ADVANCES IN ATMOSPHERIC SCIENCES, VOL. 24, NO. 5, 2007, 907–914

Modeling of Trophospheric Ozone Concentrations Using Genetically Trained Multi-Level Cellular Neural Networks

H. Kurtulus OZCAN*[1], Erdem BILGILI[2], Ulku SAHIN[1], O. Nuri UCAN[3], and Cuma BAYAT[4]

[1] *Istanbul University, Engineering Faculty, Environmental Eng. Dept.* 34320, *Avcilar, Istanbul, Turkey*

[2] *Tuzla Marine Education Center, Tuzla, Istanbul, Turkey*

[3] *Istanbul University, Engineering Faculty, Electrical-Electronics Eng. Dept.* 34320, *Avcilar, Istanbul, Turkey*

[4] *Beykent University* 34500, *Buyukcekmece, Istanbul*

(Received 12 September 2006; revised 2 February 2007)

ABSTRACT

Tropospheric ozone concentrations, which are an important air pollutant, are modeled by the use of an artificial intelligence structure. Data obtained from air pollution measurement stations in the city of Istanbul are utilized in constituting the model. A supervised algorithm for the evaluation of ozone concentration using a genetically trained multi-level cellular neural network (ML-CNN) is introduced, developed, and applied to real data. A genetic algorithm is used in the optimization of CNN templates. The model results and the actual measurement results are compared and statistically evaluated. It is observed that seasonal changes in ozone concentrations are reflected effectively by the concentrations estimated by the multilevel-CNN model structure, with a correlation value of 0.57 ascertained between actual and model results. It is shown that the multilevel-CNN modeling technique is as satisfactory as other modeling techniques in associating the data in a complex medium in air pollution applications.

Key words: genetic algorithm, cellular neural networks (CNN), ozone, meteorological data

DOI: 10.1007/s00376-007-0907-y

1. Introduction

Air pollutants exert a wide range of impacts on biological, physical and economic systems. Their effects on human health are of particular concern. The decrease in respiratory efficiency and impaired capability to transport oxygen through the blood caused by a high concentration of air pollutants may be hazardous to those having pre-existing respiratory and coronary artery disease (Sharma et al., 2003).

Ozone (O_3) is a reactive gas that forms naturally on a limited scale in the Earth's atmosphere and is the most important of the oxidizing agents. Ozone residing in the stratosphere (a layer 12–48 m above the Earth) acts as a shield to protect the Earth's surface from the Sun's harmful ultraviolet radiation. Closer to the Earth, in the troposphere, ozone is not a pollutant thrown from pollutant sources into the atmosphere, but is formed with the help of factors such as sunlight and heat, and with the adverse effects of various pollutants such as VOCs and NO_x. As ozone is a secondary pollutant, it is directly associated with the other factors affecting air pollution and meteorological agents (Wahab-Abdul and Al-Alawi, 2002).

Tropospheric ozone (O_3) is the most common photochemical oxidant in the air. While stratospheric ozone (12–48 m above the Earth) is necessary to curtain solar ultraviolet radiation, high concentrations of ozone residing closer to the Earth has negative effects on living beings. It causes coughs, dyspnea, trachea contractions, headaches, chest contractions and burns, pulmonary disfunction, changes in the cellular structure of erythrocytes, angina, as well as eye, nose and larynx irritations. Furthermore, it penetrates into plant fibers, damaging plant cell metabolism, and generates spots and stains on the leaves (Tecer, 2000).

A wide variety of operational warning and forecasting systems based on empirical, causal, statistical and

*Corresponding author: H. Kurtulus OZCAN, hkozcan@istanbul.edu.tr

hybrid models have been developed in order to begin preventive action before and during episodes (Niska et al., 2004). Cellular neural networks (CNNs)—special kinds of artificial neural networks (ANNs)—are contemporary methods recommended for the first time by Chua and Yang (1988). Their effectiveness is greater than ANNs for two-dimensional data processing and they fulfil the ANN's function by using a limited coefficient. The most important distinction between a CNN and an ANN is that, in a CNN, a cell with this structure has links only with its own adjacent cells, according to the adjacency definition, and the linkage weights between the components of the two-dimensional cell constitute a constant connection network. Due to their two-dimensional structures, CNNs are mostly applied practically in the fields of image processing and image definition (Danaci, 2002).

