-chernobyl. Sci Tot Environ 157:383–386
173. Tabone E, Poinsot-Balaguer N (1987) Dynamique de la matière organique dans un sol de forêt miste méditerranéenne (chênes verts-chênes blancs) soumise à une irradiation gamma chronique. Pedobiologia 30:112 [cited in IAEA, 1992]
174. Marshall JS (1962) The effects of continuous gamma radiation on the intrinsic rate of natural increase of *Daphnia pulex*. Ecology 43:598–607
175. Marshall JS (1966) Population dynamics of *Daphnia pulex* as modified by chronic radiation stress. Ecology 47:561–571
176. Gehrs CW, Trabalka JR, Bardill EA (1975) Sensitivity of adult and embryonic calanoid copepods to acute ionizing radiation. Radiat Res 63:382–385
177. Alonzo F, Gilbin R, Bourrachot S, Floriani M, Morello M, Garnier-Laplace J (2006) Effects of chronic alpha radiation on physiology, growth and reproductive success of *Daphnia magna*. Aquat Toxicol 80:228–236
178. Alonzo F, Gilbin R, Zeman FA, Garnier-Laplace J (2008) Increased effects of internal alpha radiation in *Daphnia magna* after chronic exposure over three successive generations. Aquat Toxicol 80:228–236
179. Abbott DT, Mix M (1979) Radiation effects of tritiated seawater on development of the goose barnacle, *Pollicipes polymerus*. Health Phys 36:238–287
180. Jha AN, Dogra Y, Turner A, Millward GE (2005) Impact of low doses of tritium on the marine mussel, *Mytilus edulis*: Genotoxic effects and tissue-specific bioconcentration. Mutation Res 586:47–57
181. Knowles JF, Greenwood LN (1994) The effects of chronic irradiation on the reproductive performance of *Ophryitrocha diadema* (Polychaeta, Dorvilleidae). Mar Environ Res 38:207–224
182. Cooley JL, Nelson DJ (1970) Effects of temperature and chronic irradiation on populations of the aquatic snail *Physa heterostropha*. Oak Ridge National Laboratory, Oak Ridge (ORNL-4612)
183. Cooley JL, Jr Miller FL (1971) Effects of chronic irradiation on laboratory populations of the aquatic snail *Physa heterostropha*. Radiat Res 47:716–724
184. Cooley JL (1973) Effects of chronic environmental radiation on a natural population of the aquatic snail *Physa heterostropha*. Radiat Res 54:130–139
185. Ravera O (1968) Molluscs. In: Proceedings of the symposium on mollusca, part II. Euratom Biology Division, European Atomic Energy Community, Brussels, p 295 (Contribution No. 295)
186. Cruz-Ramos B, Carino VS (1989) Effects of tritiated water on the digestive tract of *Pila luzonica* embryos. Nucleus 27:63–70
187. Tsytsugina VG, Polikarpov GG (2003) Radiological effects on populations of Oligochaeta in the Chernobyl contaminated zone. J Environ Radioact 66:141–154
188. Florou H, Tsytsugina V, Polikarpov GG, Trabidou G, Gorbenko V, Chaloulou CH (2004) Field observations of the effects of protracted low levels of ionizing radiation on natural aquatic population by using a cytogenetic tool. J Environ Radioact 75:267–283
189. Dugle JR, Guthrie JE (1970) Some observations on algae invading a ^{137}Cs contaminated pond. Atomic Energy of Canada Limited, Pinawa (AECL-3463)
190. Mericle LW, Phelps VR, Wheeler AC (1955) Dose-effect relationships in X-irradiated barley embryos. Genetics 40:585 [cited in Luckey, 1980]
191. Woodhead DS (1997) The report on "Effects of Radiation on the Environment" from UNSCEAR (United Nations Scientific Committee on the Effects of Atomic Energy). In: Symposium on radiological impacts from nuclear facilities on non-human species, Ottawa, 1–2 Dec 1996. Canadian Nuclear Society, Toronto, pp 109–119
192. Sparrow AH, Rodgers AF, Schwemmer SS (1968) Radiosensitivity studies with wood plants. I Acute gamma irradiation survival data for 28 species and predictions for 190 species. Radiat Bot 8:149–186
193. Gunckel JE, Sparrow AH (1961) Ionizing radiations: biochemical, physiological and morphological aspects of their effects on plants. In: Ruhland W (ed) External factors affecting growth and development, encyclopedia of plant physiology. Springer, Berlin, pp 555–583
194. UNSCEAR (United Nations Scientific Committee on the Effects of Atomic Radiation) (2008) Sources and effects of ionizing radiation. Report to the general assembly, with

scientific annexes. Annex E. Effects of ionizing radiation on non-human biota
195. Chandorkar KR, Szachrajuk RB, Clark GM (1978) Effect of extremely low radiation dosages on synchronized cultures of *Chlorella pyrenoidosa*. Health Phys 34:494–498
196. Amiro BD (1994) Response of boreal forest tree canopy cover to chronic gamma irradiation. J Environ Radioactivity 24:181–197
197. Amiro BD, Sheppard SC (1994) Effects of ionizing radiation on the boreal forest: Canada's FIG experiment, with implications for radionuclides. Sci Tot Environ 157:371–382
198. Sparrow AH, Schairer LA, Woodwell GM (1965) Tolerance of *Pinus rigida* trees to a ten-year exposure to chronic gamma irradiation from cobalt-60. Radiat Bot 5:7–22
199. Whicker FW, Fraley L Jr (1974) Effects of ionizing radiation on terrestrial plant communities. Adv Radiat Biol 4:1–317 [cited in IAEA, 1992]
200. Geras'kin SA, Zimina LM, Dikarev VG, Dikareva NS, Zimin VL, Vasiliyev DV, Oudalova AA, Blinova LD, Alexakhin RM (2003) Bioindication of the anthropogenic effects on micropopulations of *Pinus sylvestris* L. in the vicinity of a plant for the storage and processing of radioactive waste and in the Chernobyl NPP zone. J Environ Radioact 66:171–180
201. Geraskin SA, Dikarev VG, Ya Zyablitskaya Ye, Oudalova AA, YeV S, Alexakhin RM (2003) Genetic consequences of radioactive contamination by the Chernobyl fallout to agricultural crops. J Environ Radioact 66:155–169
202. Zelena L, Sorochinsky B, von Arnold S, van Syl L, Clapham DH (2005) Indications of limited altered gene expression in *Pinus sylvestris* trees from the Chernobyl region. J Environ Radioact 84:368–373
203. Taskaev AI, Frolova NP, Popova ON, Shevchenko VA (1992) The monitoring of herbaceous seeds in the 30-km zone of the Chernobyl nuclear accident. Sci Tot Environ 112:57–67
204. Rudolph TD (1974) The enterprise, Wisconsin, radiation forest. Pre-irradiation ecological studies. USAEC Report, TID-26116-P1
205. Zavitkovski J (1977) The enterprise, Wisconsin, radiation forest Radioecological studies. TID-26113-P2. National technical Information Service. US Department of Commerce, Springfield
206. Fabries M, Grauby A, trochain JL (1972) Study of a Mediterranean type phytocenose subjected to chronic gamma radiation. Rad Bot 12:125–135
207. Woodwell GM (1963) Design of the Brookhaven experiment on the effects of ionizing radiation on a terrestrial ecosystem. Rad Bot 3:125–133
208. Odum HT, Pigeon RF (1970) A tropical rain forest. A study of irradiation and ecology at El Verde, Puerto Rico. T1D-24270, USAEC Division of Technical Information Extension, Oak Ridge
209. Fraley L Jr, Whicker FW (1973) Response of short-grass plains vegetation to gamma radiation. Rad Bot 13:331–353
210. Miller GL (1968) Influence of season on the radiation sensitivity of an old field community. Thesis, University of North Carolina, Chapel Hill, North Carolina, p 262
211. McCormick JF, Golley FB (1966) Irradiation of natural vegetation. An experimental facility, procedures and dosimetry. Health Phy 12:1467–1474
212. Guthrie JE, Dugle JR (1983) Gamma-ray irradiation of a boreal forest ecosystem: the Field Irradiator-Gamma (FIG) facility and research programs. Can Field Nat 97: 120–128
213. Turner BN, Iverson SL (1976) Project ZEUS: a field irradiator for small-mammal population studies. Atomic Energy of Canada Limited Report, AECL-5524
214. Alexakhin RM, Karaban RT, Prister BS, Spirin DA, Romanov GN, Mishenkov NN, Spiridonov SI, Fesenko SV, Fyodorov YA, Tikhomirov FA (1994) The effects of acute irradiation on a forest biogeocenosis; experimental data, model and practical applications for accidental cases. Sci Tot Environ 157:357–369
215. McCormick JF (1969) Effects of ionizing radiation on a pine forest. In: Neson DJ, Evans FC (eds) Symposium on radioeclogy. United States Atomic Energy Commission, CONF-670503, Washington, DC
216. Poinsot-Balaguer N, Castet R, Tabone E (1992) Impact of chronic gamma irradiation on a Mediterrainean forest ecosystem in Cadarache (France). J Environ Radioact 14:23–36
217. Sheppard SC, Guthrie JE, Thibault DH (1992) Germination of seeds from an irradiated forest: implications for waste disposal. Ecotoxicology Environ Safety 23:320–327
218. Dugle JR, Mayoh KR (1984) Responses of 56 naturally-growing shrub taxa under chronic gamma radiation. Environ Exp Bot 63:1458–1468
219. Godward MBE (1960) Resistance of algae to radiation. Nature 185:706
220. Conter A, Dupouy D, Planel H (1983) Demonstration of a biological effect of natural ionizing radiations. Int J Radiat Biol 43:431–432
221. Garnier-Laplace J, Della-Vedova C, Andersson P, Copplestone D, Cailes C, Beresford NA, Howard BJ, Howe P, Whitehouse P (2010) A multi-criteria weight of evidence approach for deriving ecological benchmarks for radioactive substances. J Rad Prot 30:215–233
222. Woodhead DS, Zinger I (2003) Radiation effects on plants and animals *Deliverable 4* FASSET project. http://www.fasset.org

Books and Reviews

Chipman WA (1972) Ionizing radiation. In: Kinne O (ed) Marine ecology, vol I, part 3. Wiley/Interscience, New York, pp 1579–1657 [cited in Anderson and Harrison, 1986]

Hinton TG, Coughlin DP, Yi Y, Marsh LC (2004) Low dose rate irradiation facility: initial study on chronic exposures to medaka. J Environ Radioact 74:43–55

IAEA (2001) Present and future environmental impact of Chernobyl accident. IAEA-TECDOC-1240

Ophel IL (1976) Effects of ionizing radiation on aquatic organisms. In: Effects of ionizing radiation on aquatic organisms and ecosystems. International Atomic Energy Agency, Vienna, pp 57–88 (Technical Reports Series No. 172)

Polikarpov GG (1966) Radioecology of aquatic organisms. Reinhold Book Division, New York

Smith JT, Beresford NA (2005) Chernobyl – catastrophe and consequences. Springer, Published in association with Praxis Publishing, Chichester

Templeton WL, Nakatani RE, Held EE (1971) Radiation effects. In: Radioactivity in the marine environment. National Academy of Sciences, Washington, DC, pp 223–239

Irrigation Management for Efficient Crop Production

Elías Fereres, Margarita García-Vila
Institute for Sustainable Agriculture, IAS-CSIC and University of Cordoba, Cordoba, Spain

Article Outline

Glossary

Application efficiency Relationship between the target irrigation depth (depth of water stored in the root zone to be used by the crop) and the depth of water applied to meet this target during a single irrigation event.

Conservation agriculture (CA) An agricultural production system aimed at achieving a sustainable and profitable agriculture through the application of three principles: minimal soil disturbance, permanent organic soil cover, and diversification of crop species in rotations or associations.

Decision support systems (DSS) Interactive information systems (not limited to computerized systems) that aid decision makers to identify and solve problems, and make decisions, which may be rapidly changing and are not easily specified in advance.

Deficit irrigation (DI) An irrigation strategy based on applying irrigation depths that are less than the full crop water requirements (ET), either throughout the crop life cycle (continuous or sustained deficit irrigation) or during specific stages that are insensitive to water stress (regulated deficit irrigation).

Distribution uniformity A measure of the spatial evenness with which irrigation water is distributed across a field.

Evapotranspiration (ET) The combination of two separate evaporation processes whereby water is lost to the atmosphere, on the one hand from the soil surface and crop surfaces (canopy interception) and on the other hand, from inside the leaves and other organs through pores called stomata, a process termed "transpiration."

Irrigation return flows The combination of surface and subsurface water flows resulting from the runoff and drainage following the application of irrigation water which may be available for subsequent appropriation from either a stream or an aquifer downstream of the original use.

Leaching requirements The depth of water needed to displace the excess salt accumulation in the soil profile resulting from irrigation, and aimed to maintain the salt balance in the crop root zone.

Soil water balance The state of soil water in the crop root zone resulting from the balance between water inputs from precipitation and irrigation and the water losses to evapotranspiration, runoff, and drainage below the root zone.

