



PM_{2.5} concentration forecasting using ANFIS, EEMD-GRNN, MLP, and MLR models: a case study of Tehran, Iran

Jamil Amanollahi¹ · Shadi Ausati¹

Received: 31 October 2019 / Accepted: 4 December 2019 / Published online: 18 December 2019
© Springer Nature B.V. 2019

Abstract

Tehran, the capital city of Iran, is among the world's most polluted cities. Tehran is exposed to different types of pollutants, one of which is the suspended particles of PM_{2.5}. One of the steps that should be taken to reduce hazardous effects of this pollution on the health of society is timely prediction and announcement of its increased levels. Different methods can be used for predicting PM_{2.5} concentration. This study used a variety of models for predicting PM_{2.5} concentrations, including linear, nonlinear, and hybrid models. More specifically, the models which were used consisted of multiple linear regression, multi-layer perceptron (nonlinear model), and a combination of ensemble empirical mode decomposition and general regression neural network (EEMD-GRNN) and Adaptive Neuro-Fuzzy Inference System (ANFIS) (hybrid of nonlinear models). The independent variables in the current study were air quality parameters, which were measured in reference to PM_{2.5}, PM₁₀, SO₂, NO₂, CO, and O₃ and meteorological data which included average atmospheric pressure (AP), average maximum temperature (Max T), average minimum temperature (Min T), daily relative humidity level of the air (RH), daily total precipitation (TP), and daily wind speed (WS) in 2016 in Tehran. The results indicated that the ANFIS model exhibited the most accurate prediction in the training phase ($R^2 = 0.99$, RMSE (root mean square error) = 0.4794 and MAE (mean absolute error) = 0.1305) and in the testing phase ($R^2 = 0.82$, RMSE = 3.2979 and MAE = 2.1668). As it can be concluded, in comparison with a linear model, hybrid models are of higher precision in predicting PM_{2.5} concentration.

Keywords Air quality · Meteorological data · Validation · Hybrid models · Linear model

Introduction

Background

A remarkable number of cardiovascular and respiratory diseases can be caused by air pollution (Pope et al., 2004). As one of the major environmental problems these days, suspended particulate matter (SPM) in high concentrations can cause climate change (Haywood and Boucher 2000) and growth

stunting or mortality of plant species (Bench 2004). SPM has negative effects on housing market (Kim and Yoon 2019), tropopause height (Wu et al. 2013), surface temperature and energy budget (De Menezes Neto et al. 2017; Tzanis and Varotsos 2008), and childcare facility (Oh et al. 2019). Due to its small size, SPM can penetrate the lower and upper parts of our respiratory system (Liu et al. 2019), and thus, it can harm human health (Sahu et al. 2019; Wang et al. 2016; Yadav et al. 2019). Exposure to high levels of PM_{2.5} causes 3.15 million premature deaths worldwide every year, and overall outdoor air pollution causes 3.3 million mortality annually (Lelieveld et al. 2015). Although in the developing countries, most cities have similar air pollution problems, each city has different sources of air pollution and its own particular geographical and climatic features.

Motivation

Tehran is located in a developing country where rapid urbanization and population growth have resulted in its continuously expanding residential area, considerable changes in its land

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11869-019-00779-5>) contains supplementary material, which is available to authorized users.

✉ Jamil Amanollahi
J.amanollahi@uok.ac.ir

Shadi Ausati
shadiaosati@gmail.com

¹ Department of Environmental Sciences, Faculty of Natural Resources, University of Kurdistan, Sanandaj, Iran

cover, and land use (Alizadeh-Choobari et al. 2016). Reportedly in 2012, air pollution caused premature deaths of a considerable number of people ($N = 4500$) in Tehran (Ministry of Health and Medical Education, 2012). There is empirical evidence that indicates Tehran is one of the cities in the world in which high mortality is caused by long-term exposure to fine particular matter (Lelieveld et al. 2015).

Literature

PM_{2.5} concentrations can be predicted using various forecasting models. Artificial intelligence (Ventura et al. 2019), chemical transport (Sun et al. 2013), linear regression, (Vlachogianni et al. 2011), nonlinear regression (Baker and Foley 2011), and time series (Wang et al. 2012) are some of the commonly used types of forecasting models. Additionally, by combining some of these models, researchers have been able to provide more accurate prediction results (Ausati and Amanollahi 2016; Zhou et al. 2014). One such example is the Adaptive Neuro-Fuzzy Inference System (ANFIS), which has a hybrid algorithm and was proposed by Jang et al. (1997). Research evidence indicates that as a powerful method ANFIS can be used for modeling dust storm occurrences (Kaboodvandpour et al. 2015), air quality forecasting (Ghasemi and Amanollahi 2019), predicting ambient CO concentration (Jian et al. 2010), and predicting PM_{2.5} based on GTWR model and remotely sensed data (Mirzaei and Amanollahi 2019). Ghasemi and Amanollahi (2019) showed that integrated forward selected method and ANFIS model increased the accuracy of air quality forecasting. Mirzaei and Amanollahi (2019) compared the artificial neural network (ANN), linear regression, general regression neural network (GRNN), and ANFIS models to improve the correlation coefficient between output (PM_{2.5}) of GTWR model and ground measurement PM_{2.5} concentration. They concluded that ANFIS model had a better performance than other models. In a study aimed at predicting PM_{2.5} concentration 1-day-ahead, a hybrid ensemble empirical mode decomposition (EEMD) and GRNN were utilized by Ausati and Amanollahi (2016). They compared the prediction accuracy of the results obtained by a principal component regression (PCR) model, an ANFIS, and a hybrid EEMD-GRNN model with a multiple linear regression (MLR) by using the values of mean absolute error (MAE) and root mean square error (RMSE) obtained from each model. Their results indicated that the hybrid EEMD-GRNN model exhibited the highest accuracy in predicting PM_{2.5} in Sanandaj, Iran. Using EEMD-GRNN model, Zhou et al. (2014) predicted the 1-day-ahead PM_{2.5} pollution in Xian, China. Zhu et al. (2018) proposed EEMD and endpoint condition mirror method to predict the time series of air quality index in Hefei, the hybrid forecasting model. ANN and hybrid models, such as EEMD-GRNN and ANFIS, appear to be capable of predicting PM_{2.5}

more accurately. Therefore, the objective of the current study was the comparison of PM_{2.5} prediction accuracy of a linear model, such as MLR, and nonlinear models, such as EEMD-GRNN, ANFIS, and ANN in Tehran.