In this study, 24-hour estimations of tropospheric ozone values with CNN modeling and a genetic algorithm technique are made using air pollutant parameters recorded in 2003 at the Istanbul, Aksaray air pollution measurement station and the meteorological parameters of the same year obtained from the Florya meteorology station. Following estimation of the statistical parameters between the model results and actual measurement results, the appropriate model performance is determined.

2. Materials and methods

2.1 *Description of the investigated area*

The city of Istanbul, located at 41°N and 29°E, with an area of 5700 km^2, is a metropolis. Air pollutant parameters recorded in 2003 at Istanbul, Aksaray air pollution measurement station and obtained from the Istanbul Metropolitan Municipality Environment Protection and Control Directorate are used in this study. The meteorological parameters used in the model are data obtained from the Istanbul Meteorology Regional Directorate, belonging to the Florya meteorology station. Daily averages in 2003 of the eight meteorological and six pollutant parameters utilized, with minimum and maximum values and the units and abbreviations of each variable are given in Table 1.

The O_3 concentration of the following day is estimated by the CNN modeling technique using the parameters stated in Table 1. The data are arranged between the dates from 1 January 2003 to 30 December 2003, and, excluding any missing data of these obtained variables, the total sum of all has been used in 321 data training and test processes. The size of the input matrix for the training is 14×172, and for the test is 14×149.

2.2 *Basic CNN structure*

A basic 2-D CNN can be viewed as an array of basic processing units, as shown in Fig. 1. In a conventional ANN, all cells communicate with each other, whereas in a CNN only cells within a prescribed neighborhood do so. The r-neighborhood of cell $C_{i,j}$ is defined (Chua and Yang, 1988) as:

$$N_r(i,j) = [C(k,l)\max\{|k-i|,|l-j|\} \leqslant r \;,$$
$$1 \leqslant k \leqslant M; 1 \leqslant l \leqslant N] \;,$$

and is shown in Fig. 1 for $r = 1$ and $r = 2$. As each cell communicates with its neighbors, the effect of a cell propagates to cells farther away than r.

2.3 *Multi-level CNN*

The CNNs introduced above have a well suited structure for image processing. Their normalized differential state equations, which are nothing but a compact matrix representation, can be described via the matrix convolution operator defined by

$$\frac{dx}{dt} = -\boldsymbol{X} + \boldsymbol{A} \times \boldsymbol{Y} + \boldsymbol{B} \times \boldsymbol{U} + \boldsymbol{I} \;, \qquad (1)$$

where $\boldsymbol{U}, \boldsymbol{X}, \boldsymbol{Y}$ are the $M \times N$ input, state, and output matrices, respectively; $\boldsymbol{A}$ and $\boldsymbol{B}$ represent the feedback and feed-forward connections, respectively; and $\boldsymbol{I}$ is a $M \times N$ offset matrix representing the bias currents (Bilgili et al., 2005).

According to Eq. (1), CNN output changes until the derivative of the state variable of the CNN is zero; so, the last stable output is defined as $Y_{ij}^{\infty} = Y_{ij}$, when $dx/dt = 0$ for all t. For designing a stable CNN, $\boldsymbol{A}$ and $\boldsymbol{B}$ should be symmetrical and A_{22} must be greater than one, if the size of $\boldsymbol{A}$ has been selected as 3×3. CNNs are used for different special signal/image processing applications with various templates (Bilgili et al., 2005). The multi-level CNN introduced in this paper consists of the serial cascaded connection of similar type CNN structures. The same templates are used in each level, and the output of each CNN level is the input of the next CNN level in the cascade connection.

2.4 *Genetic algorithms*

Genetic algorithms are based on the mechanisms of natural selection and genetics and have proven to be effective in a number of applications (Al-Ahmad et al., 2004; Aleixandre et al., 2004; Montastruc et al., 2004). They work with a binary coding of the parameter set and search from a number of points of the parameter space for the best one; they use only a cost function during the optimization and do not need derivatives of the cost function or other information. In genetic algorithms, reproduction and mutations may cause the

Table 1. Minimum, maximum and mean values of parameters used in the model.

