Air Qual Atmos Health (2017) 10:873–883
DOI 10.1007/s11869-017-0477-9

Modeling of air pollutants using least square support vector regression, multivariate adaptive regression spline, and M5 model tree models

Ozgur Kisi[1] · **Kulwinder Singh Parmar**[2] · **Kirti Soni**[3] · **Vahdettin Demir**[4]

Received: 8 February 2017 / Accepted: 4 April 2017 / Published online: 13 April 2017

Abstract This study investigates the applicability of three different soft computing methods, least square support vector regression (LSSVR), multivariate adaptive regression splines (MARS), and M5 Model Tree (M5-Tree), in forecasting SO_2 concentration. These models were applied to monthly data obtained from Janakpuri, Nizamuddin, and Shahzadabad, located in Delhi, India. The models were compared with each other using the cross validation method with respect to root mean square error, mean absolute error, and correlation coefficient. According to the comparison, LSSVR provided better accuracy than the other models, while the MARS model was found to be the second best model in forecasting monthly SO_2 concentration. Results indicated that the applied models gave better forecasting accuracy in Janakpuri station than the other stations. The results were also compared with previous studies and satisfactory results were obtained from three methods in modeling SO_2 concentrations.

Keywords Soft computing techniques · Regression methods · Prediction modeling · Environmental management

✉ Ozgur Kisi
okisi@ibsu.edu.ge

✉ Kulwinder Singh Parmar
kulmaths@gmail.com

Kirti Soni
2006.kirti@gmail.com

Vahdettin Demir
vahdettin.demir@karatay.edu.tr

[1] Center for Interdisciplinary Research, International Black Sea University, Tbilisi, Georgia

[2] Department of Mathematics, IKG Punjab Technical University, Jalandhar -, Kapurthala, India

[3] Apex level and Industrial Metrology Division, CSIR-National Physical Laboratory, Delhi, India

[4] Engineering Faculty, Civil Engineering Department, Karatay University, Konya, Turkey

Introduction

Soft computing consists of different techniques, which are helpful to solve uncertain and complex problems (Corchado et al. 2011; Corchado and Herrero 2011; Vaidya et al. 2012, Kisi and Parmar 2016). It is used to investigate, simulate, and analyze complex issues and phenomenon in an attempt to solve real-world problems. Soft computing is useful where the precise scientific tools are incapable of giving analytic, low cost, and complete solution. The problem of air pollution is one of the most important problems among all, and it had come into play since the beginning. Air pollution affects both the developing and the developed countries alike. Air pollutants consist of gaseous pollutants (SO_2, NO_2, CO, etc.), odors, and suspended particulate matter (SPM) such as fumes, dust, smoke, and mist. The high concentration of air pollutants in and near the urban region causes severe pollution to the surroundings. Sulfur dioxide is a pungent, toxic gas that is in the atmosphere. Moreover, it harms the society, as it causes acid rain which affects the environment (Rizwan et al. 2013). Sulfur dioxide reacts in the atmosphere to form aerosol particles, which can create outbreaks of haze and other climate problems. The main sources of SO_2 are volcanic and anthropogenic emissions from burning sulfur-contaminated fossil fuels and the refinement of sulfide ores (Seinfeld and Pandis 2006). According to the new analysis of data from NASA's Aura satellite, the emissions of sulfur dioxide from power plants in India increased by more than 60% between 2005

and 2012 (Krotkov et al. 2016). In 2010, India surpassed the USA as the world's second largest emitter of SO_2 after China (EPA 2015a, b). The capital of India, Delhi, is considered among the most polluted megacities of the world (Gurjar et al. 2010). In the past, some studies were undertaken for air quality assessment of Delhi (Aneja et al. 2001; Goyal 2003; Gurjar et al. 2004; Mohan and Kandya 2007; Soni et al. 2014). Recently, Krotkov et al. (2016), studied ozone layer and major atmospheric pollutant gases (nitrogen dioxide (NO_2) and sulfur dioxide (SO_2)) by using the Ozone Monitoring Instrument (OMI) onboard NASA's Aura satellite and examined changes in SO_2 and NO_2 over some of the world's most polluted industrialized regions during the first decade of OMI observations and observed that in India, SO_2 and NO_2 levels from coal power plants and smelters are growing at a fast pace, increasing by more than 100 and 50%, respectively, from years 2005 to 2015.

The advanced soft computing techniques such as artificial neural networks (ANNs), adaptive-network-based fuzzy inference system (ANFIS), genetic algorithm (GL), fuzzy inference system (FIS), decision trees, and support vector machines have been successfully applied for modeling from the last decade (Kisi 2009a, b; Guven and Kisi 2011; Voukantsis et al. 2011; Kisi and Cengiz 2013; Antanasijević et al. 2013; Kisi and Tombul 2013; Gennaro et al. 2013; Goyal et al. 2014; Parmar and Bhardwaj 2014; Wanga et al. 2015; Kisi et al. 2016). Etemad-Shahidi and Mahjoobi (2009) used M5 algorithm for prediction of wave height, and the results of model trees were also compared with those of artificial neural networks. In some applications, generalized regression neural networks (GRNN), multilayer perceptron-neural networks (MLP), and support vector machine (SVM) are used to calculate the predicted value (Kim et al. (2012). In comparison to empirical and MLR models, ANN models performed better. In order to check the accuracy of the these models, Kisi (2015) applied multivariate adaptive regression splines (MARS), least square support vector regression (LSSVR), and M5 Model Tree (M5Tree) in Pan evaporation at Antalya and Mersin sample stations in Turkey. The LSSVR model performs more accurately in the case of using local input and output; in the second case, the MARS model is better. Kim et al. (2015) reported daily pan evaporation prediction by using soft computing models. Recently, Shafaei and Kisi (2016) employed three WANFIS (wavelet-ANFIS), WSVR (wavelet-SVR), and WARMA (wavelet-ARMA) hybrid methods for estimating monthly lake-level changes and found that three hybrid models forecasted more accurate than the single models.