Material and methods

Study area

The location of this study is a metropolitan area, called Tehran. This city is surrounded by the high Alborz Mountain range to its north and east, and to the south, it meets the *Kavir* Desert (Fig. S-1). The wind directions in Tehran are greatly affected by these topographical features; during the day, prevailing southwesterly winds blow from the desert toward the mountains while during the night, heading from the mountains toward the plains are the prevailing northwesterly-westerly winds which dominate especially the western half of Tehran. The Department of Environment in Tehran provided the data including PM_{2.5}, PM₁₀, SO₂, NO₂, CO, and O₃. The Bureau of Meteorology of Tehran provided the following meteorological data related to the year 2016: the average atmospheric pressure (AP), average maximum temperature (Max T), average minimum temperature (Min T), daily relative humidity level of the air (RH), daily total precipitation (TP), and daily wind speed (WS). These data were classified into two separate datasets which included 335 datasets for simulation purposes and 30 datasets for the purpose of testing the models.

Multiple linear regression model

Statistical Package for Social Sciences (SPSS, version 16) was used for analyzing the data. Multiple linear regression was used for determining the significance of correlations between independent variables and a dependent variable. The MLR model is presented as follows (Eq. 1):

$$Y = B_0 = B_1X_1 = B_2X_2 + \dots + B_kX_k + \varepsilon \quad (1)$$

In this equation, the dependent variable is signified by Y; the independent variables are signified by X₁, X₂, ...X_k and the error term by ε . An important assumption of multiple linear regression is that the independent variables should have linear relationships. In MLR, the method which is used for testing linearity is called the variance inflation factors (VIF) (Table 1). VIF values greater than 10 indicate that the assumption of linearity is met.

According to Table 1, some of the variables (AP, Min T, ...) had VIF values more than 5, indicating the existence of multicollinearity among these variables. Stepwise regression was used in order to overcome this problem, by determining the most effective set of independent variables that would

Table 1 VIF values for the independent variables

Model	Unstandardized coefficients		Standardized coefficients Beta	<i>t</i>	Sig.	Collinearity Statistics	
	B	Std. error				Tolerance	VIF
1 (Constant)	-367.2	222.2		-1.652	0.100		
SO ₂	0.145	0.101	0.070	1.443	0.150	0.868	1.1520
NO ₂	-0.057	0.047	-0.074	-1.205	0.229	0.545	1.8360
CO	-0.394	0.646	-0.035	-0.610	0.542	0.635	1.5760
PM ₁₀	0.115	0.025	0.241	4.691	0.000	0.786	1.2720
PM _{2.5 P}	0.337	0.048	0.352	6.968	0.000	0.812	1.2310
O ₃	-0.154	0.075	-0.125	-2.057	0.041	0.564	1.7740
Min T	0.156	0.406	0.090	0.383	0.702	0.038	26.503
Max T	0.116	0.380	0.080	0.305	0.760	0.030	33.085
AP	0.111	0.039	0.204	4.212	0.021	0.435	1.5120
RH	0.203	0.086	0.234	2.345	0.020	0.208	4.7990
PR	-0.269	0.413	-0.037	-0.652	0.515	0.651	1.5360
WS	-0.996	0.399	-0.118	-2.498	0.013	0.921	1.0860

predict the dependent variable. Table 2 shows the results of stepwise regression analysis.

Based on the results of Table 2, after stepwise method was run, none of the VIF values was larger than 10 for any of the independent variables.

Adaptive Neuro-Fuzzy Inference System model

ANFIS model has a hybrid algorithm. Its learning algorithm was initially created by Jang et al. (1997) who applied the least squares method and gradient descent. Based on a feed forward network, ANFIS is capable of optimizing parameters of a fuzzy system in order to achieve accurate results. ANFIS model consists of two components which are called primary and inference parts. These two are connected with fuzzy rules by a network. The fuzzy inference system (FIS) of this structure, which develops

in an adaptable network, is composed of directly connected nodes (Matlab 2018). The output of ANFIS is dependent on its input parameters. The input data of ANFIS are normalized for minimizing the error rate by the learning algorithm. FIS framework, on the other hand, has three major parts, which include (i) a fuzzy rule (if-then), (ii) a database (its membership functions defined according to the fuzzy rule), and arguments mechanism that follows the IF and THEN theory (Matlab 2018). For example, if X and Y are the two inputs of an FIS framework and if Z is its output which follows a fuzzy if-then rule, then:

- Rule 1. if X is A₁ and Y is B₁ then $f_1 = p_1x + q_1y + r_1$.
- Rule 2. if X is A₂ and Y is B₂ then $f_1 = p_2x + q_2y + r_2$.

In these rules, $f(x, y)$ is a polynomial, and the name of the created model is Sugeno Fuzzy (Guneri et al. 2011). ANFIS

Table 2 VIF values after stepwise regression

Model	Unstandardized coefficients		Standardized coefficients Beta	<i>t</i>	Sig.	Collinearity statistics	
	B	Std. error				Tolerance	VIF
(Constant)	8.263	2.737		3.020	0.003		
PM _{2.5 P}	0.361	0.047	0.376	7.641	0.000	0.862	1.159
RH	0.203	0.043	0.234	4.766	0.000	0.868	1.153
PM ₁₀	0.110	0.023	0.231	4.873	0.000	0.930	1.075
WS	-1.210	0.386	-0.144	-3.136	0.002	0.993	1.007
O ₃	-0.127	0.059	-0.103	-2.141	0.033	0.910	1.099
AP	0.610	0.066	0.286	5.562	0.000	0.811	1.331

model has a five-layer network (Wei et al. 2007). Its first layer is connected to a fuzzy model (Fig. S-2) which follows Eq. 2:

$$O_i^1 = \mu_{A_i}(x) \tag{2}$$

in which i and A_i constitute the linguistic variables, x indicates the input node, and O_i^1 stands for the membership function of A_i . The function of the second layer of the model is the implementation of “AND” (Fig. S-2). The second layer consists of ring layers which are multiplied by the input layers while the output is obtained by Eq. 3:

$$w_i = \mu_{A_i}(x) \times \mu_{B_i}(y), i = 1, 2 \tag{3}$$

Normalization is the function of the third layer (Fig. S-2), in which the mean score of the i^{st} created rule is calculated for each node using Eq. 4:

$$\bar{w}_i = \frac{w_i}{w_1 + w_2}, i = 1, 2 \tag{4}$$

In the fourth layer, the fuzzy rules are used (Fig. S-2) in which every node of i is a square node consisting of a membership function (Eq. 5):

$$O_i^4 = \bar{w}_i f_i = \bar{w}_i(p_i x + q_i y + r_i) \tag{5}$$

where \bar{w}_i shows the third layer’s output, while p_i , q_i , and r_i indicate the final parameters. The fifth layer involves the defuzzification process whereby all input signals are added to compute a single node of total output (Fig. S-2). Equation 6 is employed in this process for transforming the output of every fuzzy rule to the defuzzification output (Guner et al. 2011):

$$O_i^5 = \sum_i \bar{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \tag{6}$$

