Environ Sci Pollut Res (2013) 20:3395–3404
DOI 10.1007/s11356-012-1271-8

RESEARCH ARTICLE

# Ozone levels in the empty quarter of Saudi Arabia—application of adaptive neuro-fuzzy model

**Syed Masiur Rahman · A. N. Khondaker · Rouf Ahmad Khan**

Received: 17 July 2012 / Accepted: 19 October 2012 / Published online: 31 October 2012


**Abstract** In arid regions, primary pollutants may contribute to the increase of ozone levels and cause negative effects on biotic health. This study investigates the use of adaptive neuro-fuzzy inference system (ANFIS) for ozone prediction. The initial fuzzy inference system is developed by using fuzzy C-means (FCM) and subtractive clustering (SC) algorithms, which determines the important rules, increases generalization capability of the fuzzy inference system, reduces computational needs, and ensures speedy model development. The study area is located in the Empty Quarter of Saudi Arabia, which is considered as a source of huge potential for oil and gas field development. The developed clustering algorithm-based ANFIS model used meteorological data and derived meteorological data, along with NO and $NO_2$ concentrations and their transformations, as inputs. The root mean square error and Willmott's index of agreement of the FCM- and SC-based ANFIS models are 3.5 ppbv and 0.99, and 8.9 ppbv and 0.95, respectively. Based on the analysis of the performance measures and regression error characteristic curves, it is concluded that the FCM-based ANFIS model outperforms the SC-based ANFIS model.

Responsible editor: Gerhard Lammel

**Highlights**
✓ Investigation of ozone levels in an arid region of Saudi Arabia
✓ Development of initial fuzzy inference system using fuzzy C-means algorithm
✓ Development of initial fuzzy inference system using subtractive clustering algorithm
✓ Development of adaptive neuro-fuzzy inference system
✓ Analysis of model performance using regression error characteristic curves

S. M. Rahman (✉) · A. N. Khondaker · R. A. Khan
King Fahd University of Petroleum & Minerals,
P. O. Box 713, Dhahran 31261, Saudi Arabia
e-mail: smrahman@kfupm.edu.sa

**Keywords** Fuzzy C-means · Subtractive clustering · Ozone modeling

## Introduction

The Empty Quarter is considered as a source of huge potential for oil and gas field development, and recently, a number of gas exploration activities are going on in that area by different international and local companies. It is an arid region with only about 35 mm of annual rain and the temperatures during summer may reach 55 °C at noon. 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 causes negative effects on biotic health. Air quality models can play a significant role in assessing atmospheric quality, simulating the atmospheric environment system, increasing the domain knowledge on the environmental phenomenon, and producing reliable forecasts (Karatzas and Kaltsatos 2007). These can also provide early warnings to the population and reduce the number of required measuring stations. Unfortunately, the task of modeling ozone levels is considered very difficult due to the complex interactions between pollutants and meteorological variables (Borrego et al. 2003).

The machine learning model provides a flexible and adaptive modeling approach for air pollutants including ozone. It does not require making many assumptions on the modeled phenomenon. Artificial neural network (ANN) models are already widely investigated to model the concentrations of air pollutants (Prakash et al. 2011; Karatzas and Kaltsatos 2007). The capability of ANN models was demonstrated for many air pollutants including ozone (Abdul-Wahab and Al-Alawi 2002; Sousa et al. 2007). In order to ensure better performance, many approaches are adopted in developing ANN models such as wavelet neural network (Zhang and

Benveniste 1992), multitasking neural network (Caruana 1997), and evolutionary neural network (Hassoun 1995; Braun and Weisbrod 1991). Pires et al. (2012) proposed genetic algorithm-based ANN model for ozone prediction. They used genetic algorithms to define the activation function in hidden layer and the number of hidden neurons of the ANN.

The major advantages of ANN over traditional statistical models include self-learning, self-adaptation, faster computation (due to parallel processing), and noise rejection (Kao and Huang 2000; Dunea et al. 2008). The performance of ANN is influenced by the network training, the amount and quality of training data, and network parameters. 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 have to be appropriately selected. A conventional ANN may work efficiently when the modeled function is relatively monotonic with only a few dimensions of the input features (Haykin 1994; Musavi et al. 1994). The ANN may experience difficulties in approximating functions when the input features are not linearly separable, which implies that the approximated function has a higher complexity (Park et al. 1999).

The ANN is not suitable for the system in which the knowledge is represented in the form of linguistic data. On the other hand, the fuzzy logic theory allows the accurate representation of a given system behavior using a set of simple "if–then" rules but it is unable to tackle knowledge stored in the form of numerical data (Fakhreddine and de Silva 2004). Due to the linguistic variable handling capability, the fuzzy logic model is investigated for ozone prediction (Nebot et al. 2008). Jorquera et al. (1998) developed fuzzy models for forecasting the maximum daily levels of ozone. Heo and Kim (2004) used fuzzy logic and ANN model consecutively to forecast daily maximum ozone concentrations. Inspired by the combined strength of ANN and fuzzy logic model, a hybrid neuro-fuzzy system known as adaptive neuro-fuzzy inference system (ANFIS) is proposed by Jang (1993). A few ANFIS-based attempts are already made for CO prediction (Jain and Khare 2010) and $SO_2$ prediction (Yildirim and Bayramoglu 2006). Despite their advantages and wide applicability area, the neuro-fuzzy logic-based solutions for ozone concentration prediction have not been investigated in the literature adequately (Johanyák and Kovács 2011).

This study attempted to use clustering algorithm-based ANFIS for modeling ozone levels in the arid region. The initial fuzzy inference system which is obtained by using the clustering algorithm determines the important rules, increases generalization capability of the fuzzy inference system, reduces computational needs, and ensures speedy model development.