However, not many scientific literature discuss a number of robust forecasting methods using soft computing techniques for air pollution modeling. The present paper includes MARS, the LSSVR, and M5 Model Tree (M5Tree) techniques. Each one of these algorithms is discussed

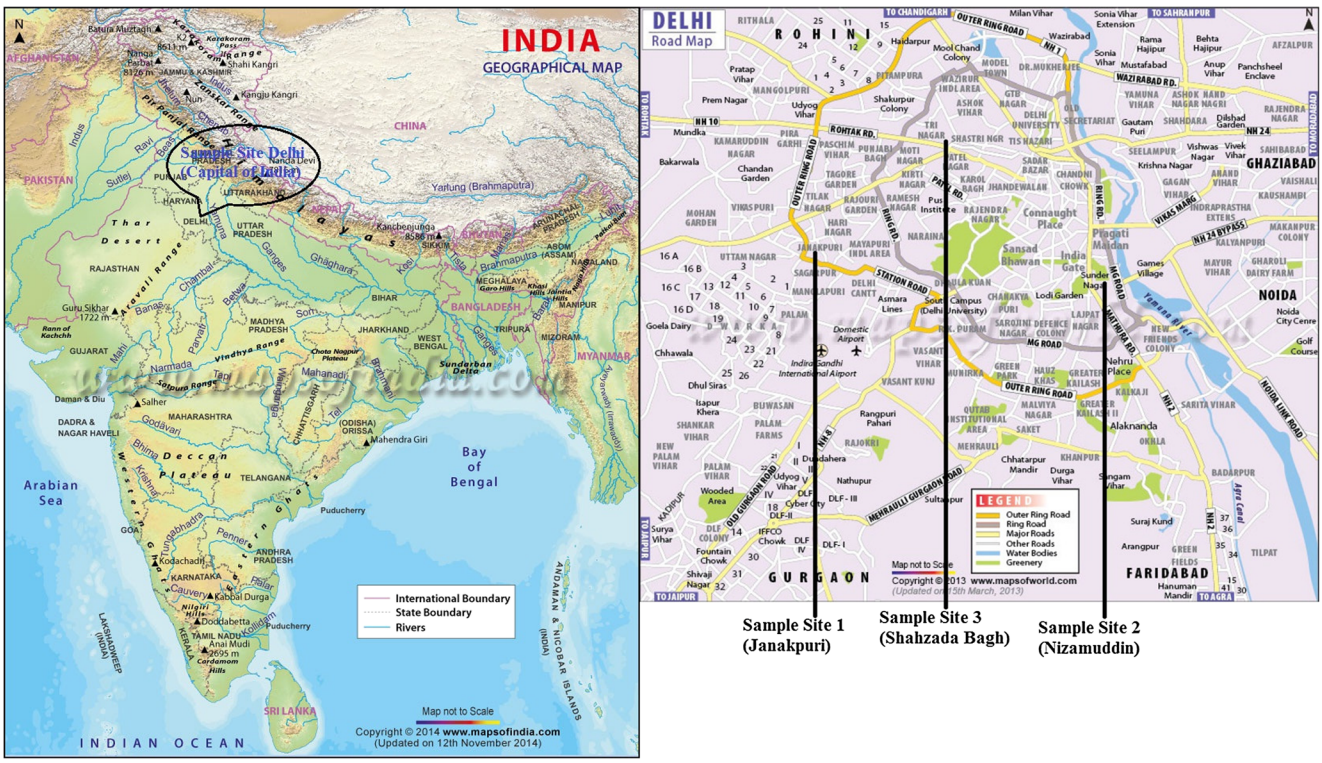

Fig. 1 The sample site area map

Table 1 The monthly statistical parameters of data set for Janakpuri, Nizamuddin, and Shahzadabad stations

	Data set	x_{mean} (ppm)	Sx (ppm)	Csx (ppm)	x_{min} (ppm)	x_{max} (ppm)	r1	r2	r3
Janakpuri	2006–2010	6.13	2.89	1.66	4	14.7	−0.217	−0.200	−0.097
	2000–2005	13.3	3.80	0.37	6.4	21.1	0.106	0.003	−0.116
	1996–1999	17.1	2.04	0.10	12.7	22.1	0.199	0.164	0.141
	1987–1995	12.7	5.06	−0.20	3	24.6	0.644	0.580	0.502
Nizamuddin	2006–2010	6.07	2.87	1.77	4	14.2	0.933	0.881	0.796
	2000–2005	12.9	3.94	0.05	2.1	20.9	0.796	0.720	0.654
	1996–1999	17.0	1.81	−0.15	12.9	20.8	0.238	0.168	0.275
	1987–1995	14.3	5.51	1.15	3.5	36.6	0.395	−0.069	−0.014
Shahzadabad	2006–2010	5.86	2.79	2.36	4	17	0.776	0.616	0.458
	2000–2005	10.1	3.23	0.99	4.4	20.5	0.771	0.674	0.615
	1996–1999	20.3	4.80	0.11	10.7	30.3	0.709	0.517	0.294
	1987–1995	19.2	9.95	0.10	3.2	42.7	0.268	0.104	−0.011

separately and the results discussed. In addition, a comparison of all methods is made to emphasize their advantages as well as their disadvantages. To the best of the authors' knowledge, it is the first time that such an analysis related to the LSSVR, MARS, and M5Tree is being performed for air pollutants of Delhi. This constitutes a real challenge as the urban pollution gets mixed with the desert dust aerosols during pre-monsoon and summer seasons over Delhi (Singh et al. 2005; Prasad et al. 2007; Soni et al. 2015; Parmar et al. 2016), whereas the winters are extremely polluted with high concentrations of black carbon aerosols from vehicular and other anthropogenic pollution sources leading to the formation of foggy and haze conditions over Delhi (Ganguly et al. 2006; Singh et al. 2010).

Delhi is the second most populous urban agglomeration in India and third largest urban area in the world. NASA's report creates the importance to investigate the pollutant levels at different sites (residential and industrial) in Delhi. Air pollutants directly affect the health of residents. This research has an importance as 19 million people have to breathe in this air and quality of air is directly related to health.

Data and methodology

In the Northern region, Delhi is located in central India and 715 ft. above the sea level (Fig. 1). The region has a semi-arid or steppe climate, with extremely hot summers, heavy rain falls

Table 2 The parameters of the optimal LSSVR models for each combination Janakpuri, Nizamuddin, and Shahzadabad stations

Cross validation	Training data set	Test data set	Input combination		
			(i)	(ii)	(iii)
Janakpuri					
M1	1987–1999	2006–2010	(100, 5)	(97, 5)	(100, 5)
M2	1987–1995 and 2006–2010	2000–2005	(100, 1)	(100, 1)	(100, 1)
M3	1987–1995 and 2000–2010	1996–1999	(37, 1)	(100, 1)	(77, 1)
M4	1996–2010	1987–1995	(1, 2)	(1, 2)	(100, 3)
Nizamuddin					
M1	1987–1999	2006–2010	(100, 59)	(100, 98)	(85, 100)
M2	1987–1995 and 2006–2010	2000–2005	(1, 4)	(24, 100)	(1, 9)
M3	1987–1995 and 2000–2010	1996–1999	(63, 100)	(100, 18)	(12, 99)
M4	1996–2010	1987–1995	(1, 100)	(1, 1)	(29, 8)
Shahzadabad					
M1	1987–1999	2006–2010	(100, 59)	(100, 96)	(90, 100)
M2	1987–1995 and 2006–2010	2000–2005	(1, 4)	(1, 8)	(1, 9)
M3	1987–1995 and 2000–2010	1996–1999	(100, 52)	(100, 97)	(80, 100)
M4	1996–2010	1987–1995	(65, 1)	(2, 25)	(3, 34)

in the monsoon months, and cold winters. There are dust storms in summer and foggy mornings in winter. Temperatures gradually rise to 46 °C in the summer and falls to 4 °C in winter. In winter months, temperature inversion and low wind speed are the main cause of accumulation of airborne pollutants in Delhi. In Delhi, industries, vehicular activities, power plants, and frequent dust storms are majorly contributing in the high concentration of the pollutants. The Central Pollution Control Board (CPCB) SO_2 data over four sites, in which two are residential, Janakpuri and Nizamuddin, and one industrial, Shahazada Bagh, are utilized for the present study. The ambient air quality and long-term data used in the present study covers the period 1993–2012 which is obtained from the CPCB. The monthly statistical parameters of the used data set for Janakpuri, Nizamuddin, and Shahzadabad stations are given in Table 1.

