Atmospheric Environment 102 (2015) 239–248

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Artificial intelligence based approach to forecast $PM_{2.5}$ during haze episodes: A case study of Delhi, India

Dhirendra Mishra*, P. Goyal, Abhishek Upadhyay

Centre for Atmospheric Sciences, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India

HIGHLIGHTS

- Neural network and fuzzy logic are combined for forecasting of $PM_{2.5}$ during haze conditions.
- The haze occurs when the level of $PM_{2.5}$ is more than 50 $\mu g/m^3$ and relative humidity is less than 90%.
- Neuro-fuzzy model is capable for better forecasting of haze episodes over urbanized area than ANN and MLR models.

ARTICLE INFO

Article history:
Received 17 September 2014
Received in revised form
29 October 2014
Accepted 23 November 2014
Available online 29 November 2014

Keywords:
Artificial Neural Network
Neuro-Fuzzy logic
Multiple Linear Regression
Haze episode
Statistical analysis

ABSTRACT

Delhi has been listed as the worst performer across the world with respect to the presence of alarmingly high level of haze episodes, exposing the residents here to a host of diseases including respiratory disease, chronic obstructive pulmonary disorder and lung cancer. This study aimed to analyze the haze episodes in a year and to develop the forecasting methodologies for it. The air pollutants, e.g., CO, O_3, NO_2, SO_2, $PM_{2.5}$ as well as meteorological parameters (pressure, temperature, wind speed, wind direction index, relative humidity, visibility, dew point temperature, etc.) have been used in the present study to analyze the haze episodes in Delhi urban area. The nature of these episodes, their possible causes, and their major features are discussed in terms of fine particulate matter ($PM_{2.5}$) and relative humidity. The correlation matrix shows that temperature, pressure, wind speed, O_3, and dew point temperature are the dominating variables for $PM_{2.5}$ concentrations in Delhi. The hour-by-hour analysis of past data pattern at different monitoring stations suggest that the haze hours were occurred approximately 48% of the total observed hours in the year, 2012 over Delhi urban area. The haze hour forecasting models in terms of $PM_{2.5}$ concentrations (more than 50 $\mu g/m^3$) and relative humidity (less than 90%) have been developed through artificial intelligence based Neuro-Fuzzy (NF) techniques and compared with the other modeling techniques e.g., multiple linear regression (MLR), and artificial neural network (ANN). The haze hour's data for nine months, i.e. from January to September have been chosen for training and remaining three months, i.e., October to December in the year 2012 are chosen for validation of the developed models. The forecasted results are compared with the observed values with different statistical measures, e.g., correlation coefficients (R), normalized mean square error (NMSE), fractional bias (FB) and index of agreement (IOA). The performed analysis has indicated that R has values 0.25 for MLR, 0.53 for ANN, and NF: 0.72, between the observed and predicted $PM_{2.5}$ concentrations during haze hours invalidation period. The results show that the artificial intelligence implementations have a more reasonable agreement with the observed values. Finally, it can be concluded that the most convincing advantage of artificial intelligence based NF model is capable for better forecasting of haze episodes in Delhi urban area than ANN and MLR models.

1. Introduction

People with health problems are being warned to avoid strenuous activity as the air pollution, increased to high levels in many parts of the world. The poor air quality is the result of gaseous and

* Corresponding author.
E-mail addresses: dhirendra.mishra@cas.iitd.ac.in (D. Mishra), pramila@cas.iitd.ernet.in (P. Goyal), abhiupadhyay777@gmail.com (A. Upadhyay).