## Fundamentals of ANFIS

ANFIS is developed to serve as a basis for constructing fuzzy inference system (FIS) with suitable membership functions, and its architecture is obtained by embedding the FIS into a framework of ANN (Jang 1993). A simple Takagi–Sugeno-type ANFIS model developed by Takagi and Sugeno (1985) for two inputs ($x$ and $y$) and one output is given in Fig. 1. The architecture and functions of each layer are described below.

*First layer* All nodes of this layer generate membership grades of input variables, which vary between 0 and 1. The node stores the parameters to define a bell-shaped membership function ($\mu$). Its function can be written as follows:

$$O_i^1 = \mu_{Q_i}(x) \tag{1}$$

$$O_{i+2}^1 = \mu_{R_i}(y) \tag{2}$$

$$\mu(x) = \exp\left\{-\frac{1}{2}\left(\frac{x-c}{\sigma}\right)^2\right\} \tag{3}$$

where $i$=1 and 2,

$x$ and $y$ are the inputs,

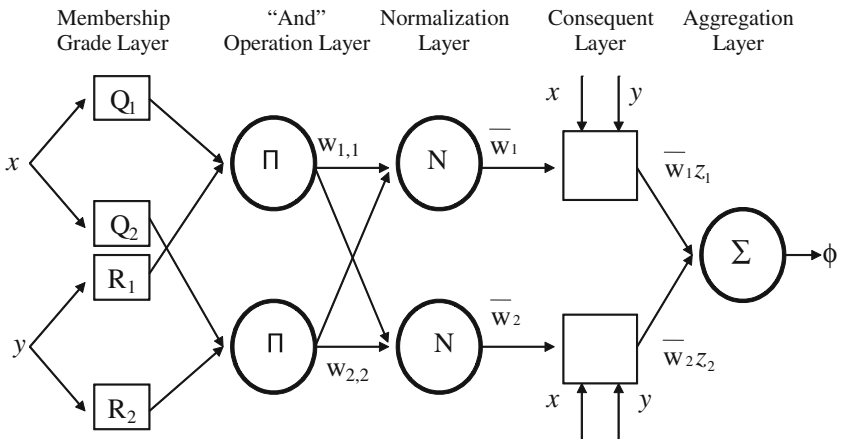


**Fig. 1** ANFIS architecture

$Q_i$ and $R_j$ are the linguistic label, and
$c$ and $\sigma$ are the mean and variance of the membership function, respectively, which are also known as premise parameters.

*Second layer* The node of this layer performs connective operation "AND" and any other T-norm within the rule antecedent to determine the corresponding firing strength. The node function follows.

$$O_{1,i}^2 = \mu_{Q_i}(x) \times \mu_{R_i}(y) = w_{1,i} \tag{4}$$

and

$$O_{2,i}^2 = \mu_{Q_i}(x) \times \mu_{R_i}(y) = w_{2,i} \quad \text{for i} = 1,\ 2. \tag{5}$$

*Third layer* The node of this layer performs normalization to determine relative strength of each rule. The output follows.

$$O_{\text{i}}^3 = \frac{w_i}{\sum w_i} = \overline{w_i} \quad \text{for } i = 1,\ 2. \tag{6}$$

*Fourth layer* In this layer, the output is obtained by multiplying the normalized firing strength of the rule with the rule output of Takagi–Sugeno type. The output of the node follows.

$$O_{\text{i}}^4 = \overline{w_i} z_i \quad \text{for } i = 1,\ 2 \tag{7}$$

where $z_m = a_k + b_k x + c_k y$ and $a$, $b$, and $c$ are constants.

*Fifth layer* The final node represents an addition node and the output ($\varphi$) is calculated as follows:

$$O^5 = \sum\nolimits_{i=1}^{2} \overline{w_i}(a_i + b_i x + c_i y) = \varphi. \tag{8}$$

If there are $n$ entries in a given dataset, the overall error measure can be defined by

$$E = \sum\nolimits_{i=1}^{n} E_i = \sum\nolimits_{i=1}^{n} (T_i - \varphi_i)^2 \tag{9}$$

where $T_i$ and $\varphi_i$ are the target and model output of the ANFIS, respectively, for the $i$th entry.

In ANFIS architecture, if the premise parameters are fixed, then the output of the whole system is a linear combination of the consequent parameters (Ying and Pan 2008). Finally, the output can be expressed as the following matrix format for $n$ number of training samples.

$$\varphi = AP \tag{10}$$

where $\varphi = \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_n \end{bmatrix}$, $P = \begin{bmatrix} a_1 \\ b_1 \\ c_1 \\ a_2 \\ b_2 \\ c_2 \end{bmatrix}$, $A = \begin{bmatrix} \overline{w}_1 & \overline{w}_1 x_1 & \overline{w}_1 y_1 & \overline{w}_2 & \overline{w}_2 x_1 & \overline{w}_2 y_1 \\ \overline{w}_1 & \overline{w}_1 x_2 & \overline{w}_1 Y_2 & \overline{w}_2 & \overline{w}_2 & \overline{w}_2 y_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \overline{w}_1 & \overline{w}_1 x_n & \overline{w}_1 y_n & \overline{w}_2 & \overline{w}_2 x_n & \overline{w}_2 y_n \end{bmatrix}$

The unknown matrix $P$ can be estimated with the help of least-squares method. The gradient descent technique is usually considered for tuning the architecture (Jang 1993). It indicates that the premise and consequent parameters are learnt with the help of gradient descent and least-squares method, respectively. The formula for the premise parameter $\delta$ by the gradient descent method can be expressed by

$$\nabla\delta = -\eta \frac{\partial E}{\partial \delta} \tag{11}$$

in which $\eta$ is the learning rate.