Least square support vector regression

Vladimir Vapnik and his co-workers developed this least square support vector machine models at AT&T Bell Laboratories in 1995, which are applied to calculate the nonlinear relationship between input variables and output variables with least error (Cortes and Vapnik 1995; Suykens 2001; Smola 2004). LSSVR generated from SVR (support vector regression), which is a great technique to solve the real-life problems by a combination of regression, function estimation, and classification. This SVR is developed on the ground of structural risk minimization (SRM), which provides the least error in forecasting problems. It is mostly suitable for signal processing, pattern recognition, and nonlinear regression estimation.

Firstly, the LSSVR model was projected by Suykens and Vandewalle in 1999 (Suykens and Vandewalle 1999), which is applied on chaotic time series forecasting. The main difference between LSSVR and SVR is consideration of the equations; during the training phase, LSSVR uses linear equations while SVR uses quadratic optimization. The other conventional models like back propagation neural networks (BPNN), partial least square regression (PLS), and multivariate linear regression (MLR) are computationally more extensive than LSSVR. So it is easy to apply this model as compared to others.

Consider a given training set $\{p_k, q_k\}_{k=1}^{N}$ with input data $p_k \in R^n$ and output data $q_k \in R$ with class labels $q_k \in \{-1, +1\}$ and linear classifier

$$q(p) = \mathrm{sign}\left[w^T p + b\right] \tag{1}$$

When the data of the two classes are separable, one can say

$$\left\{ \begin{array}{ll} w^T p_k + b \geq +1, & \text{if } q_k = +1 \\ w^T p_k + b \leq -1, & \text{if } q_k = -1 \end{array} \right\} \tag{2}$$

These two sets of inequalities can be combined into one single set as follows

$$q_k\left[w^T p_k + b\right] \geq 1, \quad k = 1, 2, 3, ..., N \tag{3}$$

The convex optimization theory is used to formulate SVR. In this methodology, firstly, it starts formulating the problem as a constrained optimization problem. In the second step, it

Table 3 Comparison of LSSVR models

Statistics	Cross validation	Test data set	Input combination			
			(i)	(ii)	(iii)	Mean
Janakpuri						
RMSE	M1	2006–2010	2.18	1.35	1.35	1.63
	M2	2000–2005	3.62	2.36	1.89	2.62
	M3	1996–1999	3.03	2.29	2.11	2.48
	M4	1987–1995	3.10	2.87	2.74	2.90
		Mean	2.98	2.22	2.02	2.41
MAE	M1	2006–2010	1.96	0.90	0.83	1.23
	M2	2000–2005	2.49	1.71	1.27	1.82
	M3	1996–1999	2.19	1.74	1.51	1.81
	M4	1987–1995	2.19	2.05	1.96	2.07
		Mean	2.21	1.60	1.39	1.73
R	M1	2006–2010	0.924	0.912	0.927	0.921
	M2	2000–2005	0.705	0.883	0.927	0.839
	M3	1996–1999	0.718	0.846	0.873	0.812
	M4	1987–1995	0.707	0.746	0.774	0.742
		Mean	0.764	0.847	0.875	0.829
Nizamuddin						
RMSE	M1	2006–2010	2.91	1.88	1.90	2.23
	M2	2000–2005	2.36	2.30	2.32	2.33
	M3	1996–1999	2.04	1.90	1.84	1.93
	M4	1987–1995	5.21	5.46	5.41	5.36
		Mean	3.13	2.89	2.87	2.96
MAE	M1	2006–2010	2.73	1.56	1.59	1.96
	M2	2000–2005	1.73	1.69	1.69	1.70
	M3	1996–1999	1.59	1.48	1.44	1.50
	M4	1987–1995	3.45	3.30	3.48	3.41
		Mean	2.38	2.01	2.05	2.14
R	M1	2006–2010	0.933	0.929	0.928	0.930
	M2	2000–2005	0.803	0.814	0.813	0.810
	M3	1996–1999	0.241	0.263	0.292	0.265
	M4	1987–1995	0.397	0.324	0.405	0.375
		Mean	0.594	0.582	0.609	0.595
Shahzadabad						
RMSE	M1	2006–2010	3.21	2.39	2.41	2.67
	M2	2000–2005	2.11	2.07	2.06	2.08
	M3	1996–1999	3.42	3.51	3.53	3.49
	M4	1987–1995	6.73	6.48	6.50	6.57
		Mean	3.87	3.61	3.63	3.70
MAE	M1	2006–2010	2.97	1.91	1.94	2.27
	M2	2000–2005	1.59	1.59	1.59	1.59
	M3	1996–1999	2.59	2.70	2.74	2.68
	M4	1987–1995	4.96	5.02	5.00	4.99
		Mean	3.03	2.81	2.82	2.88
R	M1	2006–2010	0.780	0.760	0.758	0.766
	M2	2000–2005	0.767	0.773	0.773	0.771
	M3	1996–1999	0.706	0.679	0.673	0.686
	M4	1987–1995	0.748	0.764	0.764	0.758
		Mean	0.750	0.744	0.742	0.745

formulates the Lagrangian and then takes the conditions for optimality and finally solves the problem in the dual space of Lagrange multipliers. With the resulting classifier

$$q(p) = \text{sign}\left[\sum_{k=1}^{N} \alpha_k q_k p_k^t p + b\right] \quad (4)$$

Cortes and Vapnik (1995) extended this linear SVR classifier to a non-separable case by using an additional slack variable in the problem formulation. Now, after applying additional slack variable, the set of inequalities is as

$$q_k\left[w^T p_k + b\right] \geq 1 - \xi_k, \quad k = 1, 2, 3, ..., N \quad (5)$$

In classic SVR, inequality type constraints are considered, but in LSSVR equality type of constraints are used. This equality type of constraints simplifies the problem as the solution of LSSVR, received directly after solving a set of linear equations instead of solving a convex quadratic program. In this LSSVR classifier, in the primal space is as follow,

$$q(p) = \text{sign}\left[w^T p + b\right] \quad (6)$$

where b is a real constant. In the nonlinear classification, the LSSVR classifier in the dual space is like below

$$q(p) = \text{sign}\left[\sum_{k=1}^{N} \alpha_k q_k K(p, p_k) + b\right] \quad (7)$$

In Eq. (7), α_k is the +ve real constants and b a real constant, in general, $K(p_k, p) = \langle \phi(p_k), \phi(p) \rangle$, $\langle \bullet, \bullet \rangle$ is the inner product, and $\phi(p)$ the nonlinear map from the original space to a high dimensional space. In the function inference, the LSSVR model is in the below form

$$q(p) = \sum_{k=1}^{N} \alpha_k K(p, p_k) + b \quad (8)$$

In radial basis function (RBF), kernels are in use, two alteration parameters (γ, α) are inserting. Where γ is the regularization constant and α the width of the radial basis function kernel.