Water use efficiency (WUE) The ratio between the water volume used for a specific purpose and the water volume derived from a source to accomplish that purpose.

Definition of the Subject

The anticipated population growth in the coming decades will place large worldwide demands to increase

the global production of food, animal/fish protein, livestock feed, fiber, and biofuels. Such an increase must come primarily from enhancing productivity, given the constraints in further land expansion for agriculture. Irrigated agriculture currently produces more than 40% of total production on 17% of the land. It is therefore imperative that irrigated agriculture not only sustains its current rates of productivity but that they be increased in the future. Irrigation expansion has taken place over the last 60 years, and is currently under pressure from other sectors to reduce its share of the freshwater resources. Efficient crop production under irrigation in the future would be essential to produce more food with less water than is used today. This goal is a challenge that will not be easy to achieve without new and innovative approaches in irrigation management and in crop productivity. These innovations would also have to address the problems created by the return flows from irrigation which will threaten its sustainability, unless solutions are found to resolve permanently the environmental impacts of irrigation.

Introduction: Background on the Sustainability of Irrigation

The practice of irrigation started soon after agriculture was discovered thousands of years ago. Near the main rivers in the arid zones, water was diverted to the fields where crops were grown in areas and times of lack of rainfall, in an attempt to obtain a stable food supply. The Sumerian civilization that inhabited the Mesopotamia plains is believed to be among the first that used irrigation for crop production. Other locations where major irrigation developments took place early in the history of modern agriculture include the Yellow River basin of China, the Indus River Valley in Pakistan, and the Nile River Valley of Egypt. Many civilizations developed successfully in the past centuries, their food security heavily dependent on irrigation. History is filled with cases, however, where irrigated agriculture failed after some time, leading to the decline and even the disappearance of civilizations (e.g., Mesopotamian civilizations). The causes for failure included technical as well as economical, social, and political factors, but all combined suggest that irrigated agriculture may not be sustainable indefinitely. On the contrary, cases where irrigation has been practiced successfully for millennia, such as the Nile River basin in Africa and many areas of Southeast Asia, prove that it is possible to sustain irrigated agriculture in the long run [1].

Nowadays, irrigation is by far the largest consumer of developed freshwater on a global basis. It is estimated that irrigated agriculture currently uses about two thirds of water diversions, while industry and urban diversions amount to around 20 and less than 10%, respectively [2]. Such a high level of consumption in agriculture is due to the fact that crop plants require a continuous supply of water to replace the water transpired from their leaves and other aerial organs. The water demand arises because the crop is exposed to strong evaporative demand (due to the fluxes of solar and thermal radiation and warm, dry air). For carbon dioxide to enter the leaves, the microscopic leaf pores (stomata) must be open. But when the pores are open, water vapor freely escapes from the interior of the leaves which are nearly saturated with water. Because of the differences in concentration of carbon dioxide and water vapor between the interior of the leaf and the air, 50–100 molecules of water are lost for every molecule of carbon dioxide taken up. Crop water consumptive use is thus an unavoidable consequence of wet crop surfaces being exposed to dry air. Nevertheless, despite the large amounts of water that are transported through the plants, if they are not capable of taking up soil water to replace the losses, water deficits develop which can be detrimental to yield [3]. Therefore, if irrigated agriculture is to be sustainable, it needs sufficient water to meet its requirements at present and in the future.

Presently, more than 260 million ha of irrigated lands exist, representing 17% of the cultivated area but more than 40% of food production worldwide [2]. Future increases in population combined with changes in dietary habits toward the consumption of more animal protein will require sustained increases in crop production in the next decades well above present levels. Additionally, there may be demands on agricultural products for uses other than food production, such as energy from biomass. The increase in production cannot come from a significant expansion of the area devoted to agriculture for at least two reasons. On one hand, the area best suited for agricultural use is already under production, and only in a number of

regions in the American and African continents there is additional land that has not been put under cultivation. Furthermore, that expansion would alter further the balance between agricultural and natural ecosystems, which much of the world population would oppose, arguing for the environmental preservation of the remaining areas that are not in cultivation.

If land expansion will not be possible, the goal would then be to increase productivity, the production per unit of cultivated land. Crop productivity of the main cereals has been increasing steadily since the 1960s, albeit at rates that are apparently declining in recent years [4]. Given that the average productivity of irrigated lands is more than twice that of rainfed areas, it appears that preserving irrigated agriculture would be essential for future food security. The issue is whether it would be possible to expand irrigated agriculture beyond its present level by transforming rainfed into irrigated areas. While some expansion may be possible in the foreseeable future, it is difficult to see how a major world expansion of irrigation would occur, given the present commitments of freshwater (supplies already overcommitted in many arid and semiarid areas), the perceived strong opposition from urban societies, and the uncertainties that climate change brings to future water supplies. It is therefore essential that the productivity of irrigated lands be increased to cope with future demands, and this will have to be done at efficiency levels higher than those achieved at present. The reduction in water used in irrigation per unit production emerges then as a critical issue in the area of crop production in the future.

Irrigated agriculture has to manage large amounts of water that must be utilized with a high level of efficiency. This is not the only requisite of good management, however. All irrigation waters contain salts, and crops transpire pure water; thus, the irrigation process concentrates the salts in the soil profile at a rate which mostly depends on the salt content of the irrigation water. In this respect, the salinity that develops under irrigation would make cropping unsustainable unless the salts are evacuated to prevent the salinization of the crop root zone. In many areas, natural rainfall is sufficient to leach out the salts from the potential root zone. However, when waters of high salinity are used for irrigation in arid areas of limited rainfall, salt leaching must be performed by applying irrigation in excess of the crop needs. The control of salinity is one important requisite of a sustainable irrigated agriculture. However, the water applied in excess of the consumptive use must be disposed of, and this creates a whole host of potential environmental problems associated with water pollution [5]. Many of the problems caused by the return flows from irrigation occur outside the farms, at the level of the irrigation district or at the basin level, and are discussed below.

The Process of Irrigation at Different Scales: From the Field to the Basin

When a field is irrigated, the water applied may be lost via surface runoff or by deep percolation, or may be stored in the crop root zone for subsequent uptake by the crop and lost as ET. However, the focus of irrigation management for efficient crop production extends well beyond a field, up to the farm, the irrigation district, and the watershed. The water balance is the unifying concept that connects the water disposition in the different scales. In any field, farm, or watershed, it is possible to quantify a water balance in which the incoming water in the form of rain or irrigation must be balanced by the water lost as evapotranspiration (ET), runoff, deep percolation, and that stored in the soil profile.

When scaling up from an irrigated field, one needs to consider the situation upstream and downstream of that field. Irrigation water originates from storage reservoirs, flowing streams, or from the groundwater. In all cases it must be conveyed and distributed among the different areas, their farms, and individual fields. The process of distribution entails a number of losses due to direct evaporation from reservoirs and open conduits, leakages from canals, and management losses in the handling of distribution networks that deliver the water to individual farms [6]. Farmers on collective networks do not always have access to water when they need it, and that may cause imbalances between the supply of irrigation water and the crop demand. Those that pump directly from the groundwater have greater flexibility at the expense of additional energy usage. At any rate, by the time the water is delivered to a particular field, there have been distribution and management losses upstream that are often substantial.

The classical work of Bos and Nugteren [7] quantified the efficiency of irrigation along the path from the source of water to the field and found very low values, of the order of 40–50%. Hsiao et al. [8] have also analyzed the efficiency losses in the network, from the dam to the crop, and have highlighted the dramatic differences between good and bad situations, given the multiplicative effects of the chain of efficiencies from the source of water until it is transpired to the atmosphere [8].

At the field scale, water may be used beneficially for ET and for salinity control or may not have a beneficial use when it is lost through runoff or drainage. However, at higher scales, water that evaporates from a watershed is considered a loss or consumption, while water running off a field can be recovered downstream and may not be lost to the system. Equally, water that percolates below the root zone may reach the groundwater from which it can be pumped and recovered. Thus, there are consumptive and nonconsumptive uses of irrigation water. Water applied as irrigation may be used consumptively in the ET process, while the network and runoff losses may be recovered downstream and used by others within the basin. Water used in one farm within a basin is not always consumed in that basin and can be used severa1 times before it leaves the basin.

The difference between water use and consumption is important to understand whether water conservation efforts will result in net water savings [9]. If *all* the water lost outside the ET can be recovered, then efficiency improvements via reducing runoff or percolation losses would not lead to net water savings. On the contrary, if the losses are partially or not recoverable at all, because either the water quality is deteriorated or the losses end in a saline sink, then the water saved at the field scale also represents (partial) savings at the basin scale. Thus, knowing the fate of water all along its path, it is critical to determine whether the water saved on the farm may be part of the recoverable or of the unrecoverable losses.

The basin perspective of water conservation could change the emphasis on where to act when irrigation improvements are sought, and whether such improvements are really needed. Nevertheless, the picture would not be complete if two other issues are not included in the overall assessment of basin irrigation management. One is the fact that field and farm losses, while they could be recovered downstream, often have negative environmental consequences. Surface runoff may carry sediments and chemicals that act as pollutants in streams, lakes, and dams. Drainage waters pick up salts, fertilizers, and other chemicals in the soil profile and may contaminate the groundwater. The other issue is the energy requirements for recovering the losses. Whether is surface runoff recovered at the end of a field or groundwater pumped from the aquifers, additional energy is always needed to recover the losses. Needless to say, the water quality of these return flows would always be worse than that of the original irrigation water.

Irrigation Management Goals for the Improvement of Sustainability

An agriculture that aims toward sustainability should be economically viable, efficient in the use of the natural resource base, socially equitable, and must have a minimal environmental impact. These requisites are essential when defining the goals of irrigation management for efficient crop production. Economic viability in irrigated agriculture is usually associated with *high production levels,* because the investments needed for irrigation development would not be justified under low levels of production. Thus, irrigation must be managed in such a way as to avoid water deficits that reduce economic yield. Not only high production should be sought but also, *high productivity* in relation to the use of production inputs, including water, should be an important goal. Contrary to the common belief, de Wit [10] demonstrated that production inputs are used at their highest efficiency when yields approach the maximum potential. Therefore, it is possible to combine both goals with judicious management.

Equity is an important goal in irrigated agriculture because, first of all, irrigation development represents a quantum leap in terms of increased income relative to rainfed agriculture. Also, sometimes water supplies are managed rigidly by water authorities or are insufficient relative to the potential demand; thus, *equitable distribution* of the available water supply is definitely an important goal of sustainable irrigation management. How this goal is pursued would depend on a number of socioeconomic, political, and cultural factors discussed

below. Finally, irrigation can have a number of detrimental effects on the environment that must be minimized. Firstly, without *control of salinity* agriculture cannot be sustainable [11]. However, the leaching of salts and other chemicals from the soil profile end up as part of the return flows and contribute to the nonpoint pollution generated by irrigated agriculture. To minimize these negative effects on the environment, irrigation must be managed in such a way as to *avoid runoff* and *minimize deep percolation.* To achieve this goal, the engineering of irrigation systems (to distribute water uniformly) and the scheduling of irrigations (correct timing and application amounts) are the main instruments to be optimized when walking the fine line of achieving high production and productivity while reducing the environmental impact of irrigation.

Management Options: Strategic, Tactical, and Operational

In this entry, we are assuming that farmers have already made a choice among the different irrigation methods, basing their decision on their access to capital, the availability of water and its distribution mode, their management skills, and the labor and energy costs. There are frequent interactions between engineering and management in irrigation that are frequently ignored but are covered here, even though the focus is on management.

The temporal scale determines the nature of irrigation management decisions. Operational decisions are those that must be made in the short term, within days; for instance, the advancing or delaying of one irrigation application. Tactical decisions are those that are taken within the irrigation season once it has started, and have a time scale of days to weeks. One example would be the adjustment of the number of irrigation applications within the season, if the water supply has been reduced after planting. Strategic decisions have a time scale of months to years, and are normally taken before the season starts. The preseason decisions pertaining to the allocation of water to the different fields and crops are examples of strategic decisions.

Crop and Cultivar Selection

Farmers' choice of crops in commercial agriculture is a complex decision that is, above all, based on factors related to production economics and marketing, while other socioeconomic and biophysical factors are considered afterward. The water-related factors such as crop water usage are most important when the supply available is less than the anticipated demand. Crop consumptive use depends primarily on the evaporative demand, the length of season, and the fraction of incoming radiation intercepted by the crop canopy. There are ample differences among the crop ET of different species. These differences are also modulated by the type of climate in which crops are grown. In that respect, there are major differences between tropical and temperate climates. In tropical climates, reference ET (the ET from a standard grass surface: ETo) is relatively constant throughout the year and irrigation is used during the dry season; here, the critical issue is the duration of the growing season, with the goal of producing multiple crops in 1 year. Production per day is the best indicator of efficiency in tropical climates, where crops must follow a sequence that uses best the land and water available.

