Modelling ozone levels in an arid region – a dynamically evolving soft computing approach

Syed Masiur Rahman*, A.N. Khondaker and Rouf Ahmad Khan

Center for Environment and Water, Research Institute,
King Fahd University of Petroleum and Minerals,
P.O. Box 713, Dhahran 31261, Saudi Arabia
Fax: +966-138603220
E-mail: smrahman@kfupm.edu.sa
E-mail: nasserk@kfupm.edu.sa
E-mail: roufkhan@kfupm.edu.sa
*Corresponding author

Abstract: The primary pollutants may contribute to the increase of ozone levels in the arid regions. Complex interactions between the pollutants and the meteorological variables make the study of this phenomenon more exigent. The dynamically evolving neural fuzzy inference system (DENFIS), as an example of soft computing models, allows the online evolution of both the knowledge and the inference mechanism. It is suitable for real-time applications in producing fairly reliable forecasts. The proposed DENFIS model for two sites in the Empty Quarter (Rub Al-Khali Desert) of Saudi Arabia was developed using the meteorological data collected during the winter and the summer seasons, and the transformed meteorological data. The concentrations of nitrogen oxide (NO_x) and their transformations were incorporated as additional inputs for model performance analyses. The mean absolute percentage errors of the model vary from 9.52% to 11.84% with discretion and appreciation of the limitations of the overall model predictions and its performance analyses indicate the viability of application of the adopted online DENFIS modelling approach in short-term modelling of zone levels in arid regions.

Keywords: ozone modelling; DENFIS; Empty Quarter of Saudi Arabia; soft computing.

Reference to this paper should be made as follows: Rahman, S.M., Khondaker, A.N. and Khan, R.A. (2013) 'Modelling ozone levels in an arid region – a dynamically evolving soft computing approach', *Int. J. Environment and Pollution*, Vol. 52, Nos. 3/4, pp.155–171.

Biographical notes: Syed Masiur Rahman is a Research Engineer in the Center for Environment and Water at the Research Institute of King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia. He obtained his PhD in Civil Engineering from KFUPM in 2010. He has contributed to a wide range of environmental studies including environmental and social impact assessment, air quality and meteorological monitoring and modelling, and greenhouse gas emission estimation and control. His main research interests include air quality modelling, greenhouse gas emission estimation and modelling, climate modelling, and advanced modelling for traffic forecasting and planning. He is the author of a number of technical papers published in international refereed journals and conference proceedings. He also co-authored several books and book chapters.

A.N. Khondaker is a Lead Research Engineer and Coordinator of the Environmental Modelling and Atmospheric Monitoring Group in the Center for Environment and Water at the Research Institute of KFUPM. He obtained his PhD in Civil Engineering in 1991 from KFUPM. He has a 30-year post-graduate professional career with a specialisation in environmental and social impact assessment, national and international environmental legislation and standards, environmental pollution monitoring and control, greenhouse gas emission estimation and control, and advanced modelling of flow and pollutant transport in air, groundwater, and marine environment. He is a winner of the Distinguished Applied Researcher, Research Project Management, and Research Leadership Awards of KFUPM for several times. He has published a number of technical papers in international refereed journals and conference proceedings.

Rouf Ahmad Khan is a Computer Engineer in Center for Environment and Water (CEW) at the Research Institute of KFUPM. He obtained his MS in Computer Science from Manipal University, India in 2003. He received his BS in 1999 from Kashmir University, India. He has contributed to a wide range of environmental studies including development of comprehensive database management systems for air quality and meteorological data and national inventory of greenhouse gas emissions. His main research interest includes scientific computing in environmental science and engineering including development of advanced computational tools. He is the author of several technical papers published in international refereed journals and conference proceedings.

1 Introduction

Air quality models play an important role in assessing atmospheric quality, simulating the atmospheric environmental system, increasing the domain knowledge of environmental phenomena, and producing reliable forecasts (Karatzas and Kaltsatos, 2007). The real-life implementation of such models can also provide early warnings to the population and reduce the number of required measuring stations. However, modelling ozone levels in particular is considered a very difficult task due to the complex interactions between the pollutants and the meteorological variables (Borrego et al., 2003). In an arid region, nitrogen oxides and non-methane hydrocarbons participate in photochemical reactions due to high temperature and solar radiation, and contribute to the increase of ozone levels, which has negative effects on biotic health.

A soft computing model provides a flexible and adaptive modelling approach. Typically, it does not require making many assumptions on the modelled phenomenon. The artificial neural network (ANN) model is an example of a soft computing model, which is widely used to predict the concentrations of air pollutants including ozone (Abdul-Wahab and Al-Alawi, 2002; Sousa et al., 2007). The typical ANN models are modified to the wavelet neural network (Zhang and Benveniste, 1992), the multi-tasking neural network (Caruana, 1997), and the evolutionary neural network (Hassoun, 1995; Braun and Weisbrod, 1991) to improve the performance of model prediction. Pires et al. (2012) used genetic algorithms to define the activation function in hidden layer and the number of the hidden neurons of the ANN for ozone prediction.

The main advantages of ANN over typical statistical models include self-learning, self-adaptation, faster computation, and noise rejection (Kao and Huang, 2000; Dunea et al., 2008). But the performance of ANN is affected by the network training, the amount and quality of training data, and the network parameters such as the number of hidden layers, the number of neurons in the hidden layers, the neuron transfer function, the initial weights of connections between neurons, the learning rate, and the number of training epochs. A conventional ANN may experience difficulties in approximating functions if the input features are not linearly separable, which implies that the approximated function has a higher complexity (Park et al., 1999). It is not suitable for dealing with linguistic data.

The fuzzy logic model is not suitable to handle knowledge stored in the form of numerical data. But it allows the accurate representation of a given system behaviour using a set of simple 'if-then' rules. Heo and Kim (2004) adopted fuzzy logic and the ANN model consecutively to forecast daily maximum ozone concentrations. Inspired by the combined strength of the ANN and fuzzy logic model, Kasabov (1998) introduced dynamically evolving neural fuzzy inference system (DENFIS), which is a Takagi-Sugeno type fuzzy inference system with a back propagation algorithm. It can also be considered as an ANN in which the processing units are added to their structures, and the connection weights are modified as the system evolves based on input data stream in an adaptive, life-long, and modular manner (Kasabov and Song, 2002). It allows the evolution of both the knowledge and the inference mechanism with more examples presented to the systems. The fuzzy inference system is developed using the clustering algorithm, which identifies similar characteristics of data points and develop a rule for each group. There are two types of DENFIS modelling approaches with an offline or an online learning algorithm using the Takagi-Sugeno type fuzzy inference system. A few hybrids of ANN and fuzzy logic models are already available for carbon monoxide (CO) prediction (Jain and Khare, 2010) and sulphur dioxide (SO_2) prediction (Yildirim and Bayramoglu, 2006). The neuro-fuzzy logic-based solutions for ozone concentration prediction, more specifically the online models have not yet been investigated adequately (Johanyák and Kovács, 2011).