In this current research work, LSSVR is used for modeling of air pollutants level in Delhi. By using the LSSVR model, the output prediction error is least. As compared to

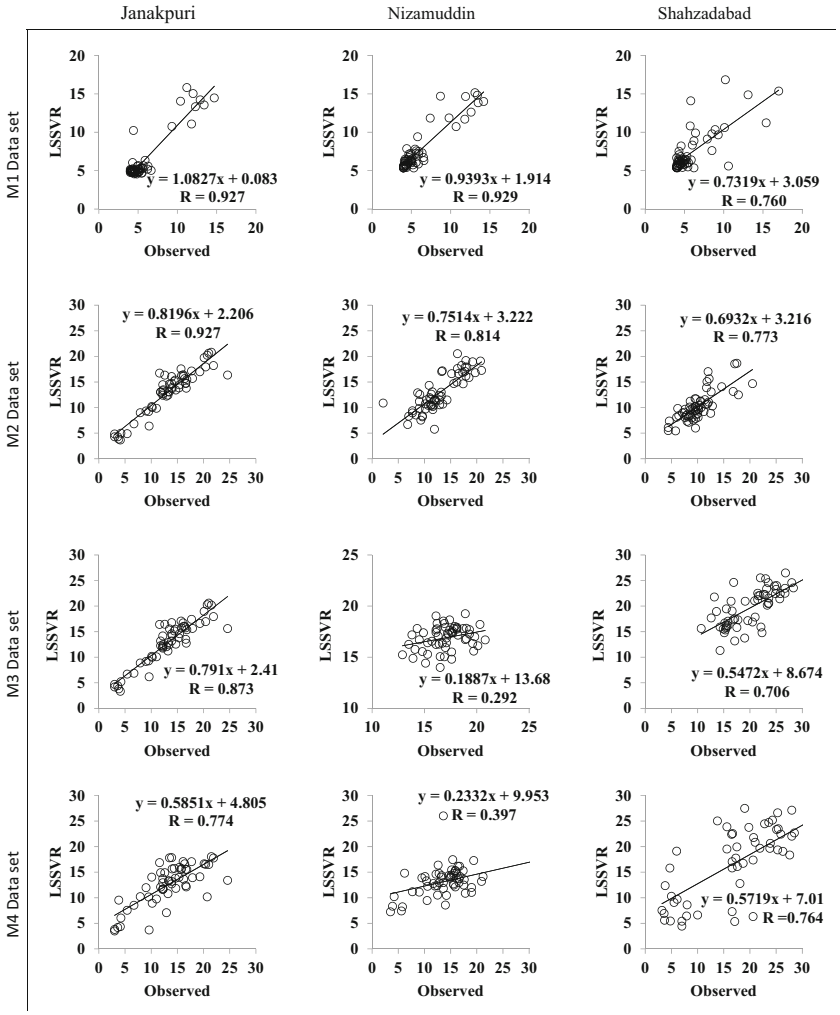

Fig. 2 The observed and forecasted SO_2 by the LSSVR model

conventional models, the LSSVR model is best to remove noises and reduces the computational labor. So, because of these benefits, the conventional models can be replaced with LSSVR. This will be more useful in application of forecast modeling in different areas of research.

Multivariate adaptive regression splines

Multivariate adaptive regression splines model is a non-parametric regression model, which is applied to predict continuous numeric outcomes. MARS was developed by Friedman (1991), which is a flexible procedure to organize relationships that are nearly additive or involve interactions with other variables. MARS makes no assumptions about the underlying functional relationship between dependent and independent variables in order to estimate the general functions of high dimensional arguments given sparse data, which is the main beauty of this model (Friedman 1991). The MARS model explains the complex nonlinear relation between predictor and response variables; this is the major beauty of this model. Apart from other conventional models, it can work by both backward and forward stepwise procedures. By using the backward stepwise procedure, it is helpful to remove preventable variables from the earlier selected set and this elimination improves prediction accuracy (Andres et al. 2010). The forward stepwise procedure helps to choose the appropriate input variables.

The value of other variables can be defined by using two basis functions or by inflection point along the range of inputs; then, we will have the new variable Y, which is mapped from variable X as below:

$$Y = \max(0, X - c) \quad (9)$$

$$Y = \max(0, c - X) \quad (10)$$

where c represents the threshold value. There are two adjacent splines, which are intersecting at knot to maintain the continuity of the basis functions (Sephton 2001; Bera et al. 2006). The MARS model has differently many applications in research, like prediction modeling, financial management, time series analysis, etc. Here, the MARS model is applied to calculate the level of air pollution at different sites in Delhi, India.

M5 model tree

Quinlan (1992) developed the model for continuous class learning, which is named as an M5 model tree. The main strength of this model is a binary decision tree. To find the relation between independent and dependent variables, linear regression function is applied to terminal leaf nodes (Mitchell 1997).

The M5 model tree is commonly used for categorical data; this is better than other conventional decision tree models. The main advantage of this model is to handle quantitative data, which makes it different from other tree models. It has two steps; in the first step, the data is divided into subsets to generate a decision tree (Solomatine and Xue 2004). In the second step, the standard deviation of the class value reached at a

