Air pollution modelling with the aid of computational intelligence methods in Thessaloniki, Greece

Kostas D. Karatzas a,*, Stamoulis Kaltsatos b

a Department of Mechanical Engineering, Informatics Applications and Systems Group, Aristotle University, GR-54124 Thessaloniki, Greece
b Physics Department, Aristotle University, GR-54124 Thessaloniki, Greece

Received 13 June 2007; received in revised form 7 September 2007; accepted 13 September 2007
Available online 29 September 2007

Abstract

Air pollution modelling is necessary for simulating the atmospheric environment system in terms of pollutants and meteorological conditions, taking into account the nonlinearities of the underlying phenomena. In the current paper, Artificial Neural Networks are used for modelling ozone, and for simulating its behaviour in relation to other atmospheric parameters of interest, for the city of Thessaloniki, Greece. This behaviour is also investigated with the aid of Principal Component Analysis (PCA). Results suggest the operational capabilities of such models, and the research potential in the application of computational intelligence methods for the environmental sector.

© 2007 Elsevier B.V. All rights reserved.

Keywords: Artificial Neural Networks; Principal component analysis; Air pollution modelling

1. Introduction

The urban atmosphere’s quality is regulated by a number of EU directives against high air pollution concentration levels that aim to protect human health and to minimise environmental degradation. These directives define the pollutants to be addressed and the related criteria to be applied for assessing atmospheric quality, while in addition they specify the methods to be used for the assessment. One of these methods is air pollution modelling for the simulation of the atmospheric environment system, towards a twofold goal: increased domain knowledge and reliable forecasting. The latter is to be used for both strategic planning and environmental decision making (and thus has a long term nature), or for short term warning at a citizen and administration level. On this basis, computational methods that allow for domain modelling and quality of life parameter forecasting are becoming more and more important.
Air pollution concentration levels are a combination of pollutant emissions, chemical and physical processes in the atmosphere, and earth surface properties and geometry. For the purposes of modelling and simulation, the urban atmosphere may be considered as a system that under varying meteorological conditions (i.e. “input”), will “respond” by producing different sets of output, i.e. concentration levels of the pollutants of interest [21]. Yet, one of the factors of complicity concerning the simulation of the urban atmospheric quality is related to the nonlinear behaviour of secondary pollutants, i.e. pollutants that are not directly emitted but are rather formed within the atmosphere as a result of chemical reactions. Such a substance is ozone \( \text{O}_3 \), a typical photochemical pollutant created from the reaction of hydrocarbons and nitrogen oxides in the presence of sunlight that acts as a catalyst.

The need for modelling and forecasting of ozone concentrations has been raised due to the mandates of the EU environmental legislation like the air quality (AQ) framework directive (1996/62/EC) and the so called ozone directive (2002/3/EC). As a first step towards environmental knowledge modelling, it is of primary interest to identify the major parameters influencing ozone formation in the vicinity of human activities and everyday life. Principal component analysis (PCA) is a proper method for multidimensional data analysis and may support the investigation of “hidden” relationships between the pollutant examined and the factors that favour its formation. The outcomes of PCA analysis are also used as driving parameters for ozone modelling, with the aim to simulate its behaviour and thus produce reliable forecasts that may be used for operational air quality management. Such modelling requires methods that are capable of adapting to the requirements of the knowledge domain of interest, while learning from changes within the system modelled. Artificial Neural Networks (ANNs) are selected for this purpose.

ANNs are computational data structures that try to mimic the biological processes of the human brain and nervous system. ANN models have been widely used for modelling air pollutant concentrations with the aim to forecast them. Due to the fact that ANNs can capture nonlinear relationships, their performance is superior when compared to statistical methods such as multiple linear regression [2,6,10,12,29], while it is supported from the usage, in parallel, of PCA [26].

The aim of this paper is to examine the relation between ozone \( \text{O}_3 \), nitrogen dioxide \( \text{NO}_2 \) and meteorological variables for the purposes of domain knowledge investigation and modelling, and to develop on this basis ANN models that are capable of simulating the ozone formation phenomenon, and then evaluate their operational forecasting capabilities with the air of appropriate statistical indexes.

