ELSEVIER

Chemosphere 63 (2006) 1575–1582

CHEMOSPHERE

www.elsevier.com/locate/chemosphere

Adaptive neuro-fuzzy based modelling for prediction of air pollution daily levels in city of Zonguldak

Yilmaz Yildirim [a,*], Mahmut Bayramoglu [b]

[a] Zonguldak Karaelmas University, Department of Environmental Engineering, Incivez Street, 67100 Zonguldak, Turkey
[b] Gebze Institute of Technology, Department of Chemical Engineering, 41420 Kocaeli, Turkey

Received 13 June 2005; received in revised form 24 August 2005; accepted 30 August 2005
Available online 28 November 2005

Abstract

Air pollution is a growing problem arising from domestic heating, high density of vehicle traffic, electricity production, and expanding commercial and industrial activities, all increasing in parallel with urban population. Monitoring and forecasting of air quality parameters in the urban area are important due to health impact. Artificial intelligent techniques are successfully used in modelling of highly complex and non-linear phenomena. In this study, adaptive neuro-fuzzy logic method has been proposed to estimate the impact of meteorological factors on SO_2 and total suspended particular matter (TSP) pollution levels over an urban area. The model forecasts satisfactorily the trends in SO_2 and TSP concentration levels, with performance between 75–90% and 69–80 %, respectively.

Keywords: SO_2 pollution; Total suspended particle (TSP) pollution; Neural networks; Fuzzy logic; Modelling

1. Introduction

In recent years, artificial intelligence (AI) based techniques have been proposed as alternatives to traditional statistical ones in many scientific disciplines. Artificial neural networks (ANN), one of the most popular AI methods, are considered to be simplified mathematical models of brain-like systems. Neural networks are generally trained by means of "training data", and due their property of generalization, they can learn new associations, new functional dependencies and new patterns. Due to these properties, they have been widely used for modelling and forecasting. Especially, the "multilayer perceptron" has been applied within the field of air quality prediction in the last decade. A summarized review of the applications of ANN in the atmospheric sciences has been carried out by Gardner and Dorling (1998). ANN models have been studied by various investigators for SO_2 (Boznar et al., 1993; Mlakar and Boznar, 1997; Reich et al., 1999; Andretta et al., 2000; Perez, 2001; Chelani et al., 2002), for NO, NO_2 and NO_x (Gardner and Dorling, 1999; Perez and Trier, 2001), ozone (Ryan, 1995; Jorquera et al., 1998; Gardner and Dorling, 2000) and PM2.5 (Perez and Trier, 2000; Perez and Ryes, 2001; Ordieres et al., 2005) concentration forecasting.

On the other hand, in recent years, new AI techniques have been developed, under the name of "soft computing", which aims at integrating powerful artificial intelligence methodologies such as neural networks

* Corresponding author. Tel.: +90 25 74010/1540; fax: +90 372 2574023.
E-mail addresses: yildirim61@yahoo.com, yildirim@karaelmas.edu.tr (Y. Yildirim).

0045-6535/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.chemosphere.2005.08.070

and fuzzy inference systems. While fuzzy logic performs an inference mechanism under cognitive uncertainty, neural networks possesses exciting capabilities such as learning, adaptation, fault-tolerance, parallelism and generalization. To enable a system to deal with cognitive uncertainties in a manner more like humans, one may incorporate the concept of fuzz logic into the neural networks. The resulting hybrid system is called a neuro-fuzzy network (Fuller, 1995). ANFIS (adaptive neuro-fuzzy inference system) is claimed as a universal approximator to represent highly non-linear functions more powerfully than conventional statistical methods (Jang et al., 1997).

Jorquera et al. (1998) successfully used neuro-fuzzy method for forecasting ozone daily maximum levels. Meanwhile, it has been seen from literature survey that there are not many studies using this method to forecast SO_2 and TSP pollution levels although other modelling approaches such as time series and neural networks have been tried by many investigators. This study aims to estimate SO_2 and TSP pollution levels in Zonguldak city (Turkey) depending on meteorological parameters by using ANFIS modelling approach.

