*Journal of Land Use Science*
Vol. 6, No. 4, December 2011, 245–259



# A new hybrid land cover dataset for Russia: a methodology for integrating statistics, remote sensing and in situ information

Dmitry Schepaschenko[a,b]*, Ian McCallum[a], Anatoly Shvidenko[a,c], Steffen Fritz[a], Florian Kraxner[a] and Michael Obersteiner[a]

[a]*IIASA, Laxenburg, Austria;* [b]*Moscow State Forest University, Moscow, Russian Federation;* [c]*Institute of Forest, Siberian Branch, Russian Academy of Science, Krasnoyarsk, Russian Federation*

(*Received 18 March 2009; final version received 24 March 2010*)

Despite being recognized as a key baseline dataset for many applications, especially those relating to biogeochemical cycles, land cover products in their current form are limiting. Typically they lack the thematic detail necessary for driving the models that depend upon them. This study has demonstrated the ability to produce a highly detailed (both spatially and thematically) land cover/land use dataset over Russia – by combining existing datasets into a hybrid information system. The resulting dataset contains detailed subclasses of land cover and attributes necessary for biogeochemical modeling. In lieu of suitable validation data, a confidence map was produced creating six classes of confidence in the agreement between the various remote sensing and statistical datasets. In specific regions, a significant difference between the remote sensing products and the official statistics was observed. For example, in the northwest of Russia the statistics appear to be underreporting the amount of forest land which has likely been increasing in recent decades because of encroachment of forests on abandoned marginal agricultural land.

**Keywords:** land cover; land use; remote sensing; GIS; inventory statistics; Russia

## 1. Introduction

Land cover is recognized as one of the fundamental terrestrial datasets required in biospheric studies across the globe. In many aspects, land cover information provides the foundation for environmental monitoring (FAO 2002) and diverse ecosystem studies (e.g., land use competition, food security). In recent years, major advances have taken place and researchers now have several global 1 km products available to choose from, each with approximately 20 thematic classes (McCallum, Obersteiner, Nilsson, and Shvidenko 2006). Additionally, a 500 m global product exists from the moderate-resolution imaging spectroradiometer (MODIS) (Friedl *et al*. 2002) and a 300 m global product (GlobCover) from the medium-resolution imaging spectrometer (MERIS) has recently been made available (Bicheron *et al*. 2008). In addition to global datasets, continental and regional products exist around the globe (Bartalev, Belward, Erchov, and Isaev 2003).

However, for many applications, especially those relating to biogeochemical cycles, these products in their current form are limiting. In particular, practically all existing land cover datasets temporally and thematically inhibit progress in accurate modeling of the

*Corresponding author. Email: schepd@iiasa.ac.at

ISSN 1747-423X print/ISSN 1747-4248 online

http://dx.doi.org/10.1080/1747423X.2010.511681
http://www.tandfonline.com

terrestrial biosphere. Temporally, the majority of datasets are restricted to a certain time period with no updates planned (MODIS being the exception). Thematically, the lack of detail among the land cover classes identified prevents the use of valuable inventory and statistical data. Tree species in particular, are difficult to classify from remote sensing data alone and therefore the majority of land cover products produced to date stop short of delineating tree species and related attributes. However, forest science has accumulated many data and semi-empirical models of forest growth and productivity. These models contain such attributes as tree species, age, site index, etc., but cannot be readily combined with remote sensing products.

The objective of this study is to develop a methodology to create a hybrid land cover dataset for Russia (the world's largest country) as an information base for a terrestrial biota full (and verified) greenhouse gas account (Shvidenko and Nilsson 2007). Such requirements necessitate the detailed quantification of land classes (e.g., for forests – dominant species, age, growing stock, net primary production, etc.). The major idea of the hybrid dataset approach involves integration of all relevant information to explore synergies, in particular the merging and harmonization of land and forest inventories, ecological monitoring data, remote sensing data, and in situ information. The base year was adopted as 2005 owing to available data sources. Additionally, we consider this methodology as suitable for application at the global level, dependent upon the availability of required input data.

## 2. Methodology

### *2.1. Data description*

Initial efforts involved a survey of the available data to determine which datasets would be suitable for inclusion in this process (Table 1). All spatial datasets collected for the study were processed in a geographic information system (GIS) and converted to 1 km raster resolution.

#### *2.1.1. Remote sensing datasets*

Choice of a land cover dataset for this exercise was crucial. After careful deliberation, it was decided to use the global land cover 2000 (GLC) dataset (Bartholome and Belward 2005) for several reasons: (1) it was produced by regional experts, specifically for northern Eurasia (Bartalev *et al*. 2003); and (2) 1 km resolution was deemed sufficient, both in terms of detail and computational speed. Other products, including MODIS and GlobCover, were not selected at this point as they were not developed specifically for northern Eurasia and their higher resolution was not required for this study. It would be possible, however, to apply this methodology using either of these datasets.

The GLC dataset is based on daily data from the VEGETATION sensor onboard the SPOT4 satellite. A total of 19 regional assessments were combined to produce the global dataset. The GLC utilizes a global classification based on the land cover classification system (LCCS) legend of 23 classes with a 1 km resolution. Based on 1265 sample sites interpreted over the globe, an overall accuracy of 68.6% was determined (Mayaux *et al*. 2006). The accuracy of the regional product for northern Eurasia is assumed to be better than the global estimate (Bartalev *et al*. 2003).

