Atmospheric Environment 43 (2009) 5588–5596

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Neural network forecasting of air pollutants hourly concentrations using optimised temporal averages of meteorological variables and pollutant concentrations

Lovro Hrust [a,*], Zvjezdana Bencetić Klaić [b], Josip Križan [a], Oleg Antonić [c], Predrag Hercog [d]

[a] Oikon Ltd., Institute for Applied Ecology, Avenija Dubrovnik 6-8, 10000 Zagreb, Croatia
[b] Andrija Mohorovičić Geophysical Institute, Faculty of Science, University of Zagreb, Horvatovac 95, 10000 Zagreb, Croatia
[c] Ruđer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia
[d] Institute of Public Health dr. Andrija Štampar, Mirogojska c. 16, 10000 Zagreb, Croatia

ARTICLE INFO

Article history:
Received 27 April 2009
Received in revised form
25 July 2009
Accepted 29 July 2009

Keywords:
Multi-layer perceptron neural networks
Air quality forecasting
Model input selection

ABSTRACT

The new method for the forecasting hourly concentrations of air pollutants is presented in the paper. The method was developed for a site in urban residential area in city of Zagreb, Croatia, for four air pollutants (NO_2, O_3, CO and PM_{10}). Meteorological variables and concentrations of the respective pollutant were taken as predictors. A novel approach, based on families of univariate regression models, was employed in selecting the averaging intervals for input variables. For each variable and each averaging period between 1 and 97 h, a separate model was built. By inspecting values of the coefficient of correlation between measured and modelled concentrations, optimal averaging periods for each variable were selected. A new dataset for building the forecasting model was then calculated as temporal moving averages (running means) of former variables. A multi-layer perceptron type of neural networks is used as the forecasting model. Index of agreement, calculated for the entire dataset including the data for model building, ranged from 0.91 to 0.97 for the respective pollutants. As suggested by the analysis of the relative importance of the input variables, different agreements for different pollutants are likely due to different sources and production mechanisms of investigated pollutants. A comparison of the new method with more traditional method, which takes hourly averages of the forecast hour as input variables, showed similar or better performance. The model was developed for the purpose of public-health-oriented air quality forecasting, aiming to use a numerical weather forecast model for the prediction of the part of input data yet unknown at the forecasting time. It is to expect that longer term averages used as inputs in the proposed method will contribute to smaller input errors and the greater accuracy of the model.

1. Introduction

Forecasting the concentrations of air pollutants represents a difficult task due to the complexity of the physical and chemical processes involved. Several approaches have been used, branching into two main streams: deterministic approaches, which involve numerically solving a set of differential equations, and empirical approaches, where different functions are used in order to approximate the concentrations of the pollutants depending on the external conditions.

The first type of approach does not require a large quantity of measured data, but it demands sound knowledge of pollution sources, the temporal dynamics of the emission quantity, the chemical composition of the exhaust gasses and physical processes in the atmospheric boundary layer. This crucial knowledge is often limited and also requires computational resources. Thus, approximations and simplifications are often employed in the modelling process. On the other hand, applications of such deterministic models are limited to a lesser extent regarding the selection of domain. A recent example of such an approach is the work of Finardi et al. (2008).

On the contrary, the second type of approach usually requires a large quantity of measured data collected under a large variety of atmospheric conditions. By applying regression and machine learning techniques, a number of functions can be used to fit the pollution data in terms of selected predictors. One drawback of this technique is that the model is usually confined to the area and conditions present during the measurements (e.g., Kukkonen et al., 2003; Niska et al., 2005). Nevertheless, this approach is

* Corresponding author. Tel.: +385 981932404.
E-mail address: lhrust@inet.hr (L. Hrust).

1352-2310/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.atmosenv.2009.07.048

generally more suitable for the description of complex site-specific relations between concentrations of air pollutants and potential predictors, and consequently, it often results in a higher accuracy, as compared to deterministic models. Gardner and Dorling (1999), for example, pointed to the complex human, weather and air pollution interaction, which is impossible to include in deterministic models without building a separate empirical model. They assumed that subtle influences determining the nature of emissions, such as an increase in people driving to work when it is cold and wet, cause the neural networks to outperform linear regression.

