Journal of ELECTRICAL ENGINEERING, VOL. 57, NO. 2, 2006, 105–110

COMMUNICATIONS

ESTIMATION OF AIR POLLUTANT CONCENTRATIONS FROM METEOROLOGICAL PARAMETERS USING ARTIFICIAL NEURAL NETWORK

Hamdy K. Elminir — Hala Abdel-Galil *

The lack of environmental data is a common feature of many developing countries. This is fact in Egypt, where air quality is beginning to be systematically monitored in some places of the country. To overcome these problems, the need for accurate estimates of air quality levels becomes evermore important. To achieve such prediction tasks, the use of artificial neural network (ANN) is regarded as a cost effective technique superior to traditional statistical methods. In this paper, ANN trained with a back propagation algorithm is used to estimate the well known pollutants, from readily observable local meteorological data. The results indicate that the ANN model predicted air pollutant concentrations with good accuracy of approximately 96 %.

K e y w o r d s: air quality, pollutants, artificial neural network

1 INTRODUCTION

For estimation of the flow of energy and the performance of systems, analytic computer codes are often used. The algorithms employed are usually complicated, involving the solution of complex differential equations. These programs usually require a large computer power and need a considerable amount of time to give accurate predictions. One approach to predict atmospheric air quality is to use a detailed atmospheric diffusion model. Such models aim to resolve the underlying physical and chemical equations controlling pollutant concentrations and therefore require detailed emissions data and meteorological fields. Collet and Oduyemi [1] provide a detailed review of this particular type of model. The second approach is to devise statistical models which attempt to determine the underlying relationship between a set of input data and targets. Regression modelling is an example of such a statistical approach and has been applied to air quality modelling and prediction in a number of studies [2, 3]. One of the limitations imposed by linear regression models is that they will underperform when used to model nonlinear systems.

Instead of complex rules and mathematical routines mentioned above, artificial neural networks are able to learn the key information patterns within a multidimensional information domain. Furthermore, the neural approach is particularly suitable to solve the problem of identification in the presence of noisy data [4, 5]. Gardner and Dorling [6, 7] concluded that ANNs generally give better results compared with statistical linear methods, especially where the problem being analysed includes nonlinear behaviour. The ANN can also be used in combination with traditional deterministic modelling techniques [8]. The drawback of the neural approach is that no deep understanding on the physical phenomena is gained using the neural network, since it resembles the behaviour of a black-box method.

On the other hand, the problems of weather and climate forecasting offer a unique area for testing and developing nonlinear algorithms. In this sense ANNs have been recently established as a reliable tool for time series analysis and some promising results have been reported [9–11]. Heymans and Baird [12] have used network analysis to evaluate the carbon flow model built for the northern Benguela upwelling ecosystem in Namibia. Antonic et al. [13] have estimated the forest survival after building the hydroelectric power plant on the Drava River, Croatia by means of a GIS constructed database and a neural network. Kolehmainen *et al* [14] evaluated various computational models using hourly concentration time series of NO_2 and basic meteorological variables collected for the city of Stockholm in 1994–1998. They concluded that the multilayer neural network yielded more accurate regression analysis and forecasting of air quality, compared with the results obtained using the self organising map or a linear time series method. Schlink et al.[15] have performed a model inter-comparison exercise within the APPETISE project for the statistical regression of tropospheric ozone, using 14 different statistical modelling techniques and a deterministic Lagrangian trajectory model including chemistry. Ten measurement sites were selected, located in Germany, Italy, United Kingdom and Czech Republic. The authors recommended those methods that are able to model static non-linearities; these include ANN and generalized additive models. The best predictions were obtained for multivariate approaches using observed meteorological data.