Parameter	Notation	Unit	Max.	Min.	Mean
Sulphur dioxide	SO_2	μg m^{-3}	82	0	15
Particular matter	PM	μg m^{-3}	198	14	66
Carbon monoxide	CO	μg m^{-3}	9575	36	1646
Nitrogen oxide	NO	μg m^{-3}	690	7	104
Nitrogen dioxide	NO_2	μg m^{-3}	127	9	59
Ozone	O_3	μg m^{-3}	86	0	16
Temperature	TEMP	°C	28.8	−0.8	14.4
Relative humidity	RH	%	95.7	43.3	72.2
Pressure	P	hPa	1025.3	100.3	1006.6
Sun shines	SUN	h	13.8	0	6.7
Cloudy	C	–	10.0	0	4.49
Wind direction	WD	–	–	–	NW
Wind speed	WS	m s^{-1}	6.2	0.4	2.6
Rainfall	R	mm	31.8	0	1.7

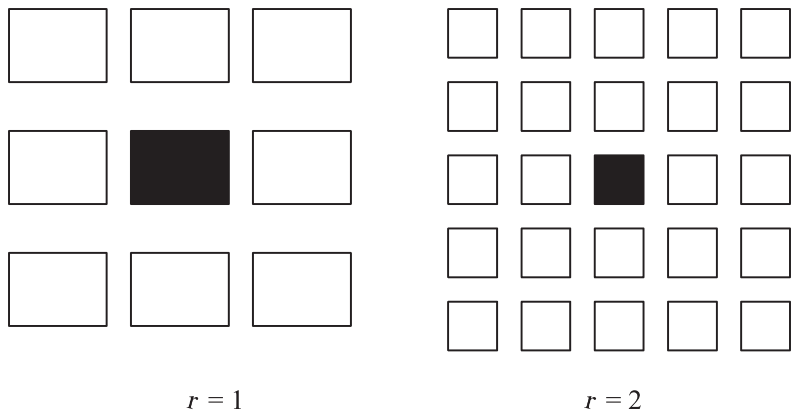

Fig. 1. 1- and 2-neighborhoods of the central cell.

chromosomes of children to be different from those of their biological parents, and crossing-over processes create different chromosomes of children by interchanging some parts of the parent chromosomes. Like in nature, the genetic approach solves the problem of finding good chromosomes by manipulating the chromosomes blindly without any knowledge about the problem they are trying to solve (Davis, 1991; Kozek et al., 1988; Holland, 1975). A general outline of the genetic approach used in this paper is as follows:

Step 1. Construction of the initial population. A matrix called a population matrix is constructed. Each row of the population matrix represents chromosomes and each column represents the bits in chromosomes, and its size is $M \times N$. At the beginning, this matrix is constructed randomly.

Step 2. Extraction of the CNN templates. Chromosomes represent the binary codes of the elements of the CNN templates, $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{I}$. In this step, each chromosome is decoded and the elements of the CNN are computed in a chosen interval. These elements are shown in vector form as

$$\begin{aligned}\boldsymbol{S} =&[a_{11}, a_{12}, a_{13}, a_{21}, a_{22}, a_{23},\\ &a_{31}, a_{32}, a_{33}, b_{11}, b_{12}, b_{13},\\ &b_{21}, b_{22}, b_{23}, b_{31}, b_{32}, b_{33}, \boldsymbol{I}]\ ,\end{aligned} \tag{2}$$

With each of the elements of $\boldsymbol{S}$ being coded in binary, the chromosome S_0 used in the algorithm is obtained from $\boldsymbol{S}$ as follows: The first five bits in S_0 represent the first five bits of the template elements, the second five bits represent the second five bits of the template elements in each chromosome, and so on; the length of each chromosome will be denoted by $LengthS$.