It must be emphasized that, by this modelling approach, it is intended to forecast the daily trend of pollutant concentration as correctly as possible, with a minimum number of false alerts. Thus, it might be possible, in the context of an environmental warning system, to reschedule urban activities in case of critical estimations above air quality standards.

2. Theoretical survey

2.1. Fuzzy modelling and ANFIS: Adaptive neuro-fuzzy inference system

Fuzzy inference is a method that interprets the values in the input vector and assigns values to the output by means of some set of fuzzy "IF-THEN" rules

$$\text{IF } x \text{ is } A \quad \text{THEN } y \text{ is } B, \tag{1}$$

where A and B are labels of fuzzy sets, e.g, "low", "high". Each fuzzy set is characterized by appropriate membership functions that map each element to a membership value between 0 and 1. The IF part (antecedent) and THEN part (consequent) of a rule can have multiple parts linked by Boolean operators (AND, OR) which have counterpart fuzzy operators (MIN, MAX).

A fuzzy inference system consists of a "rule base" containing fuzzy rules, a "database" defining the membership functions of the fuzzy sets, and a "reasoning mechanism" which performs the inference procedure. Among various fuzzy inference systems, Tagaki–Sugeno's system is more suitable for sample-data based fuzzy modelling (Tagaki and Sugeno, 1985; Sugeno and Kang, 1988), in which the output of each rule is predetermined function of input variables. To give an example, in a first-order Sugeno model with two inputs (x_1, x_2), the ith rule is described as

$$\text{IF } x_1 \text{ is } X_{1,i} \text{ AND } x_2 \text{ is } \quad X_{2,i} \quad \text{THEN}$$
$$y_i = p_{i,0} + p_{i,1}x_1 + p_{i,2}x_2, \tag{2}$$

where uppercase variable X; stands for the fuzzy sets corresponding to the domain of each linguistic label, and p_i is a set of adjustable parameters. The final output, y, is the weighted average of each rule's

$$y = \Sigma W_i y_i, \tag{3}$$

where W_i is the weight of the ith rule.

On the other hand, a neural network structure consists of a number of nodes connected through directional links. Each node is characterized by a node function with fixed or adjustable parameters. The training phase of a neural network is a process to determine optimum parameters values to sufficiently fit the training data. The basic learning rule is the well-known back-propagation method which seeks to minimize some measure of error, usually a sum of squared differences between a network's outputs and desired outputs. Meanwhile, "over-training" diminishes the forecasting capability of the network due to its structure which is excessively adapted to the training data. The model performance is always checked by means of distinct test data, and a relatively good fitting is expected, especially in the testing phase.

The functionality of nodes in ANFIS, as a five-layered feed-forward neural structure layers may be summarized as follows:

- Layer 1: Nodes are adaptive; membership functions of input variables are used as node functions, and parameters in this layer are referred to as antecedent parameters.
- Layer 2: Nodes are fixed with outputs representing the firing strengths of the rules.
- Layer 3: Nodes are fixed with outputs representing normalized firing strengths.
- Layer 4: Nodes are adaptive with node function given by Layer 1 for a first-order model, and with parameters referred to as defuzzifier or consequent parameters.
- Layer 5: The single node is fixed with output equal to the sum of all the rules outputs.

Jang et al. (1997) developed a hybrid-learning rule for ANFIS, faster than the classical back-propagation method, by combining the gradient method and the least squares estimate to identify antecedent and consequent parameters. The details are given by Jang et al. (1997).

3. Materials and methods

Zonguldak is a coastal city located in the West Black Sea region of Turkey, situated on a coast surrounded by mountains to the south, east and west. It has a current population of about 106 000. Although the city is located on the shore, a hilly landscape surrounds the city centre from SE and SW. The area surrounding the city is mainly forest.