This study adopted the online DENFIS model for prediction of ozone concentrations with a focus on the real-time application. Online algorithms focus on fast processing speed and minimal memory usage and typically process the input data piece by piece (Kasabov, 2001). It minimises the complexity of the algorithm as the input data are discarded after they are processed (Hwang and Song, 2009). This study attempted to use the clustering algorithm-based DENFIS for modelling ozone levels in the Empty Quarter (also known as the Rub Al-Khali Desert) of Saudi Arabia. The Empty Quarter is considered a source of huge potential for oil and gas field development where a number of oil and gas exploration projects are active.

2 Fundamentals of DENFIS

2.1 Basic principles of DENFIS

The inference system used in DENFIS is a Takagi-Sugeno type of fuzzy inference system and composed of *n* fuzzy rules which follow:

If x_1 is R_{n1} and x_2 is R_{n2}…. and x_p is R_{np}, then

$$y = f_n\left(x_1, x_2, \ldots, x_n\right)$$

where 'x_i is R_{ik}', $i = 1,2,\ldots, n$; $k = 1,2,\ldots, p$, are $n \times p$ fuzzy propositions; x_i, $i = 1,2,\ldots,n$, are antecedent variables defined over universes of discourse X_i, $i = 1,2,\ldots,n$, and $R_{ij,}$ $j = 1,2,\ldots,p$, are fuzzy sets represented by corresponding membership functions μ_{Rji}: $X_i \rightarrow [0,1]$, $i = 1,2,\ldots,n$; $j = 1,2,\ldots,p$. The consequent parts consists of the consequent variable y and crisp linear functions $f_{i,}$ $i = 1,2,\ldots, n$.

In the DENFIS modelling approach, the considered fuzzy membership functions are triangular type functions defined as follows (Kasabov and Song, 2002):

$$\mu(x) = mf(x, p, q, r) = \max\left(\min\left(\left(\frac{x-p}{p-q}\right), \left(\frac{r-x}{r-q}\right)\right), 0\right) \tag{1}$$

where q = the value of the cluster centre on the x dimension, $p = q\text{-}d \times D_{thr}$ and $r = q + d \times D_{thr}$, d = 1.2 ~ 2.0. The predefined threshold value, D_{thr} is a clustering parameter.

The result of inference y^1 for an input vector $x^1 = [x_1{}^1\ x_2{}^1 \ldots\ x_p{}^1]$ is the weighted average of each rule's output (Takagi and Sugeno, 1985):

$$y^1 = \frac{\sum_{i=1}^{n} w_i f_i\left(x_1^1 x_2^1, \ldots, x_p^1\right)}{\sum_{i=}^{n} w_i} \tag{2}$$

where

$$w_i = \prod_{k=1}^{p} \mu_{ik}\left(x_k^1\right); i = 1,2,\ldots,n; k = 1,2,\ldots,p.$$

2.2 Learning algorithm of DENFIS

The online model of DENFIS uses first-order Takagi-Sugeno fuzzy rules. It generates and updates the linear functions in the consequence parts by adopting a linear least-square estimator using training data (Hsia, 1977). The linear function can be expressed as follows (Goodwin and Sin, 1984):

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_p x_p.$$

A training dataset includes q samples of input and output:

$$\left\{\left(X_i y_i\right), Xi = \left[x_{i1}, x_{i2}, \ldots, x_{ip}\right], i = 1,2,\ldots,p\right\}.$$

The least-square estimators are calculated as the coefficient $a = [a_0\ a_1\ a_2 \ldots a_p]^T$, by using the following relationship (Kasabov and Song, 2002):

$$a = \left(B^T W B\right)^{-1} B^T W y \tag{3}$$

where

$$B = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{q1} & x_{q2} & \cdots & x_{qp} \end{bmatrix}, y = \begin{bmatrix} y_1 y_2 y_3 \ldots y_p \end{bmatrix}^T,$$

and

$$W = \begin{bmatrix} w_1 & 0 & \cdots & \cdots & 0 \\ 0 & w_2 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & \cdots & w_q \end{bmatrix}, w_i = 1 - d_i,$$

d_i = distance between the i^{th} sample and the corresponding cluster centre, $i = 1,2,\ldots,q$.

The equation (3) can be rewritten as follows using a recursive least square estimator formula (Kasabov and Song, 2002).

$$Q = \left(B^T W B\right)^{-1}, a = Q B^T W y \tag{4}$$

Let the j^{th} row vector of matrix B is represented by b_j^T and the j^{th} element of y be y_j, then a can be iteratively calculated using the following relationship (Goodwin and Sin, 1984):

$$a_{j+1} = a_j + w_{j+1} Q_{j+1} b_{j+1} \left(y_{j+1} - b_{j+1}{}^T a_j\right), Q_{j+1} = \frac{1}{\gamma}\left(Q_j - \frac{w_{j+j} Q_j b_{j+1} b_{j+1}{}^T Q_j}{\gamma + b_{j+1}{}^T Q_j b_{j+1}}\right); \tag{5}$$

where $j = n, n+1, \ldots, q-1$; $w_{j+1} = 1 - d_{j+1}$; γ(forgetting factor) = 0.8 to 1(typical range of values).

Generally, the forgetting factor gives less weight to older error samples. The initial values of Q_n and a_n are determined with the help of equation 4 using first n samples from the training dataset.

3 Study site and dataset description

This study uses primary sources of air quality and meteorological data of a remote site in the Empty Quarter, Saudi Arabia. The Empty Quarter is considered as one of the largest sand deserts in the world (Vincent, 2008). It encompasses most of the southern third of the Arabian Peninsula and covers the area between longitude 44°30′ to 56°30′ and latitude 16°30′–23°00′N (Clark, 1989). It is an arid region with only about 35 mm of annual rainfall. The temperature during summer may reach 55°C at noon. The selected sites Mulayhah (Site 1) and Murayt (Site 2) are located in remote areas inside two drilling sites. A location map of the selected sites is provided in Figure 1.