Step 3. Evaluation of the cost function value for each chromosome. In this step, an image that was selected as the training image is inputted to the CNN, which works with the templates belonging to the first chromosome. After the CNN output appears to be stable, the cost function is computed between this output image and the desired target image. This process is re-

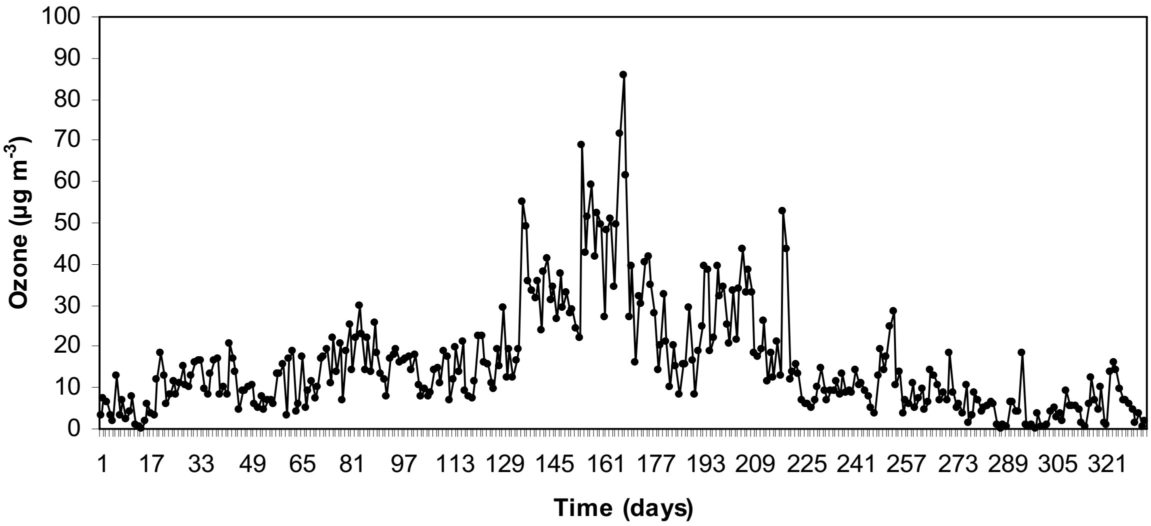

Fig. 2. Daily average level of O_3 concentration (μg m^{-3}) (January 2003–December 2003).

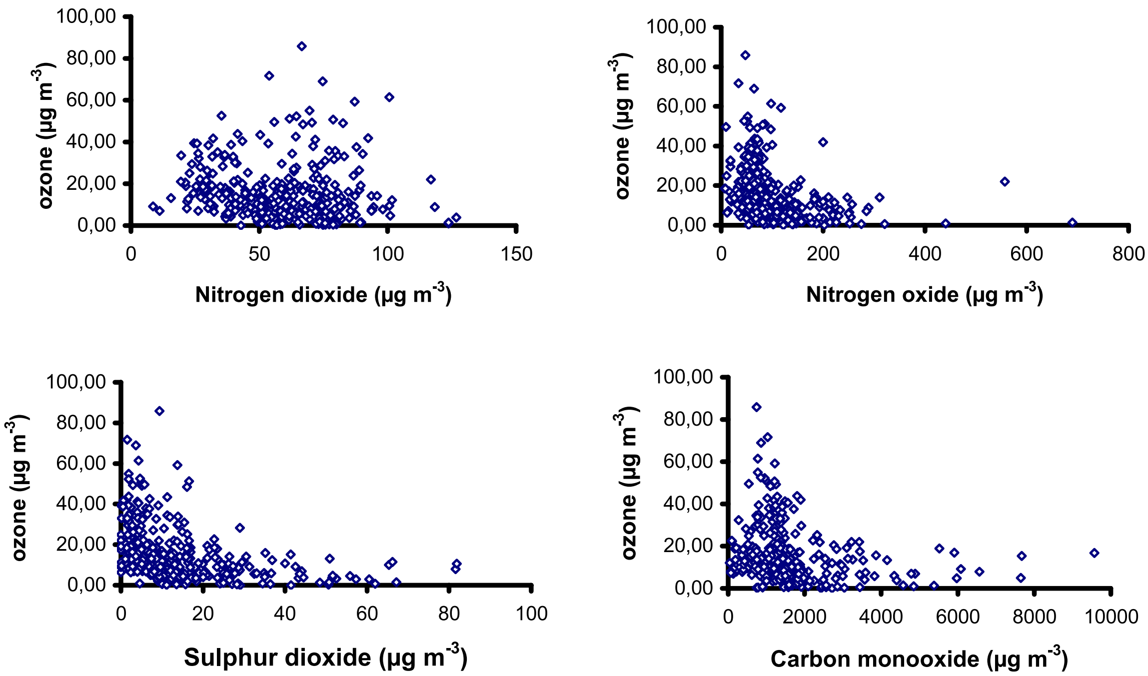

Fig. 3. Correlation between input air pollution parameters (NO_2, NO, SO_2 and CO) and ozone.