The vegetation continuous fields (VCF) product is an annual representation of percent tree, herbaceous/shrublands, and barren cover. The three layers combine to represent 100% ground cover. It was generated from monthly composites of 500 m MODIS data (Hansen *et al*. 2002). This dataset was aggregated to 1 km resolution using a median filter. Product accuracy of the percent tree product has been estimated from: (1) accuracy of training data;

Table 1. Datasets used in creation of the hybrid Russian land-cover dataset.

| Dataset | Resolution | Date | Reference |
|---|---|---|---|
| Remote sensing | | | |
| Global land cover (GLC2000) | 1 km | 2000 | Bartholome and Belward (2005) |
| Vegetation continuous fields (forest and herbaceous) | 500 m | 2001 | Hansen *et al*. (2002) |
| Vegetation fire (AVHRR and LANDSAT) | 35 m and 1 km | 2000–2007 | Sukhinin (2008) |
| GIS | | | |
| Soil | 1:2.5 Mil | 1988 | Dokuchaev Soil Science Institute, Moscow |
| Administrative regions | 1:2.5 Mil | 1993 | Stolbovoi and McCallum (2002) |
| Forest enterprises | 1:2.5 Mil | 2005 | IIASA in-house database |
| Vegetation | 1:4 Mil | 1990 | Stolbovoi and McCallum (2002) |
| Bioclimatic zones | 1:4 Mil | 1990 | Stolbovoi and McCallum (2002) |
| Rivers/lakes and roads/railways | 1:1 Mil | 1990 | IIASA in-house database |
| Statistics | | | |
| State Land Account | 87 regions | 2005 | FACRE'RF (2006) |
| State Forest Account | 1914 forest enterprises | 2003 | FFS'RF (2003) |
| Disturbances in forests | 87 regions | 1991–2005 | FFS'RF (2006) |

and (2) from limited in situ field validation datasets. Overall accuracy yielded a standard error of estimate of 15.6% from training data and 11.5% from two field test areas.

The MODIS VCF product is the second crucial satellite-based dataset required for this methodology. The VCF provides the necessary flexibility, allowing us to prioritize the assignment of statistical data to land-cover data. Data can first be assigned to pixels falling into a certain land cover class, with high percent tree values. If more area appears in the statistics than in the land cover dataset, this can be assigned using the VCF.

Additionally, natural disturbance plays a large role in shaping the landscape of northern Eurasia. In particular, wildfire is responsible for large areas of annual land cover change and needs to be included in such a dataset. Wildfire data were acquired based on the advanced very high resolution radiometer (AVHRR) (hot spots) with control of burnt area by the LANDSAT thematic mappper (Sukhinin 2008). These data were aggregated by administrative region and year and are assumed substantially more reliable than official fire statistics (Shvidenko and Goldammer 2001).

### *2.1.2. GIS datasets*

One of two key GIS datasets used to assign data from different statistical inventories, the administrative region coverage contains 91 polygons (87 regions and 4 water bodies). The original dataset dates from 1993 (Stolbovoi and McCallum 2002) and was rasterized to 1 km. The second key dataset in the assignment of statistics is the Forest Enterprise dataset for the year 2005. This was created at IIASA with the aid of hardcopy and digital products, with a total of about 2000 polygons.

A soil database was one of the key components for selecting the appropriate area of arable land and wetlands. It contains a total of 292 unique soil types across the country with 21,988 polygons. The digitized soil map was developed by V.V. Dokuchaev Soil Science Institute (Moscow) in 1996 based on a hard copy of the soil map of Russia (1:2.5 Mil scale), edited by (Fridland 1989).

A vegetation dataset was also utilized to provide broad vegetation classes and bioclimatic zones (derived from the dataset titled Vegetation of the former USSR), produced at a scale of 1:4 Mil (Stolbovoi and McCallum 2002). The dataset includes georeferencing of 101 vegetation classes (e.g., 'Spruce, fir–spruce and spruce–fir forest with mosaic grass-low bush and grass–spruce cover' or 'Northern semi-shrub and bunchgrass steppe'). Bioclimatic zones were derived from the vegetation database. A total of eight zones (polar desert; tundra; forest tundra, northern and sparse taiga; middle taiga; southern taiga; temperate forests; steppe; and deserts and semi-deserts) were identified, and rasterized to 1 km.

In order to account for data not captured in the above datasets because of small areas of individual polygons, but which could have substantial areas aggregated by administrative regions (e.g., linear features; small water bodies; harvested areas, etc.), we relied on the Russian 1:1 Mil Planimetric dataset to account for '*virtual polygons*.' This dataset is based on original cartographic work from the State Committee for Geology and Cartography, USSR, and other sources. All existing railway lines and roads were buffered with 15 m, creating 30 m wide polygons. Because of the enormous amount of small rivers, only rivers $>$10 km in length were considered, and buffered 20 m to create 40 m wide polygons. Lakes smaller in size than 400 ha were also included. All of these *virtual polygons* (not presented on the map, but taken into account for the area balance and further calculations) were then tabulated per administrative region. If we did not account for *virtual polygons*, we would overestimate the areas of some land-cover classes.

### *2.1.3. Data of different inventories and statistics*

The State Forest Account (SFA) is the only source of aggregated forest inventory data for Russia. The last available account dates from 2003. It contains statistics for approximately 2000 forest enterprises. The SFA data contain areas and growing stock by dominant forest species distributed by age, site index, and relative stocking. There are approximately 45 sets of records for each enterprise on average. More information is available online (Shvidenko, Schepaschenko, Mccallum, and Nilsson 2007).

The State Land Account (SLA) is provided annually by the State Committee of Land Resources of Russia based on land statistics. Originally, the SLA is provided by administrative districts (about 3000 for Russia) and contains areas by (approximately 50) land classes. Publicly available data are by administrative regions. The SLA contains a two-dimensional official Russian land-use and land-cover hierarchical classification. The latter includes seven primary land-use categories: (1) agricultural lands; (2) populated areas; (3) lands for industry, energy, transport, communications, aerospace activities, defense, etc.; (4) special protective territories; (5) forest fund; (6) water fund; and (7) state reserves. Land-cover classes are defined by their dominant use and are based on natural and historical characteristics. They include agricultural classes (arable; fallow; forage production (hay-fields and pastures); and perennial vegetation) and non-agricultural land classes (lands under surface water including bogs; forest lands and lands under tree and shrub vegetation; built-up land; lands under roads; disturbed land – mining operations, earthmoving, etc.; and other land – ravines, sand, dumps, etc.) (FACRE'RF 2006).