Air pollution measurements carried out by the Zonguldak Public Health Centre for the last 15 years have shown that there has been a high level of pollution in the city during winter season between November and March (Yildirim and Uzun, 2000). Two citywide air quality measurement stations were established by local authorities to observe air quality trends. While station 1 was set around the hospital, houses and some social clubs, station 2 was placed near the city's highest traffic area and around other activities such as schools, and private and government offices. The meteorological station is also very close to station 1 (the distance is about 300 m). Acidimetric and gravimetric methods were applied for analysis of sulphur dioxide and TSP, respectively (WHO, 1976). The daily arithmetic averages of SO_2 and TSP concentrations, collected from station 1 and station 2 were used in the model as training and testing data. The daily meteorological data were provided by the Department of Meteorology in Zonguldak as 8-h average values. Arithmetic averages were used to represent daily means of meteorological values. A satellite view of the city including two dimensional features of the city, air quality stations, a meteorological station and a coal treatment plant are given in Fig. 1.

4. Results and discussion

Model building, training and testing are performed by means of graphical user interface supplied in "MATLAB Fuzzy Toolbox". The ANFIS structure is generated by the subtractive clustering method. First-order Sugeno model is preferred as inference system for its simplicity. Neural networks are trained by the hybrid method as suggested by Jang et al. (1997). Considering the statistical aspect of the air pollution modelling, Gaussian type membership functions are used in this study, described by the following equation:

$$\mu_{i_i}(x) = \exp\left\{-\left(\frac{x-c_i}{a_i}\right)^2\right\}, \tag{4}$$

where a_i and c_i are memberships function parameters.

Various statistical indices are proposed in the literature to check the predictive performance of the models (Perez and Trier, 2001; Kolehmainen et al., 2001). In this study, two statistical indices are employed: the root-mean-square error (RMSE) and the index of agreement (IA) defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(o_i - p_i)^2}, \tag{5}$$

$$\mathrm{IA} = 1 - \frac{\sum_{i=1}^{N}(o_i - p_i)^2}{\sum_{i=1}^{N}[|o'_i| + |p'_i|]^2}, \tag{6}$$

Fig. 1. Satellite view of the city of Zonguldak and location of meteorological station, the coal treatment plant and air quality measurement stations. MET stands for meteorological station, COT stand for coal treatment plant, and AQS1 and AQS2 stand for air quality measurement station 1 and station 2, respectively.

where o_i and p_i are the observed and predicted SO_2 and TSP concentration values on day i, N is the number of days in the test set, $p'_i = p_i - o_m$ and $o'_i = o_i - o_m$, with o_m the average observed SO_2 and TSP concentration. The index of agreement (Eq. (6)) is a dimensionless index bounded between 0 (showing no agreement at all) and 1 (perfect agreement of the time series). The number of successful forecast and the number of false positives are also used for this purpose (Jorquera et al., 1998; Yildirim et al., 2003).

In our previous study (Yildirim et al., 2003), it was found that there is no distinct pattern of SO_2 and TSP pollution between weekends and weekdays for station 1 and station 2. This may be due to the presence of hospital, social clubs, and private education schools in the neighbourhoods of the stations, which show continues activity for the whole week. It was found that domestic heating is the cause of increase in SO_2 and TSP ambient levels during the winter season. Meanwhile, there may also be some contamination of station 2 from other sources such as traffic (approximately 16000 cars run around the station daily) and the coal treatment plant. It was found that the concentration of daily SO_2 or TSP values in station 2 is about 50% higher than station 1 (Yildirim, 2005). So, daily arithmetic averages from the two stations were employed with daily averages of meteorological parameters for training and testing performances of the model. Approximately 151 data from each winter season (November–March) for SO_2 or TSP were used for model estimation. Fifty training epochs were sufficient for good training and testing performances.