http://dx.doi.org/10.1016/j.atmosenv.2014.11.050
1352-2310/

particulate emissions from various natural and anthropogenic sources. A huge amount dust is being added to the ambient air from natural activity. Combustion activities in vehicles, power plants, wood burning and industrial processes, etc. are major source for anthropogenic emissions. One of the most harmful pollutants is grouped in a category known as particulate matter (PM), which is found in the ambient air including dust, soot, dirt, smoke and liquid droplets. The size of particles varies widely. Small particles remain suspended in the air for a longer period avoids their washout. Particles less than 2.5 μm in diameter ($PM_{2.5}$) are referred to as "fine particles" and expose high health risks (EPA, 2012). Because of their small size, these particles (approximately 1/30th the average width of a human hair) can lodge deeply into the lungs. Roughly one out of every three people in the world is at a higher risk of experiencing $PM_{2.5}$ related health effects. Exposure of $PM_{2.5}$ for eight hours daily for 10 years can cause 2170 people early deaths per million populations due to acute diseases such as respiratory ailments (Chandra, 2013). The long-term exposure of $PM_{2.5}$ is associated with an increase in the long-term risk of cardiopulmonary mortality by 6–13% as per 10 $\mu g/m^3$ increases of $PM_{2.5}$ concentrations in the ambient atmosphere as well as the haze formation criteria (Watson, 2002). The haze episode occurs when the hourly $PM_{2.5}$ concentration exceeds the National Ambient Air Quality Standards (NAAQS) for some consecutive hours. Thus, the potentially deadly pollutant, $PM_{2.5}$, is one of the reasons for the haze-filled air around appears murkier and more acrid-smelling (Chameides et al., 1999). Haze is defined as the phenomenon that leads to atmospheric visibility less than 10 km due to suspended particles, smoke, and vapor in the atmosphere (Zhang et al., 2013) and is caused when sunlight encounters tiny pollution particles in the air, which reduce the clarity and color of what we see, and particularly during humid conditions. Haze is increased, or visibility is reduced, by the absorption and scattering of gases and aerosols (particles) in the atmosphere. Therefore, haze involves both the physical interaction of light with the gases and particles as well as psychophysical processes. Hazy conditions have been studied intensively for its impact on air quality, visibility, climate, and public health (Kim et al., 2001; Chameides et al., 1999; Okada et al., 2001). The atmospheric condition is relatively more suitable during haze episodes, which results in worse accumulation of atmospheric particulate and light extinction. It often occurs when dust and smoke particles accumulate in relatively dry air. Basically, the sources for the formation of haze include traffic, industry, power plants, and wildfires. When weather conditions block the dispersal of smoke and other pollutants they concentrate and form a usually low-hanging shroud that impairs visibility and may become a respiratory health threat. Industrial pollution can result in dense haze, which is known as smog. It is estimated that more than three million deaths occur globally every year due to air pollution, mainly by particulate matter (WHO, 2010). It has been observed that humans are harmed more by long-term exposure to levels of $PM_{2.5}$ approaching 60 $\mu g/m^3$ rather than short-term exposure over 150 $\mu g/m^3$ (NEA, 2012). This consideration suggests that a haze episode can be optimally defined as a case in which the hourly $PM_{2.5}$ concentrations exceed a long term standard. Thus, in the present study, we consider a "haze episode" as a case in which the hourly $PM_{2.5}$ concentrations exceed 50 $\mu g/m^3$ level. People exposed to air pollutants for longer durations may run an increased risk of cancer or serious health disorders. Therefore, the forecasting of Haze episode is much needed in any urban area.

Generally, forecasting models can be developed through two ways, physical and statistical approaches. Physical approach models the underlying physical processes related to variables directly, whereas statistical approaches determine relationships between historical data sets. Statistical forecasting starts with certain assumptions based on the experience, knowledge and judgment. The advantages of a statistical forecasting are that it helps to attempts the uncertainty of the future, relying mainly on data from the past to present and analysis of trends. It has become the fundamental basis for informed decision-making related to forecasting in many areas. In the past, there have been many attempts to forecast $PM_{2.5}$ concentrations as well as the concentrations of other pollutants in both urban and no urban area. Specifically, Kumar and Goyal (2013) presented analytical modeling relating the air pollutant concentrations in Delhi area with various meteorological variables. Mishra and Goyal (2015) developed the reliable PCA-ANN forecasting model at TajMahal, Agra. The evaluation of the results of the developed ANN models, indicates that the models surpassed in comparison with the regression models viz. MLR. Kumar and Goyal (2011) developed a prognostic model in order to predict the daily air quality index by using PCA and auto regression integrated moving average (ARIMA). Corani (2005) applied feed forward neural network, pruned neural network, and lazy learning in order to predict at 9 AM the concentration estimated for the current day for ozone and PM_{10} in Milan, Italy. Lack of experience data, entangled cause-and-effect relationships and imprecise data make it difficult to assess the degree of exposure to certain forecasting types using only traditional statistical models e.g. MLR. But if theses data are available, then it may be beneficial to build and implement more appropriate operational risk models using a newer approach such as Neuro-Fuzzy logic. Heo and Kim (2004), described the method of forecasting daily maximum ozone concentrations at four monitoring sites in Seoul, Korea. The forecasting tools are developed by the combination of fuzzy expert and neural network systems. The forecasting of the concentrations of pollutants can be considered as a non-linear regression problem between predictors (such as meteorological and air quality variables) and predictand (in the present study, hourly concentration of $PM_{2.5}$). The time series forecasting by using different methods, including ANN approaches due to the significant properties of handling nonlinear data with self-learning capabilities (Hornik, 1991) can forecast the concentrations, but it cannot be known the degree that an input influence the output (Pao, 1989). While, fuzzy logic is an effective rule-based modeling in soft computing that not only tolerates imprecise information, but also makes a framework of approximate reasoning. The disadvantage of fuzzy logic is the lack of self-learning capability. But, the combination of fuzzy logic and neural network can overcome the disadvantages of the above approaches. In the NF model, both the learning capabilities of a neural network and reasoning capabilities of fuzzy logic is combined in order to give enhanced prediction capabilities, as compared to using a single methodology alone. The NF model is the basically a combination of neural network with back propagation network and fuzzy logic with crisp values.