2. Area of interest and data used

Thessaloniki is the second largest city of Greece, where air emissions come mainly from traffic, while formation and transportation of pollutants is heavily influenced by the local meteorological and topographic characteristics. Focusing on traffic, it should be noted that recent studies in Thessaloniki have estimated that travel demand has climbed up almost 70% over the period 1988–1998. In addition, the improvement of fuel quality and the renewal of the vehicle fleet during the last decade contributed significantly to the decreasing trend of some primary air pollutants like CO (carbon monoxide), \( \text{SO}_2 \) (sulphur dioxide), and Pb (lead), but they did not similarly influence the trend of PM10 (suspended particulates), \( \text{NO}_2 \) (nitrogen dioxide) and \( \text{O}_3 \) (ozone) [16,17].

Data used in the present paper were obtained from the monitoring stations located in Kalamaria (urban station in the east side of the city), and Eleftherio-Kordelio (urban station in the west side of the city). The selection of the stations emphasized on locations that are not in the city centre, as the latter is mostly influenced by traffic, this resulting in ozone depletion. On the other hand, ozone is known to accumulate at the city edges, and then to travel with the aid of air movement over urban and rural areas alike.

The data set comprised of hourly observations for the time period of 2001–2003 for Eleftherio-Kordelio, and for the period of 2001–2004 for Kalamaria. The atmospheric parameters monitored included hourly values of ozone \( \text{O}_3 \) and nitrogen dioxide \( \text{NO}_2 \) concentrations, temperature, humidity, wind speed and wind direction. For the latter a transformation was applied for replacing the cyclic nature of the parameter with a linear one, as follows:

\[
\sin \text{WD} = \frac{\sin(2\pi(v - \text{min}(v)))}{\text{max}(v) - \text{min}(v)}, \quad \cos \text{WD} = \frac{\cos(2\pi(v - \text{min}(v)))}{\text{max}(v) - \text{min}(v)}, \quad v \in [0, 360]
\]
where sinWD and cosWD are the results of the linearized wind direction (WD), calculated as a function of the sine and the cosine of the difference of the WD from the minimum value monitored, divided by the difference between the maximum and minimum WD values that have been monitored.

As the investigated time-series presented in many cases the problem of missing values, the latter were imputed via applying the Bayesian Principal Component Analysis method, (BPCA), implemented with the MATLAB software package (www.mathworks.com). The missing data percentages for each station were: Kalamaria (10.64%), and Eleftherio-Kordelio (4%).

3. Principal component analysis

3.1. Introduction

Air quality modelling and forecasting is a problem that employs multidimensional datasets. In the case of Kalamaria and Eleftherio-Kordelio, each data set consists of 8760 records corresponding to the hourly values for each one of the following seven parameters: O3, NO2, wind direction (two parameters), wind speed, humidity, air temperature, thus leading to a number of 61,320 parameters per year. In addition, each record has a time stamp consisting of the year, month, day, and hour of the day that the value has been recorded. One of the first questions to be answered is related to the extraction of the most characteristic within the studied dataset. PCA is computational intelligence method originating from multivariate statistical analysis that allows for the identification of the major drives within a certain multidimensional data set and thus may be applied for data compression, identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. By applying the PCA, we result in a new set of uncorrelated variables, the principal components (PCs), sorted by the percentage of the original variance they account for. Each PC is perpendicular to each other so there is no redundant information in the new dataset. By selecting the most significant PCs (on the basis of appropriate criteria), we succeed in reducing the original dataset’s dimensions and in selecting the most characteristic and influential parameters for further steps of analysis and investigations.

3.2. PCA method implementation for environmental time-series analysis in Thessaloniki, Greece

The PCA method was applied for pollutant and meteorological datasets derived from the air quality monitoring stations located in the city of Thessaloniki, Greece and operated by the Prefecture of Central Macedonia. If PCA is applied upon the initial dataset, then its results depend upon the measurement unit of the parameters investigated. As the method aims at identifying data structures and parameter “importance” in terms of sufficient description of the whole dataset, initial data should be normalised. For this reason, each variable is centred by subtracting the mean value from each time-series investigated. The resulting data are divided by the relevant standard deviation, this leading to a final dataset with zero mean and standard deviation of unity. Results for each monitoring location are presented in the next sub-chapters.