4.1. Input variable selection

The SO_2 and TSP concentrations cannot be attributed to a single cause, but it may be the result of the main sources of SO_2 and TSP (domestic heating, industrial combustions and traffic vehicles) and some meteorological variables. For a coastal region, the behaviour and characteristics of the mixing height might be important, but it could not be considered in this study due to the lack of some data. In a previous study, cross-correlation coefficients were employed and relative humidity, wind speed, precipitation and temperature were decided as input meteorological parameters (Yildirim et al., 2003). In this study, all available meteorological data including relative humidity, solar radiation, temperature, precipitation, wind speed and pressure are employed as input parameters in the model for training and testing. In addition to these parameters, the SO_2 and TSP concentration of the previous day were also taken into account; it is obvious that pollution emission is a continuous process, and therefore this fact is intentionally reflected in the model which then becomes equivalent to a second-order ARX model (Yildirim et al., 2002).

4.2. Training and testing ANFIS

In order to minimize test error given by Eq. (5), all combinations of input variables have been considered. Typical results are represented in Table 1. It is clear that wind speed, relative humidity, temperature and previous day's SO_2 or TSP concentration are parameters required for an acceptable model performance. In a previous study, it was found that the accumulated data of the past year's winter seasons may be used for training, and each data from the following winter seasons may be used separately for testing purposes (Yildirim et al., 2003). So, the input set consists of seven parameters, namely: relative humidity, solar radiation, temperature, precipitation, wind speed, pressure and the previous day's SO_2 or TSP concentration. An ANFIS structure with seven input variables is given in Fig. 2 for SO_2 and TSP. Each input variable is characterized by four Gaussian membership functions, thus, the total number of antecedent parameters is 28. The rule base contains four rules of first-order Sugeno type and the networks are trained by the hybrid method.

Data accumulated over the past years offer a great potential as training data in order to obtain a well-trained ANFIS. It is clear that the more data used in the training phase, the more adapted ANFIS is to non-linear functional dependency between input variables and the output. Various alternatives benefit from a large number of data. The obvious way is to use accumulated data of the past years for training and those of the following year for testing purposes. The results are shown in Table 2. Training and test performances are close in magnitudes, which mean that the ANFIS structure is not over-trained, but optimally trained. On the other hand, test errors, which assess the variance between measured and predicted values according to Eq. (5), are of the same order in magnitude when com-

Table 1
Effect of input set on the training and test performances of ANFIS

Model input set	Training error (RMSE)	Average test error (RMSE)
AP, RH, WS, SR, P, T, $SO_{2,j-1}$ or TSP_{j-1}	18	18
WS, RH, T, $(SO_2)_{j-1}$ or $(TSP)_{j-1}$	19	21
AP, RH, WS, SR, P, T	25	32
AP, RH, WS, SR, P	30	38

AP: atmospheric pressure, RH: relative humidity, WS: wind speed, SR: solar radiation, P: precipitation, T: temperature, $(SO_2)_{j-1}$ and $(TSP)_{j-1}$: previous day's SO_2 or TSP concentration, respectively.