In temperate climates, evaporative demand during winter is a small fraction than that in the summer. For instance, in Mediterranean-type climates, ETo oscillates between 1–2 mm/day in winter and 6–8 mm/day in summer. Thus, winter crops require much less water than summer crops per unit time. Additionally, seasonal rainfall occurs primarily from fall to spring in those climates, thus reducing the irrigation needs of winter and spring crops. One environmental factor that indirectly affects the crop water requirements in temperate climates is air temperature. Temperatures in winter are low in such climates and that slows down the rate of crop growth and development, thus lengthening crop duration. For instance, the season of winter cereals may last 6–7 months while maize, a summer crop, may be grown in 4 months. These differences in season length balance out some of differences in ETo between winter and summer, but still winter crops generally use less water than summer crops. As an example in some interior valleys of Mediterranean climate, wheat ET is between 350 and 400 mm while maize ET ranges between 600 and 650 mm. Perennial crops have higher ET rates, alfalfa ranging between 800 and 1,200 mm, and deciduous trees between 700 and 1,000 mm. Evergreen tree crops should have the highest ET, however, both citrus and olive have strong stomatal

control of transpiration and their seasonal ET does not exceed values that range between 700 and 1,000 mm, depending on the specific climate.

Crop choice is therefore the most important factor in determining irrigation water requirements. Cultivar differences in water use are considerably smaller than the differences among crops, and are directly related to season length. Early cultivars of maize in temperate climates may use 10–20% less water than late-maturing cultivars. Similar differences have been observed in other crops. Even though such differences are relatively small, they can be important when the crop sequence is such that only short-season cultivars fit in the rotation or when the water supply available is limited. Genetic improvement of the intrinsic transpiration efficiency (TE, g CO_2/g H_2O) [12] has not been successful until now. The limited variability encountered within crop species has not led to cultivars that have higher productivity in irrigated agriculture, although some yield advantage (of the order of 10% at around 1 t/ha yield levels) has been achieved when selecting wheat cultivars for high TE in rainfed environments [13]. The use of genetic engineering in the future [14] may offer new opportunities for enhancing other basic responses that can indirectly improve crop WUE (e.g., maintenance of harvest index under water stress).

Planting dates also have some effects on crop ET in temperate climates. Early plantings of spring or summer crops have lower ET by avoiding the times of peak ETo. One example is that of winter plantings of sunflower in Mediterrranean climates. By planting early, the growing season is displaced away from the summer and the seasonal ET is less, even though the season may be longer due to the slow growth and development in the early crop stages. The irrigation requirements may even be lower because of the higher rainfall probabilities in early spring. In general, plantings of summer crops have been moved as early as feasible to reduce irrigation requirements; this is the case of maize in many areas where the optimal planting dates have moved more than 40 days over the last 30 years. To displace the growing season any further the temperature limitation to growth and development must be overcome. Progress has already been made in crop improvement; for example, maize has expanded significantly into colder areas in the last decades, but more breeding efforts are needed to achieve higher growth rates under low temperatures in the principal irrigated crops.

Contrary to the measurable effects of varying planting dates on crop ET, the variation in planting density within commercial practices has little influence on the ET of annual crops. This is because the differences in radiation interception among different planting densities are small and restricted to the short period of early canopy development. It is important, however, in the case of perennials, as tree or vine densities have a strong influence in the intercepted radiation for several years after the initial planting, and may be carried over the entire life of the orchard or vineyard. Biomass production and thus yield of most crops is directly related to the seasonal intercepted radiation. Any agronomic practice that favors quick canopy development and complete radiation interception should increase production and will have a smaller effect on crop ET, thus increasing the efficiency of water use.

Optimal Use of Rainfall and of Stored Soil Water

The goal of irrigated crop production is to use all sources of water supply as effectively as possible. Soil water storage from rainfall or from preplanting irrigation is an effective way of using the water resource. To maximize stored soil water, infiltration should be enhanced so that surface runoff is minimized. Low infiltration rates decrease the effectiveness of rainfall to the point that only a small fraction of it may be stored in the potential crop root zone for subsequent use by the crop. Many irrigated soils have problems of slow infiltration, either inherent to their physical makeup or caused by the application of irrigation water for a long time that alters negatively the surface infiltration properties of the soil [15].

Excessive tillage of some irrigated soils and, more importantly, the traffic required in crop intensification under irrigation has exacerbated the problems related to low infiltration, which not only reduces the effectiveness of rainfall but also affects the distribution of irrigation water. The use of minimum tillage, no-tillage, surface residues, permanent beds, and of controlled traffic is all being incorporated into what is now known as conservation agriculture [16]. These soil management systems have been used successfully in many world areas for sometime now, primarily under

rainfed agriculture, but they are now increasingly being adopted for irrigated agriculture as well. The benefits of conservation agriculture (CA) include soil surface protection by the crop residues, the maintenance of soil organic matter, increased infiltration, and better distribution uniformity. The limitation of CA is that it needs to be tailored to the specific situation thus requiring field experimentation before it is introduced in a new area or system. The widespread expansion of CA in the main agricultural areas of the world [17] suggests that it will be adapted to many irrigated systems in the near future.

The critical role of rainfall, while being obvious in rainfed agriculture, is nearly as important in irrigated agriculture because it leads to a reduction in irrigation demand. The best strategy for optimal use of stored soil water is the conjunctive use of both, the applied irrigation water and the soil reserve. It is desirable that the soil is partially depleted to allow the storage of anticipated rainfall in the root zone, although such depletion has to be managed to avoid yield-reducing water deficits (See below). As the crop approaches maturity, it is recommended to rely more on the stored reserve and use as much of it as possible, thus reducing irrigation water use. Ideally, the soil reserve should be almost totally depleted when the crop is harvested, assuming it will be replenished by rainfall. Obviously, to manage the soil reserve, it must be quantified from planting to harvest. Growers need to know the level of soil water at planting and the rate of water use (ET) relative to the depths of irrigation and rainfall. Therefore, making best use of the rainfall and of the water reserve requires carrying out a seasonal water budget for each crop. It is also important to evaluate the risk of basing a limited irrigation strategy on using a large fraction of the stored rainfall, because in the event of a drought, irrigation would be insufficient to meet the crop demand and this strategy may not be sustainable.

Technical Irrigation Scheduling

Decisions on when to irrigate and how much water to apply – the irrigation scheduling process – are commonly made by irrigators around the world solely based on experience. There are, however, a collection of technical procedures and tools developed to forecast the timing and amount of irrigation applications. Some sort of irrigation on demand is a prerequisite for the application of these technical procedures, because when the delivery method of the network is on rotation, the farmers have no flexibility to vary the irrigation interval and they tend to use all the water they receive as insurance for uncertain conditions.

Among water-sensing devices, soil water sensors were perhaps the first instruments that were introduced for irrigation scheduling in the 1950s, and it is remarkable that they have enjoyed a certain degree of success until recently. There is now a new generation of soil water sensors that track soil water status continuously, rather than providing point measurements as the traditional instruments such as the tensiometer offer. Unfortunately, the new developments have not resolved the quantification of volumetric soil water content with depth, a parameter that is still most reliably measured with the neutron probe since the early 1970s. The regulatory constraints on this nuclear instrument have limited its use even for research in many countries, with the result that reliable data on soil water balance and crop ET are difficult to obtain with soil water measurement methods. The information from current sensors is treated as trends and these tendencies, often observed at more than one depth, are the basis for making decisions, rather than using a threshold value of soil water at a given depth as the indicator of irrigation timing. Protocols have been developed to automate irrigation scheduling on the basis of soil water status [18].

The use of plant water sensors for irrigation management lagged behind that of soil water sensors by about 2 decades. The pressure chamber was the first portable instrument that was rugged enough to be used under field conditions, although it is manually operated and its readings cannot be automated. It is now used commercially in some areas for irrigation scheduling of tree crops and vines. Other plant sensors have not been as successful for their use in practical scheduling, although they have been around for sometime. The most notable example is that of dendrometers, sensors that detect the variations in stem diameter, which were used since the 1950s and have today the same degree of precision they have had since the 1970s, but they became popular for research 20 years later. Protocols to be used with trunk diameter sensors have

also been proposed [19] although its use, while attractive, has been mostly limited for research purposes. One of the most promising plant-based indicators is the canopy temperature as measured by infrared thermometry. Jackson and coworkers (summarized in [20]) developed several indicators based on canopy temperature, the Crop Water Stress Index (CWSI) being the most popular. Threshold values of CWSI have been found for several crops and its use as a water stress indicator is gaining acceptance. Plant-based parameters are best used as pre-visual indicators of water stress, rather than being indicative of irrigation timing and amount. Their strength resides in providing a specific, crop-based calibration for other methods that use either soil water sensors or the water balance procedure. One important limitation of all soil- and plant-based sensors is the variability in the measurements, as discussed in Coping with Spatial Variability: Precision Irrigation.

The most robust technique for wide use in irrigation scheduling is the one using the soil water budget. Here, irrigation timing is computed by adding the crop ET losses minus effective rainfall until a soil water level termed "the allowable depletion" is reached. The basic information in this method is that of crop ET, computed as the product of a crop coefficient times ETo. After a method for computing ETo has been standarized and widely accepted [21], agrometeorological weather stations provide the information needed for calculating ETo from meteorological variables. In some developed countries, networks of weather stations now provide the ETo information routinely. The pan evaporation is a viable alternative for estimating ETo to the more sophisticated automated weather stations. Computer programs have been developed for calculating the water balance of fields, and irrigation scheduling services have been developed, mostly by public agencies and by consultants as well. These services have been around for several decades now but most farmers have been reluctant to pay for them. Nevertheless, irrigation advisory services are becoming more popular in areas of scarce or expensive water. One pattern that has been observed is that farmers subscribe to these technical procedures or services for a few years and then, they no longer use them. Perhaps they perceive that they have acquired sufficient knowledge during that time period, a reason that may also explain why the use of sensors is discontinued after some time by many growers.

Interactions Between Irrigation Methods and Their Management

Irrigation has been practiced for thousands of years by flooding the soil surface and keeping the water standing until it infiltrates. This method is named surface irrigation and it is still the most popular method worldwide. In surface irrigation, the soil intake rate determines the depth of water that infiltrates and if its properties are spatially variable, the farmer will not have good control of the amount of water applied. Pressurized irrigation methods (sprinkler and microirrigation) were invented much more recently, about 70 years ago. Since then, they have enjoyed increasing popularity in areas where farmers have access to sufficient capital to shift from surface to pressurized systems. In other, newly developed areas where topography and/or water infiltration properties made the use of surface methods impractical, pressurized methods have been preferred over surface irrigation. One advantage of pressurized systems is that the depth of applied water does not depend on soil properties but is determined directly by the run time. Farmers' preferences for pressurized systems are based on the need for better control and the greater skills needed for effective surface irrigation management. The higher capital and energy requirements are two limitations of pressurized systems relative to surface irrigation.

The key feature of irrigation systems in relation to their management is the degree of uniformity of water distribution. Regardless of the method, high distribution uniformity prevents excessive percolation losses in some areas of the field and the development of water deficits in others. To emphasize the importance of high distribution uniformity suffices here to summarize an example in which the performance of two systems with low (70%) and high (90%) distribution uniformity (DU) were compared for maize irrigation [22]. The additional depth of water needed under low uniformity to achieve maximum yields amounted to 400 mm, or 70% of the net irrigation requirements. The requisites for high DU include an appropriate design, good

maintenance, and correct operations. Nowadays, efficient irrigation cannot be practiced unless high to very high DU values are achieved.

All irrigation methods have potentially high performance that can eliminate or minimize runoff and percolation losses, but they seldom achieve their potential due primarily to lack of maintenance or to mismanagement. To optimize the operation of existing surface irrigation systems the water delivery procedures often need to be changed. There are needs relating to adjusting the flow rates delivered to fields, altering the operation of delivery networks, and sometimes land consolidation is also needed. Such changes not only require capital investments but agreements among various users as well. The introduction of pressurized systems (sprinkler and drip) in collective irrigation networks also requires changes in the physical infrastructure (e.g., reservoirs). Thus, there must be economic incentives for the farmers and access to capital to introduce the improvements needed to increase the potentia1 application efficiency and distribution uniformity at the system level.

Control of Salinity and of Return Flows

Irrigated agriculture cannot be sustainable unless salinity is properly managed. Salts that accumulate in the root zone must be leached but the amount of leaching must be kept to the minimum if the environmental impacts of drainage waters and the return flows are to be controlled. Determining the leaching fraction (proportion of the ET that must be added for salt leaching) should be based not only on the quality of irrigation and drainage waters but on the need to avoid excessive percolation. Here again, high DU is essential when the target leaching fraction is of the order of 5–10%, which is only achievable under very high DU values.