Table 4 Comparison of MARS models

Statistics	Cross validation	Test data set	Input combination			
			(i)	(ii)	(iii)	Mean
Janakpuri						
RMSE	M1	2006–2010	2.65	1.65	1.72	2.01
	M2	2000–2005	3.69	3.26	3.33	3.43
	M3	1996–1999	3.04	2.81	2.86	2.90
	M4	1987–1995	3.16	2.94	2.92	3.01
		Mean	3.14	2.67	2.71	2.84
MAE	M1	2006–2010	2.45	1.33	1.42	1.73
	M2	2000–2005	2.65	2.50	2.44	2.53
	M3	1996–1999	2.18	2.02	2.06	2.09
	M4	1987–1995	2.19	2.07	2.04	2.10
		Mean	2.37	1.98	1.99	2.11
R	M1	2006–2010	0.902	0.909	0.916	0.909
	M2	2000–2005	0.686	0.760	0.750	0.732
	M3	1996–1999	0.718	0.760	0.747	0.742
	M4	1987–1995	0.703	0.738	0.744	0.728
		Mean	0.752	0.792	0.789	0.778
Nizamuddin						
RMSE	M1	2006–2010	2.78	2.94	1.63	2.45
	M2	2000–2005	2.58	2.72	2.71	2.67
	M3	1996–1999	2.44	2.32	2.23	2.33
	M4	1987–1995	5.52	5.64	5.57	5.58
		Mean	3.33	3.41	3.04	3.26
MAE	M1	2006–2010	2.59	2.75	1.27	2.20
	M2	2000–2005	1.94	1.95	1.94	1.94
	M3	1996–1999	1.97	1.79	1.77	1.84
	M4	1987–1995	3.46	3.52	3.68	3.55
		Mean	2.49	2.50	2.17	2.39
R	M1	2006–2010	0.933	0.933	0.927	0.931
	M2	2000–2005	0.807	0.804	0.803	0.805
	M3	1996–1999	0.249	0.251	0.226	0.242
	M4	1987–1995	0.401	0.351	0.387	0.380
		Mean	0.598	0.585	0.586	0.589
Shahzadabad						
RMSE	M1	2006–2010	3.11	3.25	2.24	2.87
	M2	2000–2005	2.31	2.38	2.38	2.36
	M3	1996–1999	3.46	4.37	4.43	4.09
	M4	1987–1995	7.49	7.07	7.30	7.29
		Mean	4.09	4.27	4.09	4.15
MAE	M1	2006–2010	2.85	3.05	1.59	2.50
	M2	2000–2005	1.73	1.91	1.91	1.85
	M3	1996–1999	2.70	3.41	3.49	3.20
	M4	1987–1995	5.54	5.25	5.48	5.42
		Mean	3.21	3.41	3.12	3.24
R	M1	2006–2010	0.777	0.772	0.752	0.767
	M2	2000–2005	0.760	0.775	0.775	0.770
	M3	1996–1999	0.691	0.646	0.608	0.648
	M4	1987–1995	0.681	0.711	0.696	0.696
		Mean	0.727	0.726	0.708	0.720

node is used for splitting criterion. Here, expected reduction is measured, in order to check the error of testing each attribute at node. Then compute the SDR (standard deviation reduction) (Pal and Deswal 2009) as below:

$$\mathrm{SDR} = \mathrm{sd}(T) - \sum \frac{|T_\mathrm{i}|}{|T|} \mathrm{sd}(T_\mathrm{i}) \tag{11}$$

where *sd* expressed as standard deviation, T represents a set of examples which reaches at the node, and T_i is the i^{th} outcome of the possible set. The data's standard deviation (SD) are less than parent nodes. In this phase, large tree-like design have poor generalization and result in around appropriate. Quinlan (1992) suggests a solution for this circumstance, in the real dense sapling is actually clipped and then clipped subtrees are usually changed by using linear regression functions. By this process, accuracy and reliability of the design tree are very much improved. The M5 model tree is applied in this research work for decision making for the air quality level in Delhi, India.

Application and results

Monthly SO_2 of three different regions, Janakpuri, Nizamuddin, and Shahzadabad, located in India were modeled using three different heuristic methods, LSSVR, MARS, and M5Tree. Three previous lags were used as inputs to the models to forecast 1-month ahead SO_2 parameter. The cross validation method was used for each model by dividing data into four subsets. Table 2 reports the training and test data sets of each model. In this, table M1 indicates model 1 and vice versa. Evaluation criteria used in the applications are root mean square errors (RMSE), mean absolute errors (MAE), and correlation coefficient (R). The RMSE and MAE statistics can be given as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{SO2}_{i,o} - \mathrm{SO2}_{i,e}\right)^2} \tag{12}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\mathrm{SO2}_{i,o} - \mathrm{SO2}_{i,e}\right| \tag{13}$$

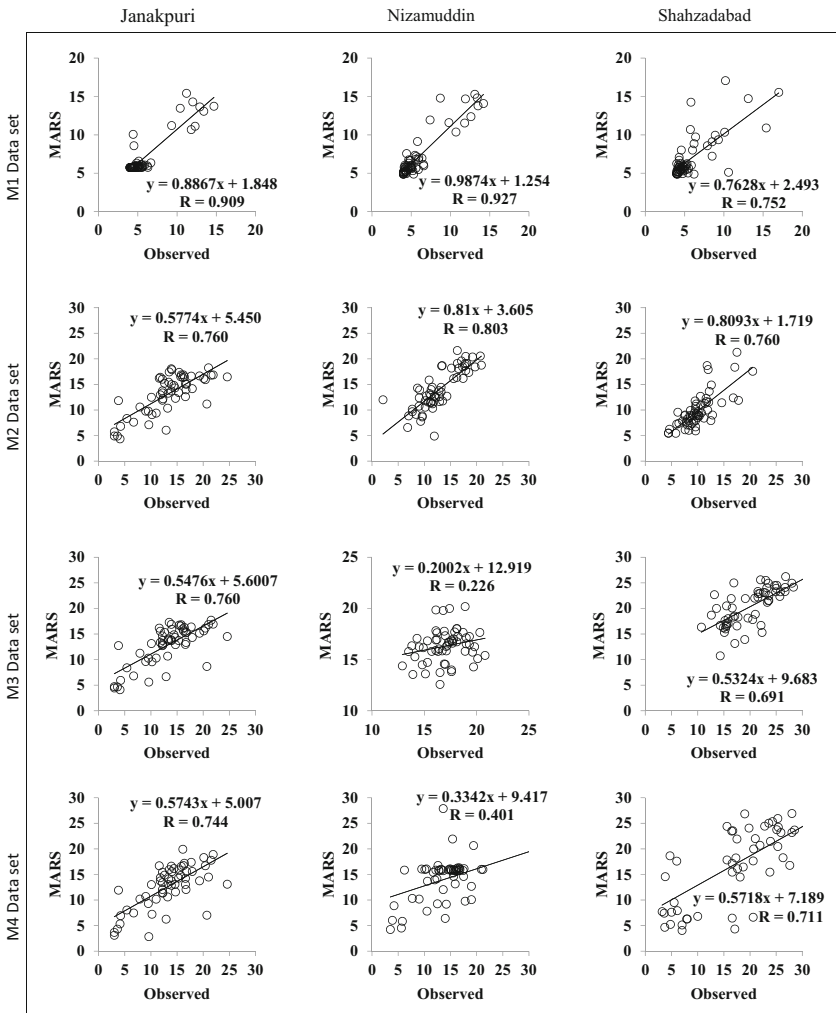

Fig. 3 The observed and forecasted SO_2 by the MARS model

where N is the number of data, $SO2_{i,o}$ is the observed SO_2 values, and $SO2_{i,e}$ is the model's estimate.