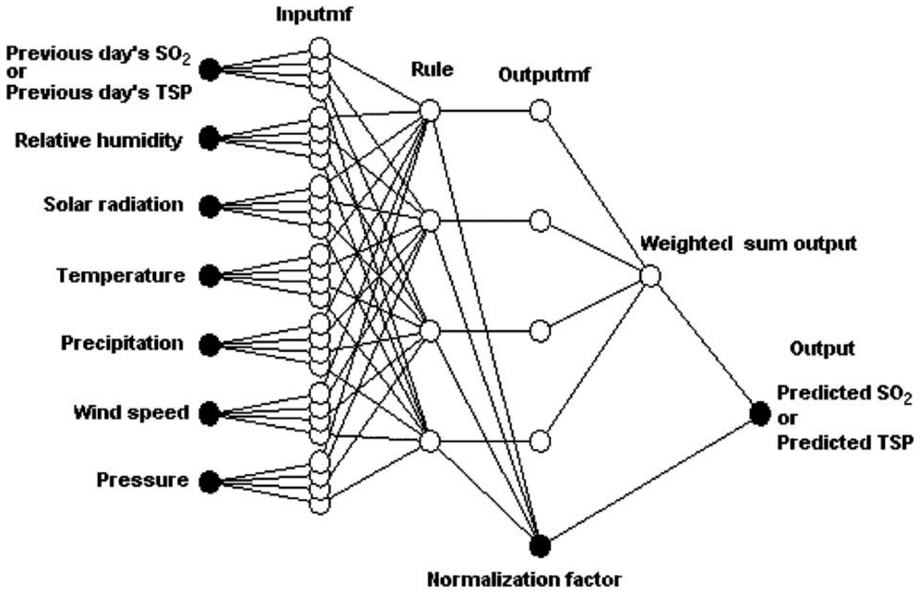

Fig. 2. The ANFIS model structure used in training and testing of the model. Seven input parameters, four Gaussian membership functions and four rules of first-order Sugeno type.

Table 2
Use of yearly progressive training sets and related performances

Training data sets for SO_2 and TSP	Test data sets for SO_2 and TSP	Training error[a]		Test error[a]	
		SO_2	TSP	SO_2	TSP
1996–1997	1997–1998	18	20	18	19
1996–1997, 1997–1998	1998–1999	19	21	19	20
1996–1997,1997–1998, 1998–1999	1999–2000	20	21	19	19
1996–1997,1997–1998, 1998–1999, 1999–2000	2000–2001	19	20	17	18

[a] Root-mean-square error (RMSE).

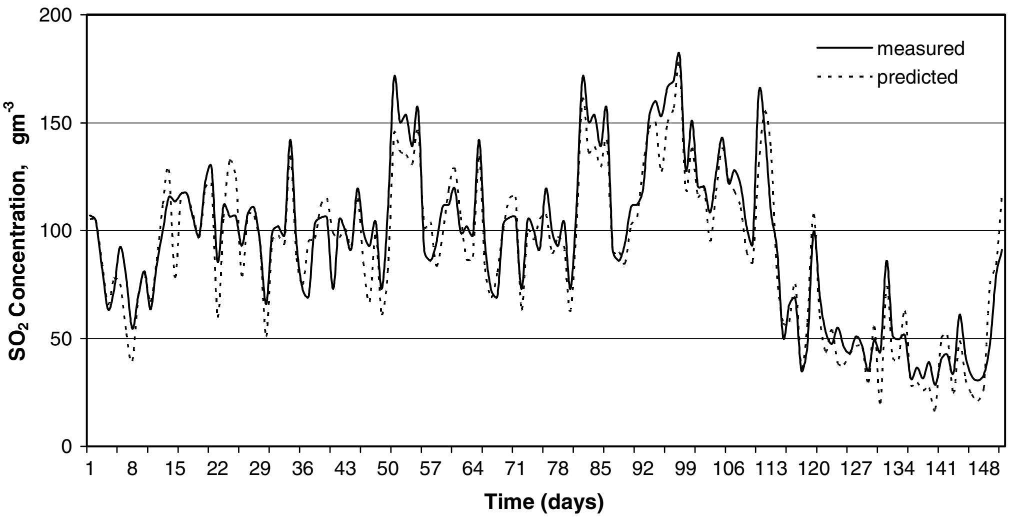

Fig. 3. Time plot of measured and predicted SO_2 concentration values for 2000–2001 winter season. Training data: 1996–1997, 1997–1998, 1998–1999 and 1999–2000 winter season; 601 data for each variable. Testing data: 2000–2001 winter season; 151 data for each variable.