The use of salt-tolerant crops is often proposed as a means of controlling salinity. It is true that there is an ample range of salinity tolerance among crop plants and that it is possible to exploit waters of low quality for irrigation of salt-tolerant crops. However, because the process of salt concentration in the profile in the absence of leaching is inexorable, the use of salt-tolerant crops should be considered as a temporary measure until excess salts can be leached out of the potential root zone. Sometimes, tolerant crops have been used for some years in the event of a drought that limits the irrigation supply available for leaching. Their long-term use in dry areas should not be considered sustainable, given the progressive salinization of irrigated soils. After some environmental problems related to toxicity caused by the selenium content of the return flows, the drainage from a large area in California, San Joaquin Valley, was interrupted in 1989 [23]. The rainfall in the area is negligible, thus salt leaching depends on artificial drainage. As of 2010, the area is still under irrigation; irrigation systems installed have high DU values and there is an extensive monitoring program of soil salinity. Perhaps long-term control of salinity would be feasible with systems that have very high DU, and where irrigations are scheduled with precision, keeping most of the salts near the bottom of the root zone, as it was first proposed by Hoffman et al. [24].

At scales beyond that of a field or farm, the concerns shift toward the quality of return flows, its reuse, and the environmental impacts of irrigation. The decline in water quality after the water has gone through the irrigation process has an associated cost that needs to be quantified. The concept that water lost to a farm is recovered downstream needs to have associated an economic analysis of the energy costs and the quality deterioration costs of recovering the return flows. If minimum leaching is practiced at the field scale, the amount of return flows is also reduced but that increases the concentration of salts and other contaminants in the drainage waters. Eventually, it may be possible to reduce the quantity of return flows so much so that they can be disposed of in evaporation ponds that become salt sinks. It would then be possible to either accumulate or export the salts and make the agriculture of that region fully sustainable.

Coping with Spatial Variability: Precision Irrigation

The major challenge that technical farm irrigation management has faced and continues to face is how to cope with variability. Under field conditions, both spatial and temporal variabilities are the norm rather than the exception. The strong spatial heterogeneity of soil water properties even in what are considered uniform soils, combined with the variations in the distribution of irrigation water applications, and the

uncertainties of rooting depth and densities, all contribute to create a heterogeneous environment that farmers have to manage as accurately as feasible. The problem has increased in magnitude over the last decades due to the increase in size of the management units, in an attempt to reduce production costs by managing uniformly larger and larger field units. The complexities involved in dealing with the variability problem are such that, until very recently, the common solution chosen by irrigators was to apply water in excess so that the risk of inducing water deficits in some parts of the field is minimized. Because of the difficulties that farmers and technicians have had in characterizing the variability, significant uncertainty is introduced and often the irrigation management decisions may be in error.

To advance solutions for coping with the variability problem in irrigation management what is needed is to be able to characterize the variation across a field, and also to have the option of applying variable amounts of water within that field. The objective would then be to apply variable water depths under non-uniform crop growing conditions to match the requirements of every area of the field, while minimizing the environmental consequences that uniform irrigation over a variable field would have. The technologies for variable water application are already available in self-propelled sprinkler systems and can lead to significant water conservation [25]. Significant efforts in the engineering of irrigation systems have been undertaken recently to offer the flexibility of applying spatially variable amounts of water (and agrochemicals) for the different pressurized methods, including microirrigation [26]. These new capabilities should enable growers to increase productivity and minimize environmental impacts of irrigation.

While the engineering solutions for precision irrigation are underway, there is still the need, not only to characterize and monitor the variability but to interpret the causes of the variations in crop growth and development. The characterization of irrigation performance through remote sensing [27] is a promising area, as it enables performance evaluation in a fast and an inexpensive way, and can also identify the areas in need of improvement. Interpreting the underlying causes of variations among and within fields is much more difficult, however. The use of remote sensing techniques has progressed substantially in recent years by developing capabilities for detecting a number of vegetation properties with very high resolution (e.g., [28]). High-resolution imagery cannot be acquired from current satellites, and a number of initiatives to obtain them from aerial vehicles flying closer to the ground have been launched recently. As an example that is relevant for irrigation management, Berni et al. [29] applied models based on canopy temperature estimated from high-resolution airborne imagery, obtained with an unmanned aerial vehicle, to calculate tree canopy conductance and the CWSI of heterogeneous canopies, such as those of tree crops.

Use of Simulation Models and of Decision Support Systems

Decision-making in crop production has been the focus of numerous studies, mainly on the description of the decision-making process and decision outcomes [30]. The numerous decisions dealing with irrigation management include, not only the scheduling and application of the available water to different crops over the irrigation season, but also strategic decisions related to crop choices and seasonal water allocation. Irrigation is a complex operation, based on technical and agronomic knowledge, and on sociological factors which may include a negotiating process among irrigators [31]. Recent advances in information and telecommunication technologies allow farmers to acquire vast amounts of site-specific data for their farms, with the ultimate goal of *reducing uncertainty* in decision-making. However, farmers face many difficulties in efficiently managing, analyzing, and interpreting the vast amount of data collected, while considering both the costs and value of the information [30].

The tactical decisions related to irrigation scheduling have been discussed above; however, strategic decisions that must take into account the complex nature of agricultural systems and changes in environmental conditions are difficult to make without tools that assist the farmer in the decision-making process. Among the tools available in the area of water management are the Decision Support Systems (DSS), which were first used in the early 1970s as a radical alternative to large-scale management information systems [32]. There are different types of DSS; many of them are expert systems [33], which have the problem of

handling multiple experts to evolve decisions and uncertainty. Linear programming, dynamic programming, and non-linear programming are the most popular modeling techniques; while recently, genetic algorithms [34] have been used to generate optimal solutions more efficiently [35] than dynamic programming (e.g., [36]). Also, DSS have been shown to help decision-making at different scales: at the field level, considering only one crop (e.g., [37]); and at the farm scale, with multiple crops (e.g., [38]).

Because irrigation decisions include many factors, the DSS combine crop simulation models with econometric models to assist farmers in optimizing irrigation management, according to environmental, socioeconomic, and political prospects [39]. The yield response to different irrigation levels is one of the inputs of DSS that traditionally has been quantified as empirical crop-water production functions [40–42]. Even though these functions have been used profusely, they are site-specific and difficult to extrapolate without costly, empirical calibration. An alternative to crop-water production functions is the use of dynamic crop simulation models [43]. Among the simulation models usable for irrigation decision-making are CropSyst [44], EPIC [45], CROPWAT [46], APSIM [47], and CERES [48]. However, most of these models require detailed information (difficult to obtain) about parameters that describe plant behavior (APSIM, CERES), or make use of simple empirical functions (CROPWAT). The models must be calibrated, validated, and be sufficiently robust to provide reliable predictions. For this reason, detailed models may be less practical than simpler but robust models [49] such as the recently published FAO water productivity model, AquaCrop [50]. AquaCrop is a model focused on simulating attainable yield in response to the water available, and it is thought to have an optimum balance between accuracy, simplicity, and robustness [50]. In water-limited situations, these models can be helpful in determining the optimal level of irrigation water that leads to maximizing income (e.g., [51]).

To make informed decisions that will enhance the efficiency of irrigation, farmers need to be able to assess how agricultural systems respond to internal (e.g., new technology) and external changes (such as those in the economic and political context, water constraints, or climate change). Therefore, DSS may be used for scenario analyses, showing the effects of alternative scenarios on irrigation management for efficient crop production [39]. The possibility of DSS implementation to assist farmers on irrigation management creates opportunities to establish a relationship that leads to the solution of problems and research feedback, redirecting the paths of research to better solve problems. Until now DSS are not commonly used directly by farmers as they are considered complex tools, which usually lack a user-friendly interface that permits easy access by the users. Irrigation advisory services of the irrigator's communities may be the right platform for the introduction of DSS, being potentially a major breakthrough in improving the use and management of irrigation water.

Management Under Water Scarcity

At present and more so in the future, irrigated agriculture will take place under water scarcity. Insufficient water supply for irrigation will be the norm rather than the exception, and irrigation management will shift from emphasizing production per unit area toward maximizing the production per unit of water consumed, the water productivity (WP).

Water Allocation Constraints and Their Impact on Management

While irrigation is an ancient technique, its expansion is very recent. The world area under irrigation has more than doubled in the last 60 years [4], in response to the increase in food demand. This expansion has required the development of additional water supply through the construction of dams for storing surface waters and exploitation of the groundwater resource. The sustainability of supply depends on the long-term rainfall and on the rate of groundwater recharge. As the irrigated areas expanded, the pressures on the finite water resources in some areas increased and the balance between supply and demand was altered (e.g., [52]). Two issues cause the imbalances; first, periodic drought cycles reduce the availability of water supply, some times during several years. Then, the increases in the demands from other sectors of society, notably the environment that has been neglected in the past, compete with irrigation demands, which are often considered the lowest in priority. The expansion of

groundwater use is also a cause of concern; it is possible that the abstraction exceeds the rate of recharge, causing a decline in the water table depth with time. In fact, groundwater may be considered a reservoir of supply that may be overexploited during drought years, when surface supplies are scarce. However, when the aquifers are depleted in the long run and do not recover after years of high rainfall, the groundwater overdraft is unsustainable and the abstraction must be reduced to sustainable levels.

Regardless of the causes for water scarcity, knowing the degree of supply reduction is essential for farmers to make rational decisions regarding how to manage the limited supplies. Preseason decisions are centered on crop choice and/or to land abandonment. Matching demand to supply is achieved by selecting low-water-use crops or by leaving some land in fallow, if the supplies are insufficient to irrigate all the developed land. Once the season starts it is much more difficult to make adjustments, if the reduction in supply is significant. Changes in the irrigation system to reduce percolation and other operational losses, changes in scheduling to reduce the number of applications thus reducing E losses, and the use of deficit irrigation (see below) are the only options left to growers once the season has started and the crops have been planted.

Water scarcity does not occur overnight, and water authorities and farmers have conservative attitudes to avoid risks. Predictions of big cuts in supply that do not materialize reduce economic opportunities, but the reverse may be even more catastrophic. Thus, planning in advance by water authorities and by irrigation districts, and knowing precisely the expected level of reduction are the two key elements to manage successfully the anticipated scarcity. Seasonal predictions of rainfall would be very useful to anticipate droughts and consequently, irrigation supply reductions. One paradox is the enormous investment in climate change research relative to that devoted to medium range weather predictions, and the apparent lack of connections between two areas that should be closely related.

Deficit Irrigation

Deficit irrigation (DI) is defined as the application of water below the crop ET requirements. Therefore, water demand for irrigation can be decreased relative to full irrigation and the water saved can be diverted for alternative uses. Even though DI is simply a technique aimed at the optimization of economic output when water is limited [53], the reduction of irrigation supply to an area imposes many adjustments in the agricultural system. Thus, DI practices are multifaceted, inducing changes at the technical, socioeconomical, and institutional levels.

In the humid and subhumid zones, irrigation supplements the rainfall as a tactical measure during drought spells to stabilize production. This practice has been called supplemental irrigation [54] and, although it uses limited amounts of water due to the relatively high rainfall levels, the goal is to achieve maximum yields and to eliminate yield fluctuations caused by water deficits. Supplementing rainfall in arid areas with one or more irrigation applications is a form of DI as maximum yields are not sought. When irrigation is applied at rates below the ET under DI, the crop extracts water from the soil reservoir to compensate for the deficit. Two situations may then develop. In one case, if sufficient water is stored in the soil and transpiration is not limited by soil water, even though the volume of irrigation water is reduced, the consumptive use (ET) is unaffected. However, if the soil water supply is insufficient to meet the crop demand, growth and transpiration are reduced and DI induces an ET reduction below its maximum potential. The difference between the two situations has important implications at the basin scale [55]. In the first case, DI does not induce net water savings and yields should not be affected. If the stored soil water that was extracted is replenished by seasonal rainfall, the DI practice is sustainable and has the advantage of reducing irrigation water use. In the second case, both water use and consumption (ET) are reduced by DI but yields may be negatively affected in cases where yields are directly related to ET [56].

There are several strategies to impose the water deficits under DI, but basically there are two alternatives [56]. One is to impose the same level of deficit over the entire irrigation season (continuous or sustained DI), while the other concentrates the deficits in certain crop growth stages believed to be the least sensitive to water stress (Regulated DI, RDI). Deficit irrigation, by reducing irrigation water use, can aid in coping with situations where supply is constrained. In field crops, a well-designed DI regime can optimize WP

over an area when full irrigation is not possible. It will reduce yield to a certain extent because of the linear relations between ET and the yield of the major field crops [40]. In many horticultural crops, such as fruit trees and vines, RDI has been shown to improve not only WP but farmers net income as well. Because of the differential responses among the different crops to water deficits it would be important to investigate the basis for the positive responses to water deficits in the cases where water deficits are not detrimental to yield. While DI can be used as a tactical measure to reduce irrigation water use when supplies are limited by droughts or other factors, it is not known whether it can be used over long time periods, given that the reduction in applied water could lead to greater accumulation of salts in the profile. It is imperative to investigate the sustainability of DI via long-term experiments and modeling efforts to determine to what extent it can contribute to the permanent reduction of irrigation water use.

Future Directions

Farmers in irrigated agriculture are confronted, at the start of every season with a critical question: How much water would be available this season, and how should I distribute it among the different crops and fields? There are many procedures and tools to answer those questions in such a way that the water allocation will be used efficiently. In fact efficiency of water use in irrigated agriculture has been steadily increasing with the improvements in science and technology, and as pressures from other sectors of society mount. Irrigation management encompasses several scales, from the network down to the individual field. One of the primary management targets at the field level is, once the amount of water needed is precisely determined, to distribute it over the field as uniformly as possible. Elimination of surface runoff and minimal percolation losses are prerequisites for optimizing irrigation water use and for limiting the environmental impacts of irrigation.