From the previous review one concludes that there is a wide range of models recommended by different inves-

* National Research Institute of Astronomy and Geophysics, Helwan, Cairo, Egypt, E-mail: Hamdy_Elminir@hotmail.com

ISSN 1335-3632

Table 1. Annual variations in climatological data for Abbassyia station.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
R.H	70	65	62	52	53	60	61	67	67	72	74	71
T	14	15	17	24	26	29	30	30	29	23	20	16
WS	2	2	3	3	4	4	4	3	3	3	3	2
WD	207°	206°	179°	171°	147°	168°	179°	203°	115°	144°	125°	176°

Table 2. Annual cycles of, PM10, O_3, SO_2, CO and NO_2 at Abbassyia station.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
PM10	199	192	188	167	127	94	111	143	132	228	325	218
O_3	21	31	44	68	76	78	72	65	51	27	24	19
SO_2	20	26	29	32	21	24	22	26	26	37	52	55
CO	10	8	7	4	4	4	4	5	6	6	6	5
NO_2	74	75	80	79	78	67	105	78	72	69	67	62

tigators for year around application. In this work, since the relationship between atmospheric aerosol concentration levels in the urban area of great Cairo and meteorology is complex and extremely nonlinear, artificial neural networks were used to model and predict air quality from readily observable local meteorological data. First, we give a brief introduction to ANN and describe their training process. Second, we apply them to the analysis of the above mentioned time series. Finally, we compare the results obtained with those measured, which allows us to establish the reliability of our approach.

2 DATABASE PREPARATION

In any model development process, familiarity with the available data is of the up most importance. Issues in relation to the statistical distribution of the input data, the effects of trends, and seasonal variation are of major importance. The center for environmental hazard mitigation CEHM at Cairo university and the institute of graduate studies and research IGSR at Alexandria university are operating, on behalf of the Egyptian Environmental Affairs Agency EEAA a total of 14 sites located in the greater Cairo area. The monitoring laboratories both at CEHM and IGSR are submitting quarterly reports as a support for the data collection. These reports briefly describe data quality, data availability and the air quality.

In general, sites characteristics are different representing industrial, traffic, urban, residential and background areas. The discussion in this work is based upon measurements performed during the period 2000 to 2002 of the different indicators measured at Abbassyia station. The variables monitored were, sulpher dioxide, (SO_2, in $\mu g/m^3$), nitrogen dioxide, (NO_2, in $\mu g/m^3$), particulate matter with diameter less than 10 micrometer (PM10, in $\mu g/m^3$), carbon monoxide (CO, in $\mu g/m^3$), ozone (O_3, in $\mu g/m^3$), wind speed (WS, in m/s), relative humidity (R.H., in %), and temperature, (T in °C). Tables 1 and 2, present the average annual cycles of climatological data used to train ANN. Following the rehabilitation techniques given by Kolehmainen *et al* [16] the collected data were examined.

We have to cope with two evident problems of the data readings: missing data and outliers. The outliers are mainly due to the incorrect working of the instruments or to incorrect methodology for collection and analysis. Generally, the maximum and minimum values can be considered as outliers and they must be examined with care, because they can cause deformation in the calibration of the predict model. In our case, we have decided to keep the outliers because the tools we use may become robust with suitable implementation.

Missing data are mainly due to failures of the measurement instruments. The presence of missing data can invalidate the statistical analysis, introducing systematic components of errors about the estimation of parameters in the prediction model. In addition, if we estimate the parameters of the model by exploiting the observed data without taking into account the presence of missing data, obtained estimations could be unreliable because much information concerning the missing data would disappear. Therefore, missing data were replaced with the expected values and suspected erroneous values were removed after careful examination.

3 FORECASTING METHODOLOGY

3.1 A capsule introduction

Neural networks are a branch of artificial intelligence developed in the 1950s aiming at imitating the biological brain architecture. They are parallel distributed systems made of many interconnected non-linear processing elements, called neurons. The network architecture is composed of many simple processing elements that are organized into a sequence of layers. The network usually consists of an input layer, hidden layers and an output layer. In its simple form, each single neuron is connected to other neurons of a previous layer through adaptable synaptic weights. Knowledge is usually stored as a set of connection weights (presumably corresponding to synapse efficacy in biological neural systems). The network uses a learning mode, in which an input is presented to the network along with the desired output and the weights are adjusted, so that the network attempts to produce the desired output. If there is a difference, the connection weights are altered in such a direction that the error is decreased.