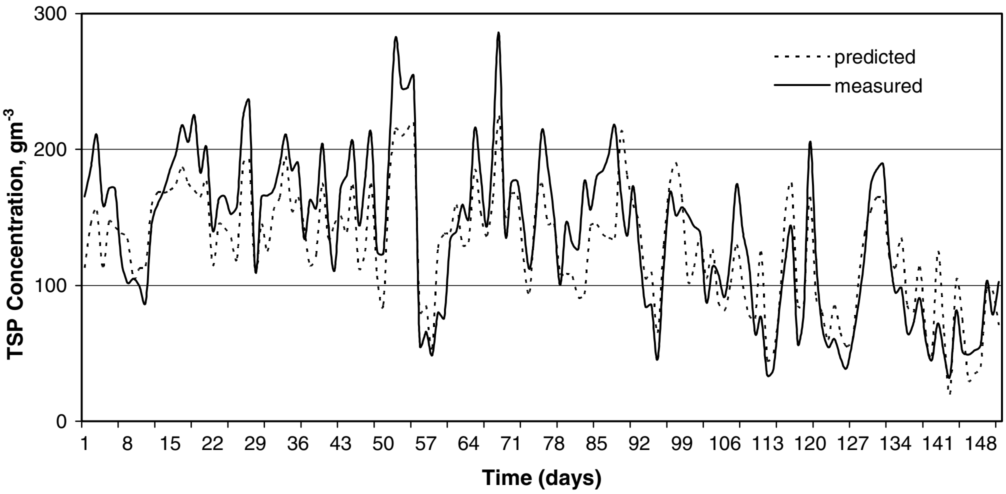

Fig. 4. Time plot of measured and predicted TSP concentration values for 2000–2001 winter season. Training data: 1996–1997, 1997–1998, 1998–1999 and 1999–2000 winter season; 601 data for each variable. Testing data: 2000–2001 winter season; 151 data for each variable.

pared with the standard error of the analytical method used in the SO_2 or TSP concentration measurements. Thus, it can be concluded that, statistically, ANFIS modelling is a valid approach of variations between model outputs and measured values, resulting greatly from measurement errors of random nature. To reinforce this conclusion, model outputs and measured data for SO_2 and TSP are given in Figs. 3 and 4 as time trends for 2000–2001 winter seasons, respectively. Predicted and observed data for SO_2 and TSP are also given in Figs. 5 and 6 as model performance respectively for 2000–2001 winter seasons. Table 3 represents the statistical evaluation of the neural fuzzy model indicating acceptable forecasting limits between 75–90% and 69–80% for SO_2 and TSP, respectively. RMSE and IA show

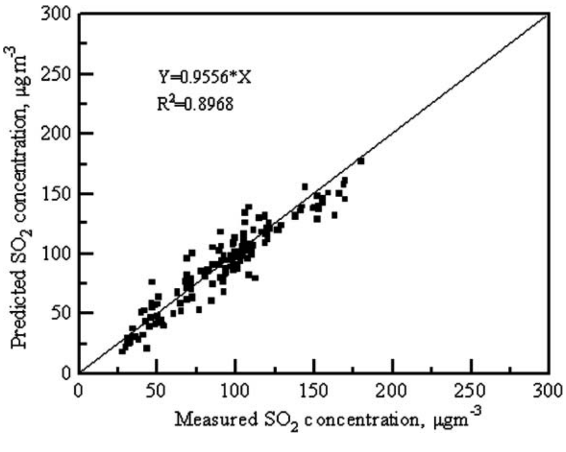

Fig. 5. Predicted and measured SO_2 concentration values for 2000–2001 winter season.