Neural network empirical approaches have been frequently used in recent atmospheric (e.g., Nath et al., 2008; de Oliveira et al., 2009) and air quality modelling studies. To our knowledge, Božnar et al. (1993) were the first to describe neural network modelling of the hourly concentrations of sulphur dioxide. Gardner and Dorling (1998) gave a very informative review of the applications of artificial neural networks in science in general and, particularly, in atmospheric sciences. They emphasised the usefulness of neural networks (NN) when dealing with non-linear systems, especially when theoretical models of the system cannot be constructed. They also accentuated the importance of understanding NN theory. NNs should not be used without an understanding of their advantages and disadvantages. A theoretical background makes it possible to build more accurate NNs by using various NN architectures and various algorithms for training. In another paper, the same authors (Gardner and Dorling, 1999) discussed neural network modelling of hourly NO_x concentrations on the basis of meteorological data. They showed a prevalence of neural networks, as compared to regression-based models, and they pointed out the ease of neural network training, without the need for external guidance. Perez et al. (2000) developed a multi-layer perceptron (MLP) type of neural network model to predict PM_{10} hourly concentrations by fitting a function of 24-hourly average concentrations from the previous day. They compared it with a linear regression and persistence models. However, they found errors ranging between 30% and 60%. In order to decrease the errors, they considered noise reduction in the data, rearrangement, an increase in the learning dataset and the inclusion of meteorological variables as input variables. They concluded that noise reduction prior to modelling is necessary. Furthermore, the relevant information is contained in the time structures of the same variable. The possible improvement can be achieved by explicitly taking into account the relevant meteorological variables. Karppinen et al. published two companion papers addressing the development of a modelling system for predicting NO_x and NO_2 concentrations in the urban environment of Helsinki. When these papers were written, modelling systems represented an important regulatory assessment tool for the national environmental authorities. The first paper (Karppinen et al., 2000a) was related to model development and its application to air quality prediction as well as traffic planning. The system includes the following models: the estimation of traffic volumes and travel speeds, the computation of emissions from vehicular sources, a model for stationary source emissions, a meteorological pre-processing model and dispersion models for stationary and mobile sources. Chemical interactions between stationary and mobile sources are allowed, which makes the entire modelling system more realistic. The companion paper (Karppinen et al., 2000b) described the comparison between the predicted and measured concentrations. According to the authors, the modelling system was fairly successful in predicting NO_x concentrations and was successful in predicting NO_2 concentrations. Kolehmainen et al. (2001) compared two popular kinds of neural networks, namely, self-organising maps and multi-layer perceptrons. They extracted periodic components out of a learning data set consisting of measured NO_2 concentrations and compared the results of a combined periodic regression method and an NN with an NN trained directly on unprocessed data. They concluded that MLP gives the best results if trained on the original data. They also argued that none of the methods are able to forecast the peak values, due to the under-representation of these cases in the total dataset. Perez and Reyes (2002) developed an MLP and linear model for predicting the maximum of 24-h average PM_{10} concentrations. As an input, they used PM_{10} hourly average concentrations at several times during the day as well as statistics of measured and forecasted meteorological variables. The statistics consisted of maxima, minima and averages for the two time intervals: first between 19:00 of the previous day and 18:00 of the present day and second for the whole next day. Differences between maximum and minimum temperature for the same time intervals were also used. They concluded that the selection of the input variables is more important than the type of model used (linear or MLP). Podnar et al. (2002) investigated the applicability of NNs as a forecasting tool of chemical tracer concentrations in complex terrains. Two kinds of models were investigated, one for daily averages and another for hourly averages. As inputs, surface as well as upper-air station meteorological measurements were used. The previous day's tracer concentration measurements at surface stations were also used as inputs, while the emissions were constant and were, therefore, omitted from the model input. They found good agreement between the measured and forecasted concentrations, where the correlation coefficients for the hourly concentrations were 0.844 and 0.896 for the training and testing sets, respectively. NN models compared to more traditional statistical methods showed significant improvements in the results. Kukkonen et al. (2003) performed an extensive evaluation of neural network models for the prediction of NO_2 and PM_{10} concentrations. The comparisons included five neural networks models and one linear and one deterministic model. The neural networks outperformed the other models, particularly the NN models that were built with the assumption of non-constant variance. The authors suggested that this resulted from the strong non-linearity between the NO_2 and PM_{10} concentrations measured at two stations and corresponding vehicular emissions and meteorological parameters. They also pointed out the negative aspects of using neural networks: they are spatially and temporally limited. They proposed using archived numerical weather forecasts to build models in which the numerical prognosis is used as an input. A special neural network MLP model for the prediction of the air quality index (API) was developed by Jiang et al. (2004). They corrected their model structure as well as the methods of training and significantly improved the results. They concluded that a simpler structure of the MLP model gives better results. Hooyberghs et al. (2005) focused on forecasting daily averages of particulate matter. They found that the boundary layer height is the most important input parameter, while an increase in the number of input parameters resulted in only slightly improved accuracy. They concluded that day-to-day fluctuations of PM_{10} concentrations in Belgian urban areas are dominantly driven by meteorological conditions. Niska et al. (2005) evaluated a combination of an MLP model and a HIRLAM prognostic model in order to predict NO_2 and PM_{10} concentrations in an urban environment. The authors used a novel method consisting of sensitivity analysis and a multi-objective genetic algorithm in order to select the optimal set of input variables. Pollutant concentration forecasts were substantially better when HIRLAM prognoses were used, as compared to measured or HIRLAM input analysis data. The performance of all models was worse in the course of air pollution episodes. Corani (2005) compared three different modelling techniques, namely, MLP, pruned NN and lazy learning (LL). No strong differences between the models were found. LL gave the best performance

indicators of average goodness of prediction, while the pruned NN was the best at predicting exceedances of defined thresholds. Perez and Reyes (2006) developed an MLP model to forecast the daily maxima of PM_{10} concentrations one day in advance. The same model was applied to five measuring stations in the city of Santiago, Chile. They compared values forecasted with MLP, linear and persistence models using the same input variables. They concluded that the MLP model performed well and that the relatively small differences between the linear and MLP models emphasised the importance of selecting the correct input variables. An interesting approach was attempted by Lu et al. (2006). A self-organising map (SOM) type of NN was first employed to classify meteorological conditions into different meteorological regimes. Then, an MLP was used to separately model ozone forecasts for each meteorological regime. The combined model explained at least 60% of the variance in the ozone concentrations. The authors compared an SOM and MLP combined model to 1) an MLP model; 2) SOM combined with multiple linear regression and, 3) multiple linear regression model. They found the combined SOM and MLP model to have the best prediction performance. Brunelli et al. (2007) tested a recurrent neural network (Elman model) for the prediction of daily maximum SO_2, O_3, PM_{10}, NO_2 and CO concentrations. Experimental trials showed that the model is appropriate, which obtained coefficients of correlation between forecasted and measured data ranging from 0.72 to 0.97. Experiments also showed somewhat better agreement between measured and forecasted daily maxima for Elman networks, as compared to MLP. However, a small number of qualitatively different elements were used as input into model: wind direction and intensity, barometric pressure and temperature. Considering that concentrations prior to the forecast were not used as input, it is to expect that recurrent Elman networks would produce better results.

The aim of this research was to develop a model that: 1) predicts hourly concentrations of CO, PM_{10}, NO_2 and O_3 at one representative location, on the basis of relevant meteorological variables and recent concentrations; 2) has acceptable accuracy in order to be applicable for public-health-oriented air quality forecasting; and 3) uses meteorological input that can be obtained from a routine weather prediction model. For details about the mentioned air pollutants, see, e.g., Brunelli et al. (2007).

Considering these requirements, it was appropriate to develop an empirical model. Among various machine learning techniques (for examples see e.g., Brunelli et al., 2007), we selected MLP type of neural networks, since several recent studies confirmed their applicability (e.g., Niska et al., 2005; Kukkonen et al., 2003; Hooyberghs et al., 2005; Perez et al., 2000). Based on the suggestion of Kolehmainen et al. (2001), special attention is given to the selection of time periods for the input data.

2. Materials and methods

2.1. Data sample

The idea behind building this model was that it is desirable for neural networks to learn as much as possible from connections between measured physical values and future pollutant concentrations, in order to achieve better accuracy. With that in mind, the results of numerical weather forecast models are not considered as an input for model building, since errors inherent to these models would be "remembered" in the NN model. The measured data are considered to be very close to the actual values (there are also measurement errors, but they should be significantly smaller).