The meteorology and air quality monitoring equipment were used to collect minute specific data for seven days (starting from 00:45:00 on December 06, 2007 to 24:00:00 on December 12, 2007) for Site 1. The data included NO, NO_2 and O_3, and wind speed, wind direction, relative humidity, temperature, and barometric pressure. The same parameters were measured for Site 2 starting from 00:30:00 on June 13, 2008 to 23:50:00 on June 19, 2008.

Figure 1 The location of the study area (22°52'12"N, 49°49'12"E) (see online version for colours)

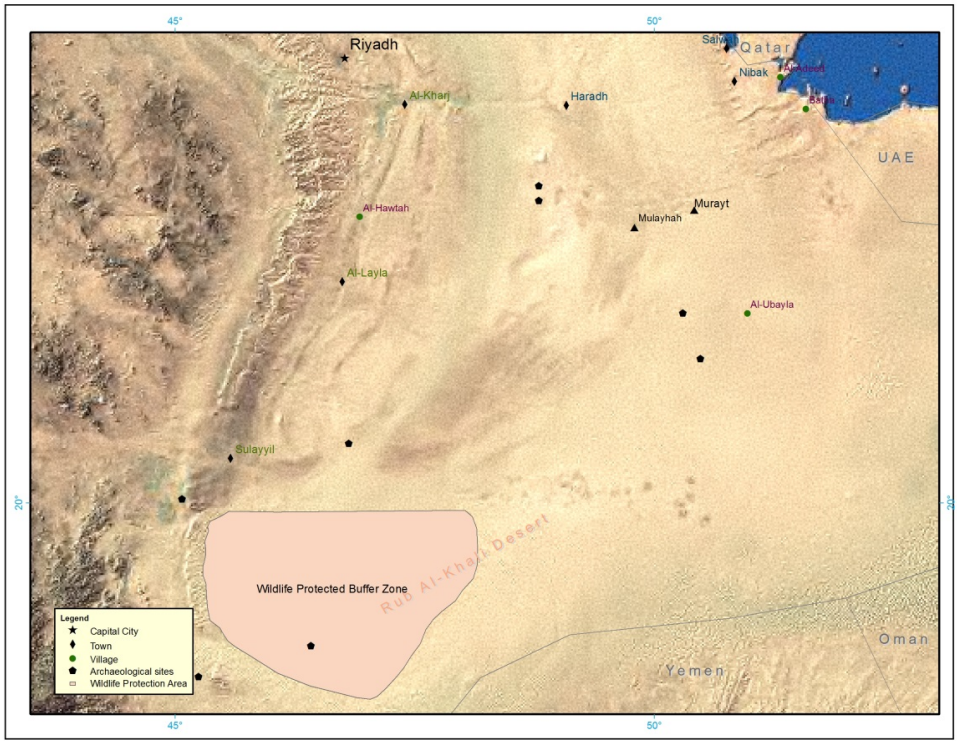


Table 1 Statistical summary of the data used in the proposed model

	Sites	*Wind speed (m/s)*	*Temperature (°C)*	*Relative humidity (%)*	*Barometric pressure (mbar)*	*NO (ppbv)*	*NO_2 (ppbv)*	*O_3 (ppbv)*
Maximum	Site 1	20.9	28.9	100	989	338.5	98.0	51
	Site 2	28.8	44.0	51	992	157.3	52.2	64
Minimum	Site 1	0	9	29	980	0.5	0.5	7
	Site 2	0	26	7	984	0.5	2.8	5
Mean	Site 1	6	19	63.4	983.7	14.5	9.2	31.2
	Site 2	14	38	16	987	6.9	4.8	32
Standard deviation	Site 1	3	5.2	16.3	1.8	26.2	10.4	7.5
	Site 2	8	4.2	8.1	1.9	6.3	2.9	13.0
Kurtosis	Site 1	1.5	–1	–0.7	–0.3	35.6	12.8	0.4
	Site 2	–1.2	0.2	5.3	0.3	387	104	–0.9
Skewness	Site 1	0.9	0.1	0.2	0.3	5.4	3.3	–0.4
	Site 2	–0.02	–0.8	2.4	0.8	18.5	8.4	–0.3

A mobile air quality monitoring system was used in this study. It is designed to measure real-time concentrations of above mentioned pollutants in ambient air. The monitoring equipment mainly includes the Monitors Labs (ML) 9800 Series ambient air analysers. The Model 9810 Ozone Analyser is a UV photometer. It measures low concentrations of O_3 using the absorption of UV radiation at 254 nm by the O_3 molecule. In order to calculate the O_3 concentration the analyser's microprocessor uses the Beer Lambert relationship. The lowest detectable limit is 1 ppbv. The ML 9841A Nitrogen Oxides Analyser measures the chemiluminescent reaction between NO and O_3. All readings were corrected for changes in sample flow rate. The lowest detectable limit of NO and NO_2 is 0.5 ppbv. The wind direction, wind speed, temperature, humidity, and barometric pressure were measured using a rotating vane, a three cup anemometer, a thermistor network, a thin-film capacitor, and a piezo resistive sensor, respectively.

The field monitoring data were collected during the winter and the summer seasons. The missing data (less than 0.02 %) were estimated through linear interpolation. In this study, ten-minute average data were used and the total number of samples were 1002 and 999 for Site 1 and Site 2, respectively. A statistical summary of the collected data is provided in Table 1. The standard deviations of wind speed, temperature, and relative humidity indicate low variability of the data. The barometric pressure rarely varies significantly from the mean value. The concentrations of NO and NO_2 show higher variability compared to O_3 for Site 1. The concentrations of O_3 for Site 2 indicate higher variability compared to NO_x. The skewness values of the data used revealed that the meteorological and air quality data except O_3 for Site 1 and wind speed, O_3, and temperature for Site 2 are spread out more above the mean. There is no clear indication that the data are generated from any perfectly symmetric distribution process except the wind speed of Site 2. The kurtosis values of the data indicate that all the data are less outlier-prone than the normal distribution except NO_x for Site 1, and relative humidity and NO_x for Site 2.