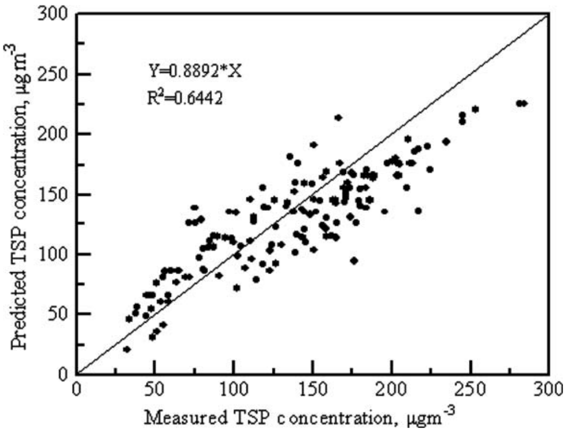

Fig. 6. Predicted and measured TSP concentration values for 2000–2001 winter season.

that the forecasting capability of the model increases gradually with time trends for 1997–1998, 1998–1999, 1999–2000 and 2000–20001 winter seasons. The model shows better performance for SO_2 in respect to TSP; serious contamination of TSP concentrations from the other sources (i.e., particles from vehicles and coal industry) in addition to domestic heating may explain this situation.

As expected, the model successfully forecasts the time trends and furthermore, the magnitudes of pollutant concentration can also be estimated within acceptable limits. It can be seen that most SO_2 and TSP peak concentrations are recognized by the model, despite the existence of slight shifts between measured and predicted values for some winter seasons. Meanwhile, large posi-

Table 3
Statistical evaluation of the ANFIS model

Test period $N = 151$	RMSE ($\mu g/m^3$)		IA (0–1)		Number of successful forecasts		Number of false positives	
	SO_2	TSP	SO_2	TSP	SO_2	TSP	SO_2	TSP
1997–1998 Winter season	38	52	0.47	0.69	113	105	38	46
1998–1999 Winter season	26	53	0.53	0.58	125	103	26	48
1999–2000 Winter season	21	39	0.71	0.53	132	112	19	39
2000–2001 Winter season	19	30	0.82	0.78	136	121	15	30

tive or negative discrepancies also exist in some situations and various causes may be cited to explain these failures. Sophisticated air quality measurement stations are under consideration in the city of Zonguldak by the Ministry of Environment and Forestry, Turkey, for more precise measurement and forecasting because of appropriate public health advice.

5. Conclusion

Air pollution models can be a very effective tool in planning strategies for management of local air quality and can provide a rational basis for the control of air pollution. If properly designed and evaluated, air pollution models play a considerable role in any air quality management system.

In this study, a new methodology based on neural fuzzy method has been proposed to estimate the concentrations of daily SO_2 and TSP pollution over an urban area. Effective input variables in the model can be ranked as temperature, pollutant (SO_2 or TSP) concentration of the previous day, wind speed, relative humidity, pressure, solar radiation and precipitation. It can be seen that the temperature and previous day's pollutant (SO_2 or TSP) concentrations are indispensable parameters for an acceptable performance of the model.

When different combinations of data sets were examined from the test performance point of view, it was found that cumulated input in five-year sets gave the best statistical results. The measured and estimated SO_2 and TSP pollutants concentrations showed peak points together. It was found that the model indicating acceptable forecasting limits between 75–90% and 69–80% for SO_2 and TSP, respectively. With a better set of training patterns, it is possible to predict the air quality levels with high accuracy.

In this study, only available pollutants (daily SO_2 and TSP) data were used for modelling and prediction studies. For better air quality management, precise daily air quality forecast with a short time interval (i.e., hourly ozone or hourly NO_x) are needed for an individual region (i.e., an urban area) when appropriate health advice is to be issued to the public using a health hazard warning system.

Acknowledgement

This study was supported by Zonguldak Karaelmas University, research Grant no. 2002-45-10-16.

References

Andretta, M., Eleuteri, A., Fortezza, F., Manco, D., Mingozzi, L., Serra, R., Tagliaferri, R., 2000. Neural net-works for sulphur dioxide ground level concentrations forecasting. Neural Computing and Applications 9, 93–100.

Boznar, M., Lesjack, M., Mlakar, P., 1993. A neural network based method for short-term predictions of ambient SO_2 concentrations in highly polluted industrial areas of complex Terrain. Atmospheric Environment 270B (2), 221–230.