For each LSSVR model in each data set, various parameters were tried and the best ones that gave the minimum RMSE error in the test period were obtained. The parameters of the optimal LSSVR models for each combination of Janakpuri, Nizamuddin, and Shahzadabad stations are provided in Table 2. In this table, (100, 5) indicates the regularization constant and RBF kernel values of the LSSVR model, respectively. Test results of the optimal LSSVR models for each station and for each data set are given in Table 3. This table obviously shows that all the LSSVR models give different accuracies for different data sets. In Janakpuri, average accuracies reveal that the best results were generally obtained for the third input combination. It is clear from Table 3 that the LSSVR model provides the worst results in forecasting SO_2 of three stations for the M4 data set (1987–1995). The basic reason for this might be the fact that the data range of this test data set is very different from those of the M1, M2, and M3 (see Table 1). The maximum values (x_{max} = 24.6, 36.6, and 42.7) of the M4 test data set are higher than those of the other test data sets for the Janakpuri, Nizamuddin, and Shahzadabad stations, respectively. Training with M1, M2, and M3 data sets causes some extrapolation difficulties for the applied LSSVR models. Standard deviation of the M4 data set is also higher than those of the others. In Janakpuri, the LSSVR model provides the best accuracy for the M1 data set and second and third input combinations while the M3 and M2 data sets with third input combinations give the best results for the Nizamuddin and Shahzadabad stations, respectively. Figure 2 illustrates the observed and predicted SO_2 values using the LSSVR model for each data set. The models' accuracies seem to be better in forecasting the SO_2 of Janakpuri station than those of the other stations. This is also confirmed by the comparison statistics reported in Table 3. This may be due to the difference of SO_2 data range of each station. Janakpuri has a lower data range (x_{min} = 3 and x_{max} = 24.6) than those of the Nizamuddin (x_{min} = 2.1 and x_{max} = 36.6) and Shahzadabad (x_{min} = 3.2 and x_{max} = 42.7) stations, respectively.

Table 4 compares the accuracy of different MARS models for different test data sets. Unlike the previous application models generally yield better results in combination ii of Janakpuri. The best MARS model was obtained from M1 and second inputs in Janakpuri while the M1 with third inputs provided the best results in Nizamuddin and Shahzadabad stations, respectively. The observed and forecasted SO_2 by the MARS models is demonstrated in Fig. 3 in the form of scatter plot. Similar to the LSSVR, here, also less scattered forecasts were obtained for the Janakpuri in relative to the other two stations. SO_2 modeling accuracy of the optimal M5-Tree models is provided in Table 5. Different from the LSSVR and MARS models, the M5-Tree model gives the best accuracy for M2 and third inputs in Janakpuri while the M3 and M2 with first input provide the best results in Nizamuddin and Shahzadabad stations, respectively. The scatter plots given in Fig. 4 clearly show that the M5-Tree model forecasts SO_2 of Janakpuri better than those of the other stations. Comparison with Figs. 2 and 3 obviously indicates that the M5-Tree model gives more scattered forecasts than the LSSVR and MARS models. The reason of this may be the fact that the linear structure of the M5-Tree model prevents it from accurately predicting highly nonlinear SO_2. Comparison

Table 5 Comparison of M5-Tree models

Statistics	Cross validation	Test data set	Input combination			
			(i)	(ii)	(iii)	Mean
Janakpuri						
RMSE	M1	2006–2010	2.95	2.88	2.89	2.91
	M2	2000–2005	3.47	3.13	2.71	3.10
	M3	1996–1999	2.93	2.79	2.79	2.84
	M4	1987–1995	3.16	3.16	3.07	3.13
		Mean	3.13	2.99	2.87	2.99
MAE	M1	2006–2010	2.81	2.75	2.76	2.77
	M2	2000–2005	2.53	2.25	1.88	2.22
	M3	1996–1999	2.23	2.09	2.03	2.12
	M4	1987–1995	2.26	2.14	2.06	2.15
		Mean	2.46	2.31	2.18	2.32
R	M1	2006–2010	0.862	0.876	0.877	0.872
	M2	2000–2005	0.722	0.782	0.841	0.782
	M3	1996–1999	0.736	0.766	0.772	0.758
	M4	1987–1995	0.696	0.711	0.729	0.712
		Mean	0.754	0.784	0.805	0.781
Nizamuddin						
RMSE	M1	2006–2010	3.84	3.65	3.65	3.71
	M2	2000–2005	3.80	5.29	4.65	4.58
	M3	1996–1999	2.96	4.51	3.44	3.64
	M4	1987–1995	5.46	6.03	5.60	5.70
		Mean	4.02	4.87	4.34	4.41
MAE	M1	2006–2010	3.68	3.51	3.51	3.57
	M2	2000–2005	3.10	4.23	3.79	3.71
	M3	1996–1999	2.14	3.41	2.34	2.63
	M4	1987–1995	3.52	3.98	3.64	3.71
		Mean	3.11	3.78	3.32	3.40
R	M1	2006–2010	0.844	0.846	0.846	0.845
	M2	2000–2005	0.694	0.675	0.724	0.697
	M3	1996–1999	0.179	0.114	0.084	0.126
	M4	1987–1995	0.396	0.327	0.369	0.364
		Mean	0.528	0.490	0.505	0.508
Shahzadabad						
RMSE	M1	2006–2010	3.83	3.69	3.69	3.74
	M2	2000–2005	3.13	4.10	3.85	3.69
	M3	1996–1999	3.70	4.86	5.36	4.64
	M4	1987–1995	7.42	7.14	7.14	7.23
		Mean	4.52	4.95	5.01	4.83
MAE	M1	2006–2010	3.54	3.46	3.45	3.49
	M2	2000–2005	2.39	3.08	2.86	2.78
	M3	1996–1999	2.74	3.87	4.29	3.63
	M4	1987–1995	5.55	5.34	5.33	5.41
		Mean	3.56	3.94	3.99	3.83
R	M1	2006–2010	0.617	0.640	0.641	0.633
	M2	2000–2005	0.703	0.523	0.581	0.602
	M3	1996–1999	0.696	0.484	0.447	0.542
	M4	1987–1995	0.686	0.705	0.706	0.699
		Mean	0.676	0.588	0.594	0.619