3.2.1. Kalamaria station

A four year data set (2001–2004) was used, on the basis of observations at the Kalamaria site. As a first step, data were normalised, as previously described. Then, the normalised covariance matrix (i.e. correlation coefficient) is calculated and PCA results are presented in Tables 1 and 2, respectively.

Table 1
Normalised covariance matrix (correlation coefficient) for hourly data concerning the Kalamaria station for the period 2001–2004

<table>
<thead>
<tr>
<th></th>
<th>NO2</th>
<th>O3</th>
<th>Temp</th>
<th>Hum</th>
<th>WS</th>
<th>sinWD</th>
<th>cosWD</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O3</td>
<td>-0.459</td>
<td>1</td>
<td>0.617</td>
<td></td>
<td>0.23</td>
<td>-0.321</td>
<td>0.109</td>
</tr>
<tr>
<td>Temp</td>
<td>-0.322</td>
<td>0.617</td>
<td>1</td>
<td>-0.494</td>
<td>0.23</td>
<td>-0.186</td>
<td>-0.243</td>
</tr>
<tr>
<td>Hum</td>
<td>0.23</td>
<td>-0.638</td>
<td>-0.494</td>
<td>1</td>
<td>-0.428</td>
<td>0.163</td>
<td>-0.028</td>
</tr>
<tr>
<td>WS</td>
<td>-0.321</td>
<td>0.462</td>
<td>0.023</td>
<td>-0.428</td>
<td>1</td>
<td>-0.245</td>
<td>-0.301</td>
</tr>
<tr>
<td>sinWD</td>
<td>0.109</td>
<td>-0.323</td>
<td>-0.186</td>
<td>0.163</td>
<td>-0.245</td>
<td>1</td>
<td>-0.018</td>
</tr>
<tr>
<td>cosWD</td>
<td>0.062</td>
<td>-0.193</td>
<td>-0.243</td>
<td>-0.028</td>
<td>0.301</td>
<td>-0.018</td>
<td>1</td>
</tr>
</tbody>
</table>
According to the covariance matrix, O₃ concentration value increase is related to NO₂ concentration value decrease, a finding that is expected, due to the fact that the one pollutant is “antagonistic” to the other. Moreover, ozone seems to increase with temperature and to decrease with increased humidity. As temperature may be associated with solar radiation, and humidity with cloud coverage, both findings verify the basic mechanisms of the simulated atmospheric system: solar radiation boosts the generation of Ozone due to photochemical reactions. Moreover, high concentration values are associated with wind speed, a finding indicating that ozone in the Kalamaria area may be linked with air mass movements. This suggests that ozone may be transported to the area via air masses, and not just generated locally. From Table 2, a number of Principal components (PC) may be selected as representative for the studied dataset, on the basis of various selection criteria. One of the commonly used criterions is the variance of each PC as a percentage of the variance of the whole data set investigated (%Var), and the cumulative variance of the component \( i \), i.e., \( \text{sum}(\%\text{Var}) \) for \( i = 1 – n \). Based on this criterion, the first six principal components account for the 97% of the total variance, a sufficient percentage of information, which, nevertheless, may include too much noise and thus requires for a more detailed PC selection criterion to be applied. Parallel analysis (PA), also known as Humphrey-Ikgen parallel analysis, is now often recommended as the best method to assess the true number of significant PCs [13]. According to the PA method, the actual data, and a set of random numbers representing the same number of cases and variables, are factor analyzed. Then, the number of each PC on the \( x \) axis and the cumulative value of eigenvalues on the \( y \) axis are plotted for both data sets. The intersection of these two lines determines the number PCs to be selected. Such a plot is presented in Fig. 1.