peated with the template sets belonging to each chromosome in the population. The cost function has been selected in this study as follows:

$$J(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{I}) = (T_{ij} - P_{ij})^2 \,, \tag{3}$$

where P and T represent the CNN output image and the target image, respectively.

Step 4. Creation of a new generation. Before creating the next generation, fitness values of the population are sorted in descending order and normalized relative to the sum of the fitness values of the population. A random number r between 0 and 1 is generated. Then, the first population member is selected whose normalized fitness, added to the normalized fitness of the preceding population members, is greater than or equal to r. This operation is repeated several times and any chromosome whose fitness is bad is deleted from the population. This above procedure is called "reproduction" in genetic algorithms. The reproduction process does not generate new chromosomes, but rather elects the best chromosomes in a population and increases the number of chromosomes whose fitness values are relatively greater than the others. After the reproduction, depending on the application, K pairs of chromosomes are selected as "parents" randomly. Two numbers, s_i, and s_{ii}, between 1 and the length of chromosomes are generated. The bit strings between s_i and s_{ii} are called the crossover site. During the crossing-over process, bit strings in the crossover sites in each pair of chromosomes are interchanged and two new chromosomes are created from a pair of old chromosomes. Finally, $2K$ new chromosomes, which are called "children", are generated to build the new population. Over these chromosomes, the mutation operation is carried out. Since the mutation probability has been set to 1%, $0.01 \times M \times N$ bits are selected randomly from the population and inverted. The chromosome whose fitness value was the best before the reproduction process is added and another randomly selected chromosome is deleted from the final generation; the purpose of this addition is to preserve the fittest chromosome of the previous step, and this new population is the next generation population. After obtaining the new generation, the search procedure goes to the second step and continues until the stopping criterion is met.

2.5 *Statistical evaluation*

Three different statistical expressions are used to evaluate the performance of the CNN model's estimations. These are: mean absolute error (MAE), root mean square error (RMSE) and correlation coefficient (R, found as a result of calculations made between the observed and estimated values (Ozcan et al., 2006). These calculations are:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|o_i - p_i| \, , \tag{4}$$

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

and

$$R = 1 - \frac{\sum_{i=1}^{n}(o_i - p_i)^2}{\sum_{i=1}^{n}(o_i - o')^2} \tag{6}$$

Here, o_i represents the observed concentration, o' the observed concentration average, n the observation number (days), and p_i the estimated concentration.

3. Results and discussion

The time change of the values of the ozone concentrations estimated between January 2003 and December 2003 used in the model is given in Fig. 2. The relationships of the pollutant parameters NO_2, NO, SO_2, and CO with O_3 are given in Fig. 3.

Table 2. Genetic training algorithm parameters for data.

Parameters	Value
Number of chromosome	30
Bits per parameter	16
Number of parameters	5
Chromosome length	152
Total bits in the population	4560
Mutation probability	1%
Number of inverted bits	16
Template parameters range	$[-5, 5]$

Table 3. Statistical evaluation of model results.

Statistical Index	Actual value	
	Training	Testing
MAE	6.30	6.32
RMSE	9.59	8.7
R	0.62	0.57

Input and output data of the same dimensions are formed and the calculation of all values as the output has been realized in the model structure of the CNN. In the dataset arrangement, effective input variables are the eight different meteorological datasets and the six different pollutant concentrations in the time t, and the output variable is the concentrations in the time $t + 1$.