Comparison was made between total forest area within the SFA and GLC (Figure 1). Overall agreement was high ($r^2 = 0.93$), however in general GLC predicts more forest area per enterprise than the SFA (approximately 20%). For individual forest enterprises ($\sim$2000), variability is rather high. Based on this assessment, the data of the SFA were basically deemed to be representative of Russian forests.

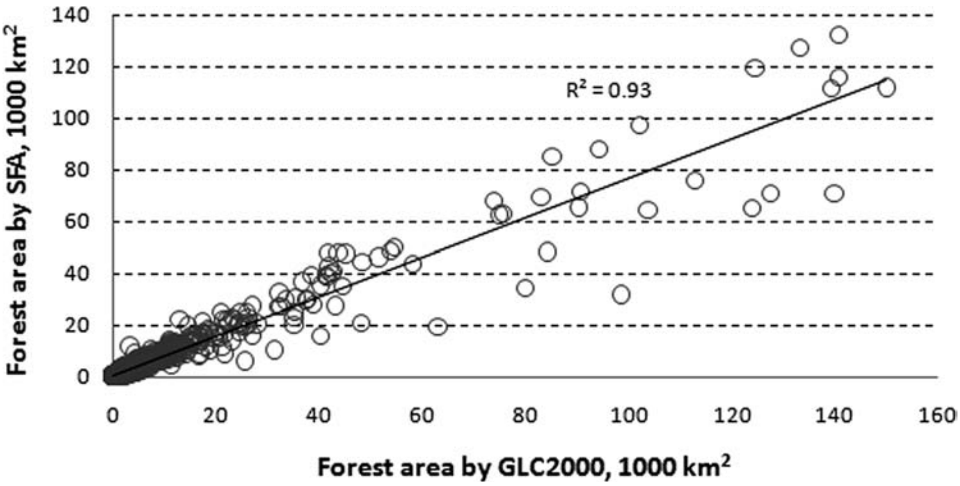


Figure 1. Comparison of forest area by individual forest enterprises from the State Forest Account and GLC2000.

### 2.2. *Land-cover assignment*

The land-cover assignment was performed on a per-pixel (1 km) basis across the entire country. The distribution of land surface by land classes is provided based upon the relevant combination of remote sensing products, GIS data, and statistics from different sources, applying the general principle that the most accurate and updated information has priority in assignment (Figure 2).

In the first stage, we allocated the total area presented in the land account by land category (forest, agriculture, wetland, shrub/grassland, burnt area, water, and unproductive), to the GIS and remote sensing resultant database by administrative units. The purpose of this first step is to ensure that broad classes (i.e., forest, agriculture, etc.) are correctly assigned, before moving to the second stage (pixel assignment). The criteria were as follows (Table 2).

The second stage involved calculating the quantitative correspondence of statistics (forest and land account) and spatial (remote sensing, GIS) data within each land-cover class. We have calculated a suitability index ($S_{ts}$) for each pixel pair (grid of territory ($t$) and statistics record ($s$)) within the territory unit (forest enterprise, administrative region) and land-cover class.

$$S_{ts} = \frac{1}{q}\left(\sum_{j=1}^{q}(x_{tj}^{\text{norm}} - x_{sj}^{\text{norm}})^2\right)^{1/2}, \text{where}$$

$q$ number of parameter;
$x_{tj}^{\text{norm}}, x_{sj}^{\text{norm}}$ normalized value of parameter $j$ for territory pixel $t$ and $j$;

$$x_j^{\text{norm}} = \frac{x_j - x_{j\,\text{min}}}{x_{j\,\text{max}} - x_{j\,\text{min}}}, \text{ where}$$

$x_{j\,\text{max}}, x_{j\,\text{min}}$ are maximum and minimum values of parameter $j$ within the certain area (forest enterprise, administrative unit).

Data on a nominal scale, that is, GLC land-cover classes were ranked with respect to a certain vegetation class in the statistics. For example, the most suitable GLC class for pine

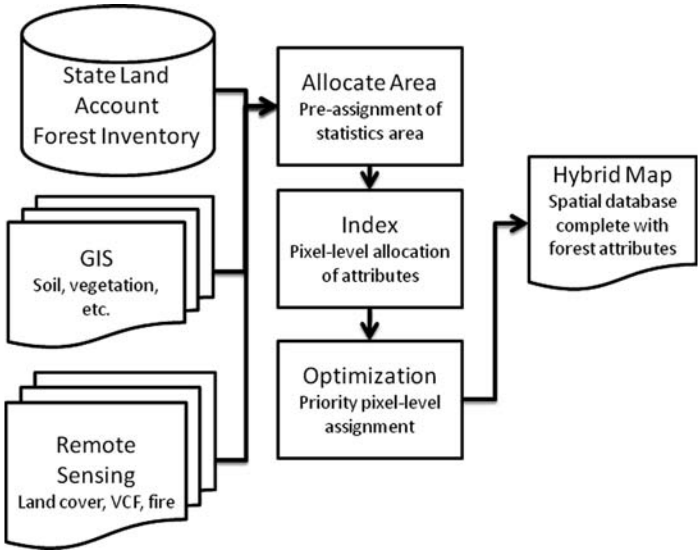


Figure 2. Flowchart outlining the process of integration of GIS, remote sensing, and statistical data to produce a hybrid land-cover dataset.