![Fig. 1. Humphrey-Ikgen Parallel analysis results for the Kalamaria data set.](image)
According to Fig. 1, the first three PCs should be selected as representative for the studied dataset. The first PC is large for large values of humidity and NO\(_2\) and for small variables of ozone (the most influencing parameter for PC1), temperature and wind speed. The situation changes for PC2, that is influenced by low values of the cosine of the wind direction, the wind vector component explaining air mass movements on the axis west to east, Kalamaria being on the east side of the city. PC2 is also large for small values of wind speed, thus verifying the influence of advective mechanisms in the concentration of photochemical pollutants in the area. Wind vector (sinus) is also the most influential parameter for PC3.

### 3.2.2. Eleftherio-Kordelo station

A three year data set (2001–2003) was used, on the basis of observations at the Eleftherio-Kordelo station. The normalised covariance matrix (i.e. correlation coefficient) calculated and PCA results are presented in Tables 3 and 4, respectively.

According to the covariance matrix (Table 3), O\(_3\) concentration value increase is related to NO\(_2\) concentration value decrease, a finding that is expected, due to the fact that the one pollutant is “antagonistic” to the other. Moreover, Ozone seems to decrease with humidity (and increase with temperature) and to increase with wind speed (increased wind speed usually causes the drop of humidity), thus indicating that ozone may be transferred to the monitoring area. Table 4 indicates that the first six principal components account for the 97.335\% of the variance. The application of the Humphrey-Ilgen parallel analysis method for the selection of PCs, (Fig. 2), results again in selecting the first three PCs.

The first PC is large for large values of Humidity and NO\(_2\) and for small variables of Ozone (the most influencing parameter for PC1), temperature and wind speed, a finding identical to the one for the Kalamaria Station. The situation changes for PC2, that is influenced by high values of the cosine of the wind direction, the wind vector component explaining air mass movements on the axis west to east, Eleftherio-Kordelo being on the west-north side of the city. PC2 is also large for large values of temperature and NO\(_2\). Lastly, PC3 is mostly influence by the sinus of the wind vector, indicating a dependency of the air pollution levels in the area by the winds blowing on the south–north axis.

### Table 3
Normalised covariance matrix (correlation coefficient) for hourly data concerning the Eleftherio-Kordelo station, for the period 2001–2003

<table>
<thead>
<tr>
<th></th>
<th>NO(_2)</th>
<th>O(_3)</th>
<th>Temp</th>
<th>Hum</th>
<th>WS</th>
<th>sin WD</th>
<th>cos WD</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO(_2)</td>
<td>1</td>
<td>−0.638</td>
<td>−0.171</td>
<td>0.249</td>
<td>−0.542</td>
<td>−0.025</td>
<td>−0.037</td>
</tr>
<tr>
<td>O(_3)</td>
<td>−0.638</td>
<td>1</td>
<td>0.57</td>
<td>−0.599</td>
<td>0.503</td>
<td>0.045</td>
<td>−0.243</td>
</tr>
<tr>
<td>Temp</td>
<td>−0.171</td>
<td>0.57</td>
<td>1</td>
<td>−0.508</td>
<td>0.091</td>
<td>0.027</td>
<td>−0.287</td>
</tr>
<tr>
<td>Hum</td>
<td>0.249</td>
<td>−0.599</td>
<td>−0.508</td>
<td>1</td>
<td>−0.403</td>
<td>0.065</td>
<td>0.157</td>
</tr>
<tr>
<td>WS</td>
<td>−0.542</td>
<td>0.503</td>
<td>0.091</td>
<td>−0.403</td>
<td>1</td>
<td>−0.013</td>
<td>−0.026</td>
</tr>
<tr>
<td>sin WD</td>
<td>−0.025</td>
<td>0.045</td>
<td>0.027</td>
<td>0.065</td>
<td>−0.013</td>
<td>1</td>
<td>0.121</td>
</tr>
<tr>
<td>cos WD</td>
<td>−0.037</td>
<td>−0.243</td>
<td>−0.287</td>
<td>0.157</td>
<td>−0.026</td>
<td>0.121</td>
<td>1</td>
</tr>
</tbody>
</table>