Chelani, A.B., Rao, C.V.C., Phadke, K.M., Hasan, M.Z., 2002. Prediction of sulphur dioxide concentration using artificial neural networks. Environmental Modelling and Software 17, 161–168.

Fuller, A.D., 1995. Neural Fuzzy Systems. Abo Akademi University, Abo.

Gardner, M.W., Dorling, S.R., 1998. Artificial neural network (the multi layer perception)—a review of applications in the atmospheric sciences. Atmospheric Environment 32, 2627–2636.

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

Gardner, M.W., Dorling, S.R., 2000. Statistical surface ozone models: an improved methodology to account for non-linear behavior. Atmospheric Environment 34, 21–34.

Jang, J.S.R., Sun, C.T., Mizutani, E., 1997. Neuro-fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence. Prentice Hall, NJ.

Jorquera, H., Perez, R., Cipriano, A., Espejo, A., Letelier, M.V., Acuna, G., 1998. Forecasting ozone daily maximum levels at Santiago, Chile. Atmospheric Environment 32, 3415–3424.

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

Mlakar, P., Boznar, M., 1997. Perceptron neural net-work-based model predicts air pollution. In: Proceeding on Intelligent Information System, IIS'97, IEEE Computer Society, 0-8186-8218-3, pp. 345–349.

Ordieres, J.B., Vergara, E.P., Capuz, R.S., Salazar, R.E., 2005. Neural network prediction model for fine particulate matter (PM2.5) on the US–Mexico border in El Paso (Texas) and

Ciudad Juárez (Chihuahua). Environmental Modelling and Software 20 (5), 547–559.

Perez, P., Ryes, J., 2001. Prediction of particulate air pollution using neural techniques. Neural Computing and Applications 10, 165–171.

Perez, P., 2001. Prediction of sulfur dioxide concentration at a site near Downtown Santiago, Chile. Atmospheric Environment 35, 4929–4935.

Perez, P., Trier, A., 2000. Prediction of PM2.5 concentrations several hours in advance using neural networks in Santiago, Chile. Atmospheric Environment 34, 1189–1196.

Perez, P., Trier, A., 2001. Prediction of NO and NO_2 concentrations near a street with heavy traffic in Santiago, Chile. Atmospheric Environment 35, 1783–1789.

Reich, S.L., Gomez, D.R., Dawidowski, L.E., 1999. Artificial neural networks for the identification of unknown air pollution sources. Atmospheric Environment 33, 3045–3052.

Ryan, W.F., 1995. Forecasting severe ozone episodes in the Baltimore Metropolitan Area. Atmospheric Environment 29, 2387–2398.

Sugeno, M., Kang, G.T., 1988. Structure identification of fuzzy model. Fuzzy Sets and Systems 28, 15–33.

Tagaki, T., Sugeno, M., 1985. Fuzzy identification of systems and its application to modelling and control. IEEE Transactions on Systems, Men and Cybernetics 15, 116–132.

WHO, 1976. Selected Methods of Measuring Air Pollutants, World Health Organization, Geneva, No. 24, pp. 17–43.

Yildirim, Y., 2005. Air quality modelling in Zonguldak region. Zonguldak Karaelmas University, Project no. 2002.45.10. 16, Zonguldak, Turkey.

Yildirim, Y., Uzun, N., 2000. An investigation of combustion efficiency in domestic housing heating systems: Zonguldak case. In: Fifth National Symposium on Combustion and Air Pollution Control, 18–21 June 2000, Firat University, Elazig, Turkey, p. 13.

Yildirim, Y., Bayramoglu, M., Hasiloglu, S., 2003. Prediction of sulfur dioxide daily levels in the city of Zonguldak using an adaptive neuro-fuzzy based method. Fresenius Environmental Bulletin 12 (10), 425–432.

Yildirim, Y., Demircioglu, N., Kobya, M., Bayramoglu, M., 2002. A mathematical modeling of sulphur dioxide pollution in Erzurum city. Environmental Pollution 118 (3), 411–417.