Table 2. Description of data combination used to assign area to land classes.

| Land category | Separation criteria | Area assignment |
|---|---|---|
| Forest | GLC: Tree cover and<br>(1) VCF_trees ≥21% for Forest Tundra and Northern taiga<br>(2) VCF_trees ≥35% for Middle taiga and other zones to the south | SFA, VCF |
| Agriculture | GLC: cropland, herbaceous cover; VCF_Herbaceous ≥50%; appropriate soils | SLA |
| Wetland | Appropriate soils; GLC: shrub and herbaceous cover; VCF_trees 13% | SLA, SFA, Soil |
| Open woodland | (1) VCF_trees: 13–20% for Forest Tundra and Northern taiga<br>(2) VCF_trees: 13–34% for Forest Tundra and Northern taiga | VCFSFA |
| Water | GLC: water; (VCF_trees + VCF_herbaceous) <50% | GLC, VCF |
| Unproductive | (1) GLC: non-vegetation; (VCF_trees + VCF_herbaceous) <50%<br>(2) (VCF_trees + VCF_herbaceous) <10% | GLC, VCF |
| Burnt | LANDSAT | SFA, LANDSAT |
| Grassland/ shrubland | Residuals area, classified by Vegetation map | |

SFA, State Forest Account; SLA, State Land Account; GLC, global land cover; VCF, vegetation continuous fields.

forest is ‘Tree Cover, Needle-leaved, Evergreen’, thus this would receive a high rank. Pine forest could also fit into ‘Tree Cover, Mixed Leaf Type’ or might be found in the ‘Tree Cover, Broadleaved, Deciduous’ GLC class (and would receive a lower rank). In accordance with SFA instructions, mixed forest is considered as coniferous if the coniferous growing stock is more than 50%.

The resultant suitability index *S* varies from 0 to 1. It can be interpreted as a distance between objects (grid of territory and statistics record) within the space of parameters. The lower the index value, the more suitable is the current piece of territory for the given statistical data. This method corresponds to the methods of land suitability assessment (Dokuchaev 1951, FAO 1976).

The final stage involved the optimization of distribution statistics data on the territory based on the suitability index (second stage) results. Each forest and land account record in the statistics was assigned to the most suitable grid within each enterprise. The following describe in detail the major land categories of land-cover assignment.

### 2.2.1. *Forest*

SFA data were assigned to the 1 km grid using the following parameters: GLC to place species from the SFA in the most appropriate GLC classes; VCF to assign the highly stocked forests to cells with a high VCF_trees; and Soil and vegetation zone maps to assign the most productive forests to the best soil and climate zones.

For example, let us calculate suitability index (Table 3) for the following: statistic record (pine forest, site index (SI) = II, relative stocking (RS) = 0.6) and the spatial data (GLC = 'Tree Cover, Mixed Leaf Type', VCF_trees = 50, Soil = 'Cambic Podzol,' Zone = 'Northern taiga').

Total area, presented in the SFA, was assigned to land cover. Both remote sensing products (VCF, GLC) sometimes identified forested area which exceeds the forested area found in the SFA, for some administrative units. Generally these areas correspond to territories with obsolete forest inventory data (time since inventory more than 15 years) or areas of abandoned intensive agriculture. We distinguish these kinds of forests using GLC 'vegetation' classes and VCF_trees $\geq$21% (for Tundra, Forest Tundra, and Sparse & Northern Taiga) and VCF_trees $\geq$35% (for forests situated southward of the middle taiga zones). For these areas, we assign the most representative SFA records within the respective administrative units and forest enterprises. The productivity parameters (i.e., growing stock, NPP) of such forests were corrected in accordance with the VCF_trees level.

### 2.2.2. *Agriculture*

The SLA contains the following agricultural land categories by administrative regions: arable land; hayfield; pasture; and fallow. An 'abandoned arable land' category was also introduced in accordance with estimates done by the Russian Academy of Agricultural

Table 3. An example of calculating suitability index.

| Parameter | Value or rank | Range (for a certain forest enterprise) | Normalized value | Index |
|---|---|---|---|---|
| VCF_trees/ | 50 | 40–80 | 0.25 | 0.25 |
| Relative stocking | 0.6 | 0.5–0.7 | 0.50 | |
| Soil | 1 (rank for forest) | 1–5 | 0 | 0.25 |
| Site index | 2 | 1.5–3.5 | 0.25 | |
| Zone | 2 (rank for forest) | 1–2 | 1 | 0.75 |
| Site index | 2 | 1.50–3.5 | 0.25 | |
| GLC | 2 (rank for pine) | 1–3 | 0.5 | 0.5 |
| Total | | | | 0.44 |

Sciences in 2007 and the Federal Service of State Statistics (Kevesh 2008). We reduced the area of arable land presented in the SLA by the abandoned land area, taking into account the area of currently cultivated land. We used the following parameters to distribute agricultural land: GLC (cropland, cultivated area, herbaceous cover, and mosaics); VCF_Herbaceous (maximum value exceeds 50%); and the Soil map (soil was ranked in order of potential agricultural productivity). The highest priority has been ascribed to arable land, then to abandoned arable, and then other agricultural land.

### *2.2.3. Wetlands*

Initially a probability table was created based on expert opinion, ranking soil types by their occurrence in major wetland types. Soils with a considerable peat layer (thickness $>30$ cm) as well as wet meadow and tundra soils were ranked with high probability. Some of the selected area according to wetland probability was classified as forest based on remote sensing data (GLC, VCF). The SLA typically overestimates wetland area because of a specific definition of this class in the Russia wetland classification, including partly shrub lands and fallow land. The SFA typically underestimates regional wetland area, because: (1) it deals with the forest fund only; and (2) by definition, the SFA identifies only treeless bogs (e.g., peatlands with tree cover stocking that satisfies the requirements of the Russian forest inventory manual, FFS'RF 1995, to be identified as forested areas). Therefore, the final wetlands area was assigned in the range of both land and forestry statistics based on the available appropriate soils. Finally, nine wetland classes were chosen and assigned according to vegetation zone, landscape peculiarities, and soil type: polygon mires; palsa mires; aapa mires; raised string bogs; pine bogs; reed and sedge fens; marshes; flood plain wetlands; and eutrophic fens.