### Table 4
PCA results for hourly data concerning the Eleftherio-Kordelo station, for the period 2001–2003

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PC5</th>
<th>PC6</th>
<th>PC7</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO(_2)</td>
<td>0.4066</td>
<td>−0.4463</td>
<td>0.1186</td>
<td>−0.2359</td>
<td>0.5114</td>
<td>−0.2526</td>
<td>0.4903</td>
</tr>
<tr>
<td>O(_3)</td>
<td>−0.547</td>
<td>0.0122</td>
<td>0.0779</td>
<td>0.0518</td>
<td>−0.2146</td>
<td>0.1745</td>
<td>0.7845</td>
</tr>
<tr>
<td>Temp</td>
<td>−0.3845</td>
<td>−0.4309</td>
<td>0.3063</td>
<td>−0.2937</td>
<td>−0.3302</td>
<td>−0.5669</td>
<td>−0.2366</td>
</tr>
<tr>
<td>Hum</td>
<td>0.4488</td>
<td>0.1685</td>
<td>−0.0135</td>
<td>0.4605</td>
<td>−0.4897</td>
<td>−0.5011</td>
<td>0.2588</td>
</tr>
<tr>
<td>WS</td>
<td>−0.3961</td>
<td>0.3981</td>
<td>−0.248</td>
<td>0.1764</td>
<td>0.5234</td>
<td>−0.5639</td>
<td>−0.0007</td>
</tr>
<tr>
<td>sin WD</td>
<td>0.0055</td>
<td>0.257</td>
<td>0.9037</td>
<td>0.2532</td>
<td>0.2145</td>
<td>0.0856</td>
<td>−0.0609</td>
</tr>
<tr>
<td>cos WD</td>
<td>0.1714</td>
<td>0.6016</td>
<td>0.0877</td>
<td>−0.7403</td>
<td>−0.1536</td>
<td>−0.1085</td>
<td>0.1324</td>
</tr>
<tr>
<td>Cumulative %Var</td>
<td>40.095</td>
<td>58.752</td>
<td>73.38</td>
<td>84.223</td>
<td>92.59</td>
<td>97.335</td>
<td>100</td>
</tr>
</tbody>
</table>

Each cell represents the weight of each of the parameters included in the data set concerning the relevant PC (column).
4. AQ modelling and simulation with the aid of Artificial Neural Networks

4.1. Introduction

The use of Artificial Neural Networks (ANNs) was introduced by McCulloch and Pitts [14] in their proposal for a mathematical model of a neuron, while Alan Touring in a report titled “Intelligent Machinery”, that he prepared in 1948 but was published after his death, describes a neural network among other artificial intelligence ideas [25]. Related literature is very extensive, providing with a deep and comprehensive view on ANNs [1,8,9,19]. Applications of ANN in AQ have been increasing in the last years in Greece [4,3,7,23,22,24]. Nevertheless, for the Thessaloniki Greater Area only one paper has been previously published [22] concerning PM10 and not ozone, suggesting that there are open issues that need to be addressed in both the ANNs application and the domain knowledge simulation and modelling for this geographic area.

4.2. The Thessaloniki ANN application: multi-layer perceptrons (MLP)

In simple neural network architecture, only unidirectional forward connections are allowed (feed-forward neural network). The simplest type of such a network is the perceptron, consisting of only one set of neural units that are connected with an input layer. Yet, such architecture is not capable of solving a large class of problems, as it was proven by Minsky and Papert [15]. Nevertheless, they also showed that many of the disadvantages of the single-layer perceptrons can be surpassed by using one hidden layer of computational nodes, although they did not present a method for adjusting the weights. This problem was solved by [20], who introduced the idea of error back-propagation from the output layers to the hidden ones. Despite the fact that error back-propagation can be applied to neural networks with any number of hidden layers, it has been proved [11] that a single hidden layer is sufficient enough for the network to represent any function as long as the activation functions of the hidden nodes are nonlinear.