At the investigated measuring site in Zagreb, the continuous measurement of several pollutants as well as some meteorological variables is performed. The station is situated in the northern, residential part of the town, at the southern slope of the Medvednica Mountain. The measured pollutants are NO_2, CO, PM_{10} (particles suspended in air having an aerodynamic diameter up to 10 μm) and O_3, while measurements of meteorological variables include relative humidity, wind direction and speed, air pressure and temperature. The monitoring site is the property of the Public Health Institute of the city of Zagreb and it is constantly maintained by a qualified staff. The data acquisition system was set to record 15-min averages. All instruments are placed in an isothermic shelter manufactured by Environment S.A. The measuring site is located at 45°50′N; 15°59′E; 175 m above sea level. It is located about 5 m from a street with moderate traffic intensity and is 400 m from a crossing with high traffic intensity. The street is a main road that leads to the main city cemetery.

Particulate matter PM_{10} is measured by the method of beta radiation absorption (Automated Equivalent Method: EQPM-0404-151) on an Environment S.A. Model MP101M PM_{10} Beta Gauge Monitor device Federal Register (2002). NO_2 and NO are measured by the method of chemiluminescence (Automated Reference Method: RFNA-0795-104) on an Environment S.A. Model AC31M chemiluminescence nitrogen oxide analyser (Federal Register, 1995). O_3 is measured by the method of UV photometry (Automated Equivalent Method: EQOA-0206-148) on an Environment S.A. Model O342M UV ozone analyser Federal Register (2002). Carbon monoxide is measured by the method of non-dispersive infrared spectroscopy (Automated Reference Method: RFNA-0206-1479) on an Environment S.A. Model CO12M gas filter correlation carbon monoxide analyser (Federal Register, 2004). The meteorological parameters are measured by automatic meteorological instruments.

The measurements began in the beginning of January 2004. The model development was based on a little less than a two–year time series of pollutant concentrations and meteorological data. Data filtering resulted in the omission of two periods: 1) March–April 2004, due to an instrument malfunction and 2) on 29 November 2004, between 13 and 22 LST, when a street chestnut salesman installed his furnace near the measuring station, which suddenly increased particulate matter concentrations. Additionally, some data were missing due to instrument calibrations or malfunctions. The total dataset contained 62 209 15-min measurement averages, with data missing in about 2–11% of the cases for a particular pollutant. For the purpose of model development, only cases with complete data (both concentrations and meteorology) were taken into account. Hourly averages for all variables were calculated from the 15-min values. If all four 15-min measured values for the respective hour were missing, the entire hour was omitted. Otherwise, the hourly average was calculated based on available 15-min values. The statistics of the measured variables are given in Table 1. Except for the wind direction, all variables were included in further analysis as measured. Due to its circular nature, which is not appropriate for the NN model, the wind direction was transformed into two components, eastern and northern. Emissions data were

Table 1
Statistics of measured values.

	Number of hourly values	Mean	St. Dev.
PM_{10} (μg m^{-3})	15 327	28.55	24.49
CO (mg m^{-3})	14 067	0.51	0.47
NO_2 (μg m^{-3})	15 200	24.18	20.73
O_3 (μg m^{-3})	15 337	53.81	34.21
Humidity (%)	13 876	72.96	19.57
Pressure (hPa)	15 337	982.49	7.50
Temperature (°C)	15 337	11.25	8.95
Wind speed (m s^{-1})	15 337	0.75	0.58
Northern wind component	15 337	0.22	0.66
Eastern wind component	15 337	0.08	0.59

not used in the model development, because they were unavailable. According to Gardner and Dorling (1999), models including or omitting NO_x emissions resulted in extremely similar results. Thus, at least for NO_2, we do not expect significant detrimental effects to the model accuracy.

2.2. Independent variables

The first step was to perform a Fourier analysis of the pollutant concentrations in order to determine which time predictors were the most appropriate as inputs. Analysis showed that the most prominent intervals were one day, half a day, approximately one month (23–29 days) and one week. While these intervals are obviously connected to human activity, we have no explanation for other intervals that emerged, such as 10 days for all pollutants and 38 days for NO_2. Based on the Fourier analysis results and human activities, three variables representing the time were selected. These are the hour of the day (UTC), the day of the week (varying from 1 to 7, where 1 corresponds to Monday) and the time of the year (hereafter TOY), which is taken as $\mathrm{TOY} = \cos(2\pi t/T)$, where t is the ordinal number of the day of the year and T is the number of days in the year. TOY therefore has one cycle over a year and reaches a maximum in the winter and a minimum in the summer.

The independent variables used as estimators of actual concentrations were relative humidity, wind speed, the northern and eastern components of the wind direction, air pressure, temperature and initial pollutant concentration for the particular day. As will be discussed in more detail in the next section, some variables were included in the model development as temporal moving averages (running means). Averaging periods were selected separately for each combination of air pollutant and meteorological variable. For this purpose, families of general linear models were used, in which the actual concentration of the particular pollutant was estimated as a function of time predictors and the specific meteorological variable using a second degree polynomial under different averaging periods (1–97 h, with a time step of 1 h). Final averaging periods were selected for each combination of air pollutant and meteorological variable by a graphical inspection of the models results.

For building the model, a multi-layer perceptron type of neural network was chosen. More details about MLPs and neural networks can be found in Haykin (1999) and Bishop (1995). For each pollutant (CO, NO_2, PM_{10} and O_3), a separate model was built. After some trial and error, network architectures with one or two hidden layers were selected. A different number of input variables for each pollutant were chosen according to the results of the general linear models. As the output variables, a time series of hourly mean pollutant concentrations was selected. All data were divided into training, verification and test sets. Verification and test sets were randomly chosen. However, the selection procedure was repeated until the difference between the variances for all three sets was less than 10%. Both sets comprised about 15% of the total data each. A large number of networks of different architectures were trained. The back propagation method of training was used, and the method of early stopping (e.g., Haykin, 1999) was applied to avoid overfitting. Based on the method of trial and