Table 2 Description of the considered input and output data

Input label	*Description*	*Input label*	*Description*
WS	Wind speed in m/s	NO	NO concentration
WD	Wind direction in degrees (0 to 360)	NO2	NO_2 concentration
BP	Barometric pressure in mbar	O3	O_3 concentration
RH	Relative humidity in percentage	XX_WA	The window average (WA) of the parameter XX considering the values of it at time t-60 min, t-50 min, t-40 min, t-30 min, t-20 min, t-10 min, and t.
TEMP	Temperature in degree Celsius	XX_T-YY	The value of the parameter, XX (WS, WD, BP, RH, TEMP, NO, or NO2) at time t-YY min (YY can be 60, 50, 40, 30, 20, or 10 min) expressed in corresponding units
TIME	*Sine* value of the time of day (normalised) expressed as a cyclic parameter	XX_WSD	The standard deviation of the parameter XX considering the values of it at time t-60 min, t-50 min, t-40 min, t-30 min, t-20 min, t-10 min, and t.

The model is developed using 80% of the available data and the remainder of the data was used for testing. The learning dataset was selected randomly. The list of the considered input data is available in Table 2. Different operations were performed to transform the inputs, including moving averages and standard deviations of a few previous values of each input, and time-lagged data. It is assumed that the transformed inputs will provide more knowledge regarding the input and improve the model performance. As the solar radiation data were not available, the time is expressed as the *sine* value of the time of the day (normalised). The transformed time input is considered as a surrogate of solar radiation. The considered inputs for the model are provided in Table 3.

Tropospheric ozone is the result of complex photochemical processes driven by two major classes of directly emitted precursors including nitrogen oxides (NOx) and volatile organic compounds (VOC) (Castellano et al., 2009). Typically, there are two regimes with different O_3-NO_x-VOC sensitivity in the relationship among O_3, NO_x and VOC. The NO_x-sensitive and VOC-sensitive regimes are characterised by relatively low NO_x and high VOC, and relatively low VOC and high NO_x, respectively. In the NO_x sensitive regime, O_3 increases with increasing NO_x and the change in O_3 in response to increase in VOC is insignificant. In the VOC-sensitive regime, O_3 decreases with increasing NO_x and increases with increasing VOC (Sillman, 2003). The ozone formation is dependent on the precursor concentrations and the characteristics of their emission sources (Castellano et al., 2009). However, the ozone distribution is mainly influenced by the meteorological conditions and the topography of the study area (Jimenez et al., 2006).

As the meteorological parameters strongly influence the ozone levels (Bloomfield et al., 1996; Gardner and Dorling, 2000; Monteiro et al., 2005), one modelling scenario was selected considering only the meteorological data as input for each site (Case 1 and Case 3). Typically, the meteorological data are available and can be collected at a relatively low cost. In order to enhance the performance of the model some other precursors such as nitrogen oxides were considered as additional inputs in the other modelling scenarios (Case 2 and Case 4).

Table 3 The considered input and output of the developed model in four cases (case 1, 2, 3 and 4)

Modeling exercise	*Inputs*	*Output*
Case 1 and 3 (Meteorological data-based model)	TIME, WS, TEMP, RH, BP, COS(WD), SIN(WD), WS_T-10, TEMP_T-10, RH_T-10, COS(WD)_T-10, SIN(WD)_T-10, WS_WA, TEMP_WA, RH_WA	O_3
Case 2 and 4 (Meteorological data and NOx-based model)	TIME, WS, TEMP, RH, BP, NO, NO2, COS(WD), SIN(WD), WS_T-10, TEMP_T-10, RH_T-10, COS(WD)_T-10, SIN(WD)_T-10, WS_WA, TEMP_WA, RH_WA, NO_T-10, NO_T-20, NO_T-30, NO_T-40, NO_WA, NO_WSD, NO2_T-10, NO2_T-20, NO2_T-30, NO2_T-40, NO2_WA, and NO2_WSD	O_3

4 Model development

The online DENFIS model generates and updates the fuzzy rules at the same time with appropriate partitioning of the input space. The DENFIS online model uses the online

evolving clustering method (ECM) for clustering the data in the input space. It is considered as a distance-based connectionist clustering method. Also, it is a one-pass algorithm for dynamically estimating the number of clusters in a dataset and for determining the centres in the input data space (Kasabov and Song, 2002). The number of clusters would depend on a predefined threshold value (D_{thr}). The steps of the ECM algorithm are provided below.

Step 1 Consider the first sample data point from the input data stream as the first cluster $C_{1,0}$ and the position of it as the first cluster centre, $Cc_{1,0}$. Set a value 0 as the cluster radius R_{u1}.

Step 2 The algorithm ends, if all the sample data points are already processed. Otherwise, the distance D_i between the current sample data x_j and the centres of all cluster centres C_{Ci} are calculated, where:

$$D_{jj} = \left\| x_j - C_{Ci} \right\|, i = 1,2,3,\ldots,n$$

Step 3 Select the cluster C_p and the corresponding cluster centre C_{Cp} with the minimum distance:

$$D_{jp} = \min\left(D_{ji}\right), \text{ when } D_{ji} \le R_{ui}, i = 1,2,3,\ldots,n.$$

If the constraint is satisfied, the algorithm returns to Step 2, otherwise to Step 4.

Step 4 Calculate the values of $S_{ji} = D_{ji} + R_{ui}$, $i = 1,2,3,\ldots,n$, and *s*elect the cluster C_b and the corresponding cluster centre C_{Cb} with the minimum value of S_{ji}:

$$S_{jb} = D_{jb} + R_{ub} = \min\left(S_{jj}\right), i = 1,2,3,\ldots,n$$

Step 5 If the minimum distance (S_{jb}) obtained in the previous step is greater than $2 \times D_{thr}$ then the sample data x_j is not a member of any existing cluster. The algorithm returns to Step 1.

Step 6 If $S_{jb} \le 2 \times D_{thr}$, the position of the cluster centre C_{Cb} is updated and its radius R_{ub} is increased as follows:

$R_{ub}^{new} = \dfrac{S_{jb}}{2}$ and the new cluster centre C_{Cb}^{new} is located on the line connecting x_j and C_{Cb}.

As a result, R_{ub}^{new} is the distance from x_j to C_{Cb}^{new}. The algorithm returns to Step 2. This algorithm does not keep any information of previously processed samples, but it ensures that the maximum distance between any cluster centre and its member data is not greater than D_{thr}.

The typical steps required for the development of online DENFIS model follow.

Step 1 Select n samples from the training dataset.

Step 2 Use ECM to determine c cluster centres for the selected samples.

Step 3 Find s_i samples which are closest to the centre in the input space, where $i = 1,2,\ldots,m$.

Step 4 Generate the antecedent parts of the fuzzy rule using the location of the cluster centre and equation (1). The consequents are calculated using the values of Q and b. These values of the consequent function are obtained using equation (4) on s_i samples.