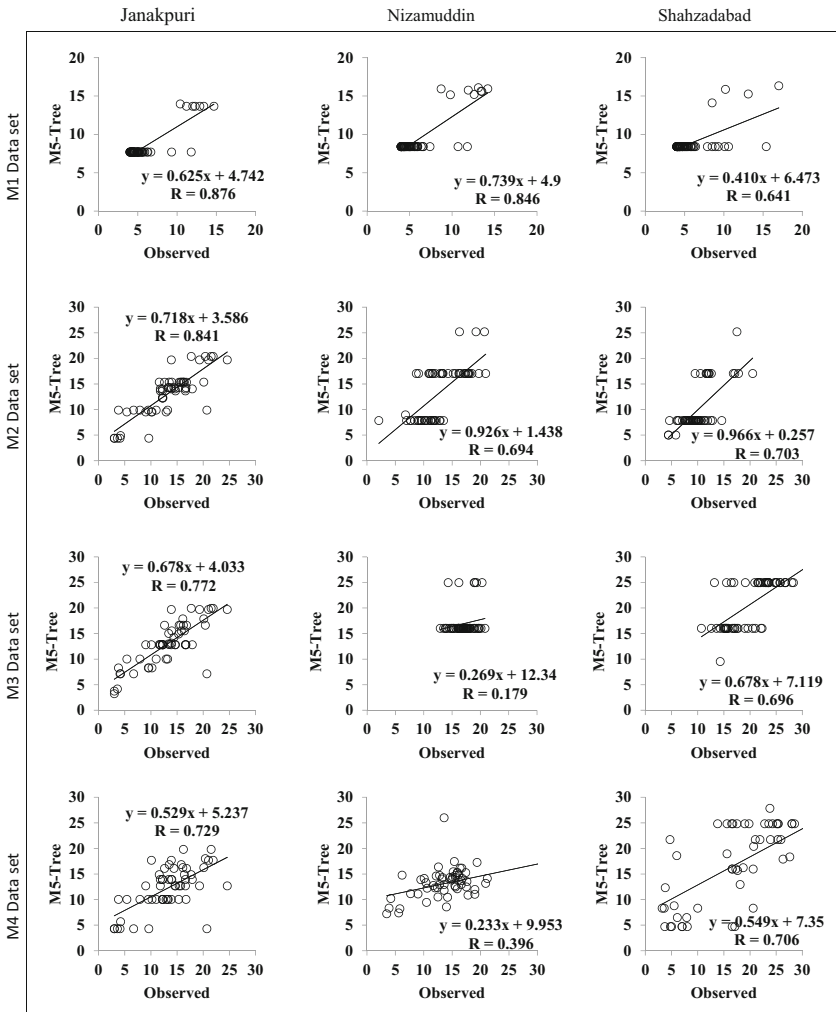

Fig. 4 The observed and forecasted SO_2 by the M5-Tree model

of average statistics provided in Tables 3, 4, 5 says that the LSSVR models are generally more successful than the MARS and M5-Tree models in forecasting SO_2.

Sahin et al. (2005) modeled SO_2 distribution in Istanbul using artificial neural networks (ANNs) and non-linear regression (NLR), and they found that the optimal ANNs and NLR provided RMSE = 23.13 μg/m^3 and 22.35 μg/m^3, MAE = 14.97 μg/m^3 and 18.41 μg/m^3, and R = 0.528 and 0.638, respectively. Akkoyunlu et al. (2010) used the ANN-based approach for the prediction of urban SO_2 concentrations and found correlation coefficients of about 0.770, 0.744, and 0.751 for the winter, summer, and overall data, respectively. Sahin et al. (2011) used cellular neural network (CNN) and the statistical persistence method (PER) to model SO_2 concentrations of Istanbul, and they found RMSE = 14.2 and 13.9, MAE = 9.9 and 7.8, and R = 0.85 and 0.83 for the CNN and PER models, respectively. It is clear from Tables 3, 4, and 5 that the applied LSSVR, MARS, and M5-Tree models in this study generally provide satisfactory results in modeling SO_2 concentrations.

Conclusions

The ability of three different soft computing methods, LSSVR, MARS, and M5-Tree, in forecasting SO_2 concentration is evaluated. Data from three stations, Janakpuri, Nizamuddin, and Shahzadabad, located in Delhi, India, were used in the applications. The cross validation method was employed for presenting generality of the applied models. LSSVR performed superior to the other models in forecasting monthly SO_2 concentration. MARS was found to be the second best method. Because of its linear nature, M5-Tree provided worse results than the nonlinear LSSVR and MARS models. All the models provided better accuracy in forecasting SO_2 of Janakpuri station than those of the other stations because of its lower data range. Comparison with previous studies showed that soft computing models applied in this study generally provided satisfactory results in modeling SO_2 concentrations.

Acknowledgements The authors are thankful to the Central Pollution Control Board (CPCB), Government of India, for providing the research data and Dr. B. R. Ambedkar National Institute of Technology, Jalandhar (Government of India) and IKG Punjab Technical University (Government of Punjab) for providing research facilities. The second author is also thankful to Prof. Rashmi Bhardwaj, Guru Gobind Singh Indraprastha University, for her motivation and astute guidance. The third author (KS) is grateful to the Director, CSIR-NPL.

References

Akkoyunlu A, Yetilmezsoy K, Erturk F, Oztemel E (2010) A neural network-based approach for the prediction of urban SO2 concentrations in the Istanbul metropolitan area. Int J Environ Pollut 40(4): 301–321

Andres JD, Lorca P, Juez FJ, Sánchez-Lasheras F (2010) Bankruptcy forecasting: a hybrid approach using Fuzzy c-means clustering and Multivariate Adaptive Regression Splines (MARS). Expert Syst Appl 38:1866–1875

Aneja VP, Agarwal A, Roelle PA, Phillips SB, Tong Q, Watkins N, Yablonsky R (2001) Measurements and analysis of criteria pollutants in New Delhi, India. Environ Int 27(1):35–42

Antanasijević DZ, Pocajt VV, Povrenović DS, Ristić MD, Perić-Grujić AA (2013) PM_{10} emission forecasting using artificial neural networks and genetic algorithm input variable optimization. Sci Total Environ 443:511–519

Bera P, Prasher SO, Patel RM, Madani A, Lacroix R, Gaynor JD, Tan SC, Kim SH (2006) Application of MARS in simulating pesticide concentrations in soil. Trans Am Soc Agric Eng 49:297–307

Corchado E, Herrero E (2011) Neural visualization of network traffic data for intrusion detection. Appl Soft Comput 11(2):2042–2056 CrossRef

Corchado E, Arroyo A, Tricio V (2011) Soft computing models to identify typical meteorological days. Logic J IGPL 19(2):373–383 MathSciNetCrossRef

Cortes C, Vapnik V (1995) Support vector networks. Mach Learn 20: 273–297

Etemad-Shahidi A, Mahjoobi J (2009) Comparison between M5′ model tree and neural networks for prediction of significant wave height in Lake Superior. Ocean Eng 36:1175–1181. doi:10.1016/j.oceaneng.2009.08.008

EPA: Environmental Protection Agency Clean Air Interstate Rule. Accessed December 10, 2015a. (http://earthobservatory.nasa.gov/IOTD/view.php?id=87182)

EPA: Environmental Protection Agency Acid Rain Program. Accessed December 10, 2015b. (http://earthobservatory.nasa.gov/IOTD/view.php?id=82626)

Friedman JH (1991) Multivariate adaptive regression splines. Ann Stat 19:1–67

Ganguly D, Jayaraman A, Rajesh TA, Gadhavi H (2006) Wintertime aerosol properties during foggy and nonfoggy days over Urban Center Delhi and their implications for shortwave radiative forcing. J Geophys Res 111:D15217. doi:10.1029/2005JD007029

Gennaro G, Trizio L, Gilio A, Pey J, Pérez N, Cusack M, Alastuey A, Querol X (2013) Neural network model for the prediction of PM_{10} daily concentrations in two sites in the Western Mediterranean. Sci Total Environ 463-464:875–883

Goyal P (2003) Present scenario of air quality in Delhi: a case study of CNG implementation. Atmos Environ 37(38):5423–5431