## Study site and dataset description

This study uses primary sources of air quality and meteorological data of a site at the Empty Quarter, Saudi Arabia, which is also known as Rub' Al Khali. It is one of the largest sand deserts in the world (Vincent 2008), which encompasses most of the southern third of the Arabian Peninsula. The desert covers the area between longitude 44°30′–56°30′ E and latitude 16°30′–23°00′ N (Clark 1989). The site is selected due to the huge potential for oil and gas field development at and around the site in the near future. The site is located in a remote area having very limited access. A location map of the site is provided in Fig. 2.

The weather and air quality stations were used to collect minute specific data for 7 days (starting from 00:45:00 on December 6, 2007 to 24:00:00 on December 12, 2007). The air quality and meteorological data included NO, $NO_2$, and $O_3$ and wind speed, wind direction, relative humidity, temperature, and barometric pressure.

A mobile air quality monitoring system was used in this short-term study which is designed to measure real-time concentrations of mentioned pollutants in the ambient air. It is housed in an environmentally controlled shelter mounted on a trailer, which contains storage and working space as well as the monitoring equipment, mainly the Monitors Labs (ML) 9800 Series ambient air analyzers. All analyzers have been designated as reference or equivalent methods for measuring ambient concentrations of the specified air pollutants by the US EPA. The Model 9810 Ozone Analyzer is a UV photometer which measures low concentrations of $O_3$ by measuring the absorption of UV radiation at 254 nm by the $O_3$ molecule. The analyzer's microprocessor uses the Beer–Lambert relationship to calculate the $O_3$ concentration. The lowest detectable limit of this analyzer is 1 ppbv. The ML 9841A Nitrogen Oxides Analyzer measures the chemiluminescent reaction between NO and $O_3$. Special software and pneumatic system ensures that accurate

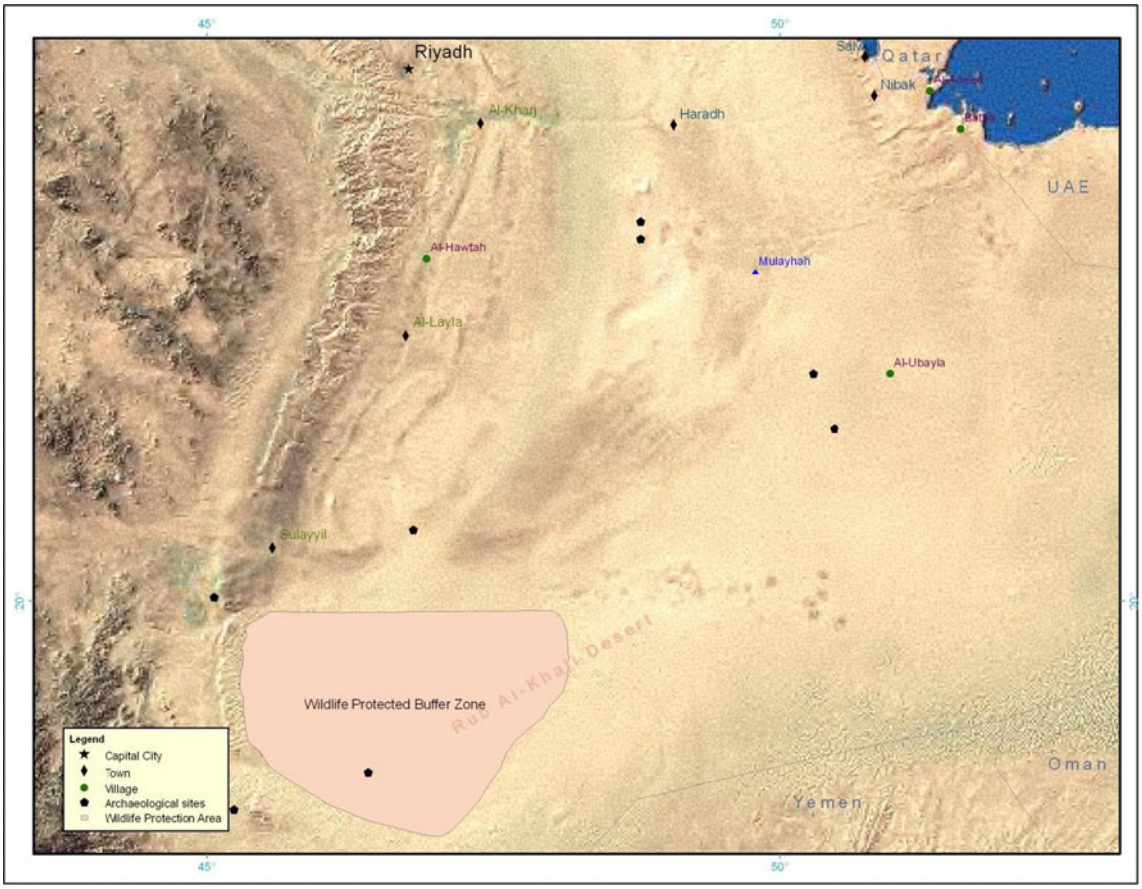


**Fig. 2** The location of the study site (22°52′12″ N, 49°49′12″ E)

$NO_2$ measurements are made, even in areas with rapidly changing NO. All readings were corrected for changes in sample flow rate, measuring cell vacuum, temperature, and atmospheric pressure, removing the major sources of calibration variations. 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 rotating vane, three-cup anemometer, thermistor network, thin-film capacitor, and piezo-resistive sensor, respectively.