In this work, an ANN model was used to predict air pollutant concentrations based on different climatological variables. The neurons in the input layer receive four input signals representing ambient air temperature, relative humidity, wind speed and wind direction; hence four

neurons were used for input in the ANN architecture. The output layer on the other hand, consists of three output neurons representing particulate matter with diameter less than 10 micrometers, carbon monoxide, and nitrogen dioxide. Since, there is no direct and precise way of determining the number of hidden layers to use and exact number of neurons to include in each hidden layer, we decided to use only one hidden layer in the ANN architecture.

3.2 Model building

For the determination of the architecture, we used a pruning approach, starting from a relatively large network and then removing connections in order to arrive at a suitable network architecture. Several approaches to network pruning are based on the following general procedure (Bishop, 1995). First, a relatively large network is trained using one of the standard training algorithms. Then the network is examined to assess the relative importance of the weights, and the least important are deleted. Typically this is followed by some further training of the pruned network, and the procedure of pruning and training is repeated for several cycles. Clearly, there are various choices to be made concerning how much training is applied at each stage, which fractions of the weights are pruned, and so on. These choices are usually made on a heuristic basis. The most important consideration, however, is how to decide which weights should be removed. For that purpose, some measure of the relative importance or saliency of weights has to be defined. The Optimal Brain Damage OBD method [17] provides such a measure. This method is briefly recalled here.

The method is based on the computation of the change ∂E in the error function E due to small changes in the values of the weights. If the weight w_i is changed to $w_i + \delta w_i$ the corresponding change in the error function E is given by:

$$\partial E = \sum_i \frac{\partial E}{\partial w_i}\delta w_i + \frac{1}{2}\sum_i\sum_j H_{ij}\delta w_i \delta w_j + O\left(\delta w^3\right) \quad (1)$$

$$H_{ij} = \frac{\partial^2 E}{\partial w_i \partial w_j} \quad (2)$$

Where, H_{ij} are the elements of the Hessian matrix. The subscripts, i and j indicate the layer position in each node. O are the desired output vectors. Finally, the δ terms for each node are used to compute an incremental change to each weight term. If we assume that the training process has converged then the first term in Eqn. (1) vanishes. As proposed by Lecun *et al* 1990, the Hessian matrix can be approximated by discarding the non diagonal terms. In mathematical term Eqn. (1) can be described by means of the following:

$$\delta E = \frac{1}{2}\sum H_{ii}\delta w_i^2 \quad (3)$$

Finally, the saliency values of the weights can be described by means of the following term:

$$H_{ii}\frac{w_i^2}{2} \quad (4)$$

3.3 Results of the statistical evaluation of the model performance

We used selected statistical indicators to provide a numerical description of the goodness of the estimates. One of the most common indicators used with neural networks is the root mean square error RMSE. This is calculated according to Eqn. (5).

$$\mathrm{RMSE} = \left\{\frac{\sum_{i=1}^{N}(Y_i - X_i)^2}{N}\right\}^{\frac{1}{2}} \quad (5)$$

Where, N is the number of data points, Y_i is the predicted data point and X_i is the observed data point. An error of zero would indicate that all the output patterns computed by the ANN perfectly match the expected values and the network is well trained. Then we used the mean bias error, MBE, to describe how much the ANN model underestimates or overestimates the situation. The MBE was calculated according to Eqn. (6).

$$\mathrm{MBE} = \frac{\sum_{i=1}^{N}(Y_i - X_i)}{N} \quad (6)$$

Finally, we used correlation coefficient, R, to test the linear relation between calculated and measured values. The statistical values have been presented in Table 3. While, the ANN performance is shown in Figs. 2–4.

$$R = \frac{\sum_{i=1}^{N}(Y_i - \overline{Y}_i)(X_i - \overline{X}_i)}{\left\{\left[\sum_{i=1}^{N}(Y_i - \overline{Y}_i)^2\right]\left[\sum_{i=1}^{N}(X_i - \overline{X}_i)^2\right]\right\}^{\frac{1}{2}}} \quad (7)$$

Where, $\overline{Y}$ is the predicted mean value, and $\overline{X}$ is the measured mean value.

Table 3. Monthly statistical errors between predicted and measured data.