The proposed model was developed automatically from the numerical training data through generating the fuzzy rules using the meteorological and NO_x data as inputs and the ozone concentration as the output. It is fairly established that the selected precursors (NO_x) and the meteorological parameters influence ozone formation and distribution, respectively (Castellano et al., 2009; Jimenez et al., 2006; Monteiro et al., 2005). Therefore, it is assumed that a fuzzy logic model can be developed to predict the ozone concentration using the meteorological and NOx data. The main aspects of the proposed model are structure identification of fuzzy inference system and its parameter estimation. A clustering technique (such as ECM) was used to find the appropriate fuzzy rules, determine the overall number of rules and tune the parameters on the consequent and/or antecedent parts of the fuzzy rules. The antecedents of a fuzzy rule corresponding to a cluster centre were created through using the position of the cluster centre and equation (1). The values of Q and a of the consequent function were calculated using equation (4) for q_i data points. New fuzzy rules may be created and some existing rules will be updated depending on new data points (Kasabov and Song, 2002). If the ECM finds a new cluster centre, then a new fuzzy rule will be created. The output of an input vector is calculated using equation 2.

In this study, different values of D_{thr} within the range between 0.01 and 0.2 are investigated during the model building process. The model is mainly sensitive to D_{thr}. The value of D_{thr} was fixed at 0.07 and 0.01 for Case 1 and Case 2, and Case 3 and Case 4, respectively based on the results of the numerical experiments. The required experiments were conducted in the MATLAB numeric computing environment.

5 Results and discussion

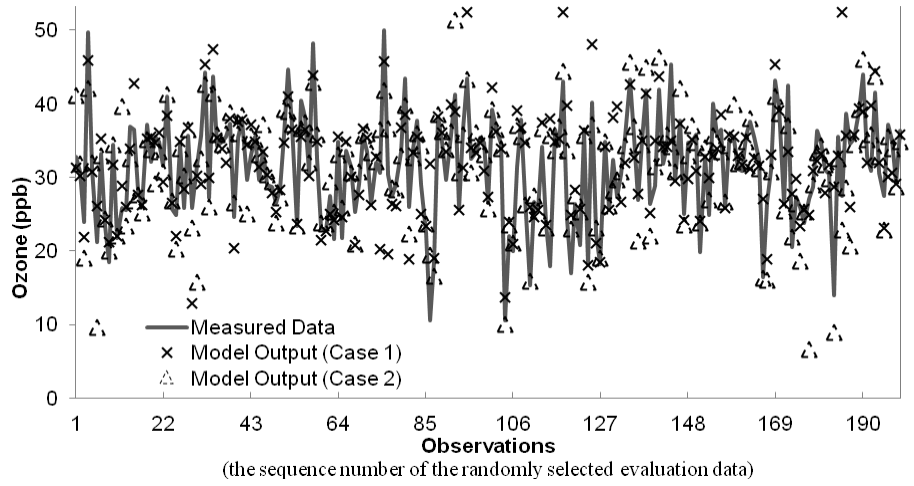
The output of the developed model for the validation data is shown along with the measured ozone concentrations in Figure 2. It appears that the outputs are in good agreement with the measured values of ozone concentrations. The model performs adequately in predicting the peak values except for a few cases.

The performance of the proposed model was evaluated with respect to a number of error measures including mean absolute percentage error (MAPE), mean absolute error (MAE), root mean square error (RMSE), Willmott's index of agreement (IA), and the coefficient of correlation (CC). The differences between the measured values and the model predictions are calculated to determine the mean difference (D) and the standard deviation (S) of the differences. A value of 1 for IA indicates a perfect match, whereas a value of 0 indicates complete disagreement (Willmott, 1981). The value of 'S' indicates the width of the confidence interval. The CC indicates the strength of statistical correlation between the predicted outputs and the measured values. The values of the selected performance measures are reported in Table 4. The Model in Case 2 performed better than that in Case 1 with respect to all the considered performance measures. The performance measures of the Model for Case 3 and Case 4 are almost same except for D, S and MAPE. It indicates that the consideration of NO and NO_2 as additional inputs

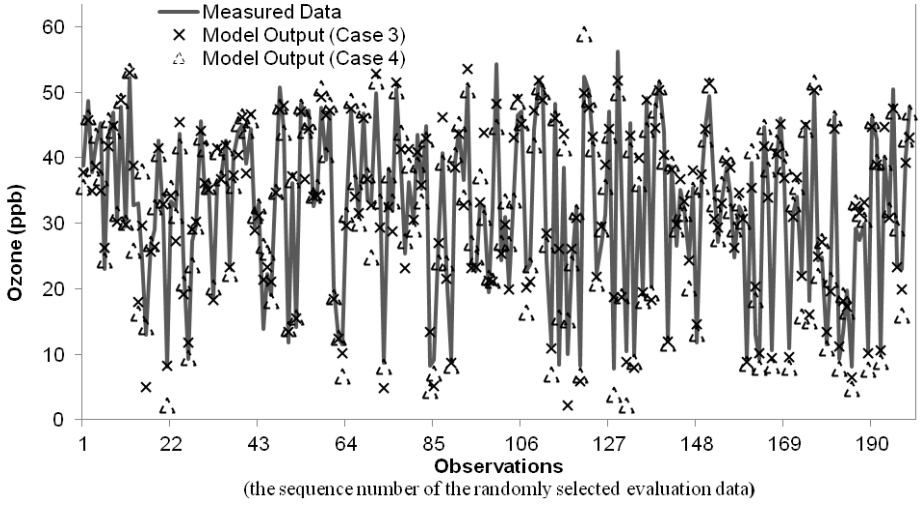
(other than the meteorological data) do not improve the performance of the Model significantly for Site 2 during the summer season. It could be attributed to the prevailing relationships of O3 with NO_x and VOC.

The scatter plot of the measured data and the model output shows the relationship between them. An identity line is often drawn as a reference. The more the datasets agree, the more the data points tend to concentrate in the vicinity of the identity line. The data points fall on the identity line exactly, if the measured data and the model output are numerically identical. The scatter plots of the measured evaluation data and the predicted outputs of the model for the four cases are shown in Figure 3. The scatter plots appear almost similar for the model in all cases.