Empirical mode decomposition and general regression neural network model

EEMD-GRNN is composed of two models, called EEMD and GRNN. First, EEMD is employed for decomposing the original time series into a certain set of IMFs. The residual r_n is assumed to be IMF. The next step involves using the GRNN model to predict each decomposed set of the IMFs, which was defined in Step 1. The value of the corresponding IMF series is forecast for the next day by using the GRNN model. As a final step, in order to obtain the final forecast, the output of the previous step is aggregated.

Ensemble empirical mode decomposition

As an adaptive method used for analyzing non-stationary and nonlinear signals, empirical mode decomposition (EMD) was initially proposed by Huang et al. (1998). EMD can be applied for decomposing a signal into several IMFs. The signal must meet two conditions before it can turn into an IMF mode: (i) the mean scores of the lower and upper envelopes should be ubiquitously zero, and (ii) the number of zero crossings and the number of extreme cases should be equal or not greater than one. A major drawback of EMD is the presence of almost identical oscillations in diverse modes or presence of oscillations of very dissimilar amplitudes in a mode, also known as “mode mixing” (Huang et al. 1998). Ensemble empirical mode decomposition (EEMD) was a possible solution offered by Wu and Huang (2009). As an updated version of the EMD, EEMD has a noise-assisted system. The mode mixing problem can be solved with the support of this white noise (Wu and Huang 2009).

In order to determine the EEMD algorithm of a signal $x(t)$, the following steps are taken. First, the amplitude of the added white noise and the ensemble number M are initialized. The m th trial is the second step that is conducted to produce the noise-added data $x_m(t)$ by adding random white noise $w_m(t)$ into $x(t)$ (Eq. 7).

$$x_m(t) = x(t) + w_m(t) \tag{7}$$

The objective of the third step is to identify all the local minima and maxima of $x_m(t)$ and using the cubic spline functions to obtain the lower and upper envelopes. The fourth step involves the computation of the mean $m_I(t)$ of the lower and upper envelopes and the calculation of the difference $h_I t$ between the mean and the signal (Eq. 8).

$$h_I t = x_m(t) - m_I(t) \tag{8}$$

In step five, Eq. 9 is used to define $r_I(t)$ providing that h_I meets the conditions of IMF, and that $h_I(t)$ constitutes the first IMF component from the signal ($h_I(t) = c_I(t)$):

$$r_I(t) = x_m(t) - c_I(t) \tag{9}$$

Steps 3 to 5 should be repeated if these conditions are not met.

To identify the residue $r_I(t)$ as a new signal and to sift out other IMFs until the stopping criteria are satisfied, steps 3 to 5 are repeated n times. If the residue $r_n(t)$ or the IMF component ($c_n(t)$) is smaller than the predetermined value, the stopping criterion has occurred. The original signal can be shown as the total of all IMFs plus the residue after sifting (Eq. 10):

$$x_m(t) = \sum_{i=1}^n c_i(t) + r_n(t) \tag{10}$$

in which n stands for the number of IMFs, $c_i(t)$ for the i th IMF and $r_n(t)$ for the final residue.

Next, $m = m + 1$ is set if $m < M$ and steps 2 to 5 above are repeated until $m = M$, but each time indicating a different white noise. As a final point, for every IMF, we calculate the ensemble mean c_i of the M trials (Lu and Shao 2012).

General regression neural network

By analyzing its past input and output data, GRNN can estimate any function. GRNN is the fastest in training and modeling nonlinear functions in comparison with all the other models. An additional distinctive feature of GRNN is its smoothing factor that enables this model to estimate the optimum value in the process of numerous performances in relation to the mean square error (Leung et al. 2000). GRNN consists of four layers. The first layer is the input layer, in which the data are keyed into the model. In a GRNN, the quantity of input neurons equals that of the variables in the input vector. The next layer is referred to as the pattern layer, the neurons of which can memorize the correlation between the proper response of the pattern layer and the input neurons. The quantity of the neurons in this layer equals that of the training cases. The following equation helps determine the Gaussian function of the pattern P_i :

$$P_i = \exp\left(\frac{(X-X_i)^T(X-X_i)}{2\sigma^2}\right) \quad (11)$$

In the equation above $i = 1, 2, \dots, n$, σ signifies the smoothing parameter (also known as a spread parameter); X shows the independent variable; and X_i signifies a training sample for the i^{th} neuron of the second layer.

The third layer is known as the summation layer to which the output of the second layer is imported. In this layer, two total values, referred to as S_s and S_w , are computed. The summation of the pattern outputs is calculated with the help of Eq. 12 below:

$$S_s = \sum_{i=1}^n P_i \quad (12)$$

Equation 13 is used for determining the weighted sum of the pattern outputs:

$$S_w = \sum_{i=1}^n W_i P_i \quad (13)$$

In this equation, W_i stands for the weight of the i^{th} neuron in the pattern layer which is linked to the third layer. The final layer is the output layer to which the results of the third layer

are sent. Equation 14 is used for determining the output that is signified by y :

$$y = S_w / S_s \quad (14)$$

Artificial neural network model

Artificial neural network (ANN) was initially proposed by McCulloch and Pitts (1943), who were inspired by neural network systems and the brain of living organisms. ANN is known as a simulating method. It is commonly employed for predicting the various methods that could replace linear regression, multivariate regression, and trigonometric functions among other statistical methods (Guneri et al. 2011). Detailed descriptions of ANN are available in the literature (for example, Nørgaard et al. 2000). Among other types of ANN, the most commonly used type is the multi-layer perceptron (MLP). MLP is composed of three distinct layers: (i) its input layer in which the data are distributed over the network; (ii) its hidden layer in which the data are processed; and finally (iii) the output layer where the results for certain inputs are extracted (Fig. S-3). Sometimes, there can be more than one hidden layers and a main parameter of the network may be set by the number of its units. In this study, Matlab, 2018 was used for running ANFIS, EEMD-GRNN, and ANN models.