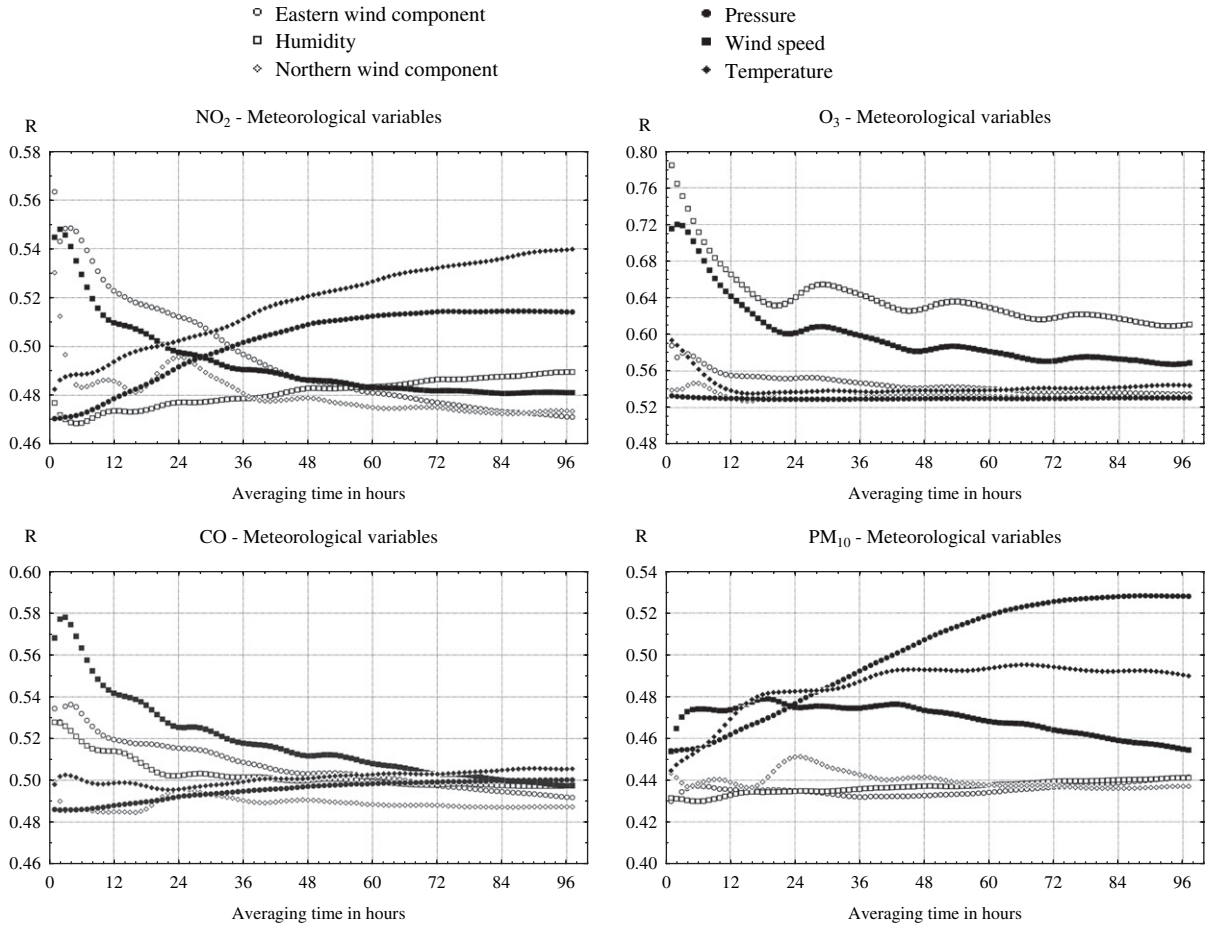

Fig. 1. Correlation coefficients between measured and modelled concentrations with respect to the averaging time interval of the input data.

error, learning rate and momentum were set to 0.1 and 0.3, respectively. Inspection of error function (sum of squared differences between measured and output values) on verification and test sets showed that maximum of 5000 epochs are needed for successful training. Each network had one or two hidden layers, where the activation function for each neuron in hidden layers was the logistic function, whereas activation function in the output layer was identity or logistic function. Experiments with tanh as an activation function were also performed. Since they resulted in similar performance measures, tanh was not used in developed models. The number of neurons in the hidden layers varied between 9 and 36. Ten networks with the smallest mean square errors for the test set were chosen as an ensemble. The average of the results of these ten networks was then taken as the final model result.

Initially, the following approach was used: for each pollutant, the model was built to predict the concentration only 1 h in advance for a given input. In order to obtain a forecast for the entire forecasting period, the predicted value for 1 h in advance was then used as an input value for the next hour and so on. However, the results of this approach show a substantial increase in error with increasing forecasting time. Therefore, another technique was used. That is, a new input variable, the time (in hours), was added. Additionally, instead of using the pollutant concentration of the previous hour, the initial pollutant concentration (taken as the hourly average between 5 and 6 UTC) was used.

3. Results and discussion

3.1. Selection of averaging periods

Correlation coefficients between modelled (obtained by general linear models) and measured pollutant concentrations for different averaging periods of input data are shown in Figs. 1 and 2.

Correlations between concentrations of different pollutants are well known (e.g., Kukkonen et al., 2001; Bešlić et al., 2005). Nevertheless, we did not use the concentration of one pollutant as an input parameter for the prediction of another pollutant, due to possible failures in concentration measurements. Thus, if there are no measurements for a particular pollutant, it is still possible to obtain forecasts for the other pollutants.

Based on visual inspection of Figs. 1 and 2, averaging periods were selected (Table 2) for each meteorological variable and separately for each pollutant. For some combinations of pollutant and meteorological variable, more than one local maximum of predictive power was recognised. In these cases, two independent estimators (based on the same meteorological variable, but under two different averaging periods) were selected as inputs for the respective pollutant. Otherwise, a particular meteorological variable was represented by only one estimator. For all pollutants except O_3, the correlation coefficient is the highest when the pressure is averaged over the preceding 97 h (Fig. 1 and Table 2). On the other hand, for O_3, the correlation coefficient is almost independent of the averaging interval. As far as the temperature is concerned, long averaging periods are related to PM_{10} (97 h) and NO_2 (67 h), while for O_3, short averaging periods are important. The correlation coefficients for the regression models built with various time-averaged values of relative humidity are almost constant, but slightly increase for PM_{10}, while they quickly fall for O_3 and CO. For NO_2, Fig. 1 shows an initial significant drop followed by a constant slight increase. Initial tests performed while building NN model for NO_2 with 1 and 97 h averages of the humidity as inputs indicated that using the latter input variable results in very small improvement in model accuracy. Thus, only 1 h average of the humidity was selected as the input variable. The same was with 25-h peak of north component of the wind for PM_{10}. However, subsequent tests showed that these were premature and false conclusions. From Fig. 1 and Table 2, it is clear that the short averaging intervals (2–3 h) are the most important for wind speed for all pollutants except for PM_{10} (19-h average). Correlations for regression models built with the eastern and northern components of the wind direction were the best for short averaging periods. There are also secondary maxima at 25 h, which are pronounced for all pollutants except O_3. These 25-h averages are likely due to periodic, up- and down-slope winds and human activities. The existence of periodic winds in Zagreb is well known (e.g., Lisac, 1984; Klaić et al., 2002, 2003), and it is also seen in Fig. 3. Fig. 2 reveals a decrease in the correlation coefficient as the averaging period increases. However, in operational use of the model, only concentrations up until the initial hour are known, while the meteorological variables in future applications are predicted by a weather forecast model. For this reason and

Table 2
Chosen averaging intervals (h). Averages are calculated backward from the forecast time. I denotes initial hour. 1 denotes hourly average for the hour of the forecast.