Monitoring, evaluation, and real-time feedbacks for benchmarking and to assess irrigation performance is essential for efficient water use. In the future, performance evaluation will be done routinely and at low cost with the use of remote sensing techniques. These surveys will allow the identification of areas within irrigation networks in need of improvement, and farmers will have the information to modify practices or to change methods, thus achieving greater productivities. Incentives are needed, however, for farmers to adopt new technologies for more efficient water use when water supplies are abundant and/or inexpensive.

The recent expansion of irrigation combined with increased water supply limitations will lead to water scarcity in many areas. In those situations, efficient use of water will be critical for the sustainability of irrigated agriculture. Deficit irrigation will be used more, and other socioeconomic measures, such as water markets, will play a more important role in water scarce situations [57]. Planning ahead in water-limited situations would be critical to achieve optimal use of water, and it is envisaged that robust DSS that include economic models will be used to allocate the limited irrigation water available among different users in networks and among crops in farms.

Finally, bridging the yield gap between potential and actual yields offers another avenue for improving the efficiency of water use in irrigated agriculture. Crops that are limited only by solar radiation and temperatures have potential yields that are several times the current world average yields. For instance, wheat world averages are reaching 3 t/ha, while the potential yield approaches 14 t/ha. In the case of maize, average world yield is about 5 t/ha while the potential yield exceeds 18 t/ha. The yield gap is not only important in rainfed conditions [58], but under irrigated conditions as well. Differences that exist between actual and potential yields are caused by many factors, water being just one of them. Therefore, it is most important to optimize crop agronomy in all of its facets, from soil to crop management, and from pest and disease management to weed control. Most of the time, yield improvement by better agronomy does not increase crop ET significantly, and it has proven over and over again to be a very effective path for enhancing the efficiency of water use now and in the near future.

Bibliography

Primary Literature

1. Fahlbusch H, Schultz B, Thatte CD (2004) The Indus basin: history of irrigation, drainage and flood management. ICID, New Delhi

2. Molden D (2007) Water for food, water for life: a comprehensive assessment of water management in agriculture. Earthscan/IWMI, London
3. Hsiao TC, Acevedo E, Fereres E, Henderson DW (1976) Water stress, growth, and osmotic adjustment. Philos Trans R Soc Lond B 273:479–500
4. FAOSTAT (2010) http://faostat.fao.org/. Verified on 26 May 2010
5. Tanji KK (1990) Nature and extent of agricultural salinity. In: Tanji KK (ed) Agricultural salinity assessment and management. ASCE, New York, pp 1–17
6. Clemmens AJ (2006) Improving irrigation water performance through an understanding of the water delivery process. Irrig Drain 55:223–234
7. Bos MG, Nugteren J (1990) On irrigation efficiencies, 4th edn. International Institute for Land Reclamation and Improvement (ILRI), Publication 19, Wageningen
8. Hsiao TC, Steduto P, Fereres E (2007) A systematic and quantitative approach to improve water use efficiency in agriculture. Irrig Sci 25:209–231
9. Jensen ME (2007) Beyond irrigation efficiency. Irrig Sci 25: 233–245
10. de Wit CT (1992) Resource use efficiency in agriculture. Agric Syst 40:125–151
11. Van Schilfgaarde J (1984) Drainage design for salinity control. In: Shainberg I, Shalhevet J (eds) Soil salinity under irrigation. Springer, New York, pp 190–197
12. Steduto P, Hsiao TC, Fereres E (2007) On the conservative behavior of biomass water productivity. Irrig Sci 25:189–207
13. Richards RA (2006) Physiological traits used in the breeding of new cultivars for water-scarce environments. Agric Water Manage 80:197–211
14. Salekdeh GH, Reynolds M, Bennett J, Boyer J (2009) Conceptual framework for drought phenotyping during molecular breeding. Trends Plant Sci 14:488–496
15. Shainberg I, Levy GJ (1996) Infiltration and seal formation processes. In: Agassi M (ed) Soil erosion, conservation, and rehabilitation. Marcel Dekker, New York, pp 1–22
16. FAO (2010) http://www.fao.org/ag/ca/. Verified on 26 May 2010
17. Derpsch R, Friedrich T (2009) Global Overview of Conservation Agriculture Adoption. Proceedings: 4th World Congress on Conservation Agriculture, 4–7 February 2009, New Delhi, pp 429–438
18. Thompson RB, Gallardo M, Agüera T, Valdez LC, Fernandez MD (2006) Evaluation of the watermark sensor for use with drip irrigated vegetable crops. Irrig Sci 24:185–202
19. Goldhamer DA, Fereres E (2001) Irrigation scheduling protocols using continuously recorded trunk diameter measurements. Irrig Sci 20:115–125
20. Jackson RD (1982) Canopy temperature and crop water stress. Adv Irrig 1:43–85
21. Allen RG, Pereira LS, Raes D, Smith M (1998) Crop evapotranspiration: guidelines for computing crop water requirements, FAO irrigation and drainage paper 56. FAO (Food and Agriculture Organization), Rome
22. Mantovani EC, Orgaz F, Villalobos FJ, Fereres E (1995) Modelling the effects of sprinkler irrigation uniformity on crop yield. Agric Water Manage 27:243–257
23. Benson SM, White AF, Halfman S, Flexser S, Alavi M (1991) Groundwater contamination at the Kesterson reservoir, California 1. Hydrogeologic setting and conservative solute transport. Water Resour Res 27:1071–1084
24. Hoffman GJ, Dirksen C, Ingvalson RD, Maas EV, Oster JD, Rawlins SL, Rhoades JD, Van Schilfgaarde J (1977) Minimizing salt in drain water by irrigation management: design and initial results of Arizona field studies. Agric Water Manage 1:233–252
25. Sadler EJ, Evans RG, Stone KC, Camp CR (2005) Opportunities for conservation with precision irrigation. J Soil Water Conservat 60:371–379
26. Evans RG, Sadler EJ (2007) New technologies to improve crop water use efficiencies [CD-ROM]. S164. Lawrence Media
27. Santos C, Lorite IJ, Tasumi M, Allen RG, Fereres E (2010) Performance assessment of an irrigation scheme using indicators determined with remote sensing techniques. Irrig Sci. doi:10.1007/s00271-010-0207-7
28. Zarco-Tejada PJ, Berni JAJ, Suárez L, Sepulcre-Cantó G, Morales F, Miller JR (2009) Imaging chlorophyll fluorescence from an airborne narrow-band multispectral camera for vegetation stress detection. Rem Sens Environ 113:1262–1275
29. Berni JAJ, Zarco-Tejada PJ, Sepulcre-Cantó G, Fereres E, Villalobos F (2009) Mapping canopy conductance and CWSI in olive orchards using high resolution thermal remote sensing imagery. Rem Sens Environ 113:2380–2388
30. Fountas S, Wulfsohn D, Blackmore BS, Jacobsen HL, Pedersen SM (2006) A model of decision-making and information flows for information-intensive agriculture. Agric Syst 87:192–210
31. Maton L, Leenhardt D, Goulard M, Bergez JE (2005) Assessing the irrigation strategies over a wide geographical area from structural data about farming systems. Agric Syst 86:293–311
32. Arnott D (2006) Cognitive biases and decision support systems development: a design science approach. Inf Syst J 16:55–78
33. Mohan S, Arumugam N (1997) Expert system applications in irrigation management: an overview. Comput Electron Agric 17:263–280
34. Goldberg DE (1989) Genetic algorithms in search, optimization and machine learning. Addison-Wesley, Madison
35. Azamathulla HMd, Wu FC, Ab Ghani A, Narulkar SM, Zakaria NA, Chang CK (2008) Comparison between genetic algorithm and linear programming approach for real time operation. J Hydro Environ Res 2:172–181
36. Kipkorir EC, Raes D, Labadie J (2001) Optimal allocation of short-term irrigation supply. Irrig Drain Syst 15:247–267
37. Bergez JE, Garcia F, Lapasse L (2004) A hierarchical partitioning method for optimizing irrigation strategies. Agric Syst 80: 235–253

38. Bazzani GM (2005) An integrated decision support system for irrigation and water policy design: DSIRR. Environ Model Softw 20:153–163
39. Stoorvogel JJ, Antle JM, Crissman CC, Bowen W (2004) The tradeoff analysis model: integrated bio-physical and economic modeling of agricultural production systems. Agric Syst 80:43–66
40. Stewart JI, Hagan RM (1973) Functions to predict effects of crop water deficits. J Irrig Drain Div 99:421–439
41. Doorenbos J, Kassam AH (1979) Yield response to water, FAO irrigation and drainage paper No. 33. FAO (Food and Agriculture Organization), Rome
42. Vaux HJ, Pruitt WO (1983) Crop-water production functions. In: Hillel DI (ed) Advances in irrigation, vol II. Academic, New York, pp 61–97
43. Loomis RS, Rabbinge R, Ng E (1979) Explanatory models in crop physiology. Annu Rev Plant Physiol 30:339–367
44. Stöckle CO, Donatelli M, Nelson R (2003) CropSyst, a cropping systems simulation model. Eur J Agron 18:289–307
45. Jones CA, Dyke PT, Williams JR, Kiniry JR, Benson CA, Griggs RH (1991) EPIC: an operational model for evaluation of agricultural sustainability. Agric Syst 37:341–350
46. Smith M (1992) CROPWAT: a computer program for irrigation planning and management, FAO irrigation and drainage paper No. 46. FAO (Food and Agriculture Organization), Rome
47. McCown RL, Hammer GL, Hargreaves JNG, Holzworth DP, Freebairn DM (1996) APSIM: a novel software system for model development, model testing and simulation in agricultural systems research. Agric Syst 50:255–271
48. Ritchie JT, Godwin DC, Otter-Nacke S (1985) CERES – wheat: a simulation model of wheat growth and development. Texas A&M University Press, College Station
49. Sinclair TR, Seligman NG (1996) Crop modeling: from infancy to maturity. Agron J 88:698–704
50. Steduto P, Hsiao TC, Raes D, Fereres E (2009) AquaCrop – the FAO crop model to simulate yield response to water: I. Concepts and underlying principles. Agron J 101:426–437
51. García-Vila M, Fereres E, Mateos L, Orgaz F, Steduto P (2009) Deficit irrigation optimization of cotton with AquaCrop. Agron J 101:477–487
52. Sophocleous M (2005) Groundwater recharge and sustainability in the high plains aquifer in Kansas, USA. Hydrogeol J 13:351–365
53. English MJ (1990) Deficit irrigation. I: analytical framework. J Irrig Drain Eng 116:399–412
54. Debaeke P, Aboudrare A (2004) Adaptation of crop management to water-limited environments. Eur J Agron 21:433–446
55. Fereres E, Goldhamer DA, Parsons LR (2003) Irrigation water management of horticultural crops. Historical review compiled for the American Society of Horticultural Science's 100th Anniversary. HortScience 38:1036–1042
56. Fereres E, Soriano MA (2007) Deficit irrigation for reducing agricultural water use. J Exp Bot 58:147–159
57. Jury AJ, Vaux HJ (2007) The emerging global water crisis: managing scarcity and conflict between water users. Adv Agron 95:1–76
58. Sadras VO, Angus JF (2006) Benchmarking water use efficiency of rainfed wheat in dry environments. Aust J Agric Res 57: 847–856

Books and Reviews

Burt CM, Clemmens AJ, Strelkoff TS, Solomon KH, Bliesner RD, Howell TA, Eisenhauer DE (1997) Irrigation performance measures: efficiency and uniformity, Journal of Irrigation and Drainage Engineering 123, No. 6. ASCE (American Society of Civil Engineers), New York, pp 423–442

Evans RG, Sadler EJ (2008) Methods and technologies to improve efficiency of water use: W00E04, Water Resources Research 44, No. 7. American Geophysical Union, Washington

Fereres E, González-Dugo V (2009) Improving productivity to face water scarcity in irrigated agriculture. In: Sadras VO, Calderini DF (eds) Crop physiology: applications for genetic improvement and agronomy. Academic, New York, pp 123–143

Lamm FR, Ayars JE, Nakayama FS (2007) Microirrigation for crop production, Developments in agricultural engineering 13. Elsevier, Amsterdam

Malano H, Burton M (2001) Guidelines for benchmarking performance in the irrigation and drainage sector. IPTRID/FAO, Rome

Molden D (2007) Water for food, water for life, A comprehensive assessment of water management in agriculture. Earthscan/IWMI, London

National Research Council (1996) A new era for irrigation. National Academy, Washington

Passioura JB, Angus JF (2010) Chapter 2 – improving productivity of crops in water-limited environments, Advances in agronomy 106. Academic, San Diego, pp 37–75

Isotope Separation Methods for Nuclear Fuel

Shuichi Hasegawa
Department of Systems Innovation, School of Engineering, The University of Tokyo, Tokyo, Japan

Article Outline

Glossary

Isotope Nuclei of a chemical element which have the same number of protons but different number of neutrons. Some isotopes are stable; some are radioactive.