Evaluation of models

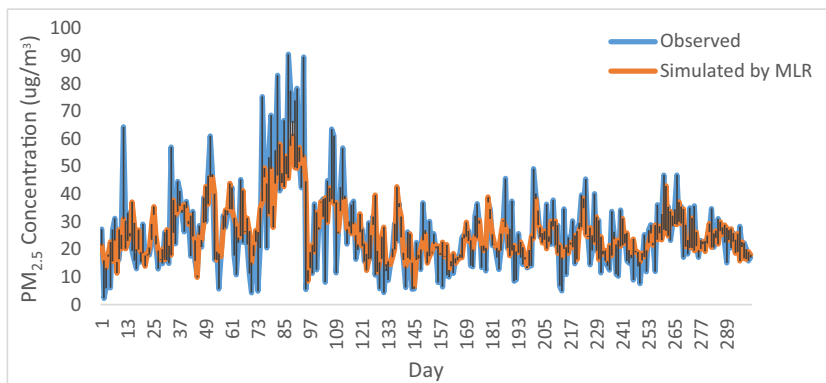
R^2 is a statistical parameter commonly used to determine the validity of the model's output. The value of R^2 which ranges between 0 and 1 is used as an index for determining the precision of the regression line. The closer the value of R^2 is to one, the better compliance is estimated for the predicted and observed data. Nevertheless, according to Legates and McCabe (1999), R^2 should be used cautiously as its value can be affected by the Perth data. Therefore, it should be employed alongside other parameters such as mean absolute error and root mean square error (RMSE) in order to determine the validity of the output (Noori et al. 2010). In order to calculate the root mean square error and mean absolute error, Eqs. 15 and 16 are used (Alimissis et al. 2018; Ding et al. 2016; Tzani et al. 2019; Willmott and Matsuura 2005), respectively.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_{dp} - Y_{da})^2} \quad (15)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |Y_{dp} - Y_{da}| \quad (16)$$

where Y_{da} denotes the real value, Y_{dp} signifies the predicted value, and n shows the sample size.

Fig. 1 MLR results for simulated and observed $PM_{2.5}$ concentrations ($\mu\text{g}/\text{m}^3$)



Results and discussion

Prediction of $PM_{2.5}$ concentrations in Tehran via multiple linear regression model training and testing

In the current study, MLR was the first model which was used for predicting the $PM_{2.5}$ concentration. MLR is used for determining the variables which have statistically significant effects on the dependent variable. Many previous studies have frequently applied this method (Golchoubian et al. 2012; Pouretedal et al. 2018). Among commonly used methods in the MLR model is the stepwise method. The predicting variables that remained in the model in the current study were O_3 , $PM_{2.5}$ on the previous day, PM_{10} , RH, and WS. This means that the concentration of $PM_{2.5}$ of Tehran is affected by the concentrations of these variables. According to MLR results (Eq. 17), the variables O_3 , $PM_{2.5}$ on the previous day, PM_{10} , RH, AP, and WS had positive associations with $PM_{2.5}$ concentrations in Tehran. Figure 1 illustrates the training results of multiple linear regression model for prediction of $PM_{2.5}$ suspended particle concentration in Tehran and what follows is the obtained equation (Eq. 17):

$$PM_{2.5F} = 8.263 + 0.203RH + 0.11PM_{10} + 0.361PM_{2.5P} + 0.127O_3 + 1.21WS + 0.61AP \tag{17}$$

In this equation, $PM_{2.5F}$ is the predicted concentration of $PM_{2.5}$ while $PM_{2.5P}$ is the $PM_{2.5}$ concentration on the previous day. The equation that follows is formulated according to the results of this section: $R^2 = 0.38$, RMSE = 11.8095, and MAE = 8.8234 (Fig. 1).

The MLR testing phase results ($R^2 = 0.44$ RMSE = 7.6402 and MAE = 5.9961) are presented in Fig. 2

Numerous assumptions have to be met before applying MLR. Considering all these assumptions makes running MLR a challenge. Therefore, this method seems to be less efficient and less practical than nonlinear models

Prediction of $PM_{2.5}$ concentrations in Tehran via artificial neural network (multiple linear regression) model training and testing

Based on the results, as compared with the other methods, MLP model was ranked second for its accuracy for estimation

Fig. 2 MLR results for predicted and observed $PM_{2.5}$ concentrations ($\mu\text{g}/\text{m}^3$)

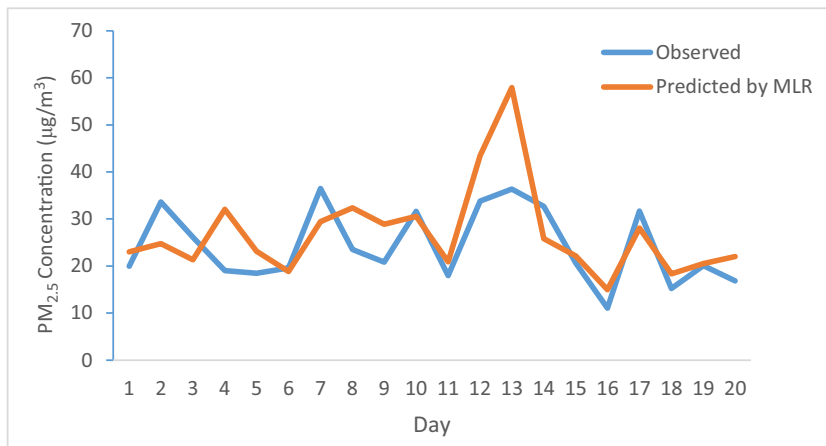
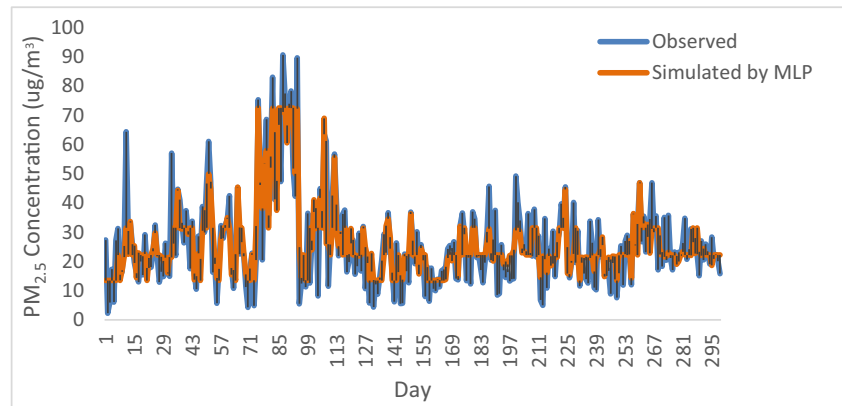