	RMSE (%)			MBE (%)			R		
	PM10	CO	NO_2	PM10	CO	NO_2	PM10	CO	NO_2
Jan	2.819	4.647	4.040	0.824	2.133	0.797	0.998	0.996	0.975
Feb	4.346	6.159	3.295	-0.213	2.145	0.167	0.996	0.994	0.986
Mar	5.399	6.147	2.184	1.080	-1.190	-0.884	0.996	0.959	0.994
Apr	5.968	8.123	4.526	0.548	3.828	-1.084	0.995	0.994	0.992
May	7.084	8.925	2.885	0.686	3.398	-0.382	0.988	0.976	0.996
Jun	6.944	8.901	4.066	2.689	2.851	-2.091	0.998	0.984	0.984
Jul	5.111	8.469	3.878	-0.814	4.435	2.089	0.998	0.982	0.989
Aug	3.169	8.716	2.906	-1.166	-0.001	-1.214	0.998	0.967	0.964
Sep	2.662	7.141	3.349	-0.478	0.295	-1.053	0.999	0.932	0.967
Oct	2.387	5.568	2.318	-0.283	1.571	-0.422	0.999	0.941	0.963
Nov	1.811	6.217	3.155	-0.456	1.776	1.592	0.999	0.956	0.951
Dec	2.667	7.815	3.443	-1.165	3.901	-0.172	0.999	0.951	0.979

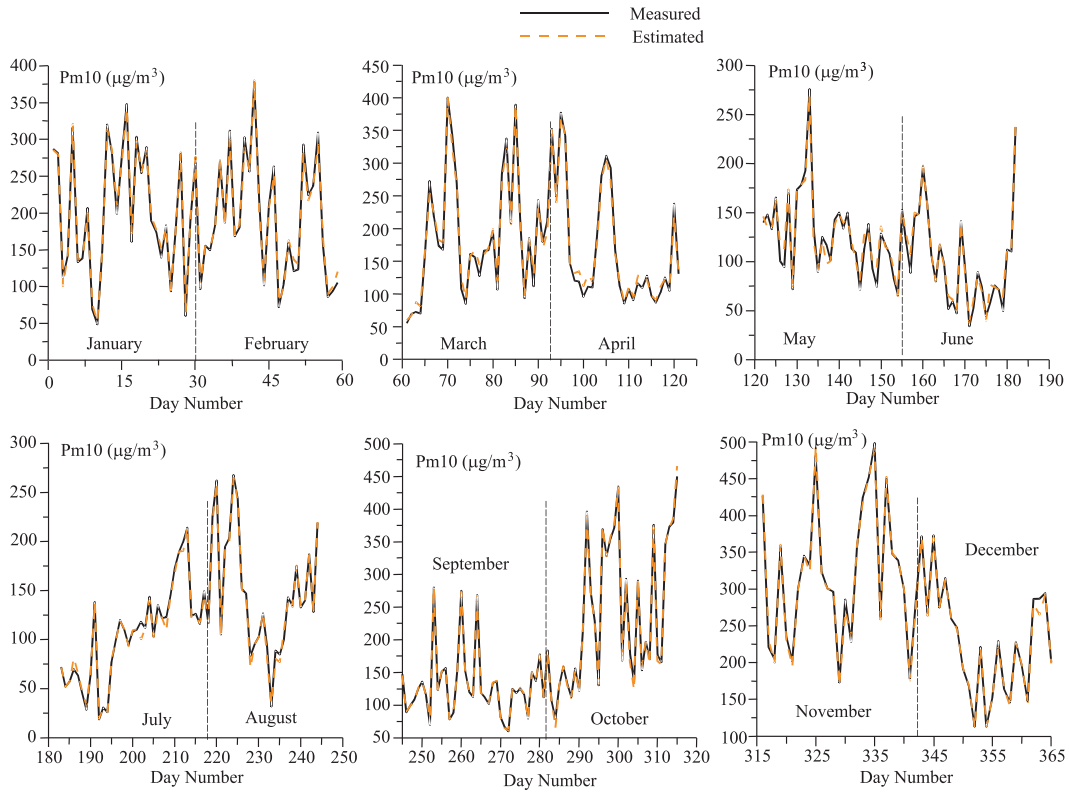

Fig. 1. Daily time series of measured and predicted concentration of PM10 at Abbassyia station.