Separation factor A ratio of a mole fraction of an isotope of interest to that of non-interest in an enriched flow divided by that in a depleted flow from a separation unit. The factor should be larger than unity for the unit to result in isotopic enrichment.

Separation capability A measure of separative work by a cascade per unit time.

Mean free path An average distance of a moving gas molecule between its collisions.

Molecular flow Low-pressure phenomenon when the mean free path of a gas molecule is about the same as the channel diameter; then a molecule migrates along the channel without interference from other molecules present.

Definition of the Subject

Isotope separation, in general, means enrichment of a chemical element to one of its isotopes (e.g., ^{10}B in B; ^{6}Li in Li, ^{157}Gd, etc). In the case of uranium, isotope separation refers to the enrichment in the isotope ^{235}U, which is only 0.711% of natural uranium; today's nuclear power plants require fuel enriched to 3–5% in ^{235}U. Uranium enrichment is the subject of this article.

Efficiencies of sorting out different isotopes of the element (separation factor) are usually very low. For practical enrichment plants, a gaseous diffusion process has been successfully employed to obtain enriched uranium. A gas centrifugation process is the preferred method of enrichment today due to reduced energy consumption. A new process using lasers, which can have a high efficiency of separation, is under development and has the potential to replace the current enrichment methods.

Introduction

The fuel used today by commercial nuclear power plants is the fissile isotope ^{235}U. Unfortunately, ^{235}U is only 0.711% of natural uranium, the rest of which is, essentially, ^{238}U. Light water reactors (LWR) operating dominantly all over the world require isotope enrichment processes because the isotopic ratio of ^{235}U for their fuels should be 3–5%. The processes used to elevate the ^{235}U content from 0.711% to 3–5% are called isotope separation or enrichment processes. Table 1 shows the current trends of isotope separation capabilities of the world. The main countries performing the process are Russia, France, US, and URENCO (Germany, Netherland, and UK). A number of separation processes have been studied so far, but the principles of the current isotope separation processes mainly use gaseous diffusion or gas centrifugation. The diffusion process was commercialized first but the centrifugation is taking over because of less energy consumption. This article following mainly [2, 3] describes the

Isotope Separation Methods for Nuclear Fuel. Table 1 World Enrichment capacity (thousand SWU/year) [1]

Country	2010	2015	2020
France (Areva)	8,500*	7,000	7,500
Germany, Netherlands, UK (Urenco)	12,800	12,200	12,300
Japan (JNFL)	150	750	1,500
USA (USEC)	11,300*	3,800	3,800
USA (Urenco)	200	5,800	5,900
USA (Areva)	0	>1,000	3,300
USA (Global Laser Enrichment)	0	2,000	3,500
Russia (Tenex)	23,000	33,000	30–35,000
China (CNNC)	1,300	3,000	6,000–8,000
Pakistan, Brazil, Iran	100	300	300
Total approx.	57,350	69,000	74–81,000
Requirements (WNA reference scenario)	48,890	55,400	66,535

Source: WNA Market Report 2009; WNA Fuel Cycle: Enrichment plenary session WNFC April 2011
*Diffusion

principles of the two processes and cascade theory, which explains why it is required to repeat the process many times (using successive stages/cascades) to obtain a certain desired enrichment fraction such as 3–5% because a single step provides only a small incremental enrichment. The new enrichment technology using lasers will be described at the end.

Principles of the Separation Processes

Gaseous Diffusion

Figure 1 shows the schematic diagram of the gaseous diffusion process. Consider a chamber divided into two compartments by a porous membrane. When dilute gases are introduced into the bottom compartment of the chamber, the pores of the membrane (membrane) make dependency of the transmission of the gases on their molecular masses.

If we have a mixture of two molecules in a gas with the same kinetic energy (kinetic energy is determined by kT, k = Boltzmann constant; T = temperature in K; $(1/2\ mv^2 \sim kT)$), the lighter molecule is faster than the heavier one. Therefore, their frequencies of hitting the membrane is higher for the lighter than for the heavier molecule. However, the mass preference phenomena occur only when the mean free path of the gas molecule is longer than the diameter of the pores $2r$ and the thickness of the membrane l. The mean free path, λ of the molecule can be written as [2]

$$\lambda = \frac{kT}{4\sqrt{2}\pi\sigma^2 p} \tag{1}$$

where k is the Boltzmann constant, T is the absolute temperature, σ is the radius of the molecule, and p is the gas pressure in the chamber. In this condition, a molecule cannot collide with others during the transmission through the membrane so that its dynamics can be considered as a single molecule process. This process is called molecular flow. The flux of the molecular flow through the flow path with circular cross section is derived by Knudsen as [3]

$$G_{mol} = \frac{8r\Delta p}{3l\sqrt{2\pi mRT}} \tag{2}$$

where G_{mol} is the molecular flow velocity, m is the molecular mass, R is the gas constant, and $\Delta p = p'' - p'$ is the pressure difference between the bottom and top compartments of the chamber. Equation 2 shows that the flow velocity depends on the mass of the gas molecules so that the ratio of the molecules in the mixture transmitted to the upper compartment of the chamber is changed compared with that of the feeding gas. The opposite condition where flows do not depend on the molecular mass is called viscous flow.

We will derive the ideal separation factor in the case of $^{235}UF_6$ and $^{238}UF_6$ [3], the gas molecules used for uranium enrichment. On the ideal condition where p'' is very small and p′ can be neglected compared with p'', when we have a binary mixture of gases which consist of $^{235}UF_6$ (molecular mass: m_{235} = 349, mole fraction: x) and $^{238}UF_6$ (molecular mass: m_{235} = 352, mole fraction: $1 - x$), the molecular flow velocities of $^{235}UF_6$ and $^{238}UF_6$ are

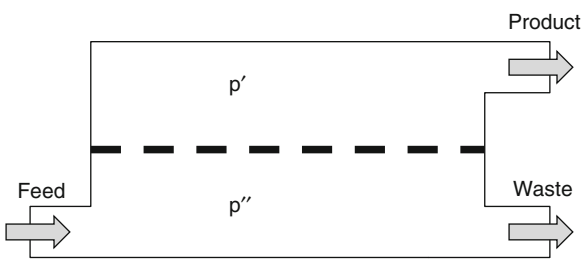
$$G_{235} = \frac{ap''x}{\sqrt{m_{235}}}, \quad G_{238} = \frac{ap''(1-x)}{\sqrt{m_{238}}} \tag{3}$$

where the constant a includes factors in Eq. 2. The ratio of the molecular flow of $^{235}UF_6$ to the whole can be written as

$$s = \frac{G_{235}}{G_{235} + G_{238}} = \frac{\frac{x}{\sqrt{m_{235}}}}{\frac{x}{\sqrt{m_{235}}} + \frac{1-x}{\sqrt{m_{238}}}} = \frac{\frac{x}{1-x}}{\frac{x}{1-x} + \sqrt{\frac{m_{235}}{m_{238}}}} \tag{4}$$

Therefore, the ideal separation factor α_0 of the gaseous diffusion process can be derived as the separation factor of the molecular flow of the porous media

$$\alpha_0 = \frac{\frac{s}{1-s}}{\frac{x}{1-x}} = \sqrt{\frac{m_{238}}{m_{235}}} = \sqrt{\frac{352}{349}} = 1.00429 \tag{5}$$



Isotope Separation Methods for Nuclear Fuel. Figure 1
A single gaseous diffusion stage

The separation factor depends on the ratio of the molecular masses so that this method is more effective for the isotope separation of lighter elements. For heavier elements, a larger number of repeated processes is required to obtain sufficiently enriched products.

However, in reality, the real value of the separation factor is smaller than that given by Eq. 5 due to reverse molecular flow from the upper compartment to the bottom one and viscous flow not depending on the molecular mass; these two phenomena work in a direction negating the enrichment process. Furthermore operating conditions (porous media performance, working pressures, etc.) affect the value of the separation factor. The energy consumption to run the process is very high due to pressure controlling of the gases, small separation factors and so on (see discussion about Separative Work Unit). Because of the relatively high energy consumption, uranium enrichment by gaseous diffusion is on the way out and is replaced by the gas centrifugation method.

Gas Centrifugation

The principle of gas centrifugation is based upon centrifugal forces that are created inside a rotating cylinder containing two different gas molecules, forces that depend on the molecular mass. Let's see how it works in detail [2]. When we have a mixture of two gas molecules in a rotating cylinder (centrifuge), pressure gradients develop with respect to the radial direction. The pressures can be written as

$$\frac{dp}{dr} = \omega^2 r\rho \tag{6}$$

where p is the pressure, r is the radial distance, ω is the angular frequency of rotation, and ρ is the density of the gases. By substituting the equation of state $\rho = pm/RT$ into the differential equation, we can derive the following equation,

$$\frac{dp}{p} = \frac{m\omega^2}{RT} rdr \tag{7}$$

When we integrate this differential equation from the radial distance r (pressure p_r) to the inner radius of the cylinder a (pressure p_a), we can obtain this expression,

$$\frac{p_r}{p_a} = \exp\left[-\frac{1}{2}\frac{mv_a^2}{RT}\left\{1-\left(\frac{r}{a}\right)^2\right\}\right] \tag{8}$$

where the speed of the outer circumference of the cylinder $v_a = \omega a$ This equation shows that the ratio of the pressure at radius r to that of radius a depends on the molecular mass of the gases.

If we have the gases which consist of $^{235}UF_6$ (molecular mass: m_{235} = 349, mole fraction: x) and $^{238}UF_6$ (molecular mass: m_{235} = 352, mole fraction: $1 - x$), their ratios of the partial pressures at the radius r to the radius a can be derived as

$$\frac{p_r x_r}{p_a x_a} = \exp\left[-\frac{1}{2}\frac{m_{235}v_a^2}{RT}\left\{1-\left(\frac{r}{a}\right)^2\right\}\right] \tag{9}$$

$$\frac{p_r(1-x_r)}{p_a(1-x_a)} = \exp\left[-\frac{1}{2}\frac{m_{238}v_a^2}{RT}\left\{1-\left(\frac{r}{a}\right)^2\right\}\right] \tag{10}$$

Therefore, the local separation factor at radial distance r of radius a is given by

$$\alpha = \frac{\frac{x_r}{1-x_r}}{\frac{x_a}{1-x_a}} = \exp\left[\frac{(m_{238}-m_{235})v_a^2}{2RT}\left\{1-\left(\frac{r}{a}\right)^2\right\}\right] \tag{11}$$

which depends on the difference of their molecular masses, $\Delta m = m_{238} - m_{235} = 3$. Values of the local separation factor of $^{235}UF_6$ and $^{238}UF_6$ with T = 300 K and v_a = 700 m/s are given in Table 2. This feature is superior to the gaseous diffusion method when the difference of the masses is large, (e.g., for heavier elements). The separation factor increases as the speed of the outer circumference increases. However, the maximum speed v_{max} is limited by stresses created to the cylinder from the force of the centrifugation and can be written as [2]

$$v_{max} = \sqrt{\frac{\sigma}{\rho}} \tag{12}$$

where ρ is the density of the material of the cylinder, and σ is the tensile strength. Although most molecular

Isotope Separation Methods for Nuclear Fuel. Table 2
The local separation factor of $^{235}UF_6$ and $^{238}UF_6$ at r/a with T = 300K, v_a = 700m/s

r/a	0	0.5	0.8	0.9	0.95	0.98	0.99	1.0
α	1.343	1.247	1.112	1.058	1.029	1.012	1.006	1.0

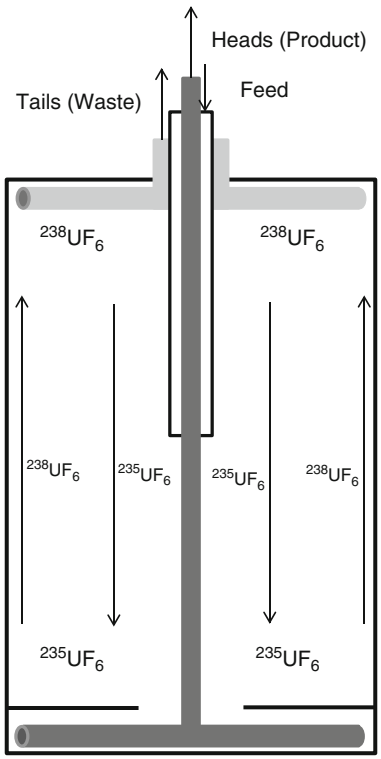
gases are localized at $\frac{r}{a} \approx 1$ because of the centrifugation, the values of the separation factor could be higher than those obtained from the gaseous diffusion method. These values can be enhanced if we make use of a countercurrent flow in the vertical direction. Figure 2 shows the schematic diagram of countercurrent centrifugation method. Gernot Zippe performed pioneering work on the development of the centrifugation first in the Soviet Union during 1946–1954, and from 1956 to 1960 at the University of Virginia. The countercurrent flow can be induced by heating and cooling centrifuges, or pipes drawing off flows in centrifuges. The temperature control can adjust the flow deliberately but the equipment becomes more complicated than that of the flow control by the pipes (Fig. 2). This countercurrent flow makes enrichment of the lighter isotopes at inner radius as the flow descending along the axis direction, and the heavier isotopes are being enriched at the circumference as the flow ascending. These enriched gases are collected at different radial positions of the both ends (at outer radius for heavier isotope and at inner radius by baffle for lighter isotope). When the centrifuge has a length L, the maximum separative power δU_{max} can be derived as [2, 4]

$$\delta U_{max} = \frac{\pi}{2} L \rho D \left(\frac{\Delta m v_a^2}{2RT} \right)^2 \tag{13}$$

where D is diffusion coefficient. The maximum separative power is proportional to the height of the centrifuge. It is preferable to have a taller centrifuge in the vertical direction, but the length is imposed on the resonant vibration of the centrifuge. The resonant conditions can be written as [3]

$$\left(\frac{L}{a} \right)_i = \sqrt{\lambda_i} \sqrt[4]{\frac{E}{2\sigma'}} \quad \lambda_i = 22.0, 61.7, 121.0, 200.0, 298.2, \cdots \tag{14}$$

where E is coefficient of elasticity. A taller centrifuge can give a larger separative power although excellent mechanical properties are required to overcome the resonant conditions.