Variable	NO_2	O_3	CO	PM_{10}
Pressure	97	97	97	97
Temperature	97	1	3	67
Humidity	1 and 96[b]	1	1	97
Speed	2	2	3	19
Eastern wind component	1 and 4	1 and 3	4	6
Northern wind component	1 and 25	5	1 and 25	1, 9[a] and 25[b]
Concentration of the same pollutant	I	I	I	I

[a] Marked intervals were incorrectly chosen as inputs for 'main' model (referred to as 'NN model optim.' in Table 3). This was corrected for other models.
[b] Marked intervals were incorrectly omitted as inputs for 'main' model. This was corrected for other models.

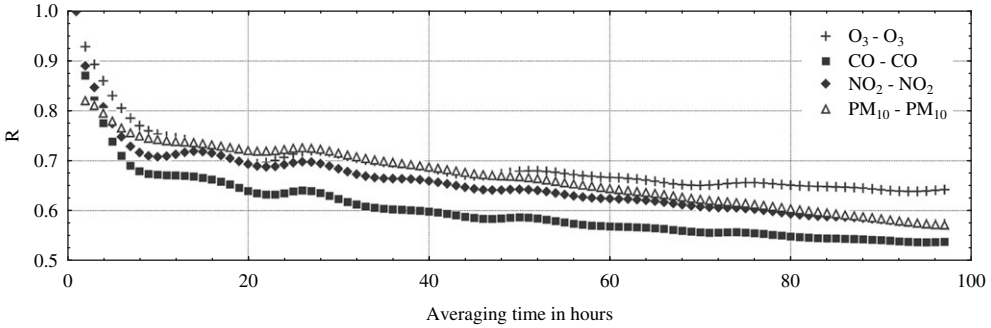

Fig. 2. Correlation coefficients obtained by general linear models. For each point on the graph, a different averaging period for the respective input variable is selected. Each pollutant is given as a function of the same pollutant and averaging time.

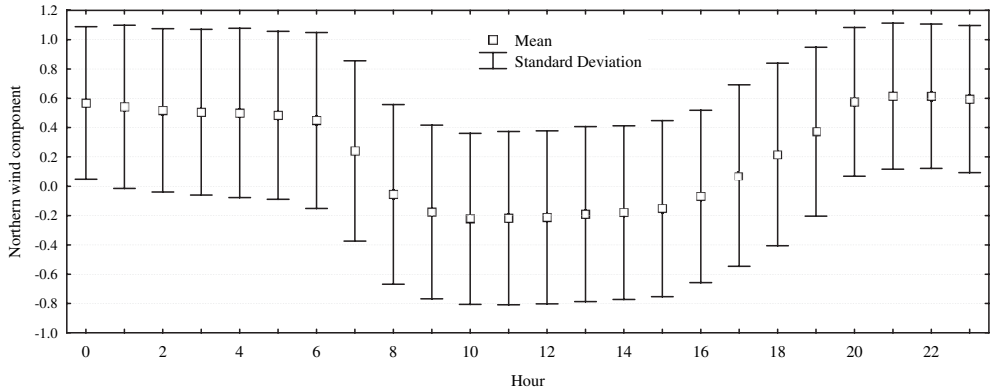

Fig. 3. Diurnal variation of the northern component of the wind direction for the year 2004.

simplicity, we chose only hourly concentration values of the same pollutant in the initial hour (5–6 UTC) as input. Different averaging intervals for pollutants would require further analysis on an hourly basis. This analysis would very likely result in different averaging intervals for each hour of the forecast, making the model much more complicated.

3.2. Model performance and comparison with other models

In order to illustrate the model performance, two multiday periods, one with good (shown for O_3) and another with poorer agreement (shown for PM_{10}), are shown in Fig. 4. The performance of the model for each pollutant and the entire modelling period is listed in Table 3 (column NN model optim). Additionally, a comparison with other models, namely, the persistence, linear and two NN models with one hidden layer containing 22 neurons, is also shown. The prediction of the persistence model for a particular hour simply corresponds to the value measured at the same time on the previous day. The linear model is a linear combination of inputs, where each continuous variable is multiplied by a factor of weight, which is determined by a method of error minimisation, the least square means in our case. Discrete variables (i.e., hour and day of the week) are accounted in such a way that each day, hour and combination of day and hour has a separate weight that is also determined by a method of error minimisation. Two additional neural network models (NN 22 last and NN 22 optim. shown in Table 3) were built separately from the 'main' model in order to test the performance of the novel method of time-averaged optimised inputs with respect to the inputs corresponding to the term of the forecast. The optimal number of neurons for the architecture consisting of one hidden layer was about 22 neurons, where the range between 6 and 25 neurons was inspected for PM_{10} and CO. Due to the time-consuming procedure of selecting the optimal number of neurons, it was assumed that the same optimal number 22 is also valid for O_3 and NO_2. Afterwards, 50 MLP networks for each pollutant and two input sets (one for NN 22 last and another for NN 22 optim.) were built, using the same number of neurons in the hidden layer (22) and the same