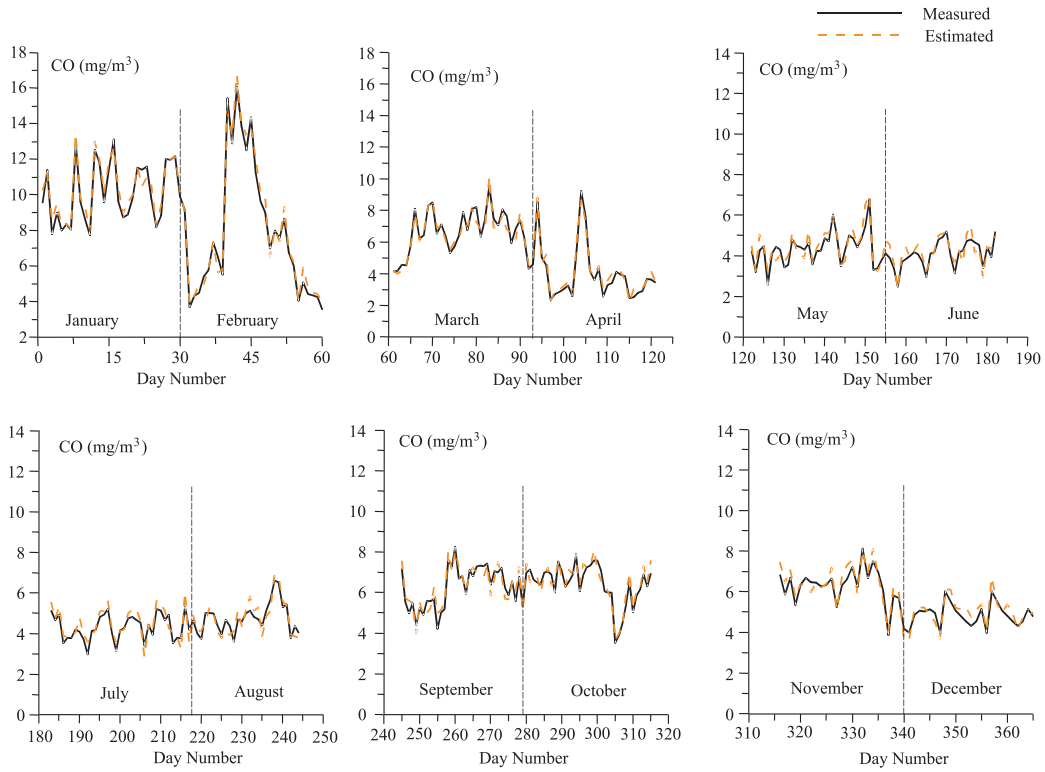

Fig. 2. Daily time series of measured and predicted concentration of CO at Abbassyia station.

The values in Table 3, are for the most suitable algorithms and for the hidden layer giving the most appropriate approach. The range of RMSE for PM10 was 1.81 to 7.08 % while the range of RMSE for CO was 4.64 to 8.92 %. For the whole year, RMSE was found to be 3.99 %, 6.77 % and 3.82 % for PM10, CO and NO_2, respectively. These results demonstrate that, ANN model can estimate air pollutant concentrations in urban area,