Fig. 3 MLP results for simulated and observed $PM_{2.5}$ concentrations

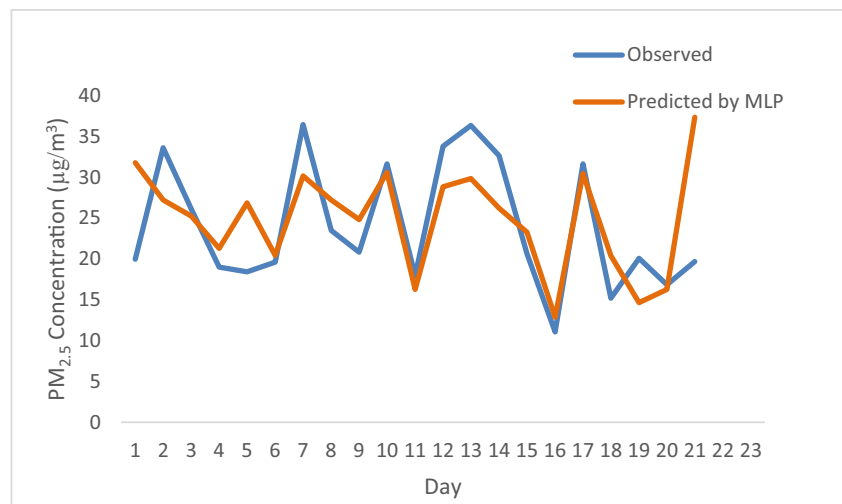


of the $PM_{2.5}$ concentrations in Tehran. As a nonlinear model, the advantage of MLP is its high tolerance for the small number of errors in the related data. Several studies (for example, Mashaly and Alazba 2016; Messikh et al. 2017; Thorkashvand et al. 2017) provide evidence for the superiority of MLP model over other models. The training results of the MLP model to forecast $PM_{2.5}$ concentrations in Tehran are shown in Fig. 3. In comparison with MLR results, the simulated accuracy of MLP turned out to be higher in terms of R^2 , RMSE, and MAE between the simulated and observed data at 0.67, 7.8849, and 6.4209, respectively.

The results of the MLP model, which was used to predict $PM_{2.5}$ concentrations in Tehran, are illustrated in Fig. 4.

In comparison with the MLR results, the consistency was higher for the predicted and observed data of MLP model (Fig. 4). Moreover, as compared with the MLR data, RMSE and MAE in MLP model were lower (RMSE = 6.2522 and MAE = 4.4781) while R^2 in MLP model ($R^2 = 0.51$) was higher than those of MLR.

Fig. 4 MLP results for predicted and observed $PM_{2.5}$ concentrations



EEMD results—predicting $PM_{2.5}$ suspended particle concentrations for Tehran using general regression neural network model training and testing

The first step in the EEMD-GRNN is breaking down the original signal ($PM_{2.5}$), and then, using the GRNN model to predict each component. In EEMD model, ranges of white noise and the number of tests are respectively 0.2 and 100 for analyzing $PM_{2.5}$ signal (Lu and Shao 2012). $PM_{2.5}$ signal consists of seven intrinsic mode functions and a residual. The residual is regarded as an intrinsic mode function (Fig. 5).

As ANFIS model inputs, GRNN model inputs are MLR model outputs. That is to say, for the purpose of training, the independent variables are used by the model. These variables influence the $PM_{2.5}$ concentrations, as the dependent variable. For estimating the output with varying smoothing factors, network training was conducted. The factor value always ranges from 0 to 1 (Gheyas and Smith 2009). Gheyas and Smith (2009) showed that

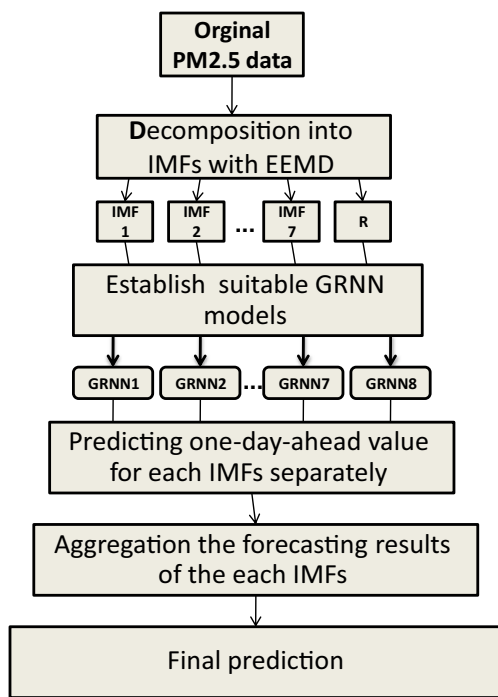
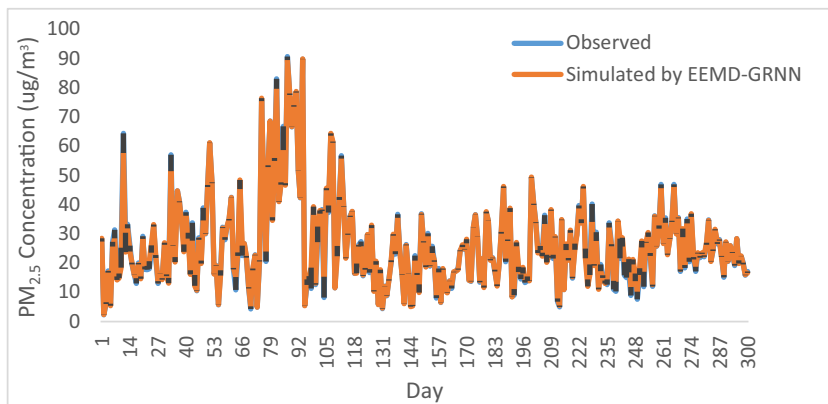


Fig. 5 EEMD-GRNN structure (Ausati and Amanollahi 2016)

GRNN model has superiority over MLP in forecasting univariate time series and suggested that GRNN performance is not very sensitive to smoothing factor. As the smoothing factor increases, the correlation coefficient value gradually decreases in data training, and yet this value slowly rises for test data. Figure 6 indicates the training results of the EEMD-GRNN model that was used to predict concentrations of PM_{2.5} suspended particles in Tehran. Here is a summary of the results presented in this section: $R^2 = 0.98$, $RMSE = 1.8622$, and $MAE = 1.1639$.