The data were collected during the winter season. The missing data (less than 0.02 %) were estimated using linear interpolation. In order to build model, 10-min average data were used and the total number of samples was 1,002. A description of the data collected is provided in Table 1. The standard deviations of wind speed, temperature, and relative humidity are not high compared to the mean value which indicates low variability of the data. It also appears that the barometric pressure rarely varies significantly from the mean value. Concentrations of NO and $NO_2$ show higher variability compared to $O_3$. The skewness values of the data used revealed that the meteorological and air quality data except $O_3$ are spread out more above the mean and there is no clear indication that the data are generated from any

**Table 1** Description of the data used in the proposed model

| | Wind speed (m/s) | Temperature (°C) | Relative humidity (%) | Barometric pressure (mbar) | NO (ppbv) | $NO_2$ (ppbv) | $O_3$ (ppbv) |
|---|---|---|---|---|---|---|---|
| Maximum | 20.9 | 28.9 | 100 | 989 | 338.5 | 98.0 | 51 |
| Minimum | 0 | 9 | 29 | 980 | 0.5 | 0.5 | 7 |
| Mean | 6 | 19 | 63.4 | 983.7 | 14.5 | 9.2 | 31.2 |
| Standard Deviation | 3 | 5.2 | 16.3 | 1.8 | 26.2 | 10.4 | 7.5 |
| Kurtosis | 1.5 | −1 | −0.7 | −0.3 | 35.6 | 12.8 | 0.4 |
| Skewness | 0.9 | 0.1 | 0.2 | 0.3 | 5.4 | 3.3 | −0.4 |

**Table 2** Description of the input and output data

| Input label | Description | Input label | Description |
|---|---|---|---|
| WS | Wind speed in m/s | XX_T-10 | The value of the parameter, XX (WS, WD, BP, RH, TEMP, NO, or $NO_2$) at time $t-10$ min expressed in corresponding units |
| WD | Wind direction in deg (0 to 360) | O3_T+60 | $O_3$ concentration at time $t+60$ min |
| BP | Barometric pressure in mbar | NO | NO concentration |
| RH | Relative humidity in % | NO2 | $NO_2$ concentration |
| TEMP | Temperature in °C | O3 | $O_3$ concentration |
| TIME | Cosine value of the time of day (normalized) expressed as a cyclic parameter | 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 $t$ |
| 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 $t$ | XX_T-YY | The value of the parameter, XX (WS, WD, BP, RH, TEMP, NO, or $NO_2$) at time $t-$YY (YY can be 60, 50, 40, 30, or 20 min) expressed in corresponding units |

perfectly symmetric distribution process. The kurtosis values of the data indicate that all the data are less outlier prone than the normal distribution.

Randomly selected 80 % of the data were used for training the model and the rest were used for testing. The descriptions of the input and output data are provided in Table 2. In order to gain more knowledge about the input and improve the model performance, different operations were performed to transform the inputs including moving averages and standard deviations of few previous values of each input, time-lagged data. The inputs for the model are provided in Table 3.

## Development of ANFIS model

In ANFIS, the least-squares method leads to fast training and the gradient descent method slowly changes the underlying membership function that generates the basis functions for the least-squares methods (Jang et al. 1996). Therefore, it can be expected that ANFIS is likely to produce satisfactory results even after a few epochs of training. The ANFIS model shows good performance but sometimes produces spurious rules, which makes little sense (Jantzen 1998). A fuzzy model with a large number of rules can reduce generalization capability (Yen and Wang 1999). This problem can be solved by using an initial FIS generated by clustering techniques such as fuzzy C-means (FCM) which was developed and improved by Dunn (1973) and Bezdek (1981), respectively. The clustering algorithm determines the important rules, increases generalization capability of FIS, reduces computational needs, and ensures speedy model development.

FCM clustering algorithm considers each cluster as a fuzzy set, while a membership function measures the degree to which each training vector belongs to a certain cluster (Tsao et al. 1994). Each training vector may be assigned to multiple clusters. In this algorithm, the following objective function is minimized:

$$J_m = \sum_{i=1}^{n} \sum_{j=1}^{K} U_{ij}^{p} \|x_i - k_j\| \text{ for } 1 \le p < \infty \tag{12}$$

where $p$ is a fuzzy exponent having the value of real number, $U_{ij}$ is the degree of membership of $x_i$ in the cluster $j$, $x_i$ is the $i$th measured data, $k_j$ is the center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center such as Euclidian, Manhattan, and Mahalanobis distance.

The clustering is carried out through an iterative optimization of the mentioned objective function along with the update of membership $U_{ij}$ and the cluster centers $k_j$ by

$$U_{ij} = \frac{1}{\sum_{c=1}^{K} \left( \frac{\|x_i - k_j\|}{\|x_i - k_c\|} \right)^{\frac{2}{p-1}}} \text{ and } k_j = \frac{\sum_{i=1}^{n} U_{ij}^{p} x_i}{\sum_{i=1}^{n} U_{ij}^{p}}. \tag{13}$$

When $\max_{ij} \left\{ \left| U_{ij}^{(m+1)} - U_{ij}^{(m)} \right| \right\}$ is less than the predefined termination value, then the iteration will stop. Here, $m$ is the number of iteration steps.

**Table 3** The input and output of the developed models

| Inputs | Output |
|---|---|
| TIME, WS, TEMP, RH, BP, NO, 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 | $O_3$ at time $t$ |

FCM is used to develop the initial fuzzy inference system. The considered fuzzy exponent, iterations, and minimum amount of improvement were 2, 300, and 1e−05, respectively. The obtained FIS of nine clusters is then fine-tuned using ANFIS modeling approach. After systematic investigation, the number of epochs and learning rate were set to 187 and 0.00001, respectively. In the ANFIS structure, there are nine Gaussian membership functions for each input. As an example, the membership functions of temperature are shown in Fig. 3. The rule base contains nine rules. This model is named as model 1. All the above computation of the ANFIS model and FCM was performed in Matlab environment. The gradient descent- and least-squares method-based training algorithms are readily incorporated in the software.