New Delhi, the capital of India, $PM_{2.5}$ levels has been observed more than at 500 $\mu g/m^3$ between November and February months at many locations. Which has been found to be exceeding the NAAQS limits (60 $\mu g/m^3$) in all areas. During the same time period in Beijing, the most polluted city, the $PM_{2.5}$ level was never higher than 400 $\mu g/m^3$, i.e. significantly less than that seen in Delhi. Residents of Delhi, who were breathing easier after the introduction of CNG vehicles in the past few years, have reason to worry about the air they inhale (Mishra and Goyal, 2014). Since, the haze or smog in Delhi is due to the concentrations of $PM_{2.5}$ pollutants in the ambient atmosphere (George et al., 2013). Therefore, the forecasting of $PM_{2.5}$ concentrations can be formulated as the search of a suitable mapping between the set of available meteorological data and the selected set of pollutant parameters. It is also observed that in many cases where external variables are not available or not reliable then this kind of forecasting through models are not easily

solvable. Therefore, the objective of the present study is to develop statistical as well as the artificial intelligence models to forecast haze episodes in Delhi. These models are also applicable for higher levels of $PM_{2.5}$ concentrations.

2. Material and methods

2.1. Study area

Fig. 1 show the two-selected monitoring location: Indira Gandhi International (IGI), Airport and Income Tax Office (ITO) in Delhi, India. To select these monitoring locations, both the geographical locations and the history of high haze episodes were considered. ITO monitoring station represents a heavy vehicular traffic (Goyal et al., 2013) area in the northeast part of Delhi. Monitoring station, IGI, Airport, located on the southwest, is well known to record the most frequent high levels of haze episodes. Several roads, including one national highway (NH8) so that the site is vulnerable to the build-up of pollutants surround this. The numbers of vehicles plying on NH8 were approximately 108918 in a day and making it the busiest highway in the subcontinent (DTCP, 2010). Also, IGI Airport is located in the downstream of semi-industrialized area, Gurgaon with factories, is a potential source of local air pollution. In this study, the forecasting models have been developed and analyzed for IGI, Airport data.

2.2. Database preparation

Delhi is a rapidly growing city with rapid changes in air pollution emissions. The last year's data may be significantly different than recent years data. Hence, this study uses the data of the past one-year for the model development. The observed concentrations of $PM_{2.5}$ have been used at both monitoring stations, with the aim to develop the forecasting models for haze episode. Firstly, the daily concentrations of $PM_{2.5}$ have been analyzed. The monthly maximum, minimum and the monthly average values the year 2012 at IGI, Airport and ITO have been shown in Fig. 2. The variations of concentrations are observed to be almost the same in all the months. The minimum values are observed during July, August and September (monsoon season) and maximum during October and November (post-monsoon season) (Goyal et al., 2014). It is noticeable that they all (maximum, minimum and average)values of $PM_{2.5}$ concentrations are always higher at IGI, Airport than ITO, which imply that IGI, Airport is most polluted region in Delhi and favoring the haze episodes. It is also noticeable that in general all the concentration values of $PM_{2.5}$ at IGI, Airport exceed the prescribed NAAQS in every month. Therefore, haze episode at IGI, Airport has been chosen for the development of $PM_{2.5}$ forecasting models.

In order to develop the domain knowledge for the intelligent haze level forecasting system, hourly air quality is acquired from government agency Central Pollution Control Board (CPCB), New Delhi and meteorological data from University of Wyoming web archive for the year 2012. These data have been processed into an hourly format, and used to construct data series for only for haze hours. The entire table consists of 4126 rows, or 4126 haze hours out of the total 8601 h of observations, which are approximately 48% of the total data. However, the rows containing any incomplete information are removed from the data series and finally only 3813 authentic hours with haze conditions from total available 8601 h in the year 2012 have been chosen for the present study. Fig. 3 shows the monthly variations of haze, RH < 90% and $PM_{2.5}$ > 50 μg/m^3