Figure 2 (a) Measured ozone concentrations and the corresponding model predictions for case 1 and case 2 (b) Measured ozone concentrations and the corresponding model predictions for case 3 and case 4

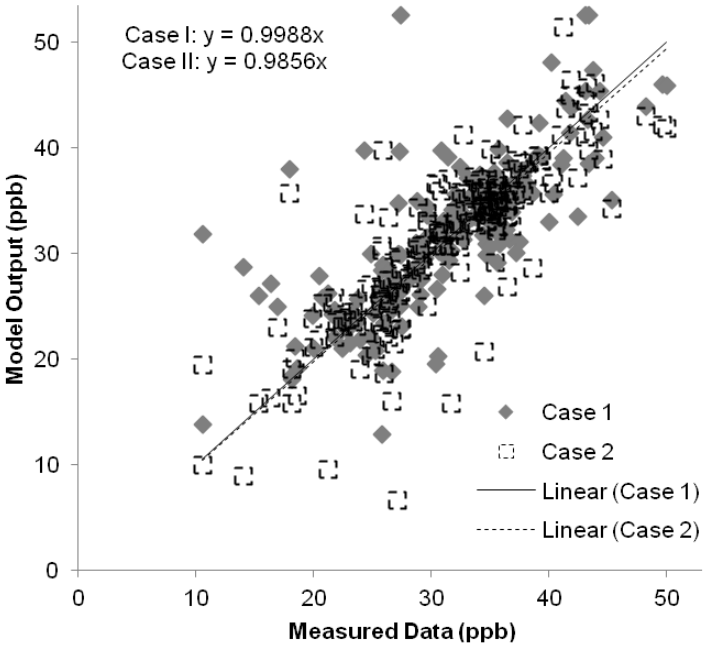


(a)

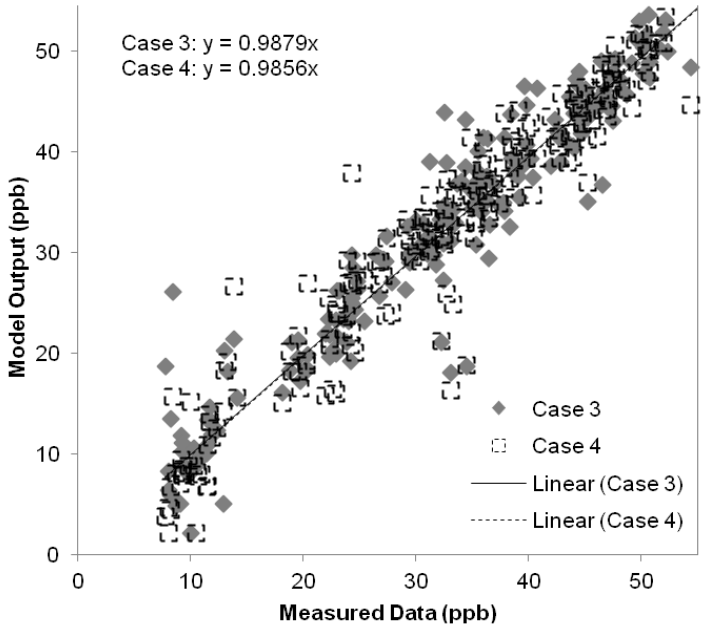


(b)

Figure 3 (a) The scatter plot of the measured ozone concentrations and the output for case 1 and case 2 (b) The scatter plot of the measured ozone concentrations and the output for case 3 and case 4



(a)

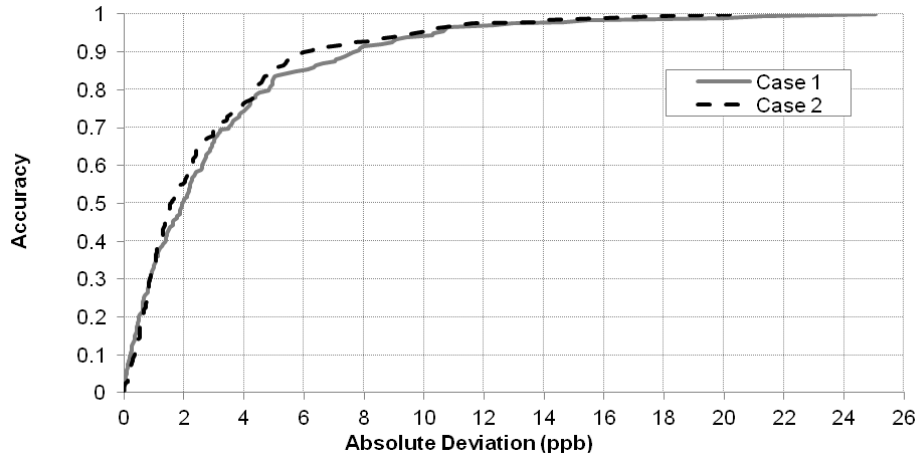


(b)

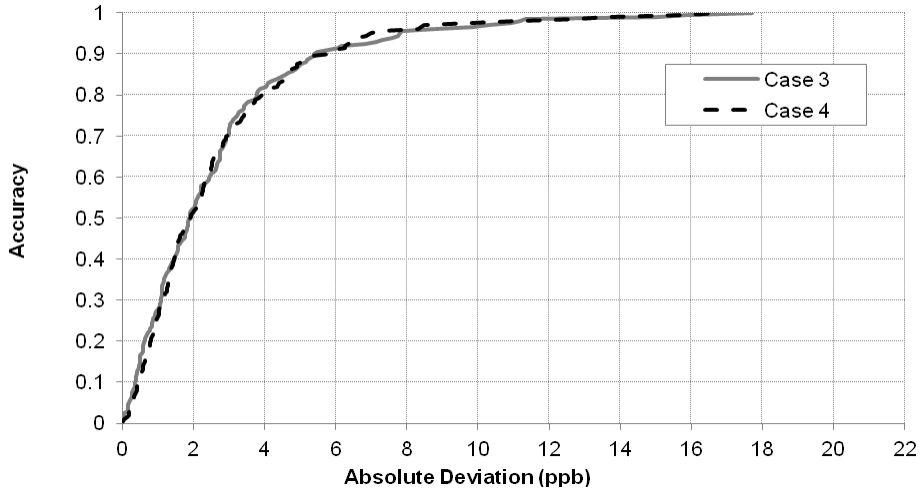
Table 4 Performance measures for the developed model

Modelling exercise	*MAE (ppbv)*	*MAPE (%)*	*RMSE (ppbv)*	*IA*	*CC*	*D (ppbv)*	*S (ppbv)*	*AOC (ppbv)*
Case 1	3.14	11.84	4.9	0.98	0.76	-0.39	4.96	3.08
Case 2	2.81	9.82	4.3	0.99	0.83	0.32	4.28	2.76
Case 3	2.66	11.51	3.9	0.98	0.95	0.15	3.88	2.62
Case 4	2.66	11.39	3.8	0.98	0.96	0.40	3.76	2.62