The model works by selecting the adjacency $r = 1$ and the coefficients $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{I}$ extracted at the end of the training, as presented below. These coefficients have been applied to the test dataset and the results are given graphically in Fig. 4. Genetic algorithm parameters used in the training phase are presented in Table 2.

$$\boldsymbol{A} = \begin{pmatrix} 0.4938 & 0.1688 & -0.2500 \\ -0.1187 & -0.7813 & 0.4688 \\ 0.1000 & 0.4562 & -0.4375 \end{pmatrix}$$

$$\boldsymbol{B} = \begin{pmatrix} -0.5250 & 0.6188 & 0.5188 \\ 0.1000 & 0.3937 & 0.0125 \\ -0.5938 & -0.0437 & 0.5750 \end{pmatrix}$$

$$\boldsymbol{I} = [-0.4875]$$

The model results have been evaluated by calculating three separate statistical parameters based on

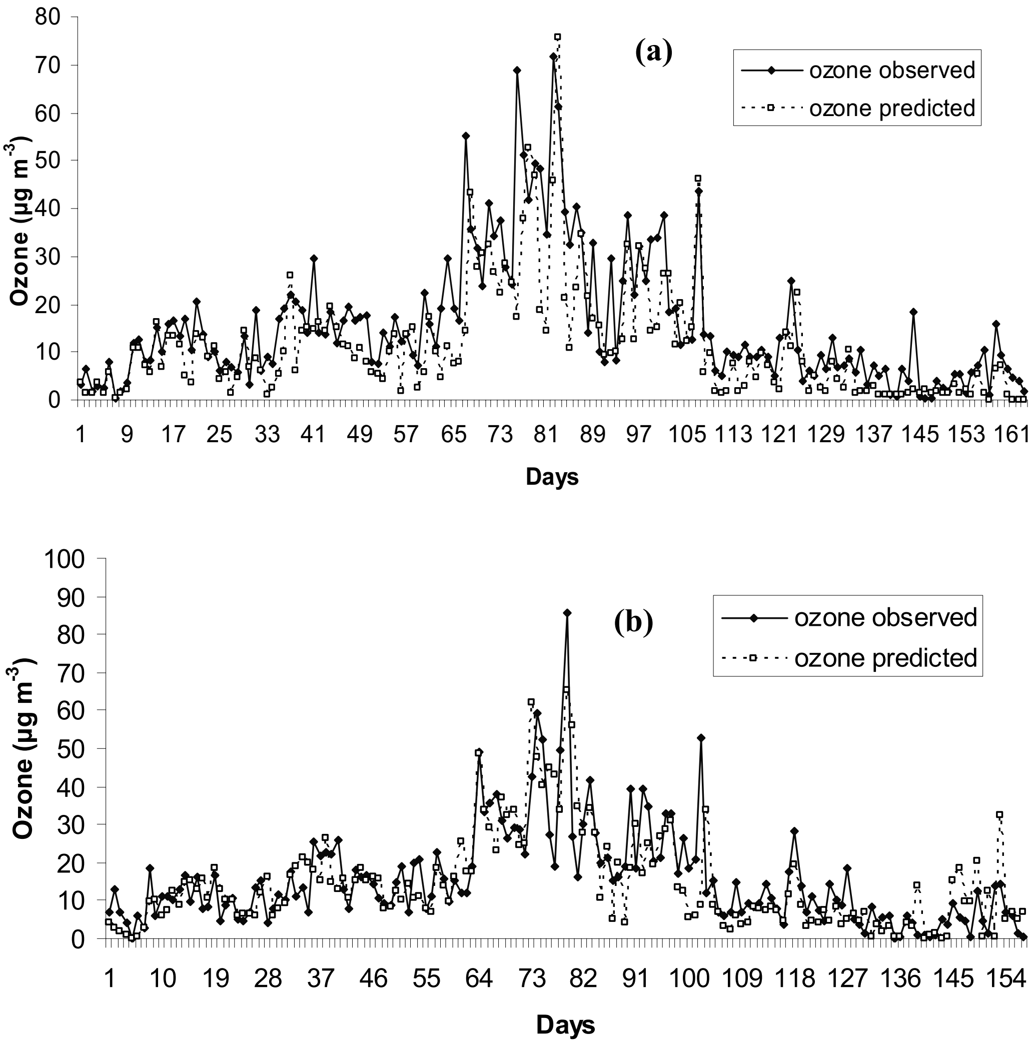

Fig. 4. Outputs for ML-CNN model structure: (a) training set and (b) testing set.

error calculation between the observed and estimated O_3 concentrations. Mean absolute error (MAE) expresses the absolute average of the difference (Eq. 4) and the root mean square error (RMSA) expresses the root of the total sum of the squares of all differences (Eq. 5). The measurement of the proximity of the observed and estimated concentrations is given by the correlation coefficient R (Eq. 6). In a well adjusted model, low MAE and RMSA values and a value of r closer to 1 represents a satisfactory modeling.