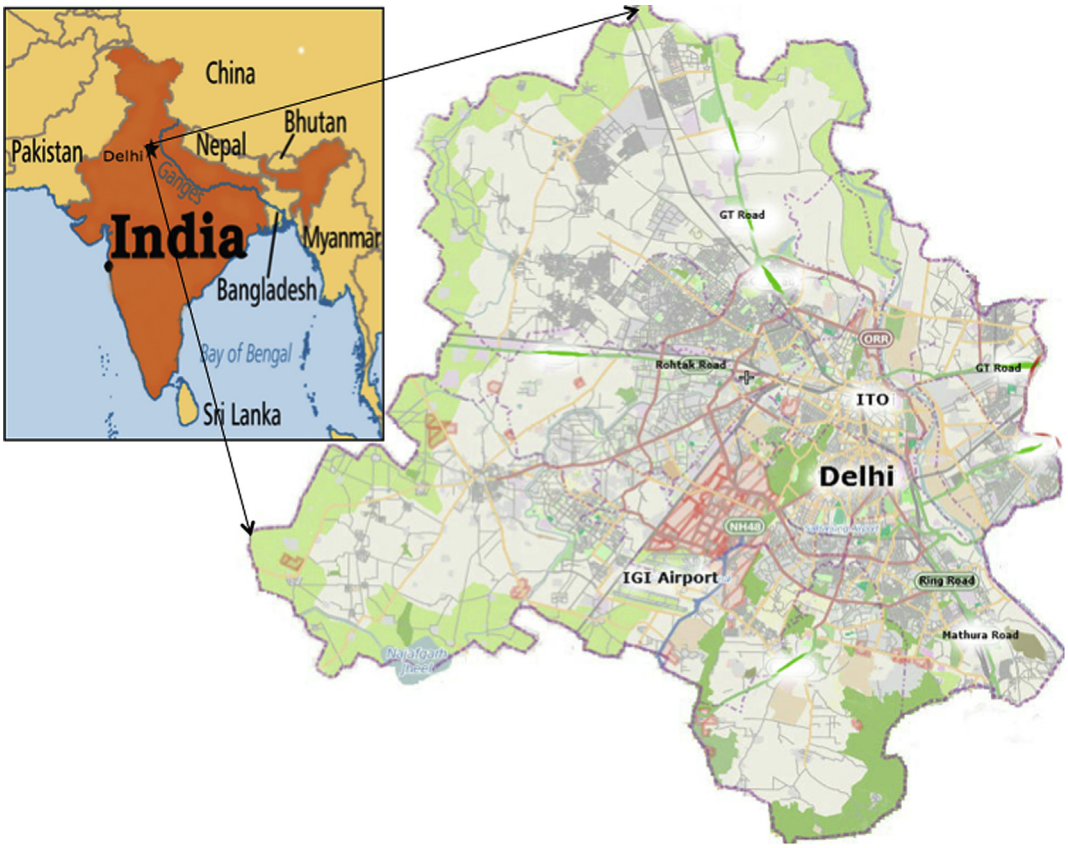

Fig. 1. Study area of Delhi with CPCB monitoring stations ITO and IGI, Airport.

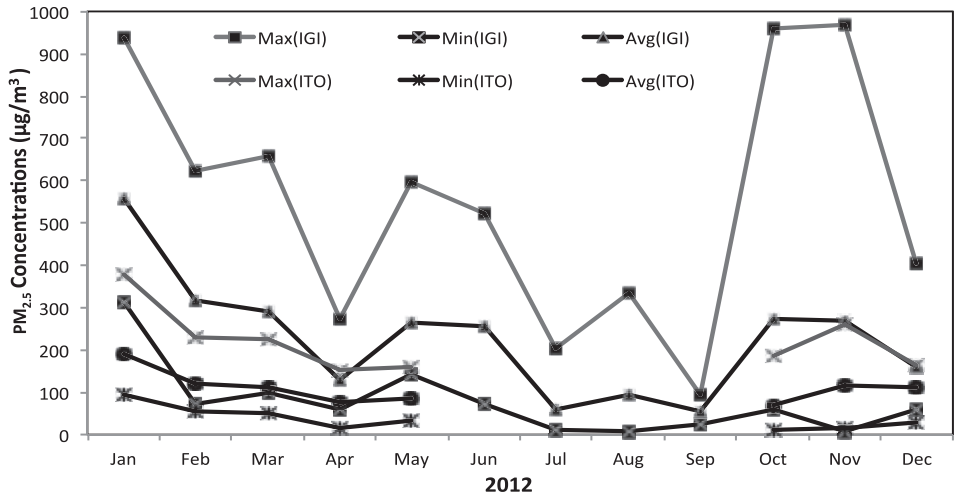

Fig. 2. Monthly variation of $PM_{2.5}$ concentrations at CPCB monitoring stations.