Figure 4 (a) The RECs curve for case 1 and case 2 (b) The RECs curve for case 3 and case 4



(a)

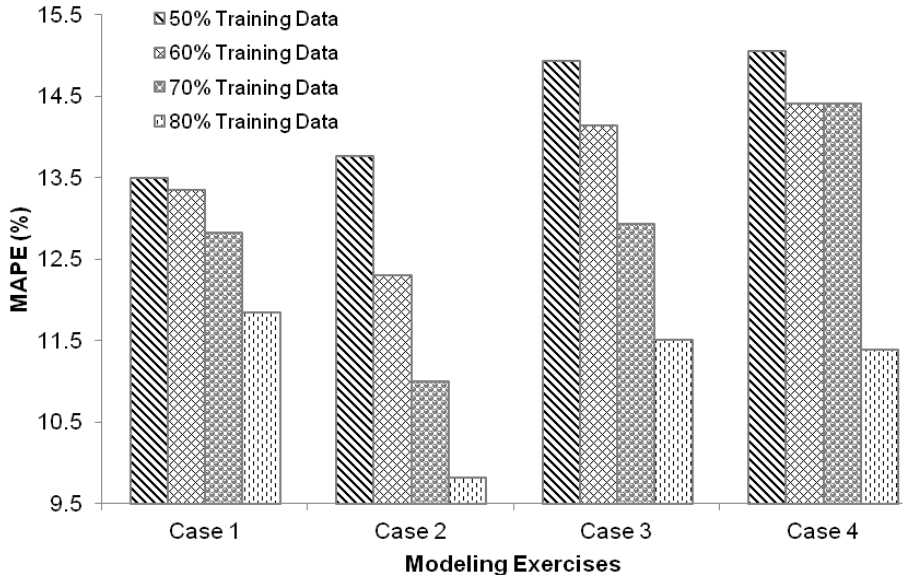


(b)

The regression error characteristic (REC) curve is used in this study for further analysis of the proposed model. It provides an approach to visualise and evaluate different regression models (Fawcett, 2003) by plotting error measures versus the percentage of points predicted within the tolerance. Accuracy indicates the percentage of points which are within the tolerance limit (Bi and Bennett, 2003). It also provides an estimation of the cumulative distribution function of the error. The area-over-curve (AOC) is a biased estimate of the expected error for a prediction. The details of REC can be found in Bi and Bennett (2003), De Pina and Zaverucha (2006), and Torgo (2005). The REC curve of the Model in Case 1 shows that the absolute deviation of around 90% of the evaluation data is less than or equal to 8 ppbv, which is around 25% of the median [Figure 4(a)]. On the other hand, in Case 2, the absolute deviation of around 90% of the evaluation data is less than or equal to 6 ppbv, which is around 18.5% of the median [Figure 4(a)]. In Case 3 and Case 4, the absolute deviation of around 90% of the evaluation data is less than or equal to 5.4 ppbv (16% of the median) and 6.0 ppbv (18% of the median), respectively [Figure 4(b)]. The values of AOC are 3.08 ppbv, 2.76 ppbv, 2.62 ppbv, and 2.62 ppbv for the Model in Cases 1, 2, 3 and 4, respectively. These values are within 7% and 10% of the median. Based on the analysis of REC curves and AOC values, it appears that the performance of the Model in Case 2 is superior compared to Case 1. But the performance of the Model for Case 3 and Case 4 is almost the same with respect to REC curves and AOC values.

In order to evaluate the sensitivity of the Model with respect to training data, 50%, 60%, 70% and 80% of the total data were randomly selected for training and the remaining data were used for evaluating the model for each case. The changes in MAPE for different percentages of training data are shown in Figure 5. It shows that the MAPE values for the evaluation data are within 9.8% to 15.1% for different percentages of training data.

Figure 5 The effect of the training data on the performance of the evaluation data for different modelling exercises



The proposed model performed adequately in modelling and predicting the ozone levels in the Empty Quarter in four cases investigated. Based on the estimated performance indicators and prediction analysis, the proposed modelling approach can be considered as a fairly viable approach for ozone modelling as a function approximation with the help of the meteorological data and the concentrations of NO and NO_2 in an arid region.

6 Conclusions

In order to provide flexible, adaptive, and less assumption-dependent real-time modelling tools, this study used an online DENFIS-based modelling approach, which is based on Takagi-Sugeno fuzzy inference system. The developed model is capable of predicting ozone concentrations using the meteorological data, the NO and NO_2 concentrations, and their statistical transformations. It changes in time with new examples presented to the system while both the knowledge and the inference mechanism evolve. In order to provide deep insights on the performance of the model, the RECs curves were used along with other performance indicators. The results and the performance analysis of the model indicate the viability of application of the adopted online DENFIS modelling approach in short-term modelling of ozone levels in the Rub Al Khali Desert during both the winter and the summer seasons. The proposed model performs well with the NO_x precursors and meteorological data as input. Future work would investigate the use of the adopted approach for other sites of the area for different seasons and compare them with traditional machine learning models. Future endeavours may also focus on improving the online learning algorithm by incorporating advanced parameter optimisation techniques.

Acknowledgements

The authors would like to gratefully acknowledge the support of King Fahd University of Petroleum and Minerals (KFUPM) in conducting this research.

References

Abdul-Wahab, S.A. and Al-Alawi, S.M. (2002) 'Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks', *Environmental Modeling & Software*, Vol. 17, No. 3, pp.219–228.

Bi, J. and Bennett, K.P. (2003) 'Regression error characteristic curves', *Twentieth International Conference on Machine Learning (ICML-2003)*, Washington, DC.

Bloomfield, P., Royle, J.A., Steinberg, L.J. and Yang, Q. (1996) 'Accounting for meteorological effects in measuring urban ozone levels and trends', *Atmospheric Environment*, Vol. 30, No. 17, pp.3067–3077.

Borrego, C., Tchepel, O., Costa, A.M., Amorim, J.H. and Miranda, A.I. (2003) 'Emission and dispersion modeling of Lisbon air quality at local scale', *Atmospheric Environment*, Vol. 37, No. 37, pp.5197–5205.

Braun, H. and Weisbrod, J. (1991) 'Evolving neural feedforward networks', *International Conference on Artificial Neural Nets and Genetic Algorithms (ANNGA93)*, pp.13–22, Innsbruck, Austria.