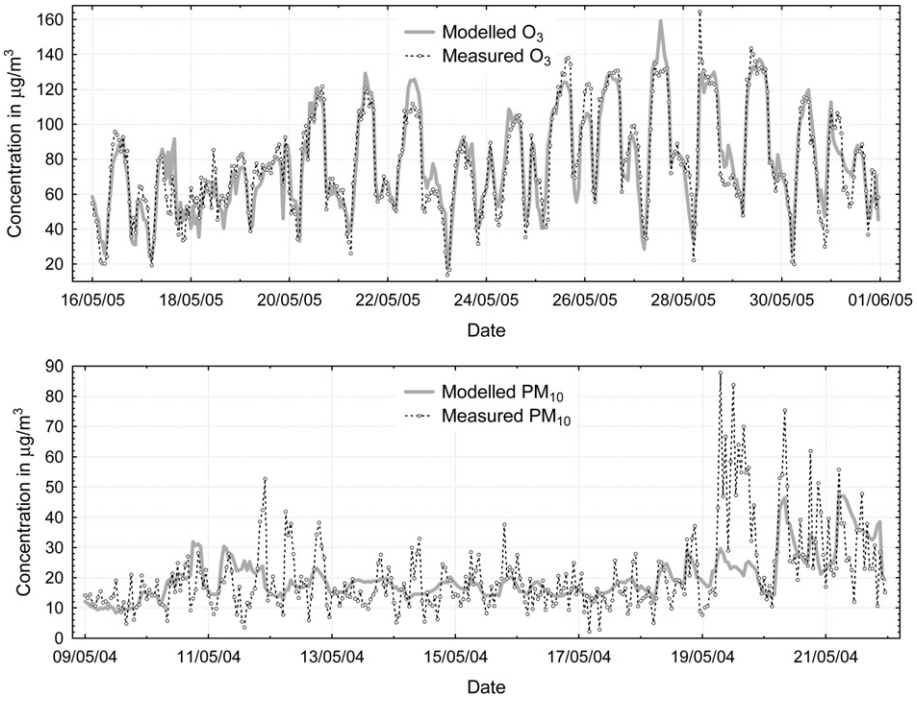

Fig. 4. Hourly mean modelled and measured concentrations for O_3 (top) and PM_{10} (bottom).

Table 3
Statistics of model performance for the first 24 h of model predictions. The statistics are determined for the whole period on which the model was built and for all subsets of data (training, testing and verification). The standard deviation of the measured data is denoted by St. Dev, the ratio of standard deviation and mean is denoted by SD/Mean. NN 22 corresponds to an ensemble of 50 neural networks with 22 neurons in the hidden layer. The designation "last" indicates that the input consists of values measured at the time of the forecast, while "optim." indicates that the input consists of optimised time averages. MAE and RMSE are given in μg m^{-3} (NO_2, O_3 and PM_{10}) and mg m^{-3} (CO).

	Mean St. dev. SD/mean	Perf. meas.	Persistence model	Linear model last	Linear model optim.	NN 22 last	NN 22 optim.	NN model optim.[a]
	24.18	MAE	13.27	9.13	9.00	6.04	5.43	5.34
NO_2	20.73	RMSE	19.33	12.94	12.77	8.86	7.95	7.56
$N = 10\,247$	0.86	IA	0.76	0.87	0.87	0.95	0.96	0.96
		R^2	0.33	0.61	0.63	0.82	0.86	0.87
	53.81	MAE	21.10	14.74	14.01	8.75	8.49	8.26
O_3	34.21	RMSE	27.92	19.08	18.16	11.60	11.24	10.86
$N = 10\,360$	0.64	IA	0.82	0.91	0.92	0.97	0.97	0.97
		R^2	0.46	0.70	0.73	0.89	0.90	0.90
	0.51	MAE	0.31	0.23	0.23	0.17	0.16	0.16
CO	0.47	RMSE	0.50	0.35	0.35	0.26	0.24	0.24
$N = 9400$	0.92	IA	0.69	0.80	0.81	0.91	0.93	0.93
		R^2	0.24	0.49	0.50	0.72	0.77	0.77
	28.55	MAE	14.86	12.44	12.42	10.34	10.11	9.24
PM_{10}	24.49	RMSE	23.23	18.60	18.54	15.34	14.70	13.26
$N = 10\,444$	0.86	IA	0.75	0.78	0.79	0.87	0.89	0.91
		R^2	0.33	0.46	0.47	0.63	0.66	0.72

[a] Due to an error in choosing optimising intervals for NO_2 and PM_{10}, it is to assume that performance measures for NN model optim. can be somewhat improved for these pollutants.

training parameters. Broyden–Fletcher–Goldfarb–Shanno algorithm (e.g., Haykin, 1999; Bishop, 1995) was used with maximum of 450 training epochs. Activation functions for neurons in hidden and output layer were the logistic function and the identity, respectively. Early stopping method (e.g., Haykin, 1999) was used to select the model with the best generalization performance. The final result of models is then constructed as an average value of output values of mentioned 50 networks. The measures of model performance shown in Table 3 are the mean absolute error (MAE), which is recommended by Wilmott (2005), the root mean square error (RMSE), the index of agreement (IA) and the linear correlation coefficient (R). Although R is not recommended as a measure of model performance (Wilmott, 1981, 1982), it is shown here in order to enable a comparison with the results of other authors. The measures were calculated as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_i |P_i - O_i|, \tag{1}$$

$$\mathrm{RMSE} = \left[\overline{(P_i - O_i)^2}\right]^{1/2}, \tag{2}$$

$$\mathrm{IA} = 1 - \frac{\overline{(P_i - O_i)^2}}{\left[|P_i - \overline{O}| + |O_i - \overline{O}|\right]^2}, \tag{3}$$

$$R^2 = \left[\sum_i (P_i - \overline{P})(O_i - \overline{O})\right]^2 \Big/ \sum_i (O_i - \overline{O})^2 \sum_i (P_i - \overline{P})^2, \tag{4}$$

where P_i and O_i are predicted and observed hourly mean concentrations. The measures (1)–(4) were tested on the original dataset (which comprises the training, verification and testing subsets).

As seen from Table 3, the best agreement is obtained for O_3, followed by NO_2, CO and PM_{10}. The best agreement for O_3 is probably due to the mechanism of O_3 production, in which solar radiation plays a major role. Namely, the pattern of solar radiation is regular both on a daily (incoming radiation is disturbed solely by clouds) and yearly basis. Thus, it is well incorporated in the model through simple functions of the variables hour and TOY, respectively. On the other hand, PM_{10}, which originates from many sources, such as soil dust, road dust due to resuspension or tire and clutch wear, construction work and plants, is produced by substantially more irregular processes. This reasoning is supported by the ratios of standard deviation and mean value for O_3 and PM_{10} (Table 3).

Table 3 shows that the developed model (NN model optim.) performed the best, while the performance of the two additional NN models (NN 22 last and NN 22 optim.) was somewhat poorer, with NN 22 optim being slightly better for PM_{10}, NO_2 and CO and very similar for O_3, compared to NN 22 last. As expected, the poorest performance was obtained by the persistence model.

Sensitivity analysis for the model forecasting from the hour following initial hour (6 UTC) to 23 UTC of the same day was performed. The variable tested was kept as constant at its average value for the entire modelling period, while the other variables were taken as described in the previous sections. RMSE values for each variable are shown in Table 4. Referent values $RMSE_0$ were calculated using original input dataset. The abbreviations used for the different input variables are as follows. The pollutant name is followed by the letter "I", which indicates that concentrations correspond to initial hour, namely the time between 5 and 6 UTC. The meteorological variables are followed by a number denoting the averaging time (in hours), which spans from the hour of forecast backwards. Here, the greater RMSE indicates that the observed pollutant has a greater impact on the accuracy of the prediction. Relative RMSE error, shown in table as "Rel." is the ratio between the RMSE and $RMSE_0$.