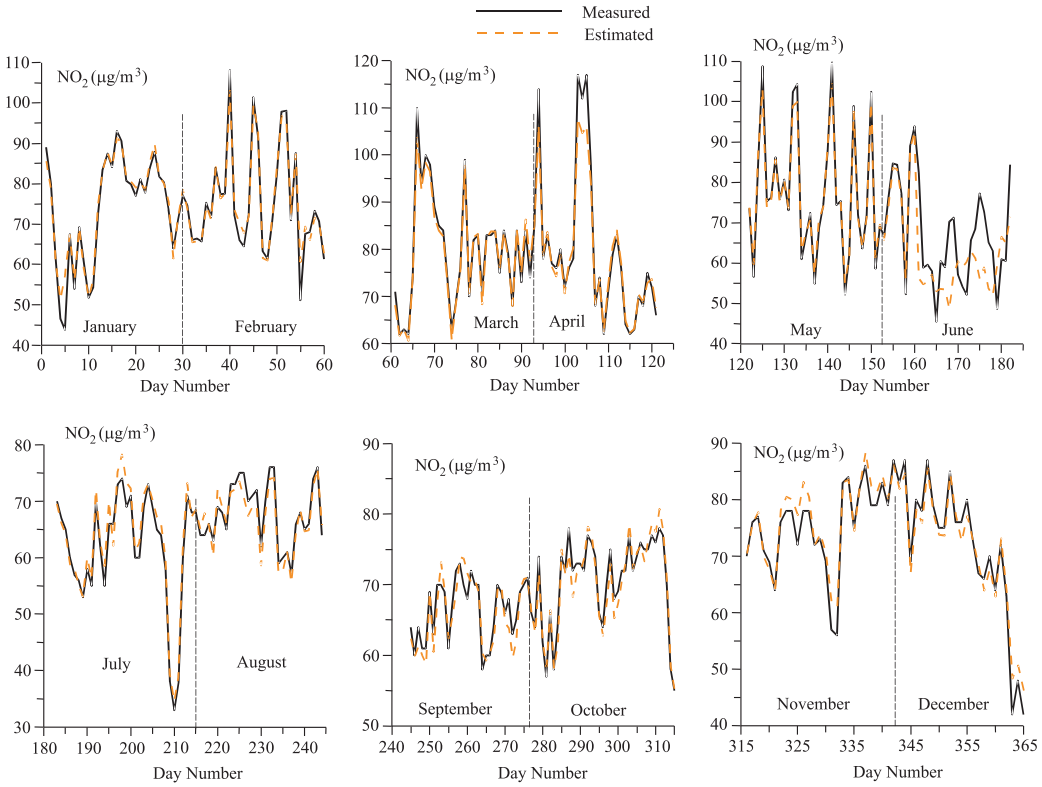

Fig. 3. Daily time series of measured and predicted concentration of NO_2 at Abbassyia station.

for the given data set with an accuracy of approximately 96 %. The generalization of the model was also tested by correlation coefficient. Based on the results obtained, annual correlation coefficient was found to be 0.997, 0.969 and 0.978 for PM10, CO and NO_2 respectively. These results show that approximately 98 % of the variation in the dependent variables (output parameters) can be explained by the independent variables (input parameters) selected. However, the MBE values range from an under-estimation of — 0.289 % for NO_2 to an over-estimation of 1.82 % for CO. On an overall basis, MBE results indicate that, ANN model always tend to over-estimate air pollutant concentrations, but remain in a domain of errors for which this model can be applied with good accuracy.

Figure 1 shows the measured and estimated PM10 using ANN model that resulted in the annual overall RMSE value of 3.99 %. Figures 2, 3 show results obtained for CO, and NO_2 which resulted in the annual overall RMSE of 6.77 % and 3.32 % respectively. Inspecting the results, model evaluation regarding CO was less extensive than that for both PM10 and NO_2. A general conclusion from these figures is that the training sets show a perfect fit as expected.

4 CONCLUSION

Although the monitored period considered (3 years) seems to be quite short and the lack of continuous hourly data occurred to many parameters, from the present study, it has appeared that the ANN with a single hidden layer based on the standard back propagation algorithm described above, using eventually only pruning approach, resulted as a very efficient model to forecast long-term air pollutant concentrations in urban area for locations not covered by the model's training data. The results indicate that ANN model predicted the PM10 with a good accuracy of approximately 96 %. In addition, the model was also tested to predict NO_2 values over a 12 months period. The monthly predicted values of the ANN produced an accuracy of approximately, 97 %. Further information is given by MBE results, which indicate that, ANN model always tend to over-estimate the air pollutant concentrations on an overall basis, but remain in a domain of errors for which this model can applied with good accuracy. However, ANN models also have inherent limitations. The main limitation is the extension of models in terms of time period and location; this always requires training with locally measured data. The ANN models cannot therefore be recommended for analysing various air pollution abatement scenarios for future years.

References

[1] COLLET, R.—ODUYEMI, K.: Air Quality Modelling: a Technical Review of Mathematical Approaches. Meteorological Applications 4, 1997.

[2] SHI, J.—HARRISON, R.: Regression Modelling of Hourly NO_x and NO_2 Concentrations in Urban Air in London, Atmospheric Environment **31** No. 24 (1997).