4.3. ANN models for modelling and forecasting ozone concentrations in Thessaloniki, Greece

The emphasis of the investigation was put on O₃ hourly concentration modelling for the purposes of forecasting. The datasets used include hourly concentration values for O₃, NO₂, and hourly records of temperature, humidity, wind speed and the transformed wind direction, as all these parameters were proven to be of importance for describing the investigated datasets via the PCA method results in the previous chapter. The observations took place at different locations within the city, and thus a different neural network model has been applied for each location, in order to increase local representativity of modelling results and the effective-
ness of simulations towards forecasts. The network architecture chosen was the multi-layer perceptron (MLP) with one hidden layer: the network consisted of one input layer, one hidden layer and the output layer. MLP models are widely applied in predicting air pollutant concentrations since they can capture the highly nonlinear relationship between the variables, [6,10,12,29]. In addition, MLP can be trained to approximate any smooth, measurable (highly nonlinear) function without prior assumptions concerning the data distribution [18].

The choice of one hidden layer was made after conducting several tests with various network structures, since it provided with lower error values and smaller convergence times. Thus, the hidden layer included 10 nodes, while the ANN required 500 runs for the training, and the activation function used was the sigmoid. The ANN simulation and modelling was performed with the aid of WEKA version 3.4, an open source machine learning software [27].

4.3.1. Kalamaria station

Data from the period 2001 to 2004 were used to build and test an ANN for forecasting $O_3$ hourly concentrations at the area of Kalamaria 24 h in advance. This means that the ANN built materialised a function $G$ where

$$\text{Forecasted } O_3(t + 24) = G[O_3(t), NO_2(t), \text{Temp}(t), \text{Hum}(t), \text{WS}(t), \sin WD(t), \cos WD(t)],$$

$t$ being the time variable

It should be noted that all parameters included in the ANN construction have been found to be influential for the pollutant studied via the PCA. The dataset available was divided so that the 2001–2003 data were used for training and developing the ANN, while the 2004 data were used as a test set for evaluating the model's performance. A 10-fold cross validation was also performed at the training dataset: the studied dataset was divided randomly into 10 subsets; training was performed on the basis of 9, and testing for the remaining one. The procedure is repeated 10 times, so that testing will be done with each one of the 10 subsets. The performance statistics are then calculated on the basis of their mean value from all 10 experiments [27]. The results are being presented at Table 5, with the aid of statistical indexes applied for the evaluation of model performance [28], as presented in detail in the relevant annex.

The performance of the model may be investigated with the aid of the index of agreement (IA) that presents an overall estimation of the agreement between modelled and actual data. It is clear that in both cases (with and without cross validation), the IA is very satisfactory, as it reaches 83%. Another useful criterion is the Critical Success Index that represents the number of emergencies that were successfully forecasted by the model. Here, in both cases, the model failed to provide with a valid forecast. The overall performance of the model is in line with the findings of [3], concerning the application of ANN for the same pollutant (ozone).

<table>
<thead>
<tr>
<th>Table 5</th>
<th>Evaluation of the ANN model for the Kalamaria station, for the forecasting of hourly ozone values 24 h in advance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Comparison with the 2004 data</strong></td>
<td></td>
</tr>
<tr>
<td>Correlation coefficient $r$</td>
<td>0.701</td>
</tr>
<tr>
<td>MAE</td>
<td>18.007</td>
</tr>
<tr>
<td>RMSE</td>
<td>22.807</td>
</tr>
<tr>
<td>RAE</td>
<td>71.668%</td>
</tr>
<tr>
<td>RRSE</td>
<td>73.286%</td>
</tr>
<tr>
<td>IA</td>
<td>0.83</td>
</tr>
<tr>
<td>CSI</td>
<td>$0 (A = 0, B = 2, C = 0)$</td>
</tr>
<tr>
<td><strong>Cross-validation</strong></td>
<td></td>
</tr>
<tr>
<td>Correlation coefficient $r$</td>
<td>0.754</td>
</tr>
<tr>
<td>MAE</td>
<td>15.526</td>
</tr>
<tr>
<td>RMSE</td>
<td>21.515</td>
</tr>
<tr>
<td>RAE</td>
<td>64.186%</td>
</tr>
<tr>
<td>RRSE</td>
<td>68.06%</td>
</tr>
<tr>
<td>IA</td>
<td>0.833</td>
</tr>
<tr>
<td>CSI</td>
<td>$0 (A = 0, B = 8, C = 0)$</td>
</tr>
</tbody>
</table>
for the forecast of hourly concentration levels 24 h in advance in Athens, Greece, ranging from 0.8 to 0.86, while they reported a RMSE of approx 20%.