### *2.2.4. Open woodland*

SFA data by administrative region were used to define the minimum area of open woodland (sparse forests according to the Forest Inventory Manual definition, FFS'RF 1995) using the correlation between canopy closure and relative stocking of stands by tree species and vegetation zone. This area was distributed over remaining land with percent tree cover (VCF) in the range of 13–35%. For some regions and vegetation zones, areas of open woodland proved to be higher than the area indicated by the SFA, particularly for the regions outside of the forest fund area (Sparse & Northern Taiga, Forest Tundra, Tundra). For such regions, the entire area with a VCF_trees level in the range of 13–20% was shifted to open woodland.

### *2.2.5. Burnt area*

Burnt area is presented in the SFA by forest enterprise. Fire statistics in Russia are notoriously uncertain (and typically underestimated) – however, burnt area measurements made from remote sensing are assumed to be rather accurate. We combined burnt area in the SFA with that from remote sensing, and assigned it to the appropriate GLC class (10 – tree cover, burnt). The remaining area was distributed according to areas of low VCF_trees value, not previously assigned to other vegetation classes.

#### 2.2.6. *Shrub and grassland*

The assignment of shrub and grassland was the final step in the assignment of vegetative land cover. There are no appropriate statistics recorded for this land class. Thus, we used remote sensing data (VCF, GLC) to define the area. The vegetation dataset was used to classify shrub and grassland types (50 classes in total).

#### 2.2.7. *Water*

The GLC water class was used to assign water to the resultant coverage. Additionally, we calculate *virtual polygons*: small water bodies and rivers which are not captured by remote sensing at the resolution of 1 km. It was necessary to balance the area of the territory and for posterior assessment of vegetation productivity. We used the 1:1 Mil Russian planimetric dataset to provide virtual polygon assessment.

#### 2.2.8. *Unproductive*

Unproductive polygons for our dataset were derived from the GLC coverage. We used the GLC non-vegetation classes: 'Bare Areas', 'Snow and Ice,' and 'Artificial surfaces and associated areas.' Infrastructure in the form of roads and railways were accounted for by administrative region with the use of *virtual polygons*.

## 3. Results and discussion

Application of the methodology described above has resulted in the new hybrid land-cover/land-use map of Russia at 1 km resolution (Figure 3). A total of six major land-cover types were identified, namely: forest, agriculture, wetlands, shrubs/grasses, water, and unproductive land. These are further subdivided into the following classes: forest – each grid links to the SFA database (the SFA data contain areas and growing stock by seven dominant forest species distributed by age, site index, and relative stocking), containing 86,613 records; agriculture – five classes, parameterized by 87 administrative units; wetlands – eight classes, parameterized by 83 zones/regions; and shrub/grassland – 50 classes, parameterized by 300 zones/regions.

A principal problem of merging and harmonizing substantially different datasets deals with the compatibility of definitions and classification schemes used. Taking into account the need for proper implementation of available data, we applied the Russian definitions wherever possible. However, in the case of Russia, many important definitions do not correspond to those agreed internationally and additionally; inconsistencies between national definitions of different datasets (e.g., SFA and SLA) are common. However, indicators that are measured on ground (e.g., by forest inventory) and by remote sensing are often not compatible.

Overall, we used decision rules and regional empirical models for harmonizing the definitions if relevant. We present our explanation for some forest definitions, while similar problems arose for other land classes. Because of Russian forest inventory manuals, forests are defined as tree communities with relative stocking $\geq$0.35 for young stands and $\geq$0.25 for other development stages (immature, middle-aged, and mature stands). Relative stocking is understood as the ratio between basal areas of individual stands and reference data for fully stocked stands. The Food and Agricultural Organization (FAO) definition of forest (which is also used in a number of global datasets) does not use relative stocking, but instead uses canopy closure (as a ratio between the sum of horizontal projections of tree crowns to the

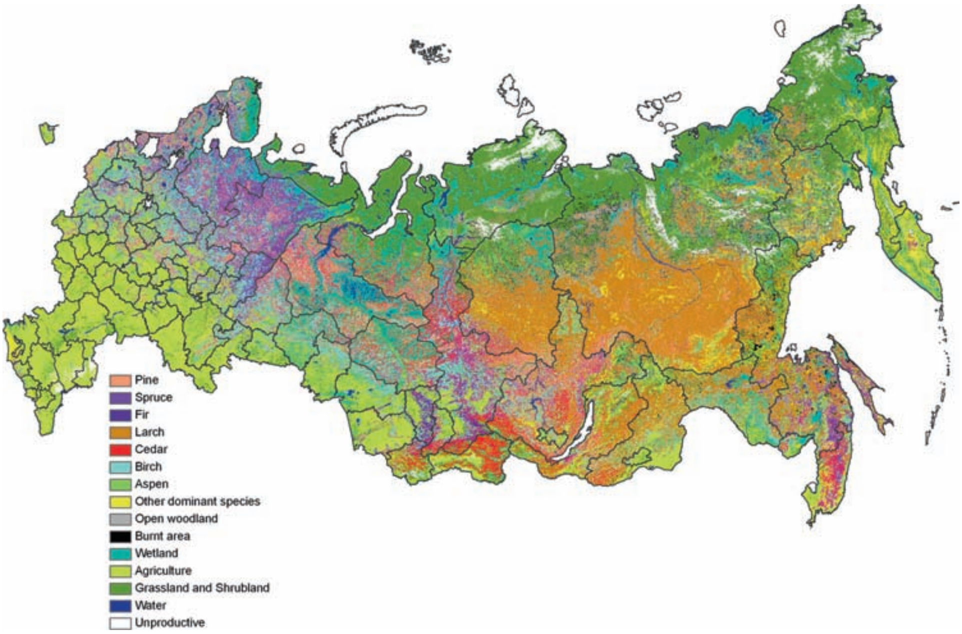


Figure 3. The new hybrid land-cover/forest species map of Russia.

area covered by forest), and the minimal threshold is set at 10% (FAO 2000). Canopy closure is used in the majority of remote sensing products, but not uniformly. For example, VCF accounts for crown gaps (Hansen *et al*. 2002), whereas the FAO definition does not. In order to merge these definitions, we used aggregated empirical models linking canopy closure and relative stocking which are available for Russian forests (Sukhikh 2005).