Goyal MK, Bharti B, Quilty J, Adamowski J, Pandey A (2014) Modeling of daily pan evaporation in sub-tropical climates using ANN, LS-SVR, Fuzzy Logic, and ANFIS. Expert 516 Syst Appl 41(11):5267–5276

Gurjar BR, Van Aardenne JA, Lelieveld J, Mohan M (2004) Emission estimates and trends (1990–2000) for megacity Delhi and implications. Atmos Environ 38(33):5663–5681

Gurjar BR, Jain A, Sharma A, Agarwal A, Gupta P, Nagpure AS, Lelieveld J (2010) Human health risks in megacities due to air pollution. Atmos Environ 44(36):4606–4613

Guven A, Kisi O (2011) Daily pan evaporation modeling using linear genetic programming technique. Irrig Sci 29(2):135–145

Kim S, Shiri J, Kisi O (2012) Pan evaporation modeling using neural computing approach for different climatic zones. Water Resour Manag 26(11):3231–3249

Kim S, Shiri J, Singh VP, Kisi Ö, Landeras G (2015) Predicting daily pan evaporation by soft computing models with limited climatic data. Hydrol Sci J 60(6):1120–1136

Kisi O (2009a) Daily pan evaporation modelling using multi-layer perceptrons and radial basis neural networks. Hydrol Process 23(2):213–223

Kisi O (2009b) Fuzzy genetic approach for modeling reference evapotranspiration. J irrig 538 drainage eng 136(3):175–183

Kisi O (2015) Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree. J Hydrol 528:312–320

Kisi O, Cengiz TM (2013) Fuzzy genetic approach for estimating reference evapotranspiration of Turkey: Mediterranean region. Water Resour Manag 27:3541–3553

Kisi, O and Parmar, K. S, 2016. Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution. J Hydrology.

Kisi O, Tombul M (2013) Modeling monthly pan evaporations using fuzzy genetic approach. J Hydrol 477:203–212

Kisi O, Genc O, Dinc S, Zounemat-Kermani M (2016) Daily pan evaporation modeling using chi-squared automatic interaction detector, neural networks, classification and regression tree. Comput Elect Agr 122:112–117

Krotkov NA, McLinden CA, Li C, Lamsal LN, Celarier EA, Marchenko SV, Swartz WH, Bucsela EJ, Joiner J, Duncan BN, Boersma KF, Veefkind JP, Levelt PF, Fioletov VE, Dickerson RR, He H, Lu Z, Streets DG (2016) Aura OMI observations of regional SO2 and NO2 pollution changes from 2005 to 2015. Atmos Chem Phys 16: 4605–4629. doi:10.5194/acp-16-4605-2016

Mitchell TM (1997) Machine learning. The McGraw-Hill Companies, Inc., New York 414

Mohan M, Kandya A (2007) An analysis of the annual and seasonal trends of air quality index of Delhi. Environ Monit Assess 131(1–3):267–277

Pal M, Deswal S (2009) M5 model tree based modelling of reference evapotranspiration. Hydrol Process 23:1437–1443

Parmar KS, Bhardwaj R (2014) River water prediction modeling using neural networks, fuzzy and wavelet coupled model. Water Resour Manage 29(1):17–33

Parmar KS, Soni K, Kumar N, Kapoor S (2016) Statistical variability comparison in MODIS and AERONET derived aerosol optical depth over Indo-Gangetic Plains using time series modeling. Sci Total Environ:553

Prasad AK, Singh S, Chauhan SS, Srivastava MK, Singh RP, Singh R (2007) Aerosol radiative forcing over the Indo-Gangetic Plains during major dust storms. Atmos Environ 41(6289–6301):2007

Quinlan JR (1992) Learning with continuous classes. In proceedings of the Fifth Australian Joint Conference on Artificial Intelligence, Hobart, Australia, 16–18 November. World Scientific, Singapore, pp 343–348

Rizwan SA, Nongkynrith B, Gupta SK (2013) Air pollution in Delhi: its magnitude and effects on health. Indian J Community Med 38:4–8 http://www.ijcm.org.in/text.asp?2013/38/1/4/106617

Sahin U, Ucan ON, Bayat C, Oztorun N (2005) Modeling of SO2 distribution in Istanbul using artificial neural networks. Environ Model Assess 10:135–142

Sahin UA, Ucan ON, Bayat C, Tolluoglu O (2011) A new approach to prediction of SO2 and PM10 concentrations in Istanbul, Turkey: cellular neural network (CNN). Environ Forensic 12(3):253–269

Seinfeld JH, Pandis SN (2006) Atmospheric chemistry and physics: from air pollution to climate change, vol 2006, 2nd edn. John Wiley & Sons, Hoboken

Sephton P (2001) Forecasting recessions: can we do better on MARS? Federal Reserve Bank of St. Louis Rev 83:39–49

Shafaei M, Kisi O (2016) Lake level forecasting using wavelet-SVR, wavelet-ANFIS and wavelet-ARMA conjunction models. Water Resour Manag 30(1):79–97. doi:10.1007/s11269-015-1147-z

Singh S, Nath S, Kohli R, Singh R (2005) Aerosols over Delhi during pre-monsoon months: characteristics and effects on surface radiation forcing. Geophys Res Lett 32:L13808

Singh S, Soni K, Bano T, Tanwar RS, Nath S, Arya BC (2010) Clear-sky direct aerosol radiative forcing variations over mega-city Delhi. Ann Geophys 28:1157–1166

Smola JA, Bernhard S (2004) A tutorial on support vector regression. Stat Comput 14:199–222

Solomatine DP, Xue Y (2004) M5 model trees compared to neural networks: application to flood forecasting in the upper reach of the Huai River in China. J Hydrol Eng 9:491–501

Soni K, Kapoor S, Parmar KS, Kaskaoutis DG (2014) Statistical analysis of aerosols over the Gangetic–Himalayan region using ARIMA model based on long-term MODIS observations. Atmos Res 149: 174–119. doi:10.1016/j.atmosres.2014.05.025

Soni K, Parmar KS, Kapoor S (2015) Time series model prediction and trend variability of aerosol optical depth over coal mines in India. Environ Sci Pollut Res 22:3652–3671

Suykens JAK (2001) Support vector machines: a nonlinear modeling and control perspective. Eur J Control 7:311–327

Suykens JAK, Vandewalle J (1999) Least square support vector machine classifiers. Neural Process Lett 9:293–300

Vaidya V, Park JH, Arabnia HR, Pedrycz W, Peng S (2012) Bio-inspired computing for hybrid information technology. Soft Comput 16(3): 367–368

Voukantsis D, Karatzas K, Kukkonen J, Räsänen T, Karppinen A, Kolehmainen M (2011) Intercomparison of air quality data using principal component analysis, and forecasting of PM_{10} and $PM_{2.5}$ concentrations using artificial neural networks, in Thessaloniki and Helsinki. Sci Total Environ 409:1266–1276

Wanga P, Liu Y, Qin Z, Zhang G (2015) A novel hybrid forecasting model for PM_{10} and SO_2 daily concentrations. Sci Total Environ 505: 1202–1212