Isotope Separation Methods for Nuclear Fuel. Figure 2 Schematics of gas centrifuge with countercurrent flow

Cascade Theory

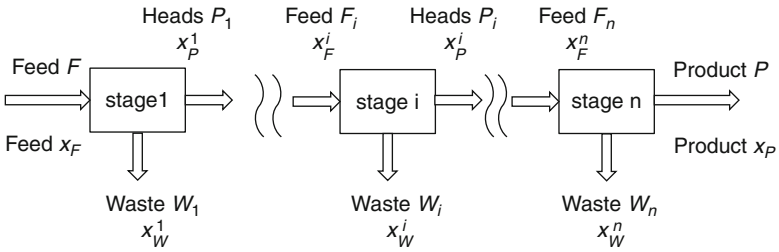
The present isotope separation plants make use of these principles of enrichment with small separation factors. In order to obtain high enrichment ratios, cascade theory is necessary [3]. According to the theory, we can enhance the ratios by iterating a single physical stage many times. Figure 3 shows a simple scheme of a cascade. An original material "feed" is provided to the system. The isotope of interest is enriched as going through many separation stages and a final output "product" is obtained. Another output which mainly contains unnecessary isotopes is called "waste." Each flow F, P and W should have the following equation

$$F = P + W \tag{15}$$

and with mole fractions of the isotope of interest in each flow x_F, x_P, x_W, we can obtain

$$Fx_F = Px_P + Wx_W \tag{16}$$

In this system, we have four independent parameters to define. In order to obtain necessary flow of Product "P" and mole fraction "x_P" of the isotope of interest, we need the design methodology to construct stages of separation units. The product of



Isotope Separation Methods for Nuclear Fuel. Figure 3
Simple scheme of the cascade

a single stage (unit) is called heads and the waste of that is tails. The ratios of the isotope of interest in the product are usually most important. If, for instance, we have two isotopes "1" and "2," and want to enrich the "1" isotope, we would focus on the variation of the mole fraction ratio of the two isotopes, $\frac{x_1}{x_2}$, which can be rewritten as $\frac{x_1}{1-x_1}$. The capability of each enrichment unit is described as separation factor α. This factor is defined as the ratios of the isotopes of interest to that of not-interest in the heads (product) divided by those in the tails (waste)

$$\alpha = \frac{\frac{x_P}{1-x_P}}{\frac{x_W}{1-x_W}} \tag{17}$$

In a similar way, we can define the ratio of the heads (product) to the feed as heads separation factor β, and that of the feed to the tails as tails separation factor γ,

$$\beta = \frac{\frac{x_P}{1-x_P}}{\frac{x_F}{1-x_F}},\ \gamma = \frac{\frac{x_F}{1-x_F}}{\frac{x_W}{1-x_W}} \text{ and } \alpha = \beta\gamma \tag{18}$$

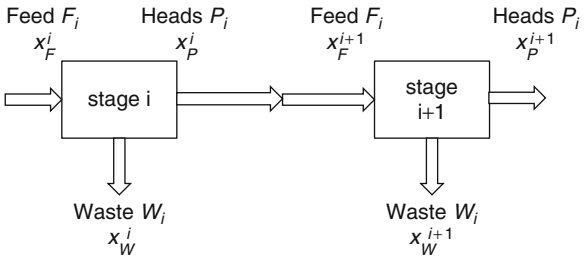
The ratio of the product to the feed is called "cut" θ and defined as

$$\theta \equiv \frac{P}{F} = \frac{x_F - x_w}{x_P - x_W} \tag{19}$$

The simplest design to accomplish enrichment is to accumulate separation stages in a single line such as Fig. 4. This scheme is called simple cascade.

Simple Cascade

In this scheme, the heads and the mole fraction of the i th stage are equal to the feed flow and the mole fraction of the $i+1$ th stage (Fig. 4).



Isotope Separation Methods for Nuclear Fuel. Figure 4
Simple cascade of the i and i + 1 th stages

$$F_{i+1} = P_i, \quad x_F^{i+1} = x_P^i \tag{20}$$

This cascade disposes of the tails of all stages so that the total amount of the isotope of interest in the waste should be given sufficient attention. This can be evaluated by means of the recovery rate of the i th stage r_i

$$r_i = \frac{P_i x_P^i}{F_i x_F^i} = \theta_i \frac{x_P^i}{x_F^i} = \frac{x_F^i - x_W^i x_P^i}{x_P^i - x_W^i x_F^i} = \frac{1 - \frac{x_W^i}{x_F^i}}{1 - \frac{x_W^i}{x_P^i}} = \frac{\alpha_i - \beta_i}{\alpha_i - 1} \tag{21}$$

When we have n stages in the cascade, the total recovery rate r can be expressed as

$$r = \frac{P x_P}{F x_F} = \frac{P_n x_P^n}{F_1 x_F^1} = \frac{P_1 x_P^1 P_2 x_P^2}{F_1 x_F^1 F_2 x_F^2} = \frac{P_n x_P^n}{F_n x_F^n} = r_1 r_2 \cdots r_n \tag{22}$$

The over-all separation factor of the cascade ω can be derived as

$$\omega = \frac{\frac{x_P^n}{1-x_P^n}}{\frac{x_F^1}{1-x_F^1}} = \beta_1 \beta_2 \cdots \beta_n \tag{23}$$

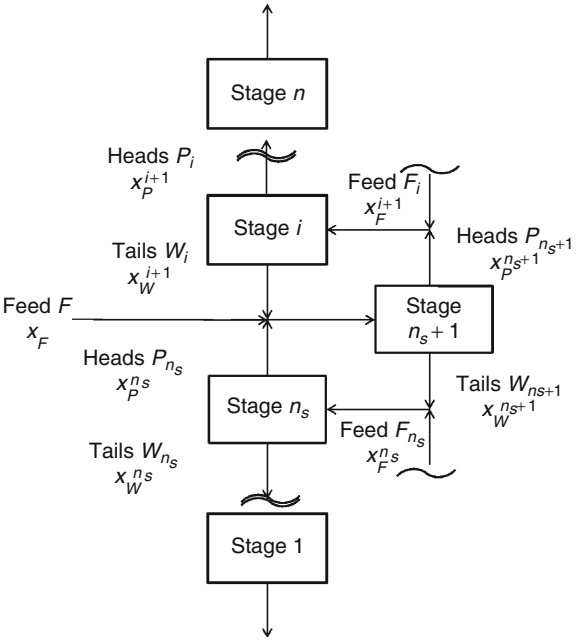
Therefore, if α, β do not depend on each stage, the total recovery rate can be rewritten as

$$r = \left(\frac{\alpha-\beta}{\alpha-1}\right)^n = \left(\frac{\alpha-\omega^{\left\{\frac{1}{n}\right\}}}{\alpha-1}\right)^n \tag{24}$$

When the feed itself is available without any special cost, the simple cascade is effective. But in case the wastes from each stage should not be disposed because, for instance, it is valuable or the recovery rate has to be increased, the waste flows are recycled as feed flow, which is called countercurrent recycle cascade (Fig. 5).

Countercurrent Recycle Cascade

Since the simple cascade cannot improve the recovery rate, the tail flow is recycled into either stage to use it efficiently, which is called recycle cascade (Fig. 5). If β (heads separation factor) is equal to γ (tails separation factor) in all stages, we can obtain $x_W^{i+2} = x_F^{i+1} (= x_P^i)$. So the tails flow of the $i+2$ th stage can be merged to the heads flow of the i th stage and fed into the $i+1$ th stage without any mixing loss. We will consider the case that the tails flow of the second upper stage is refluxed to the i th stage.



Isotope Separation Methods for Nuclear Fuel. Figure 5
Countercurrent recycle cascade

The flows and the fractions of the isotope of interest in each stage of enriching sections should have the following relationships.

$$P_i = W_{i+1} + P, \quad P_i x_P^i = W_{i+1} x_w^{i+1} + P x_P \tag{25}$$

In a similar way, those in stripping sections can be expressed as

$$W_{j+1} = P_j + W, \quad W_{j+1} x_W^{j+1} = P_j x_P^j + W x_W \tag{26}$$

Let's estimate the number of stages. From these equations, we can derive

$$x_P^i - x_W^{i+1} = \frac{x_P - x_P^i}{\frac{W_{i+1}}{P}} \tag{27}$$

At total reflux, where the reflux ratio is infinity,

$$\frac{W_{i+1}}{P} \to \infty \tag{28}$$

the mole fraction of the heads flow at the i th stage x_P^i becomes equal to that of the tails flow at the i + 1 th stage x_W^{i+1} and the number of the stages is minimal.

$$\frac{x_P^{i+1}}{1-x_P^{i+1}} = \alpha \frac{x_W^{i+1}}{1-x_W^{i+1}} = \alpha \frac{x_P^i}{1-x_P^i} = \alpha^2 \frac{x_P^{i-1}}{1-x_P^{i-1}} = \cdots \tag{29}$$

gives the following equation,

$$\frac{x_P}{1-x_P} = \alpha^n \frac{x_W}{1-x_W} \tag{30}$$

and the minimum number of the stages at total reflux can be derived as

$$n = \frac{1}{\ln \alpha} \ln\left(\frac{x_P}{1-x_P}\frac{1-x_W}{x_W}\right) \tag{31}$$

On the contrary, the reflux ratio becomes minimum when the mole fraction of the heads at the i + 1 th stage is equal to that of the heads at the i th stage $(x_P^{i+1} = x_P^i)$.

Ideal Cascade

Ideal cascade satisfies the condition that the values of β (heads separation factor) at all stages are constant and the mole fraction of the heads flow at the $i+1$ th stage is equal to those of the tails flow at the $i-1$ th stage and

of the feed flow at the i th stage ($x_p^{i+1} = x_W^{i-1} = x_F^i$). In this instance, each separation factor satisfies the following relationship.

$$\beta = \sqrt{\alpha} = \gamma \tag{32}$$

In a similar way to the previous section, we can obtain the total number of the stages for an ideal cascade

$$\begin{aligned} n &= \frac{1}{\ln \beta} \ln \left(\frac{x_P}{1 - x_P} \frac{1 - x_W}{x_W} \right) - 1 \\ &= \frac{2}{\ln \alpha} \ln \left(\frac{x_P}{1 - x_P} \frac{1 - x_W}{x_W} \right) - 1 \end{aligned} \tag{33}$$

The number of stages in stripping n_S and enriching $n_E = n - n_S$ sections can be derived as

$$n_S = \frac{1}{\ln \beta} \ln \left(\frac{x_F}{1 - x_F} \frac{1 - x_W}{x_W} \right) - 1 \tag{34}$$

$$n_E = n - n_S = \frac{1}{\ln \beta} \ln \left(\frac{x_P}{1 - x_P} \frac{1 - x_F}{x_F} \right) \tag{35}$$

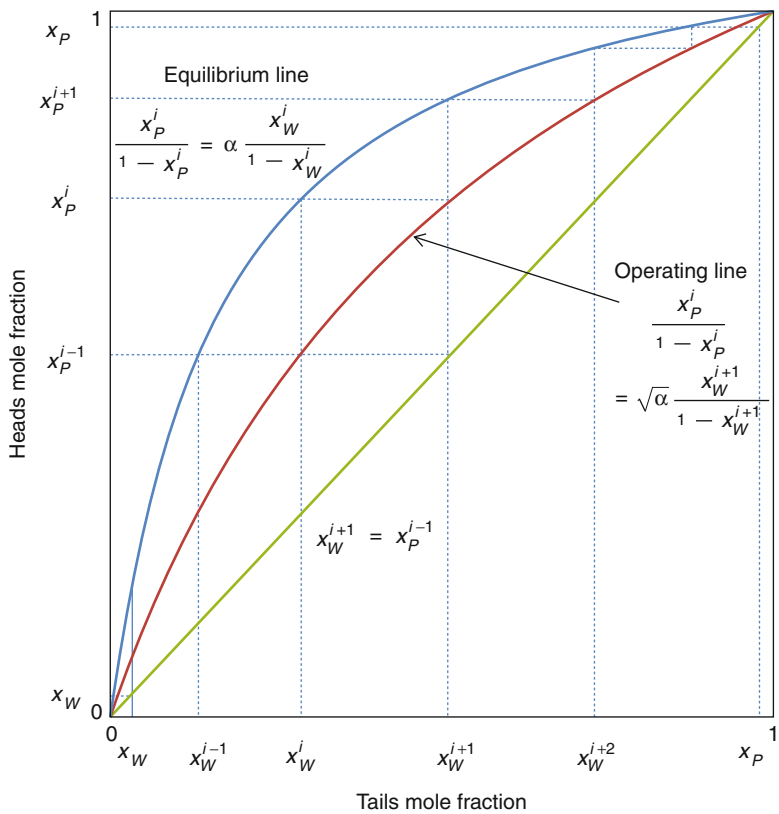
The reflux ratio Eq. 27 can be rewritten using $x_P^i = x_F^{i+1}$ and β as

$$\frac{W_{i+1}}{P} = \frac{x_P - x_P^i}{x_P^i - x_W^{i+1}} = \frac{1}{\beta - 1} \left\{ \frac{x_P}{x_W^{i+1}} - \frac{\beta(1 - x_P)}{1 - x_W^{i+1}} \right\} \tag{36}$$

Mccabe–Thiele Diagram

It is useful to draw McCabe–Thiele diagram to investigate the design of the cascade, the mole fractions of the stages and so on. Figure 6 shows a typical McCabe–Thiele diagram. In this graph, the horizontal and vertical axes correspond to the mole fractions of the heads flow x_P^i and of the tails flow x_W^i, respectively.