### *3.1. Evaluation*

Evaluation of the Russian hybrid land-cover dataset is rather difficult, owing to the fact that the majority of available material of relevance has been used in the creation of the dataset itself. In addition, the resultant coverage is highly detailed (i.e., forest species and 1 km resolution), making comparison to existing coarser datasets somewhat irrelevant. We have therefore attempted to assess spatially the level of confidence in the assignment (Figure 4), based on the assessed agreement between the input datasets.

The first class (52% of area) infers a high degree of confidence in the agreement among the remote sensing products and the statistics. A vegetated GLC class, combined with maximum VCF classes, was matched with corresponding statistical data. For example, larch forest presented in the SFA for a certain administrative unit was assigned to the 'needle-leaved deciduous' GLC class with VCF_trees level exceeding 70%. The second confidence class (18% of area) was assigned to territories where groups of GLC classes (forest, agriculture, herbaceous, shrubs) and a rather high level of VCF (>50%) were used.

For the third class (24% of area), the GLC group of classes 'vegetation' was used along with rather high levels of VCF (>50%). For example, we assign 'forest' to the GLC class 'Shrub cover' (not forest, but still vegetation) with VCF_trees exceeding 50%. In general, these represent insufficient GLC parameterization and remote areas with obsolete (outdated in situ) statistical data. Additionally, we can see this class in mountains and densely populated

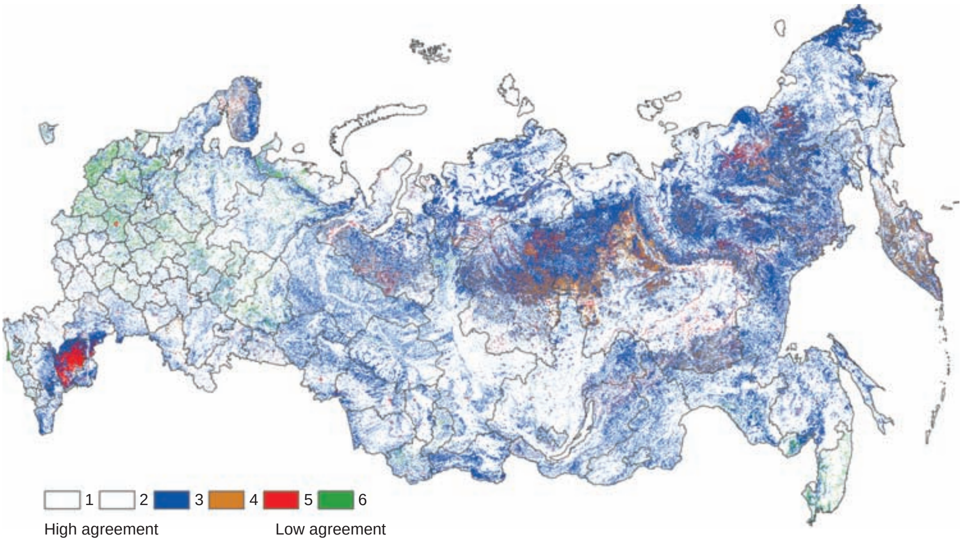


Figure 4. Agreement/confidence classes of the hybrid land-cover map (1 = higher agreement/confidence; 6 = lower agreement/confidence – classes 1 and 2 omitted for clarity).

areas with complex vegetation mosaics. In such mosaic areas, a 1 km grid cannot perform better. The fourth class (4% of area) implies 'vegetation' on the GLC coverage and low VCF levels (10–50%). This is mostly sparse vegetation with typical forms of disagreement such as wetland–grassland in West Siberia and forest-tundra ecotones in East Siberia and the Far East.

Disagreement between the two remote sensing products (GLC and VCF) is shown in the fifth class (1% of area) where GLC reported 'non-vegetation' (bare areas, snow and ice, artificial surfaces), but VCF reported 'vegetation' (VCF_trees + VCF_herbaceous > 50%). The majority of the area belongs to semi-desert or tundra zones. For example, a large portion of this class ($36 \times 10^6$ ha) is situated in Kalmyk Republic and Astrahan oblast, comprised of dry steppe and semi-dessert zone. In GLC, it is indicated as 'bare areas,' but VCF shows herbaceous cover in the range of 61–98%. In accordance with the SLA data, we assigned 'pasture' to this area. The fifth class also appears in mountains and on the outskirts of large cities.

Forest statistics incompleteness is shown in class 6 (1% of area). The remote sensing products report forest area, but the SFA contain far fewer forests for the respective regions. Most of the area lies in the European southern taiga (assumed to be abandoned agricultural area which is afforested) or in the north (outside of managed forest area) (Figure 5). Figure 6 shows an example of afforested land in Pskov oblast, illustrating a reason for the significant difference between the Russian SFA statistics and the remote sensing products in this part of the country.

## 4. Conclusions

There is a critical need for accurate land-cover information for resource assessment, biophysical modeling, greenhouse gas studies, and for the estimation of possible terrestrial responses and feedbacks to climate change (Frey and Smith 2007). Over the past two decades, evidence has accumulated of significant contributions of extratropical Northern Hemisphere land areas to the global uptake of anthropogenic $CO_2$ (Schimel *et al*. 2001).