Caruana, R. (1997) 'Multitask learning', *Machine Learning*, Vol. 28, No. 1, pp.41–75.

Castellano, M., Franco, A., Cartelle, D., Febrero, M. and Roca, E. (2009) 'Identification of NOx and Ozone episodes and estimation of ozone by statistical analysis', *Water Air Soil Pollution*, Vol. 198, Nos. 1–4, pp.95–110.

Clark, A. (1989) 'Lakes of the Rub' al-Khali', in Amdt, R. (Ed.): *Saudi Aramco World*, Vol. 40, No. 3, pp.28–33.

De Pina, A.C. and Zaverucha, G. (2006) 'Using regression error characteristic curves for model selection in ensembles of neural networks', *Proceedings of 14th European Symposium on Artificial Neural Networks*, 26–28 April, pp.425–430, Bruges, Belgium.

Dunea, D., Oprea, M. and Lungu, E. (2008) 'Comparing statistical and neural network approaches for urban air pollution time series analysis', *Proceedings of the 27th IASTED International Conference on Modeling, Identification and Control*, 11–13 February, pp.93–98, Innsbruck, Austria.

Fawcett, T. (2003) *ROC Graphs: Notes and Practical Considerations for Data Mining*, Technical Report HPL-2003-4, HP Laboratories, Palo Alto, CA.

Gardner, M.W. and Dorling, S.R. (2000) 'Statistical surface ozone models: an improved methodology to account for non-linear behavior', *Atmospheric Environment*, Vol. 34, No. 1, pp.21–34.

Goodwin, G.C. and Sin, K.S. (1984) *Adaptive Filtering Prediction and Control*, Prentice-Hall, Englewood Cliffs, N.J.

Hassoun, M.H. (1995) *Fundamentals of Artificial Neural Networks*, MIT Press, MA.

Heo, J.S. and Kim, D.S. (2004) 'A new method of ozone forecasting using fuzzy expert and neural network systems', *Science of the Total Environment*, Vol. 325, Nos. 1–3, pp.221–237.

Hsia, T.C. (1977) *System Identification: Least-Squares Methods*, Health and Company, D.C.

Hwang, Y.C. and Song, Q. (2009) 'Dynamic neural fuzzy inference system', *Advances in Neuro-Information Processing*, Vol. 5506, pp.1245–1250.

Jain, S. and Khare, M. (2010) 'Adaptive neuro-fuzzy modeling for prediction of ambient CO concentration at urban intersections and roadways', *Air Quality, Atmosphere & Health*, Vol. 3, No. 4, pp.203–212.

Jimenez, P., Jorba, O., Parra, R. and Baldasano, J.M. (2006) 'Evaluation of MM5-EMICAT2000-CMAZ performance and sensitivity in complex terrain: high-resolution application to the northeastern Iberian Peninsula', *Atmospheric Environment*, Vol. 40, No. 26, pp.5056–5072.

Johanyák, Z.C. and Kovács, J. (2011) 'Fuzzy model based prediction of ground-level ozone concentration', *Acta Technica Jaurinensis*, Vol. 4, No. 1, pp.113–126.

Kao, J.J. and Huang, S.S. (2000) 'Forecasts using neural network versus Box-Jenkins methodology for ambient air quality monitoring data', *Journal of Air Waste Manag. Assoc.*, Vol. 50, No. 2, pp.219–226.

Karatzas, K.D. and Kaltsatos, S. (2007) 'Air pollution modeling with the aid of computational intelligence methods in Thessaloniki, Greece', *Simulation Modeling Practice and Theory*, Vol. 15, No. 10, pp.1310–1319.

Kasabov, N. (1998) 'ECOS: a framework for evolving connectionist systems and the eco learning paradigm', *Proceedings. of ICONIP*, pp.1222–1235, IOS Press, Kitakyushu.

Kasabov, N. (2001) 'Evolving fuzzy neural networks for supervised/unsupervised online knowledge-based learning', *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions*, Vol. 31, No. 6, pp.902–918.

Kasabov, N. and Song, Q. (2002) 'DENFIS: dynamic evolving neural-fuzzy inference system and its application for time-series prediction', *Fuzzy Systems, IEEE Transactions*, Vol. 10 No. 2, pp.144–154.

Monteiro, A., Vautard, R., Borrego, C. and Miranda, A.I. (2005) 'Long-term simulations of photo oxidant pollution over Portugal using the CHIMERE model', *Atmospheric Environment*, Vol. 39, No. 17, pp.3089–3101.

Park, D., Rilett, L.R. and Han, G. (1999) 'Spectral basis neural networks for real-time travel time forecasting', *Journal of Transportation Engineering*, Vol. 125, No. 6, pp.515–523.

Pires, C.M., Gonçalves, B., Azevedo, F.G., Carneiro, A.P., Rego, N., Assembleia, A.J.B., Lima, J.F.B., Silva, P.A., Alves, C. and Martins, F.G. (2012) 'Optimization of artificial neural network models through genetic algorithms for surface ozone concentration forecasting', *Environmental Science and Pollution*, Vol. 19, No. 8, pp.3228–3234, DOI: 10.1007/s11356-012-0829-9.

Sillman, S. (2003) 'Tropospheric ozone and photochemical smog', in Sherwood Lollar, B. (Ed.): *Treatise on Geochemistry*, Vol. 9: Environmental Geochemistry, Ch. 11, Elsevier.

Sousa, S.I.V., Martins, F.G., Alvim-Ferraz, M.C.M. and Pereira, M.C. (2007) 'Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations', *Environmental Modelling & Software*, Vol. 22, No. 1, pp.97–103.

Takagi, T. and Sugeno, M. (1985) 'Fuzzy identification of systems and its applications to modelling and control', *IEEE Trans. Syst., Man, Cybern.*, Vol. SMC-15, No. 1, pp.116–132.

Torgo, L. (2005) 'Regression error characteristic surfaces', *Proceedings of Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp.697–702.

Vincent, P. (Ed.) (2008) *Saudi Arabia: An Environmental Overview*, Taylor & Francis, Psychology Press, UK.

Willmott, C.J. (1981) 'On the validation of models', *Physical Geography*, Vol. 2, No. 2, pp.184–194.

Yildirim, Y. and Bayramoglu, M. (2006) 'Adaptive neuro-fuzzy based modeling for prediction of air pollution daily levels in City of Zonguldak', *Chemosphere*, Vol. 63, No. 9, pp.1575–1582.

Zhang, Q. and Benveniste, A. (1992) 'Wavelet networks', *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 3, No. 6, pp.889–898.