[3] ZIOMASS, I.—MELAS, D.—ZEREFOS, C.—BAIS, A.: Forecasting Peak Pollutant Levels from Meteorological Variables, Atmospheric Environment **29** No. 24 (1995).

[4] RUMELHART, E.—HINTON, G.—WILLIAMS, R.: Learning Internal Representation by Error Propagation, In: Parallel distributed processing: Explorations in the Microstructure of Cognition, vol. . 1, MIT Press, Cambridge, MA, 1986.

[5] HERTZ, J.—KROGH, A.—PALMER, R.: Introduction to the Theory of Neural Computation, Addison Wesley, Canada, 1995.

[6] GARDNER, M.—DORLING, S.: Neural Network Modelling and Prediction of Hourly NO_x and NO_2 Concentrations in Urban Air in London, Atmospheric Environment **33** No. 5 (1999).

[7] GARDNER, M.—DORLING, S.: Artificial Neural Networks (the M Perceptron) a Review of Applications in the Atmospheric Sciences, Atmospheric Environment **32** (1998).

[8] BISHOP, A.: Neural Networks for Pattern Recognition, Oxford University Press, UK, 1995.

[9] WEIGEND, A.—HUBERMAN, A.—RUMELHART, D.: Predicting the Future: A Connectionist Approach, International Journal of Neural Systems **1** (1990).

[10] BOZNAR, M.—LESJAK, M.—MLAKAR, P.: A Neural Network Based Method for Short Term Predictions of Ambient SO_2 Concentrations in Highly Polluted Industrial Areas of Complex Terrain, Atmospheric Environment **27** No. 2 (1993).

[11] COMRIE, A.: Comparing Neural Networks and Regression Models for Ozone Forecasting, Journal of Air Waste Manage **47** (1997).

[12] HEYMANS, J.—BAIRD, D.: A Carbon Flow Model and Network Analysis of the Northern Benguela Upwelling System, Namibia. Ecol. Model. **126** (2000), 932.

[13] ANTONIC, O.—HATIC, D.—KRIAN, J.—BUKOCEV, D.: Modelling Groundwater Regime Acceptable for the Forest Survival after the Building of the Hydroelectric Power Plant, Ecol. Model. **138** (2001).

[14] KOLEHMAINEN, M.—MARTIKAINEN, H.—RUUSKANEN, J.: Neural Networks and Periodic Components Used in Air Quality Forecasting, Atmospheric Environment **35** (2001).

[15] SCHLINK, U.—DORLING, S.—PELIKAN, E.—NUNNARI, G.—CAWLEY, G.—JUNNINEN, H.—GREIG, A.—FOXALL, R.—EBEN, K.—CHATTERTON, T.—VONDRACEK, J.—RICHTER, M.—DOSTAL, M.—BERTUCCO, L.—KOLEHMAINEN, M.—DOYLE, M.: A Rigorous Inter-Comparison of Ground Level Ozone Predictions, Atmospheric Environment **37** (2003).

[16] KOLEHMAINEN, M.—MARTIKAINEN, H.—HILTUNEN, T. —RUUSKANEN, J.: Forecasting Air Quality Using Hybrid Neural Network Modelling, Environmental Monitoring and Assessment **65** (2000).

[17] LECUN, Y.—DENKER, S.—SOLLA, A.: Optimal Brain Damage. Advances in Neural Information Processing Systems 2, Morgan Kaufman, San Mateo, CA, 1990.

Received 15 September 2004

Hamdy Kamal Elminir was born in EI_Mahala, Egypt in 1968. He received the BSc in Engineering, from Monofia university Egypt in 1991 and completed master degree in automatic control system, Mansoura university, Egypt in 1996. He obtained his PhD degree from the Czech Technical University in Prague, in 2001, and currently is a researcher in the National Institute of Astronomy and Geophysics, Helwan Cairo.

Hala Abdel-Galil was born in Cairo, Egypt in 1964. She received the BSc in pure mathematics and computer science and completed her master degree in computer science in 1996. She obtained her PhD degree from Ain Shams University in 2003, and currently is a lecturer in Faculty of Computers and Information, Helwan University, Cairo.