Figure 7 shows the results of suspended particle concentrations of PM_{2.5} tests in Tehran using the EEMD-GRNN model. A summary of the results reported in this section may be presented thus: $R^2 = 0.76$, $RMSE = 4.2655$ and $MAE = 3.7427$.

Fig. 6 Using EEMD-GRNN for predicting simulated and observed PM_{2.5} concentrations ($\mu\text{g}/\text{m}^3$)



Predicting PM_{2.5} suspended particle concentrations for Tehran using Adaptive Neuro-Fuzzy Inference System model training and testing

As compared with MLR and MLP, ANFIS model performed better. The model, which functions based on Takagy-Sugeno, consists of five input and one output parameters. ANFIS follows five phase rules. Each of these rules will be influenced by input parameters; that is, any changes in the input value will result in a change in the respective output value. One of the major limitations of the models which do not follow Fuzzy logic-based methods is that they are sensitive to errors in the data. MLR model output comprises the inputs of ANFIS model. That is to say, in order to train, ANFIS uses the predicting (independent) variables that affect the concentrations of predicted (dependent) variable (PM_{2.5}). The testing phase results of fuzzy inference system according to neural network are illustrated in Fig. 8. The results reported in this section may be summarized thus: $R^2 = 0.99$, $RMSE = 0.4794$, and $MAE = 0.1305$.

The testing phase results of ANFIS model are presented in Fig. 9. These results could be summarized thus: $R^2 = 0.82$, $RMSE = 3.2979$, and $MAE = 2.1668$.

Based on the results of ANFIS (Figs. 8 and 9) that was used to predict the suspended particle concentrations of PM_{2.5} in Tehran, it was found that the value was higher than the values of MLR and MLP models. These results are in agreement with the findings reported by Shahbazi et al. (2013), Amirkhani et al. (2015), Kaboodvandpour et al. (2015), and Zendejboudi et al. (2017). As the results indicated, however, the suspended particle concentrations of PM_{2.5} in Tehran as predicted by ANFIS were close to those of EEMD-GRNN model. As it appears and Table 3 shows, the linear model could not deliver a model to take the fluctuations of time series of the PM_{2.5} concentrations into account since these fluctuations were high. It seems that nonlinear models such as ANFIS and EEMD-GRNN can be considered very practical replacements for linear models since they are able to test the nonlinear associations between the inputs and outputs. As the

Fig. 7 Using EEMD-GRNN to predicted and observed concentrations of PM_{2.5} (µg/m³)

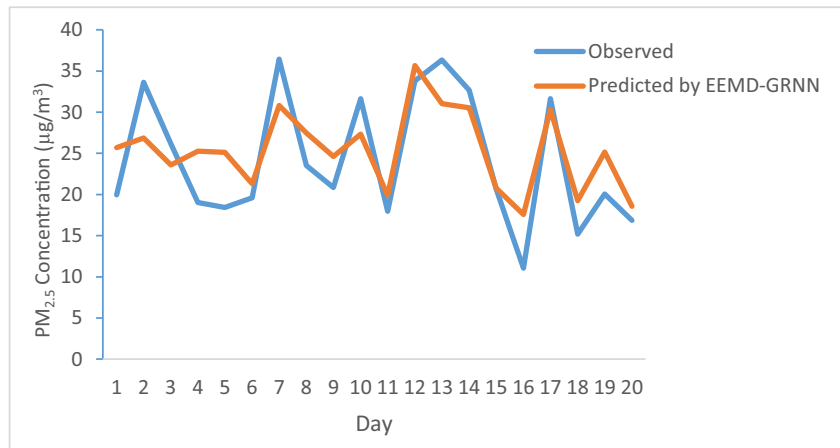


Fig. 8 Using ANFIS to predict simulated and observed PM_{2.5} concentrations (µg/m³)

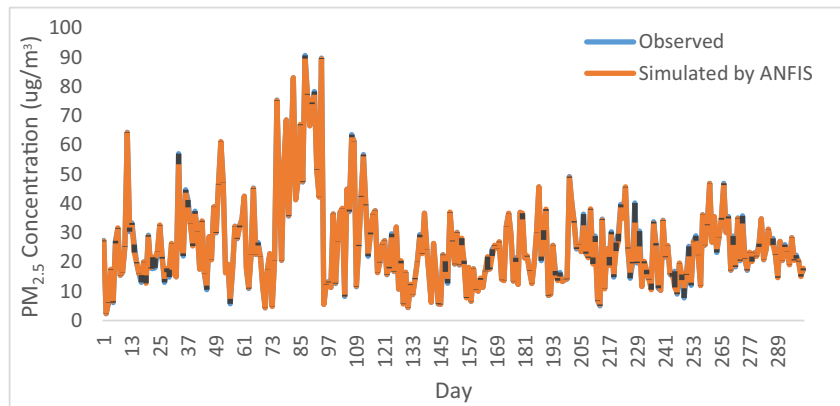


Fig. 9 Using ANFIS to analyze predicted and observed PM_{2.5} concentrations (µg/m³)