According to the statistical evaluation made between the model results and the actual O_3 concentration values, it can be seen that MAE and RMSE expressions reach the values of 6.32 μg m^{-3} and 8.70 μg m^{-3}, respectively. The value of the correlation coefficient R, that is, a coefficient giving paralleling the follow-ups for the increase and decreases in O_3 concentration, is 0.57 for the estimations. Statistical evaluation of training and testing sets are given in Table 3.

4. Conclusion

In this study, ozone, which is an important air pollutant has been modeled by the use of CNNs, and the common pollutant parameters and meteorological factors. The model outputs and statistical values in Fig. 4 show that the CNN technique used in data processing can be applied to air pollution modeling.

One-year changes of the atmospheric parameters used in the ozone estimation have been taken into consideration. It is probable that increasing the data record time is directly proportional with the sensibility of the model. However, in this study, data that can provide all atmospheric conditions and four seasons have been entered into the model to minimize the errors of the model. That all kinds of atmospheric and meteorological variables (belonging to winter, summer, spring, and fall periods) have been applied accentuates the consistency of the model.

In order to evaluate the performance of the ANN model, results were compared using measures of error. These were based on the deviations between predicted values and original observations. The MAE is the average absolute value of these residual values (Eq. 4) and the RMSE the square root of all squared residuals (Eq. 5). Evaluation can also be undertaken by considering measures of agreement, such as the Pearson product moment correlation coefficient (R) (Eq. 6) (Nunnari et al., 2004; Sahin et al., 2004, 2005). Here, they were derived using observation and model output predictions and are summarized in Table 3 to obtain the ANN model's performance. Table 3 shows the performance statistics of the trained and tested networks and regression for the same data when used to predict pollutant concentrations for 2003. In the literature, Spellman (1999) developed an ANN model which estimated ozone concentrations. Spellman's work presented a model for London, Harwell and Birmingham in the United Kingdom, and the correlation coefficients (R) of the model for these places were 0.59, 0.51 and 0.28, respectively. Gardner and Dorling (2000) also developed a MLP (Multi layer perceptron) neural network model for ozone prediction. They obtained satisfactory results and correlation coefficients of models between 0.40–0.60. In the present study, the correlation coefficient (R) between the actual values and the model's results was 0.57. It is therefore demonstrated that the ML-CNN modeling technique produces results that are equally as satisfactory as previous studies in the literature.

Prior research on modeling ozone concentrations have used ANN approaches (Ruiz-Suarez et al., 1995; Comrie, 1997; Spellman, 1999; Gardner and Dorling, 2000; Elkamel et al., 2001; Ozcan et al., 2005). As mentioned in section 1, CNNs have mainly been applied in the field of image processing and until now have not been used in studies of air pollution modeling. The present work, therefore, is a novel approach that can be used as a baseline for improvement by fellow researchers working in this area.

REFERENCES

Al-Ahmad, I. A., A. R. Azid, and K. N. Yusof, 2004: Emission control in palm oil mills using artificial neural network and genetic algorithm. *Computers and Chemical Engineering*, **28**, 2709–2715.

Aleixandre, M., I. Sayago, M. C. Horrillo, M. J. Fernandez, L. Ares, M. Garcia, J. P. Santos, and J. Gutierrez, 2004: Analysis of neural networks and analysis of feature selection with genetic algorithm to discriminate among pollutant gas. *Sensors and Actuators B: Chemical*, **103**, 122–128.

Bilgili, E., I. C. Goknar, A. M. Albora, and O. N. Ucan, 2005: Potential anomaly separation and archeological site localization using Genetically Trained Multi-Level Cellular Neural Networks. *Electronics and Telecommunications Research Institute* (*ETRI*) *Journal*, **27**(3), 294–303.