Table 4 further confirms that O_3 and PM_{10} are modelled the best and the worst, respectively. The modelled O_3 concentration shows a very strong dependence on TOY (which is directly related to solar radiation and is, therefore, very important for ozone production), temperature and air humidity. On the other hand, the concentration of PM_{10} is much less sensitive (smaller values of Rel.) to the inspected variables, where the most important among them is the average measured concentration of PM_{10} for the initial hour (5–6 UTC). It is seen that the pollutants, which are better

Table 4
Sensitivity analysis for the modelling interval from the hour following initial hour (6 UTC) to 23 UTC of the same day (until midnight in local time). Here, the greater RMSE indicates that this variable is more important to the model performance. $RMSE_0$ is a reference value calculated using original input dataset. CO_I is the hourly average value for the initial hour, namely between 5 and 6 UTC, and HUMID_1 is the humidity for the hour of the forecast. SPEED_2 is the average value of 2 h, the hour of prediction and the preceding hour. Rel. is an abbreviation for relative RMSE error = $RMSE/RMSE_0$.

NO_2			O_3			CO			PM_{10}		
Input var.	RMSE	Rel.	Input var.	RMSE	Rel.	Input var.	RMSE	Rel.	Input var.	RMSE	Rel.
NO2_I	13.38	1.68	TOY	19.84	1.87	Hour	0.35	1.46	PM10_I	19.30	1.38
Temp_97	12.45	1.56	Temp_1	18.39	1.74	TOY	0.33	1.37	TOY	18.06	1.29
TOY	11.62	1.46	Humid_1	17.69	1.67	CO_I	0.33	1.36	Press_97	17.05	1.22
Hour	11.58	1.45	Hour	14.04	1.32	Wday	0.31	1.30	Wday	16.97	1.21
East_1	10.90	1.37	Speed_2	14.02	1.32	Speed_3	0.31	1.28	North_1	16.48	1.18
Wday	10.14	1.27	East_1	13.98	1.32	Temp_3	0.29	1.21	Temp_67	16.32	1.17
Speed_2	9.86	1.24	O3_I	13.90	1.31	North_1	0.28	1.16	North_9	15.59	1.11
Humid_1	9.83	1.23	Wday	13.23	1.25	Humid_1	0.28	1.15	Speed_19	15.30	1.09
Press_97	9.62	1.21	North_5	11.73	1.11	East_4	0.27	1.14	Hour	15.25	1.09
North_1	9.42	1.18	East_3	11.55	1.09	Press_97	0.27	1.14	Humid_97	15.20	1.09
East_4	8.71	1.09	Press_97	11.38	1.07	North_25	0.26	1.08	East_6	14.89	1.06
North_25	8.46	1.06									
$RMSE_0$	7.96	1.00	$RMSE_0$	10.60	1.00	$RMSE_0$	0.24	1.00	$RMSE_0$	14.00	1.00

modelled, are also more sensitive to all input variables (i.e., have larger values of Rel. in Table 4). Thus, for O_3 with a constant TOY (i.e., a TOY value equal to the average value), Rel. for the entire data set is 1.87, while for PM_{10} with a constant PM_{10}_I, Rel. is only 1.38. The same could be observed when comparing other input variables in decreasing order of importance. This leads to the conclusion that the importance of specific input variables for the prediction of O_3 is more prominent, as compared to the importance of specific variables for PM_{10}. This could be due to the more random behaviour of PM_{10} concentrations (i.e., more noise in the signal) or due to the omission of other relevant variables/processes (such as boundary layer height, emissions, resuspension, etc.).

For NO_2, the past concentrations are the most important. These are followed by temperature, TOY and hour, thus indicating the importance of chemical and photochemical reactions for the production and decay of NO_2 (e.g., Eschenroeder, 1982). The temporal variables (hour and TOY) and past concentrations are the most important for CO. For PM_{10}, the most important variables are the concentrations measured during initial hour and TOY. These are followed by several almost equally important variables.

4. Conclusion

An MLP type of NN was used to build a prognostic model for forecasting hourly concentrations of CO, PM_{10}, NO_2 and O_3 at an urban residential location with moderate traffic. The model is built on measured meteorological data and concentrations of the pollutant concerned. A novel approach, based on general linear models, is employed in selecting the averaging interval over which each input variable is averaged. In such a way, selection of the input variables for the model is based on an objective method. Information about the average value of some parameter during the last several hours, as is shown here, is equally or more valuable for the prediction of pollutant concentrations than the value at the forecasting time.

Generally good agreement between the prognostic and observed pollutant concentrations confirms the relationship between the physical state of the atmosphere (which is described by meteorological data) and the fate of pollutants. The agreement between the modelled and measured concentrations decreased in the following order: O_3, NO_2, CO, PM_{10}. The best agreement obtained for O_3 suggests the major role of the production mechanism (i.e., solar radiation) for forecasted concentrations of O_3. In the case of CO, the temporal variables, which imply variation of human activities, such as traffic, heating, etc., are the most important. The poorer agreement between the modelled and measured PM_{10} concentrations can be attributed to the irregularity of the processes affecting particle production (such as traffic, dust storms, resuspension etc.) or due to the omission of relevant input variables (such as boundary layer height, etc.). In summary, we may conclude that the most prominent factors affecting the investigated pollutants are either time or temperature dependent.

In this study, the measured data were employed as meteorological input. However, the employed input meteorological variables are generally available from routine weather prediction models. Thus, the developed model can be used for operational forecasts of air pollution. Furthermore, the advantage of the developed model is that it uses meteorological variables averaged backwards in time from the moment $(t + \Delta t)$ to the moment $(t + \Delta t -$ averaging interval). Also, part of the data required for calculating variables averaged over longer periods, from the moment $(t + \Delta t -$ averaging interval) to the moment (t), will be already measured at the monitoring station at the time of the forecast. Thus, we expect smaller input errors since the forecast errors decrease with the decrease of the time (Δt), whereas already measured data have significantly smaller errors, under consumption of regular maintenance of the instruments. Consequently, accuracy of the model should be improved.

Acknowledgments

LH and ZBK involvement in this study is supported by the Ministry of Science, Education and Sports of the Republic of Croatia Project "Air quality over complex topography". The model was developed at the Oikon Ltd., Institute for Applied Ecology, for the purpose of forecasting hourly and daily concentrations of NO_2, CO, PM_{10} and O_3.