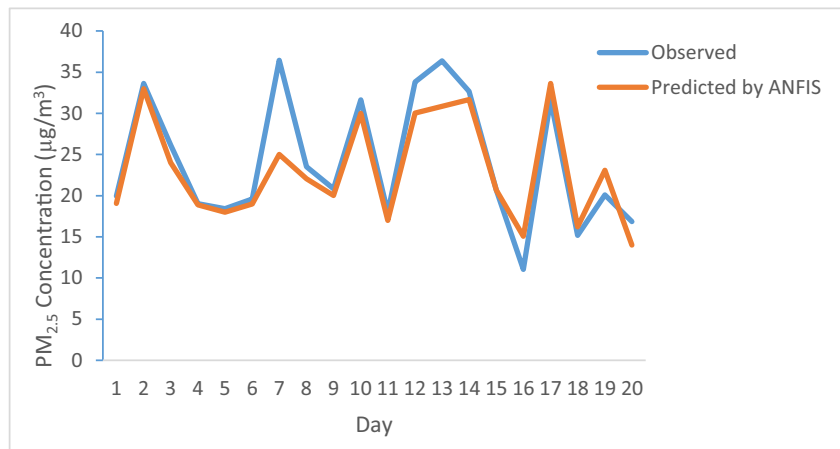


Table 3 Models error

Error	MLR		MLP		EEMD-GRNN		ANFIS	
	Train	Test	Train	Test	Train	Test	Train	Test
RMSE	11.8095	7.6402	7.8849	6.2522	1.8622	4.2655	0.4794	3.2979
MAE	8.82340	5.9961	6.4209	4.4781	1.1639	3.7427	0.1305	2.1668

results illustrated, among the models, in the training phases, the lowest R^2 was reported for MLR at 0.38 while the highest R^2 value was obtained by ANFIS at 0.99. Likewise, in the testing phases, MLR model acquired the lowest R^2 value (0.44) while the highest R^2 value (0.82) was obtained for ANFIS model. In terms of model accuracy, as compared with the other models, ANFIS and EEMD-GRNN models exhibited the best results in predicting $PM_{2.5}$ in Tehran with the lowest RMSE and MAE values and the highest R^2 values in training and testing phases

Conclusion

Public health can be affected by the prediction accuracy of $PM_{2.5}$ concentrations. This study compared the accuracy of linear model (MLR), nonlinear models (MLP) and hybrid models (EEMD-GRNN and ANFIS) in predicting $PM_{2.5}$ concentrations in Tehran. As it could be concluded based on the overall results, in comparison with the linear and nonlinear models, the hybrid of nonlinear models exhibited higher accuracy in prediction of $PM_{2.5}$ concentrations. However, the comparison of the results emphasizes that the ANFIS model obtained the highest accuracy for training ($R^2 = 0.99$, RMSE = 0.4794, and MAE = 0.1305) and testing phases ($R^2 = 0.82$, RMSE = 3.2979, and MAE = 2.1668) to predict the $PM_{2.5}$ concentrations but the results of hybrid models used in this study were close to each other. To generate the ANFIS model, the grid partition FIS (pimf) was applied. The best model generated by ANFIS consisted of three input MFs and nine fuzzy rules. To conclude, the best model, which was obtained for predicting $PM_{2.5}$ concentrations in Tehran was created by ANFIS with pimf-type input and three input MFs.

Funding information The work was supported by the Iran National Science Foundation: INSF through grant agreement 95850153.

References

- Alimissis A, Philippopoulos K, Tzani CG, Deligiorgi D (2018) Spatial estimation of urban air pollution with the use of artificial neural network models. *Atmos Environ* 191:205–213
- Alizadeh-Choobari O, Ghafarian P, Adibi P (2016) Inter-annual variations and trends of the urban warming in Tehran. *Atmos Res* 170:176–185
- Amirkhani S, Nasirivatan S, Kasaeian AB, Hajinezhad A (2015) ANN and ANFIS models to predict the performance of solar chimney power plants. *Renew Energy* 83:597–607
- Ausati S, Amanollahi J (2016) Assessing the accuracy of ANFIS, EEMD-GRNN, PCR, and MLR models in predicting $PM_{2.5}$. *Atmos Environ* 142:465–474
- Baker KR, Foley KM (2011) A nonlinear regression model estimating single source concentrations of primary and secondarily formed $PM_{2.5}$. *Atmos Environ* 45:3758–3767
- Bench G (2004) Measurement of contemporary and fossil carbon contents of $PM_{2.5}$ aerosols: results from turtleback dome, Yosemite National Park. *Environ Sci Technol* 38(8):2424–2427
- De Menezes Neto OL, Coutinho MM, Marengo JA, Capistrano VB (2017) The impact of a plume-rise scheme on earth system modeling: climatological effects of biomass aerosols on the surface temperature and energy budget of South America. *Theor Appl Climatol* 129(3–4):1035–1044
- Ding W, Zhang J, Leung Y (2016) Prediction of air pollutant concentration based on sparse response back-propagation training feedforward neural networks. *Environ Sci Pollut Res* 23:19481–19494
- Ghasemi A, Amanollahi J (2019) Integration of ANFIS model and forward selection method for air quality forecasting. *Air Qual Atmos Health* 12(1):59–72
- Gheyas I, Smith L (2009) A neural network approach to time series forecasting. *Proceedings of the World Congress on Engineering Vol II*
- Golchoubian H, Moayyedi G, Fazilati H (2012) Spectroscopic studies on Solvatochromism of mixed-chelate copper(II) complexes using MLR technique. *Spectrochim Acta A Mol Biomol Spectrosc* 85(1):25–30
- Guneri AF, Ertay T, Yücel A (2011) An approach based on ANFIS input selection and modeling for supplier selection problem. *Expert Syst Appl* 38(12):14907–14917
- Haywood J, Boucher O (2000) Estimates of the direct and indirect radiative forcing due to tropospheric aerosols: a review. *Rev Geophys* 38(4):513–543
- Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, Yen NC, Tung CC, Liu HH (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc Royal Soc Math Phys Eng Sci* 454:903–995
- Jang JSR, Sun CT, Mizutani E (1997) Neuro-fuzzy and soft computing: a computational, approach to learning and machine intelligence. *IEEE Trans Autom Control* 42(10), 1482–1484
- Jian S, Khare M (2010) Adaptive neuro-fuzzy modeling for prediction of ambient CO concentration at urban intersections and roadways. *Air Quality, Atmosphere & Health*. 3(4): 203–212.
- Kaboodvandpour S, Amanollahi J, Qhavam S, Mohammadi B (2015) Assessing the accuracy of multiple regressions, ANFIS, and ANN models in predicting dust storm occurrences in Sanandaj, Iran. *Nat Hazards* 78(2):879–893
- Kim SG, Yoon S (2019) Measuring the value of airborne particulate matter reduction in Seoul. *Air Qual Atmos Health* 12(5):549–560
- Legates DR, McCabe GJ (1999) Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. *Water Resour Res* 35(1):233–241
- Lelieveld J, Evans JS, Fnais M, Giannadaki D, Pozzer A (2015) The contribution of outdoor air pollution sources to pre-mature mortality on a global scale. *Nature* 525:367–371
- Leung MT, Daock H, Chen A (2000) Forecasting stock indices: a comparison of classification and level estimation models. *Int J Forecast* 16(2):173–190
- Liu X, Nie D, Zhang K, Wang Z, Li X, Shi Z, Wang Y, Huag L, Chen M, Ge X, Ying Q, Yu X, Liu X, Hu J (2019) Evaluation of particulate matter deposition in the human respiratory tract during winter in Nanjing using size and chemically resolved ambient measurements. *Air Qual Atmos Health* 12(5):529–538
- Lu CJ, Shao YE (2012) Forecasting computer products sales by integrating ensemble empirical mode decomposition and extreme learning machine. *Math Probl Eng* 15p