As an alternative approach, the subtractive clustering (SC) algorithm is used to determine the initial fuzzy inference system. Chiu (1994) developed the SC method, which is a modified form of the mountain method. In this method, the potential of each data point to become a cluster center is calculated using the density of the surrounding data points. A set of $m$ data points ($X_1$, $X_2$, $X_3$,…, $X_m$) with $n$ dimensions are considered, which are assumed to have fallen inside a hyper box. The density ($\rho_i$) of data point $X_i$ can be expressed as follows:

$$\rho_i = \sum_{k=1}^{m} \exp\left(-\frac{||X_i - X_k||^2}{\left(\frac{r_p}{2}\right)^2}\right) \tag{14}$$

where $r_p$ is a positive number, which defines the influence area of a data point. The data points beyond the radius are considered not to have any influence on the density of $X_i$. The data point with the highest density value is selected as the first cluster center. After this selection, the density of a data point is modified by the following formula:

$$\rho_i = \rho_i - \rho_t \exp\left(-\frac{||X_i - X_t||^2}{\left(\frac{r_q}{2}\right)^2}\right) \tag{15}$$

where $r_q$ is a positive number which defines the influence area within which the function of the density of data point will reduce and $\rho_t$ is the density of the first selected cluster center ($X_t$) among the data points. After modifying the density of data point, the new cluster center will be selected and the density of all data points will be changed accordingly. This process will continue until all of the data points are within the radius of a cluster.

In the proposed model, SC algorithm is used to determine the FIS. The considered cluster center's range of influence, acceptance ratio, and rejection ratio were 0.1, 0.5, and 0.15, respectively. The considered radius for the data space was 0.85 times the width of the data space, and it indicates that each cluster center has a spherical neighborhood of influence with the given radius. The quash factor, acceptance ratio, and rejection ratio were 1.25, 0.5, and 0.15, respectively. The quash factor is used to multiply the radii values that determine the neighborhood of a cluster center, so as to quash the potential for outlying points to be considered as part of that cluster. The acceptance ratio sets the minimum threshold potential as a fraction of that of the first cluster. Another data point is accepted as a cluster center if the potential is higher than the minimum threshold potential. The rejection ratio sets the maximum threshold potential, as a fraction of the potential of the first cluster center. A data point is rejected as a cluster center when the potential is less than the threshold potential. The considered configurations produced 23 rules. The obtained FIS is then fine-tuned using ANFIS modeling approach. After systematic investigation, the number of epochs was set to ten. Each input contains Gaussian membership functions. This model is named as model 2. All the above computation of the ANFIS model and SC was performed in Matlab environment.

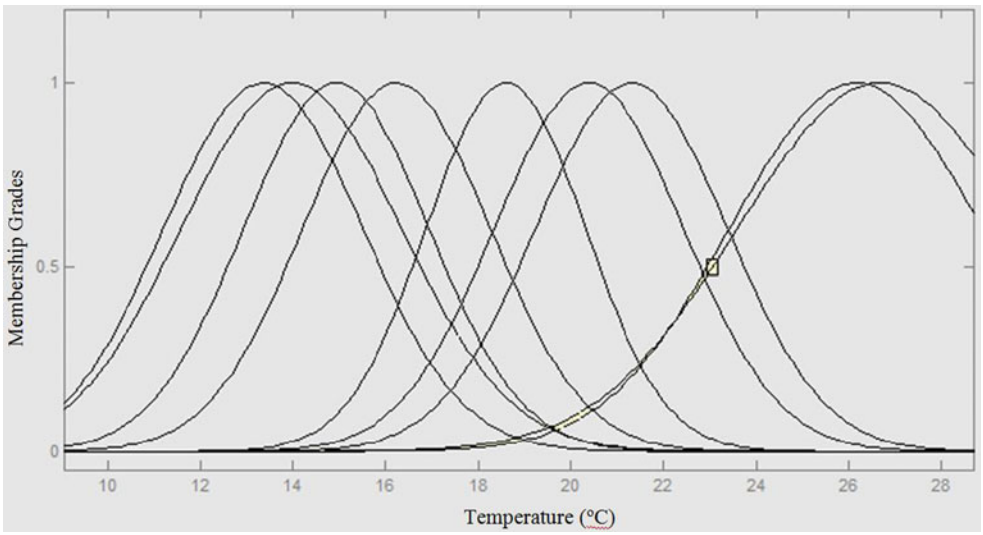


**Fig. 3** Membership functions of temperature considered in model 1

## Model evaluation and discussion

The predicted outputs of the developed models for the evaluation data are shown along with the measured ozone concentrations in Fig. 4. It seems that the outputs of both models are in good agreement with the measured values of ozone concentrations. Generally, the deviations of the output of model 1 are smaller compared to that of model 2. Model 1 performs adequately in predicting the peak values but model 2 fails to predict the peak values closely at least for a few cases.

In order to investigate the performance of the proposed models, mean absolute percentage error (MAPE), mean absolute error (MAE), root mean square error (RMSE), Willmott's index of agreement (IA), and coefficient of correlation (CC) were determined. The mean difference ($D$) and standard deviation ($S$) of the differences were also determined. If the value of IA is 1, it indicates a perfect match. A value of 0 indicates complete disagreement (Willmott 1981).The small value of $S$ indicates narrow width of confidence interval. The CC indicates the strength of statistical correlation between the predicted and actual outputs. A value of "1" indicates perfect statistical correlation and the value of "0" indicates no correlation at all. The MAPE, RMSE, MAE, IA, $S$, and $D$ are commonly used performance measure for numeric prediction. The mentioned performance measures are reported in Table 4 for both models. The MAE and MAPE values of model 2 are quite high compared to that of model 1. It seems that model 1 performs better than model 2 with respect to all the considered performance measures.