Comrie, A. C., 1997: Comparing neural networks and regression model for ozone forecasting. *Journal of the Air and Waste Management Association*, **47**, 653–663.

Chua, L. O., and L. Yang, 1988: Cellular neural networks: Application. *IEEE Transactions on Circuits and Systems*, **35**(10), 1273–1290.

Danaci, E., 2002: Image processing with cellular neural networks. Ph. D. dissertation, Selcuk University Natural Science Institute, Konya, Turkey, 166pp.

Davis, L., 1991: *Handbook of Genetic Algorithms*. Van Nostrand Reinhold, New York, 385pp.

Elkamel, A., S. A. Wahab-Abdul, W. Bouhamra, and E. Alper, 2001: Measurement and prediction of ozone levels around a heavily industrialized area: A neural network approach. *Advances in Environmental Research*, **5**, 47–59.

Gardner, M. W., and S. R. Dorling, 2000: Statistical surface ozone models: An improved methodology to account for nonlinear behavior. *Atmos. Environ.*, **34**, 21–34.

Holland, J. H. 1975: *Adaptation in Neural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI, USA, 2nd ed., 228pp.

Kozek, T., T. Roska, and L. O. Chua, 1988: Genetic algorithms for CNN template learning. *IEEE Transactions on Circuits and Systems*, **40**(6), 392–402.

Montastruc, L., C. Azzaro-Pantel, L. Pibouleau, and S. Domenech, 2004: Use of genetic algorithms and gradient based optimization techniques for calcium phosphate precipitation. *Chemical Engineering and Processing*, **43**, 1289–1298.

Niska, H., T. Hiltunen, A. Karppinen, J. Ruuskanen, and M. Kolehmanien, 2004: Evolving the neural network model for forecasting air pollution time series. *Engineering Applications of Artificial Intelligence*, **17**, 159–167.

Nunnari, G., S. Dorling, U. Schlink, G. Cawley, R. Foxall, and T. Chatterton, 2004: Modeling SO_2 concentration at a point with statistical approaches. *Environmental Modelling and Software*, **19**, 887–905.

Ozcan, H. K., U. Sahin, C. Bayat, and O. N. Ucan, 2005: ANN modeling of tropospheric ozone concentration of Istanbul City. *Proc. HITEK 2005, High Technologies of Aviation Symposium*, Turkish Air Force Academy, Istanbul, Turkey, 709–713.

Ozcan, H. K., O. N. Ucan, U. Sahin, M. Borat, and C. Bayat, 2006: Artificial neural network modeling of methane emissions at Istanbul Kemerburgaz-Odayeri Landfill Site. *Journal of Scienfitic and Industrial Research*, **65**(2), 128–134.

Ruiz-Suarez, J. C., O. A. Mayora-Ibarra, J. Torres-

Jimenes, and L. G. Ruiz-Suarez, 1995: Short term ozone forecasting by artificial neural networks. *Advances in Engineering Software*, **23**, 143–149.

Sahin, U., O. N. Ucan, B. Soyhan, and C. Bayat, 2004: Modeling of CO distribution in Istanbul using Artificial Neural Networks. *Fresenius Environmental Bulletin*, **13**(9), 839–845.

Sahin, U., O. N. Ucan, C. Bayat, and N. Oztorun, 2005: Modeling of SO_2 distribution in Istanbul using Artificial Neural Networks. *Environmental Modeling and Assessment*, **10**, 135–142.

Sharma, S., A. K. Dikshit, and S. V. Barai, 2003: Studies on neural networks based air quality predictors. *International Journal of Environment and Pollution*, **19**(5), 442–453.

Spellman, G., 1999: An application of artificial neural networks to the prediction of surface ozone concentrations in the United Kingdom. *Applied Geography*, **19**, 123–136.

Tecer, L., 2000: Modeling of photochemical smog mechanism in Istanbul City. Ph. D. dissertation Cumhuriyet University Natural Science Institute, Sivas, Turkey, 135pp.

Wahab-Abdul, S. A., and S. M. Al-Alawi, 2002: Assessment and prediction of trophospheric ozone concentration levels using artificial neural networks. *Environmental Modelling and Software*, **17**, 219–228.