- Mashaly AF, Alazba AA (2016) MLP and MLR models for instantaneous thermal efficiency prediction of solar still under hyper-arid environment. *Comput Electron Agric* 122:146–155
- Matlab (2018) Anfis and the ANFIS Editor, Available at: <http://www.mathworks.com/help/fuzzy/anfis-and-the-anfis-editor-gui.html>
- McCulloch W, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys* 5:115–133
- Messikh N, Bousba S, Bougdah N (2017) The use of a multilayer perceptron (MLP) for modelling the phenol removal by emulsion liquid membrane. *J Environ Chem Eng* 5(4):3483–3489
- Ministry of Health and Medical Education 2012. Available online at <http://www.behdasht.gov.ir/>
- Mirzaei M, Amanollahi J, Tzani CG (2019) Evaluation of linear, non-linear, and hybrid models for predicting PM_{2.5} based on a GTWR model and MODIS AOD data. *Air Qual Atmos Health* 12(10):1215–1224
- Noori R, Hoshyaripour G, Ashrafi K, NadjarArrabi B (2010) Uncertainty analysis of developed ANN and ANFIS model in prediction of carbon monoxide daily concentration. *Atmos Environ* 44(4):476–482
- Nørgaard M, Ravn O, Poulsen NK, Hansen LK (2000) Neural networks for modelling and control of dynamic systems. Springer, United Kingdom
- Oh H-J, Kim J, Sohn J-R, Kim J (2019) Exposure to indoor-outdoor particulate matter and associated trace elements within childcare facilities. *Air Qual Atmos Health* 12(8):993–1001
- Pope CA, Burnett RT, Thurston GD, Thun MJ, Calle EE, Krewski D, Godleski JJ (2004) Cardiovascular mortality and long-term exposure to particulate air pollution: epidemiological evidence of general pathophysiological pathways of disease. *Circulation* 109(1):71–77
- Pouretedal HR, Damirri S, Shahsavan A (2018) Modification of RDX and HMX crystals in procedure of solvent/anti-solvent by statistical methods of Taguchi analysis design and MLR technique. *Def Technol* 14(1):59–63
- Sahu SK, Zhang H, Guo H, Hu J, Ying Q, Kota SK (2019) Health risk associated with potential source regions of PM_{2.5} in Indian cities. *Air Qual Atmos Health* 12(3):327–340
- Shahbazi B, Rezazi B, Chehreh Chelgani S, Javad Koleini SM, Noaparast M (2013) Estimation of diameter and surface area flux of bubbles based on operational gas dispersion parameters by using regression and ANFIS. *Int J Min Sci Technol* 23(3):343–348
- Sun W, Zhang H, Palazoglu A, Singh A, Zhang W, Liu S (2013) Prediction of 24-hour-average PM_{2.5} concentrations using a hidden Markov model with different emission distributions in northern California. *Sci Total Environ* 443:93–103
- Thorkashvand AM, Ahmadi A, Layegh Nikravesh N (2017) Prediction of kiwifruit firmness using fruit mineral nutrient concentration by artificial neural network (ANN) and multiple linear regressions (MLR). *J Integr Agric* 16(7):1634–1644
- Tzani C, Varotsos CA (2008) Tropospheric aerosol forcing of climate: a case study for the greater area of Greece. *Int J Remote Sens* 29(9):2507–2517
- Tzani CG, Alimissis A, Philippopoulos K, Deligiorgi D (2019) Applying linear and nonlinear models for the estimation of particulate matter variability. *Environ Pollut* 246:89–98
- Ventura LMB, Pinto FO, Soares LM, Luna AS, Gioda A (2019) Forecast of daily PM_{2.5} concentrations applying artificial neural networks and Holt-Winters models. *Air Qual Atmos Health* 12(3):317–325
- Vlachogianni A, Kassomenos P, Karppinen A, Karakitsios S, Kukkonen J (2011) Evaluation of a multiple regression model for the forecasting of the concentrations of NO_x and PM₁₀ in Athens and Helsinki. *Sci Total Environ* 409(8):1559–1571
- Wang J, Pan Y, Tian S, Chen X, Wang L, Wang Y (2016) Size distributions and health risks of particulate trace elements in rural areas in northeastern China. *Atmos Res* 168:191–204
- Wang Y, Wang J, Zhao G, Dong Y (2012) Application of residual modification approach in seasonal ARIMA for electricity demand forecasting: a case study of China. *Energy Policy* 48:284–294
- Wei M, Bai B, Sung AH, Liu Q, Wang J, Cather ME (2007) Predicting injection profiles using ANFIS. *Inf Sci* 177(20):4445–4461
- Willmott CJ, Matsuura K (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim Res* 30:79–82
- Wu J, Xu Y, Yang Q, Han Z, Zhao D, Tang J (2013) A numerical simulation of aerosols direct effects on tropopause height. *Theor Appl Climatol* 112(3–4):659–671
- Wu Z, Huang NE (2009) Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Adv Adapt Data Anal* 1:1–41
- Yadav AK, Sahoo SK, Dubey JS, Kumar AV, Pandey G, Tripathi RM (2019) Assessment of particulate matter, metals of toxicological concentration, and health risk around a mining area, Odisha, India. *Air Qual Atmos Health* 12(7):775–783
- Zendehboudi A, Li X, Wang B (2017) Utilization of ANN and ANFIS models to predict variable speed scroll compressor with vapor injection. *Int J Refrig* 74:475–487
- Zhou Q, Jiang H, Wang J, Zhou J (2014) A hybrid model for PM_{2.5} forecasting based on ensemble empirical mode decomposition and a general regression neural network. *Sci Total Environ* 496:264–274
- Zhu J, Wu P, Chen H, Zhou L, Tao Z (2018) A hybrid forecasting approach to air quality time series based on endpoint condition and combined forecasting model. *Int J Environ Res Public Health* 15:1941–1960

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.