The scatter plot shows the relationship between the measured data and the model output. In this case, an identity line, i.e., a $y=x$ line is often drawn as a reference and the more the datasets agree, the more the data points tend to concentrate in the vicinity of the identity line. If the measured data and the model output are numerically identical, the data points fall on the identity line exactly. The scatter plots of the measured evaluation data and the output of the models are shown in Fig. 5. In the case of model 1, the CC value is high and the data points are generally very close to the identity line. It seems that the model output of model 1 has a very strong linear relationship with the measured evaluation data, which indicates the soundness of the proposed model. But the CC

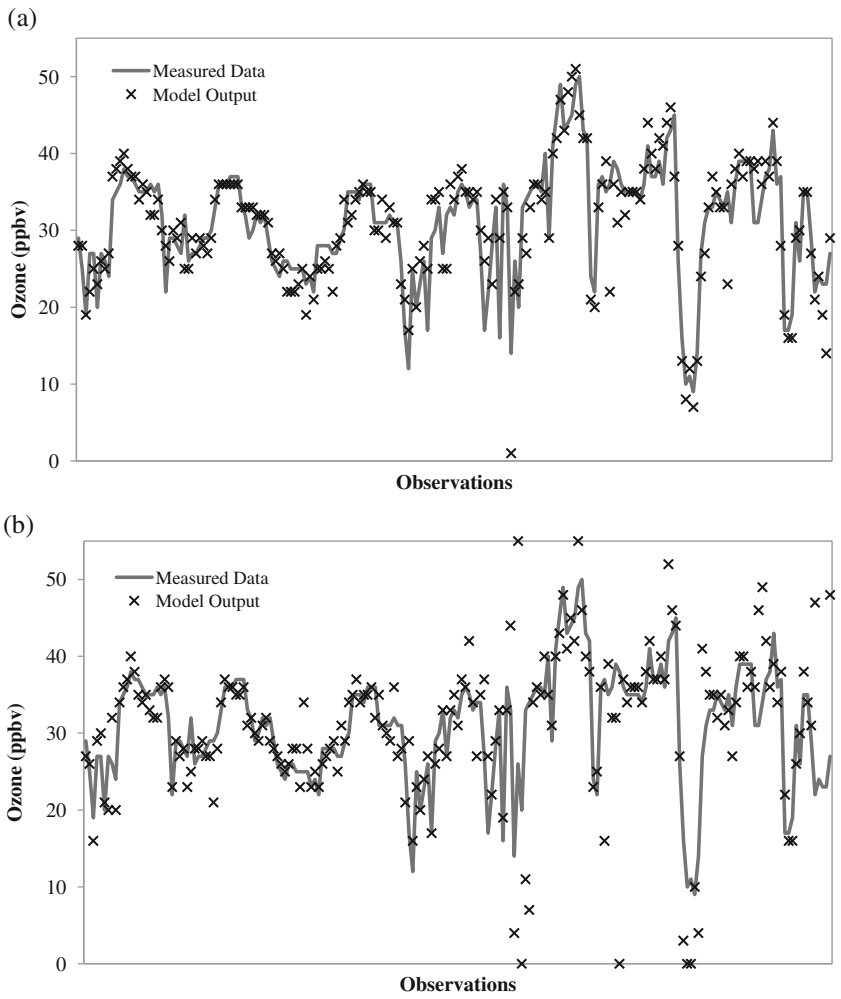


**Fig. 4** **a** Measured evaluation data and the corresponding outputs of model 1. **b** Measured evaluation data and the corresponding outputs of model 2

**Table 4** Performance measures for the developed model

| Model | MAE (ppbv) | MAPE (%) | RMSE (ppbv) | IA | CC | *D* (ppbv) | *S* (ppbv) |
|---|---|---|---|---|---|---|---|
| Model 1 (FCM-based ANFIS) | 2.5 | 9.18 | 3.5 | 0.99 | 0.89 | 0.2 | 3.5 |
| Model 2 (SC-based ANFIS) | 4.3 | 16.57 | 8.9 | 0.95 | 0.58 | −0.6 | 8.8 |

value of model 2 is small and many data points are away from the vicinity of the identity line. The CC value and the scatter plot reveal that model 1 outperforms model 2.

In order to get more insights about the performance of the models, the regression error characteristic (REC) is used for further analysis. The REC curve provides an approach to visualize and evaluate different regression models (Fawcett 2003) by plotting error measures such as absolute deviation or squared residual versus the percentage of points predicted within the tolerance, and it provides an estimation of the cumulative distribution function of the error. The area-over-curve (AOC) provides a biased estimation of the expected error for a prediction. The details can be found at Bi and Bennett (2003), De Pina and Zaverucha (2006), and Torgo (2005). The REC curve of model 1 shows that the absolute deviation of around 85 % of the evaluation data is less than or equal to 5 ppbv (Fig. 6a). On the other hand, in the case of model 2, the absolute deviation of around 90 % of the evaluation data is less than or equal to 10 ppbv (Fig. 6b). The values of AOC are 2 and 4 ppbv for model 1 and model 2,