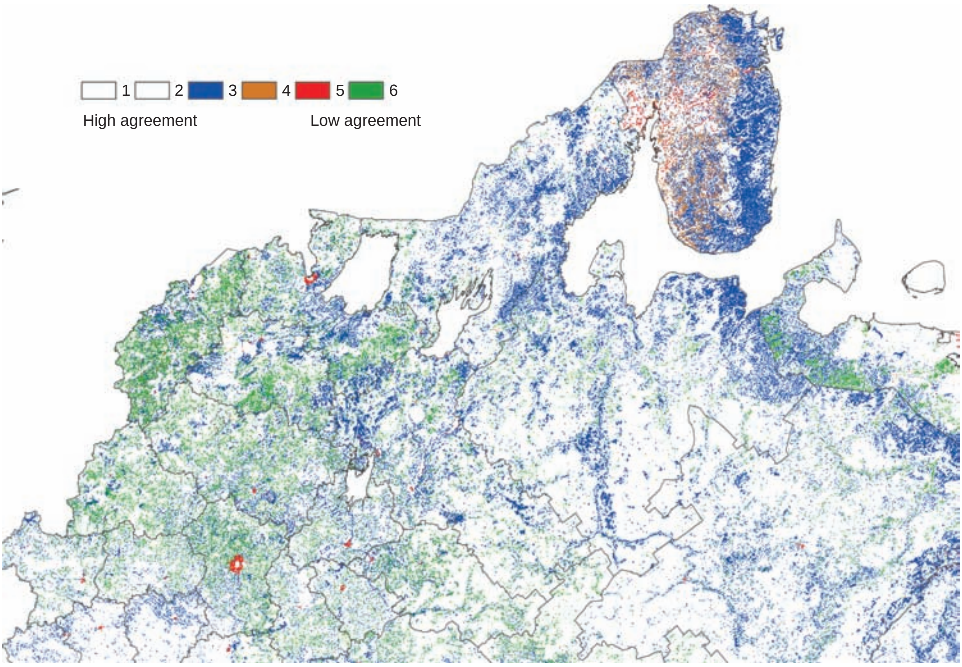


Figure 5. Areas of inconsistency between inventory statistics and remote sensing data in Northwest Russia. Areas in green assumed to be abandoned agricultural area which is naturally afforested and areas in blue assumed to represent obsolete inventory statistics or areas outside of the managed forest area.

Figure 6. © Google Earth 2009. © 2009 Geocentre Consulting Image, © 2009 Digital Globe, © Tele Atlas. A potential example of forest encroachment occurring on abandoned agricultural land in Pskov Oblast in Northwest Russia (56°18′40.24″N, 28°53′13.65″E). The old drainage system is no longer maintained and arable land shifts to wet grassland and finally to forest.

Without accurate baseline measures of crucial datasets such as land use and land cover, we will have little hope of monitoring the effects of global change on vegetation over large regions. A variety of global information sets exist with data on land use and land cover, although none of these alone satisfy the requirements of the above-mentioned aims.

This study has demonstrated the ability to produce a highly detailed (both spatially and thematically) land-cover/land-use dataset over Russia – the largest country on the planet, by combining existing datasets into a hybrid information system. Every source of information has its advantages and shortcomings. Remote sensing supplies the most up-to-date information, but a lack of parameterization and interpretation. Land statistics are the best parameterized product, but lack spatial distribution and are partially out-of-date. GIS datasets are explicit spatially, well parameterized, but also sometimes out-of-date. The new land-cover product uses the advantages of all sources, supplies up-to-date geographically explicit and well-parameterized information, thus allowing for reduction of source's uncertainty. Validation efforts are now underway. In order to fully validate the hybrid map, representative areas of forest inventory data from each of the oblasts would be necessary, allowing for the creation of confusion matrices. In lieu of such data, a confidence map was produced creating six classes of confidence in the agreement between the various remote sensing and statistical datasets.

One interesting result from this exercise was the significant difference between the remote sensing products and the Forest State Account for the Northwest of Russia. In this region, the statistics appear to be underreporting the amount of forest land, which has been likely increasing in recent decades because of forest encroachment of abandoned agricultural land.

Future efforts include further validation of the hybrid land-cover dataset for Russia, and its use for assessment of the terrestrial biota full greenhouse gas budget for northern Eurasia. The algorithm presented in this study is flexible, allowing for the inclusion of additional existing datasets or newly created datasets in the future (i.e., elevation, lidar biomass, and more). The main advantage of the methodology is the ability to link on-ground data and models to the remote sensing products. Additional dataset attributes not described here but included in the hybrid dataset include live biomass, growing stock volume, gross growth, net growth, mortality, net primary production, coarse woody debris, and others.

In addition to regional and continental applications, the possibility of applying this methodology over the globe now exists, with the majority of input datasets used being global. The method allows use of not only the fixed set of data, but all existing relevant information. We consider some other remote sensing products (e.g., GlobCover, elevation models, lidar data, and others) as very promising for use in this methodology. In addition, it would be possible to place weights on datasets (i.e., favoring data with low uncertainty) or recently updated statistics. Adoption of such techniques by the FAO might provide a useful enhancement to their provision of global forest statistics. With land-cover statistics typically updated every five years (e.g., Russia), this methodology could be repeated. In combination with annual GLC products (i.e., MODIS); new products could be operationally produced on an annual basis in an automated fashion. However, global implementation of the discussed technology requires substantial efforts: it is impossible to implement based on only a few remote sensing indicators (common practice for many global products) with insufficient use of ground truth data.

## Acknowledgments

Funding for this work was partially provided by the European Commission FP6 Project GEOBENE (www.geo-bene.eu).

## References

Bartalev, S.A., Belward, A.S., Erchov, D.V., and Isaev, A.S. (2003), "A New SPOT4-VEGETATION Derived Land Cover Map of Northern Eurasia," *International Journal of Remote Sensing*, 24, 1977–1982.

Bartholome, E., and Belward, A.S. (2005), "GLC2000: A New Approach to Global Land Cover Mapping from Earth Observation Data," *International Journal of Remote Sensing*, 26, 1959–1977.