References

Bešlić, I., Šega, K., Šišović, A., Klaić, Z.B., 2005. PM_{10}, CO and NO_x concentrations in the Tuhobić road tunnel, Croatia. International Journal of Environment and Pollution 25, 251–262.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Clarendon Press, Oxford.

Božnar, M., Lesjak, M., Mlakar, P., 1993. A neural network-based method for the short time predictions of ambient SO_2 concentrations in highly polluted industrial areas of complex terrain. Atmospheric Environment 27B, 221–230.

Brunelli, U., Piazza, V., Pignato, L., Sorbello, F., Vitabile, S., 2007. Two-days ahead prediction of daily maximum concentrations of SO_2, O_3, PM_{10}, NO_2, CO in the urban area of Palermo, Italy. Atmospheric Environment 41, 2967–2995.

Corani, G., 2005. Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning. Ecological Modelling 185, 513–529.

de Oliveira, M.M.F., Ebecken, N.F.F., de Oliveira, J.L.F., Santos, I.D., 2009. Neural network model to predict a storm surge. Journal of Applied Meteorology and Climatology 48 (1), 143–155.

Eschenroeder, A., 1982. Atmospheric dynamics of NO_x emission controls. The Science of the Total Environment 23, 71–90.

Federal Register, 1995, Office of the Federal Register, National Archives and Records Administration, 60, 38326.

Federal Register, 2002, Office of the Federal Register, National Archives and Records Administration, 67, 42557.

Federal Register, 2004, Office of the Federal Register, National Archives and Records Administration, 69, 18569.

Finardi, S., De Maria, R., D'Allura, A., Cascone, C., Calori, G., Lollobrigida, F., 2008. A deterministic air quality forecasting system for Torino urban area, Italy. Environmental Modelling & Software 23, 344–355.

Gardner, M.W., Dorling, S.R., 1998. Artificial neural networks (the multilayer perceptron) – a review of applications in the atmospheric sciences. Atmospheric Environment 32, 2627–2636.

Gardner, M.W., Dorling, S.R., 1999. Neural network modelling and prediction of hourly NO_x and NO_2 concentrations in urban air in London. Atmospheric Environment 33, 709–719.

Haykin, S., 1999. Neural Networks: a Comprehensive Foundation, second ed. Prentice Hall, Upper Saddle River, NJ, pp. 237–239.

Hooyberghs, J., Mensink, C., Dumont, G., Fierens, F., Brasseur, O., 2005. A neural network forecast for daily average PM_{10} concentrations in Belgium. Atmospheric Environment 39, 3279–3289.

Jiang, D., Zhang, Y., Xiang, H., Zeng, Y., Jianguo, T., Demin, S., 2004. Progress in developing an ANN model for air pollution index forecast. Atmospheric Environment 38, 7055–7064.

Karppinen, A., Kukkonen, J., Elolähde, T., Konttinen, M., Koskentalo, T., Rantakrans, E., 2000a. A modelling system for predicting urban air pollution: model description and applications in the Helsinki metropolitan area. Atmospheric Environment 34, 3723–3733.

Karppinen, A., Kukkonen, J., Elolähde, T., Konttinen, M., Koskentalo, T., 2000b. A modelling system for predicting urban air pollution: comparison of model predictions with the data of an urban measurement network in Helsinki. Atmospheric Environment 34, 3735–3743.

Klaić, Z.B., Nitis, T., Kos, I., Moussiopoulos, N., 2002. Modification of the local winds due to the hypothetical urbanization of the Zagreb surroundings. Meteorology and Atmospheric Physics 79, 1–12.

Klaić, Z.B., Belušić, D., Bulić, I.H., Hrust, L., 2003. Mesoscale modeling of meteorological conditions in the lower troposphere during a winter stratospheric ozone intrusion over Zagreb, Croatia. Journal of Geophysical Research 108, 0148–0227.

Kolehmainen, M., Martikainen, H., Ruuskanen, J., 2001. Neural networks and periodic components used in air quality forecasting. Atmospheric Environment 35, 815–825.

Kukkonen, J., Härkönen, J., Karppinen, A., Pohjolaa, M., Pietarila, H., Koskentalo, T., 2001. A semi-empirical model for urban PM_{10} concentrations, and its evaluation against data from an urban measurement network. Atmospheric Environment 35, 4433–4442.

Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorling, S., Chatterton, T., Foxall, R., Cawley, G., 2003. Extensive evaluation of neural network models for the prediction of NO_2 and PM_{10} concentrations, compared with a deterministic modelling system and measurements in central Helsinki. Atmospheric Environment 37, 4549–4550.

Lisac, I., 1984. The wind in Zagreb (a contribution to the knowledge of climate of the city of Zagreb, II). Geofizika 1, 47–123.

Lu, H.C., Hsieh, J.C., Chang, T.S., 2006. Prediction of daily maximum ozone concentrations from meteorological conditions using a two-stage neural network. Atmospheric Research 81, 124–139.

Nath, S., Mitra, A.K., Bhowmik, S.K.R., 2008. Improving the quality of INSAT derived quantitative precipitation estimates using an neural network method. Geofizika 25 (1), 41–51.

Niska, H., Rantamäki, M., Hiltuinen, T., Karpinen, A., Kukkonen, J., Ruuskanen, J., Kolehmainen, M., 2005. Evaluation of an integrated modelling system containing a multi-layer perceptron model and the numerical weather prediction model HIRLAM for the forecasting of urban airborne pollutant concentrations. Atmospheric Environment 39, 6524–6536.

Perez, P., Trier, A., Reyes, J., 2000. Prediction of $PM_{2.5}$ concentrations several hours in advance using neural networks in Santiago, Chile. Atmospheric Environment 34, 1189–1196.

Perez, P., Reyes, J., 2002. Prediction of maximum of 24-h average of PM_{10} concentrations 30 h in advance in Santiago, Chile. Atmospheric Environment 36, 4555–4561.

Perez, P., Reyes, J., 2006. An integrated neural network model for PM_{10} forecasting. Atmospheric Environment 40, 2845–2851.

Podnar, D., Koračin, D., Panorska, A., 2002. Application of artificial neural networks to modeling the transport and dispersion of tracers in complex terrain. Atmospheric Environment 36, 561–570.

Wilmott, C.J., 1981. On the validation of models. Physical Geography 2, 184–194.

Wilmott, C.J., 1982. Some comments on the evaluation of model performance. Bulletin American Meteorological Society 63 (11), 1309–1313.

Wilmott, C.J., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research 30, 79–82.