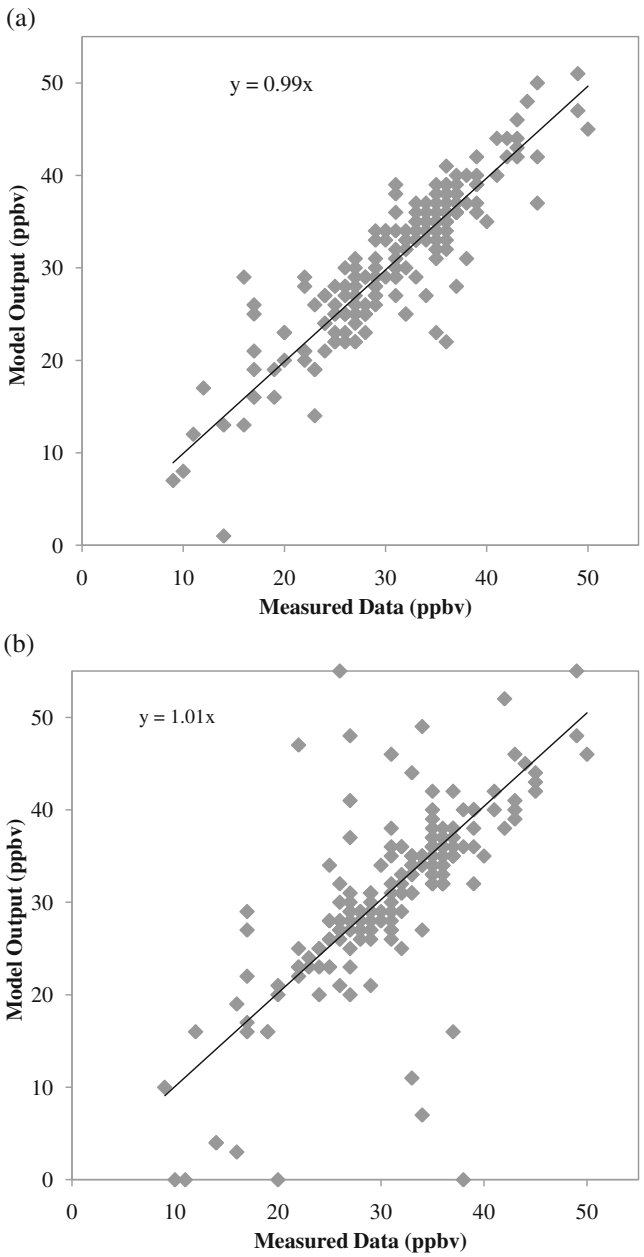


**Fig. 5** **a** The scatter plot of the measured data and the output of model 1. **b** The scatter plot of the measured data and the output of model 2

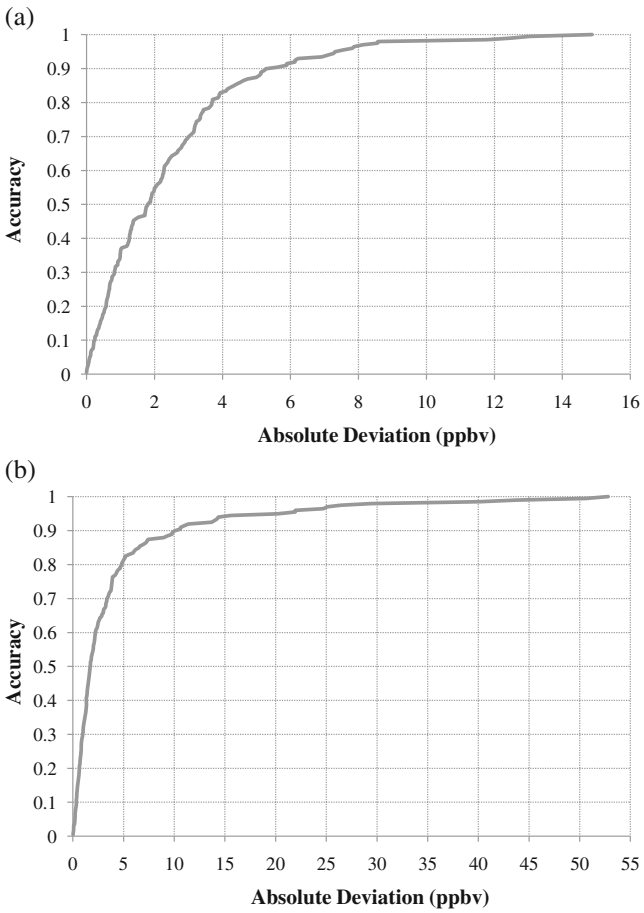


**Fig. 6** **a** The regression error characteristic curve for model 1. **b** The regression error characteristic curve for model 2

respectively. Based on the analysis of REC curves and AOC values, it seems that the performance of model 1 is superior compared to model 2.

Generally, both the proposed models performed adequately in modeling and predicting the ozone levels in the Empty Quarter. Based on the reported performance measures and prediction analysis, it can be summarized that the proposed modeling approach can be considered as a viable approach for ozone modeling in arid region as a function approximation with the help of weather data and the concentrations of NO and $NO_2$.

The developed models perform adequately considering the NOx precursors along with meteorological data and without considering volatile organic compounds (VOCs) as input, which probably indicate that the ozone level in the study area is more sensitive to NOx compared to VOCs. In this study, the solar radiation data were not available to use as input. In order to address this limitation, the time is expressed as the cosine value of the time of day (normalized), which exhibits cyclic characteristics. The transformed time input is considered as a surrogate of solar radiation.

## Conclusions

In order to provide flexible, adaptive, and less assumption-dependent models, this study proposed an ANFIS-based modeling approach, which exploits the capability of both ANN and fuzzy logic models. The developed models are used to predict ozone concentrations based on meteorological data, NO and $NO_2$ concentrations, and their statistical transformations. As a fuzzy logic model with a large number of rules can reduce generalization capability, this study determined the initial fuzzy inference system with the help of FCM and SC. The adopted clustering algorithm-based approach reduces the size of the rule base significantly, i.e., contributed to increase the generalization capability of the model. The developed models perform adequately considering only the NOx precursors (without VOCs) along with meteorological data, which probably indicate that the ozone level in the study area is more sensitive to NOx compared to VOCs. In order to provide better insights about the performance of the developed models, regression error characteristic curves were used along with traditional performance measures. It is observed that the FCM-based ANFIS model outperformed the SC-based ANFIS model depending on the considered performance measures. The obtained results and the performance analysis of the models indicate the applicability of the adopted neuro-fuzzy approach in short-term modeling of ozone levels in the Rub Al Khali Desert, specifically during winter season. The future research should focus on the use of the adopted approach for other sites of the area for different seasons and compare them with traditional machine learning models.