Bicheron, P., Defourny, P., Brockmann, C., Schouten, L., Vancutsem, C., Huc, M., Bontemps, S., Leroy, M., Achard, F., Herold, M., Ranera, F., and Arino, O. (2008), "Globcover: Products Description and Validation Report: M. France," POSTEL and Medias-France, Toulouse, France.

Dokuchaev, V.V. (1951), *Steppe Transformation. Papers on Soil Study and Land Evaluation (1888–1900)*, Moscow: Academy of Science.

FACRE'RF. (2006), "State (National) Report About the State and Use of Lands of Russian Federation in 2005," Federal Agency of Real Estate Cadastre Report, Moscow, Russia, 220pp. [in Russian].

FAO. (1976), "A Framework for Land Evaluation," *Soils Bulletin*, Rome: Author, 72.

FAO. (2000), "Forest Resources of Europe, CIS, North America, Australia, Japan, and New Zealand (Industralized Temporate/boreal Countries)", *UN-ECE/FAO Contributions to the Global Forest Resources Assessment* http://www.unece.org/timber/fra/pdf/contents/htm.

FAO. (2002), *Proceedings of the FAO/UNEP Expert Consulation on Strategies for Global Land Cover Mapping and Monitoring*, Rome, Italy: Author, 38.

FFS'RF. (1995), *Manual of Forest Inventory and Planning in the Forest Fund of Russia. Vol. 1*, *Field Work*, Moscow, Russia: Federal Forest Service [in Russian].

FFS'RF. (2003), *Forest Fund of the Russian Federation*, Moscow, Russia: Federal Forest Service [in Russian].

FFS'RF. (2006), *Major Indicators of Forestry Activities*, Moscow, Russia: Federal Forest Service [in Russian].

Frey, K.E., and Smith, L.C. (2007), "How Well Do We Know Northern Land Cover? Comparison of Four Global Vegetation and Wetland Products with a New Ground-Truth Database for West Siberia," *Global Biogeochemical Cycles*, 21, GB1016.

Fridland, V.M. (1989), *Soil Map of the USSR*, Moscow, Russia: Committee on Cartography and Geodesy.

Friedl, M.A., Mciver, D.K., Hodges, J.C.F., Zhang, X.Y., Muchoney, D., Strahler, A.H., Woodcock, C.E., Gopal, S., Schneider, A., Cooper, A., Baccini, A., Gao, F., and Schaaf, C. (2002), "Global Land Cover Mapping from MODIS: Algorithms and Early Results," *Remote Sensing of Environment*, 83, 287–302.

Hansen, M.C., Defries, R.S., Townshend, J.R.G., Sohlberg, R., Dimiceli, C., and Carroll, M. (2002), "Towards an Operational MODIS Continuous Field of Percent Tree Cover Algorithm: Examples Using AVHRR and MODIS Data," *Remote Sensing of Environment*, 83, 303–319.

Kevesh, A.L. (ed.) (2008), *Russian Regions. Social-Economic Indicators. 2008: Statistics Collection*, Moscow: Rosstat.

Mayaux, P., Eva, H., Gallego, J., Strahler, A.H., Herold, M., Agrawal, S., Naumov, S., De Miranda, E.E., Di Bella, C.M., Ordoyne, C., Kopin, Y., and Roy, P.S. (2006), "Validation of the Global Land Cover 2000 Map," *IEEE Transactions on Geoscience and Remote Sensing*, 44, 1728–1737.

Mccallum, I., Obersteiner, M., Nilsson, S., and Shvidenko, A. (2006), "A Spatial Comparison of Four Satellite Derived 1 km Global Land Cover Datasets," *International Journal of Applied Earth Observation and Geoinformation*, 8, 246–255.

Schimel, D.S., House, J.I., Hibbard, K.A., Bousquet, P., Ciais, P., Peylin, P., Braswell, B.H., Apps, M.J., Baker, D., Bondeau, A., Canadell, J., Churkina, G., Cramer, W., Denning, A.S., Field, C.B., Friedlingstein, P., Goodale, C., Heimann, M., Houghton, R.A., Melillo, J.M., Moore, B., Murdiyarso, D., Noble, I., Pacala, S.W., Prentice, I.C., Raupach, M.R., Rayner, P.J., Scholes, R.J., Steffen, W.L., and Wirth, C. (2001), "Recent Patterns and Mechanisms of Carbon Exchange by Terrestrial Ecosystems," *Nature*, 414, 169–172.

Shvidenko, A., and Goldammer, J.G. (2001), "Fire Situation in Russia," *International Fire News*, 24, 41–59.

Shvidenko, A., and Nilsson, S. (2007), "Can the Level of Uncertainties of a Regional Terrestrial Biota Full Carbon Account be Made Acceptable for Policy Makers?" *Proceedings of the 2nd International Workshop on Uncertainty in Greenhouse Gas Inventories*, 27–28 September 2007, Laxenburg, Austria: IIASA, pp. 181–194.

Shvidenko, A., Schepaschenko, D., Mccallum, I., and Nilsson, S. (2007), "Russian Forests and Forestry. International Institute for Applied Systems Analysis and the Russian Academy of Science," http://www.iiasa.ac.at/Research/FOR/forest_cdrom [Accessed 1 February 2009].

Stolbovoi, V., and McCallum, I. (2002), *Land Resources of Russia*, Laxenburg, Austria: IIASA & RAS. http://www.iiasa.ac.at/Research/FOR/ [Accessed 1 February 2009].

Sukhikh, V.I. (2005), *Remote Sensing Methods in Forestry, Forest Management and Landscape Architecture*, Yoshkar-Ola: Mary-El State Technical University, 390pp. [in Russian].

Sukhinin, A.I. (2008), *Satellite Data on Wild Fire*, Laxenburg, Austria: IIASA & RAS. www.iiasa.ac.at/Research/FOR/forest_cdrom/index.html [Accessed 01 February 2009].