First, the following equation is satisfied at the enrichment process of the i th stage because of the definition of the separation factor



Isotope Separation Methods for Nuclear Fuel. Figure 6
McCabe-Thiele diagram

$$\frac{x_P^i}{1-x_P^i} = \alpha \frac{x_W^i}{1-x_W^i} \text{(Equilibrium line)} \tag{37}$$

Second, the condition that the tail (waste) flow at the i + 1 th stage is the feed of the i th stage ($x_F^i = x_W^{i+1}$) defines the relationship between the mole fractions of the tail (waste) and head (product) flows at different stages as follows

$$\frac{x_P^i}{1-x_P^i} = \beta \frac{x_F^i}{1-x_F^i} = \sqrt{\alpha} \frac{x_W^{i+1}}{1-x_W^{i+1}} \text{(Operating line)} \tag{38}$$

And third, the feed flow at the i th stage consists of the tails flow of the i + 1 th stage and the heads flow of the i − 1 th stage and their mole fractions are the same.

$$x_W^{i+1} = x_P^{i-1} \tag{39}$$

These three formulae can be shown in the McCabe–Thiele diagram as shown in Fig. 6. We can estimate the number of necessary stages, mole fractions of the stages, and overview the total processes through the graphical construction.

Separative Work Unit

The total flow in the cascade can be derived as

$$\sum_i (P_i + W_i) = \frac{\beta+1}{(\beta-1)\ln\beta}\left[W(2x_W-1)\ln\left(\frac{x_W}{1-x_W}\right) + P(2x_P-1)\ln\left(\frac{x_P}{1-x_P}\right) - F(2x_F-1)\ln\left(\frac{x_F}{1-x_F}\right)\right] \tag{40}$$

The first term of Eq. 40 including β indicates the difficulty of the separation and increases as the value of β approaches to unity. The second term corresponds to the amount of work for separation, and it has the same dimension as flow rates and is called separative capacity or separative power. This value is important because it is considered to be proportional to the initial cost of the plant. When we use the unit of the amounts of material (mole, kg, etc.) instead of flow rates, this is called separative work. The sum of the annual investment and operation costs can be expressed by the product of the separative work SW (kg SWU/year) and unit price of separative work c_s ($/kg SWU). SWU is the abbreviation of Separative Work Unit. The separative work is defined as

$$SW = W\phi(x_W) + P\phi(x_P) - F\phi(x_F) \tag{41}$$

where $\phi(\mathrm{x_i})$ is called separation potential and written as

$$\phi(x_i) = (2x_i - 1)\ln\frac{x_i}{1-x_i} \tag{42}$$

When we use kg SWU/year for the separative work, the unit of W, P, and F should be kg/year.

For operating the plant, we need the raw materials, the amount of which is F (kg/year) and unit price of the raw materials c_F ($/kg). The total cost per year c ($) can be written as

$$c = SWc_S + Fc_F \tag{43}$$

When the amount of the product per year is P (kg), the unit cost of the product $c_P = \frac{c}{P}$. could be derived as

$$\begin{aligned} c_P &= \frac{SWc_s}{P} + \frac{Fc_F}{P} \\ &= \left\{(\phi(\mathrm{x_P}) - \phi(\mathrm{x_F})) - (x_P - x_F)\frac{\phi(\mathrm{x_F}) - \phi(x_W)}{x_F - x_W}\right\}c_s \\ &\quad + \left(\frac{x_P - x_W}{x_F - x_W}\right)c_F \end{aligned} \tag{44}$$

Example

With the ideal cascade of the gaseous diffusion method (α = 1.00429), the mole fraction of the feed flow 0.711% (x_F = 0.00711) would be enriched to 3% (x_P = 0.03) and the mole fraction of the waste is planned to be 0.3% (x_W = 0.003). In this case, the necessary moles of the feed and the waste to obtain the product of 1 [mol] are

$$F = \frac{P(x_P - x_W)}{x_F - x_W} = \frac{1\times(0.03-0.003)}{0.00711-0.003} = 6.569[mol]$$

$$\begin{aligned} \mathrm{W} &= \frac{P(x_P - x_F)}{x_F - x_W} = \frac{1\times(0.03-0.00711)}{0.00711-0.003} \\ &= 5.569(=6.569-1)[mol] \end{aligned}$$

The total number of stages n and the number of stages in stripping section n_S and in enriching section n_E are calculated as

Stripping Section

$$n_S = \frac{2}{\ln\alpha}\ln\left(\frac{x_F}{1-x_F}\,\frac{1-x_W}{x_W}\right) - 1$$
$$= \frac{2}{\ln 1.00429}\ln\left(\frac{0.00711}{1-0.00711}\,\frac{1-0.003}{0.003}\right) - 1 = 404$$

Enriching Section

$$n_E = \frac{2}{\ln\alpha}\ln\left(\frac{x_p}{1-x_p}\,\frac{1-x_F}{x_F}\right)$$
$$= \frac{2}{\ln 1.00429}\ln\left(\frac{0.03}{1-0.03}\,\frac{1-0.00711}{0.00711}\right) = 683.5$$

The total number of stages

$$n = \frac{2}{\ln\alpha}\ln\left(\frac{x_p}{1-x_p}\,\frac{1-x_W}{x_W}\right) - 1$$
$$= \frac{2}{\ln 1.00429}\ln\left(\frac{0.03}{1-0.03}\,\frac{1-0.003}{0.003}\right) - 1 = 1087.5$$

The heads flow rate in the enriching section can be written as

$$P_i = P + W_{i+1}$$
$$= P + \frac{P}{\beta-1}\{x_p(1-\beta^{i-n}) + (1-x_p)\beta(\beta^{n-i}-1)\}$$

and that in the stripping section

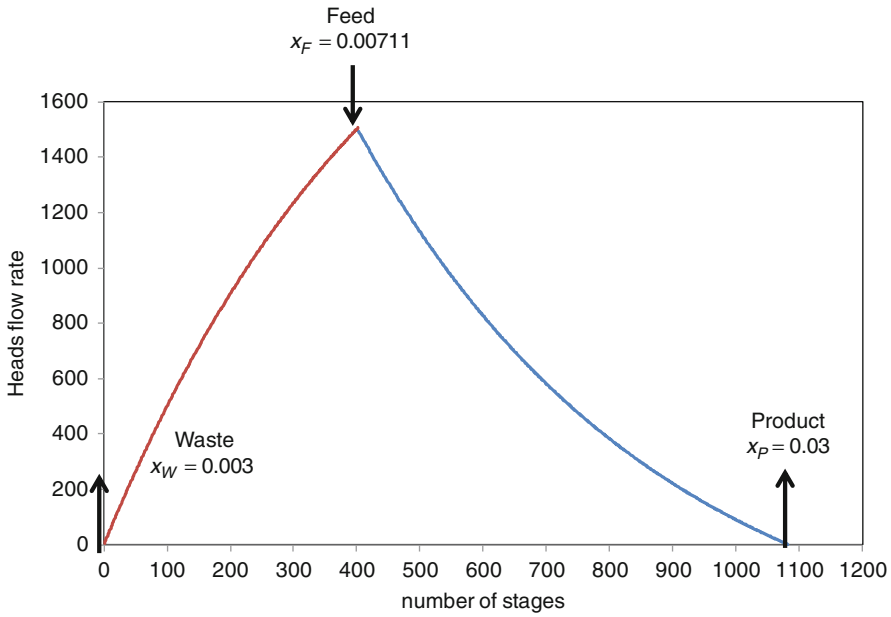
$$P_i = \frac{W}{\beta-1}\{x_W\beta(\beta^i-1) + (1-x_W)(1-\beta^{-i})\}$$

These flows as a function of the number of the stages can be shown as Fig. 7 in this example.

When we need higher concentration, such as 5%, $F = 11.436[mol]$, $W = 10.436[mol]$, $n = 1336$ and $n_E = 932$.

Laser Isotope Separation (LIS)

The photon absorbing frequencies of isotopes show small differences caused by shifts of atomic electron energies due to the differences in the number of neutrons among isotopes. This is called isotope shift. The invention and development of lasers enable to resolve the isotope shift sufficiently and make isotope-selective photo-chemical reaction possible. Laser Isotope Separation may lead to almost 100% isotope separation in a single stage. Mainly, two methods such as Atomic Vapor Laser Isotope Separation (AVLIS) and Molecular Laser Isotope Separation (MLIS) were intensively studied. AVLIS uses uranium atomic vapor that is struck by lasers of such wavelength that only ^{235}U atoms are excited and then ionized; once ionized, the



Isotope Separation Methods for Nuclear Fuel. Figure 7
Heads flow rate

^{235}U ions are collected by an electromagnetic field. MLIS uses UF_6, and vibrationally excites and multiphoton-dissociates only $^{235}UF_6$ into $^{235}UF_5$ by infrared lasers. The research to commercialize them has faded on a global scale.

A new process called Separation of Isotopes by Laser Excitation (SILEX) is under development. All details are not out in the open yet; but SILEX is considered to be a kind of molecular LIS using UF_6. The method only isotope-selectively excites but not dissociates $^{235}UF_6$. The separation factor announced by the company has been 2–20 [5]. Silex Systems Ltd was originally established as a subsidiary of Sonic Healthcare Limited of Australia in 1988. In 2007, the SILEX Uranium Enrichment project was transferred to GE's nuclear fuel plant in the United States. Global Laser Enrichment (GLE) was formed as a subsidiary of GE-Hitachi in 2008 [5]. In June 2009, GE-Hitachi submitted a license application to construct a commercial laser enrichment plant in Wilmington, NC. The NRC staff is currently reviewing that application. They announced that they succeeded the initial measurement program at Test Loop in 2010 and proceeded to evaluate the program to decide the commercialization of the process [6].

Future Directions

As of today, the gaseous diffusion and centrifuge processes have been used on a commercial scale. For the future, it seems that laser enrichment (the SILEX process) may be the successor to current enrichment methods. Preliminary results, based on enrichment by lasers, are encouraging. However, considerable improvements are needed before this method achieves commercial competitive status. Every uranium enrichment process is linked to nuclear proliferation issues. It would be very beneficial for the world if a method of enrichment is devised which inherently offers non- proliferation safeguards for nuclear materials.

Bibliography

1. World Nuclear Association, Uranium Enrichment, World Enrichment capacity - operational and planned. http://www.world-nuclear.org/info/inf28.html
2. Villani S (1976) Isotope separation. American Nuclear Society, Hillsdale
3. Benedict M, Pigford TH (1957) Nuclear chemical engineering. Mcgraw-Hill, New York; Benedict M, Pigford TH, Levi HW (1981) Nuclear chemical engineering (second edn.), (trans: by Kiyose R into Japanese)
4. Kemp RS (2009) Gas centrifuge theory and development: a review of U.S. programs. Science and Global Security 17, 1; Wood HG, Glaser A, Kemp RS (2008) The gas centrifuge and nuclear weapons proliferation. Physics Today 40
5. Silex Systems Limited home page. http://www.silex.com.au/
6. World Nuclear News (2010) Initial Success from SILEX test loop, 12 April 2010. http://www.world-nuclear-news.org/NN-Initial_success_from_SILEX_test_loop-1204104.html