**Acknowledgments** The authors would like to gratefully acknowledge the support of King Fahd University of Petroleum & Minerals in conducting this research.

## References

Abdul-Wahab SA, Al-Alawi SM (2002) Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks. Environ Model Softw 17:219–228

Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Plenum, New York

Bi J, Bennett KP (2003) Regression error characteristic curve. Proc. twentieth int. conf. machine learning (ICML'03).

Borrego C, Tchepel O, Costa AM, Amorim JH, Miranda AI (2003) Emission and dispersion modeling of Lisbon air quality at local scale. Atmos Environ 37:5197–5205

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

Caruana R (1997) Multitask learning. Mach Learn 28(1):41–75

Chiu SL (1994) Fuzzy model identification based on cluster estimation. J Intell Fuzzy Syst 2:267–278

Clark A (1989) Amdt R (ed.) Lakes of the Rub' al-Khali. Saudi Aramco World 40(3): 28–33.

De Pina AC, 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, Bruges, Belgium, April 26–28, pp. 425–430.

Dunea D, Oprea M, 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, February 11–13, Innsbruck, Austria, pp. 93–98.

Dunn JC (1973) A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J Cyber. 3–32.

Fakhreddine K, De Silva C (2004) Soft computing and tools of intelligent systems design: theory, tools and applications. Pearson, Harlow

Fawcett T (2003) ROC graphs: notes and practical considerations for data mining. Technical report, HP-2003-4.

Hassoun MH (1995) Fundamentals of artificial neural networks. MIT, Cambridge

Haykin S (1994) Neural networks: A comprehensive foundation. Prentice-Hall, Upper Saddle River

Heo JS, Kim DS (2004) A new method of ozone forecasting using fuzzy expert and neural network systems. Sci Total Environ 325:221–237

Jain S, Khare M (2010) Adaptive neuro-fuzzy modeling for prediction of ambient CO concentration at urban intersections and roadways. Air Qual Atmos Health 3:203–212

Jang JSR (1993) ANFIS adaptive-network-based fuzzy inference systems. IEEE Trans Syst Man Cyber 23(3):665–685

Jang JSR, Sun CT, Mizutani E (1996) Neuro-fuzzy and soft computing: A computational approach to learning and machine intelligence. Prentice-Hall, Upper Saddle River

Jantzen J (1998) Neurofuzzy modelling. Technical University of Denmark, Department of Automation, Tech. report no 98-H-874.

Johanyák ZC, Kovács J (2011) Fuzzy model based prediction of ground-level ozone concentration. Acta Technica Jaurinensis 4 (1):113–126

Jorquera H, Perez R, Cipriano A, Espejo A, Letelier MV, Acuna G (1998) Forecasting ozone daily maximum levels at Santiago. Chile Atmos Environ 32:3415–3424

Kao JJ, Huang SS (2000) Forecasts using neural network versus Box-Jenkins methodology for ambient air quality monitoring data. J Air Waste Manag Assoc 50(2):219–26

Karatzas KD, Kaltsatos S (2007) Air pollution modeling with the aid of computational intelligence methods in Thessaloniki, Greece. Simul Model Pract Theory 15:1310–1319

Musavi MT, Chan KH, Kalantri K (1994) On the generalization ability of neural network classifiers. IEEE Trans Pattern Anal Mach Intell 16(6):659–663

Nebot A, Mugica V, Escobet A (2008) Ozone prediction based on meteorological variables: a fuzzy inductive reasoning approach. Atmos Chem Phys Discuss 8:12343–12370

Park D, Rilett LR, Han G (1999) Spectral basis neural networks for real-time travel time forecasting. J Transp Eng 125(6):515–523

Pires CM, Gonçalves B, Azevedo FG, Carneiro AP, Rego N, Assembleia AJB, Lima JFB, Silva PA, Alves C, Martins FG (2012) Optimization of artificial neural network models through genetic algorithms for surface ozone concentration forecasting. Environ Sci Pollut Res. doi:10.1007/s11356-012-0829-9

Prakash A, Kumar U, Kumar K, Jain VK (2011) A wavelet-based neural network model to predict ambient air pollutants' concentration. Environ Model Assess 16:503–517

Sousa SIV, Martins FG, Alvim-Ferraz MCM, Pereira MC (2007) Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. Environ Model Softw 22(1):97–103

Takagi T, Sugeno M (1985) Fuzzy identification of systems and its applications to modeling and control. IEEE Trans Syst Man Cyber 15:116

Torgo L (2005) Regression error characteristic surfaces. Proc. eleventh ACM SIGKDD int. conf. knowledge discovery in data mining, pp. 697–702.

Tsao ECK, Bezdek JC, Pal NR (1994) Fuzzy Kohonen clustering networks. Pat Recog 27(5):757–764

Vincent P (2008) Saudi Arabia: An environmental overview. Taylor & Francis, London, p 141

Willmott CJ (1981) On the validation of models. Phys Geogr 2:184–194

Yen J, Wang L (1999) Simplifying fuzzy rule-based models using orthogonal transformation methods. IEEE Trans Syst Man Cybern B Cybern 29(1):13–24

Yildirim Y, Bayramoglu M (2006) Adaptive neuro-fuzzy based modeling for prediction of air pollution daily levels in City of Zonguldak. Chemosphere 63(9):1575–1582

Ying LC, Pan MC (2008) Using adaptive network based fuzzy inference system to forecast regional electricity loads. Energ Convers Manag 49:205

Zhang Q, Benveniste A (1992) Wavelet networks. IEEE Trans Neural Networks 3:889–898