4.3.2. Eleftherio-Kordelio station

Data from the air quality station of Eleftherio-Kordelio were used for training the ANN as in the case of the Kalamaria station, with the only difference that measurements from the period 2001–2002 were used for the training phase, while the year 2003 data were used as the forecasting performance test set. A 10-fold cross validation was also applied on the training dataset. The forecasting performance statistical indexes are presented in Table 6.

Again, the IA and the CSI are the two criteria that are of interest from the operational point of view of such a forecast. It is clear that the performance of the model is better in the case of the 10-fold cross validation. Reaching an IA of 83.7%, and yet failing again, as in the case of Kalamaria, to forecast episodes (CSI = 0).

5. Conclusions

In the present paper the application of Principal Component Analysis for environmental data investigations is presented, resulting in a set of parameters that were selected as input to atmospheric quality simulations. These simulations were performed with the aid of Artificial Neural Network models that are developed and tested for two locations within the Thessaloniki area, in Greece. On the basis of various tests concerning model setups and prediction, forecasting capabilities were investigated with the aid of appropriate statistical indexes. The neural network architecture that was finally adopted for use was the multi-layer perceptron (MLP) with one hidden layer of ten nodes. The performance of the models in both locations was very satisfactory, as they were able to successfully simulate the atmospheric time-series of interest (ozone concentration levels). It worthies noting that in another related study, Dutot et al. [5] reported their findings concerning the performance of an ANN for ozone, that made use of more detailed and comprehensive atmospheric data, which had an IA just reaching 92%, and a RMSE ranging from 15% to 18%. This indicates that data availability and quality are playing an important role on the development of such models for the purpose of forecasting. Overall, and on the basis of these findings, such models may be considered as appropriate for operational usage in urban air quality management. Moreover, it has been demonstrated that computational intelligence methods have a high potential for further research in the area of environmental modelling and simulation.

Acknowledgements

The authors would like to acknowledge the project www.airthess.gr, supported by the Organisation of the Master Plan of Thessaloniki and the Prefecture of Central Macedonia, Greece.
Appendix A. Statistical indexes

Model validation and performance is based on the following statistical measures [28].

- Correlation coefficient
  \[ r = \frac{S_{pA}}{\sqrt{S_pS_A}} \]
  where \( S_{pA} = \frac{\sum_i (p_i - \bar{p})(a_i - \bar{a})}{n - 1} \), \( S_p = \frac{\sum_i (p_i - \bar{p})^2}{n - 1} \) and \( S_A = \frac{\sum_i (a_i - \bar{a})}{n - 1} \)

\( P \) refers to forecasted values, while \( A \) to actual (observed) ones.

- Mean absolute error (MAE):
  \[ |p_i - a_i| \]

- Root mean squared error (RMSE):
  \[ \sqrt{\frac{\sum_i (p_i - a_i)^2 + \sum_j (p_j - a_j)^2}{n}} \]

- Relative absolute error (RAE):
  \[ \frac{\sum_i |p_i - a_i|}{\sum_j |p_j - a_j|} \]

- Root relative squared error (RRSE):
  \[ \sqrt{\frac{\sum_i (p_i - a_i)^2 + \sum_j (p_j - a_j)^2}{\sum_i (a_i - \bar{a})^2 + \sum_j (a_j - \bar{a})^2}} \]

- Index of agreement (IA): \( IA = 1 - \frac{\sum_i (p_i - a_i)^2}{\sum_j (p_j - a_j)^2 + \sum_k (a_k - \bar{a})^2} \)

The value of the IA is from 0 to 1 and is a measure of the agreement between forecasts and observations. The best performance corresponds to the value 1.

- Critical success index (CSI): \( CSI = \frac{A}{A + B + C} \)

Where \( A \) is the number of the cases for which an episode occurred (hourly \( O_3 \) concentration values higher than 180 \( \mu g/m^3 \)) and was also forecasted, \( B \) the cases where an episode occurred but was not forecasted and \( C \) the cases where an episode was forecasted but did not occur.

References