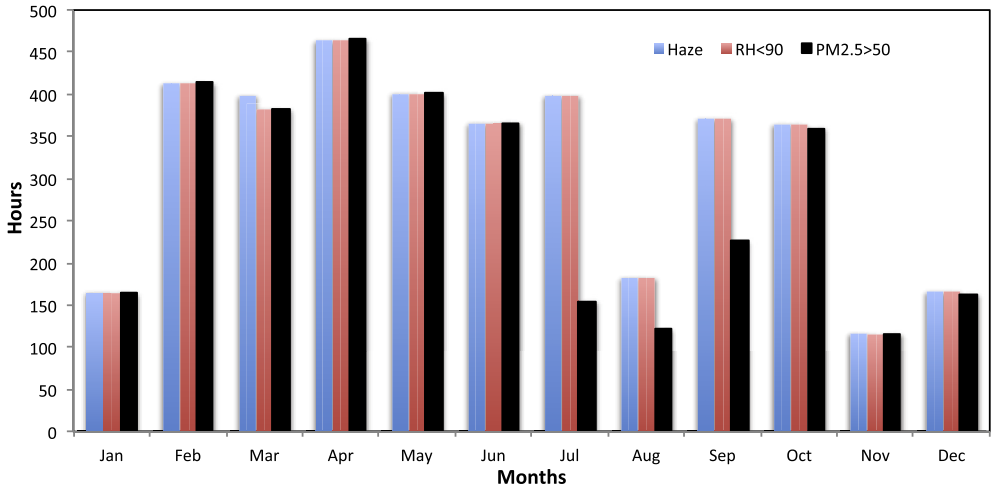

Fig. 3. Monthly observed haze hours at IGI, Airport, Delhi.

hours in the year 2012 at IGI, Airport. The maximum haze, RH < 90% and $PM_{2.5}$ > 50 µg/m^3 hours are observed in April months and minimum in November respectively. It can be concluded that the concentrations of fine particles greater than NAAQS level and relative humidity less than 90% are directly related to haze episodes over the urbanized area of Delhi in each month except monsoon seasons. The maximum haze hours were observed in April and May due to the dry atmosphere with both favorable conditions while the fog phenomenon occurs in winter months i.e. November, December and January. Haze often occurs when dust and smoke particles accumulate in relatively dry air while fog is due to suspension of water droplets in the air near to the surface. Also, fog is the local phenomenon, but the haze is no longer a domestic problem. The concentrations of $PM_{2.5}$ pollutant have also been found to be the maximum during the same month and these values are observed to be exceeded more than 5–6 times by the NAAQS. The columns represent 11 condition variables consist of last-day's same hour $PM_{2.5}$ concentration, CO, NO_x, SO_2, O_3, temperature, wind speed, wind direction index, relative humidity, visibility and dew point temperature. Further, the data series of the year 2012 have been divided for training and validation process. The initial 3164 observation hours of January to September are used for model training and rest 649 observation haze hours of October to December are used for validation. Data normalization is necessary for the artificial intelligence algorithms. All the data are normalized to the range of [0, 1] by linear scaling. In other words, the input and output data would be converted to values between zero and one.

2.3. Multiple linear regression (MLR)

A MLR technique to be used for forecasting can be expressed as a function of a certain number of factors that includes one dependent variable to be predicted and two or more independent variables. In general, multiple linear regression can be expressed as in Equation (1):

$$Y = b_1 + b_2X_2 + \ldots + b_k X_k + e \quad (1)$$

where Y is the dependent variable, X_2, X_3…,X_k are the independent variables, b_1, b_2…,b_k are linear regression parameters. In this study, $PM_{2.5}$ is the dependent variable, concentrations of air pollutant and meteorological variables are independent variables, e is an estimated error term which is obtained from independent random sampling from the normal distribution with mean zero and constant variance. The task of regression modeling is to estimate the b_1, b_2…,b_k which can be done using the least square error technique.

2.4. Artificial neural network (ANN)

The forecasting of air quality can be considered as a non-linear regression problem between predictand (here, hourly concentration) and predictors (such as meteorological and air quality variables). ANN models are capable of approximating any smooth differentiable function and used for modeling complex non-linear processes. ANN models have been utilized for several tasks within the air quality domain, such as forecasting, function approximation and pattern classification (Gardner and Dorling, 1999). Multi Layer perceptron's have been applied successfully to solve some difficult and diverse problems, by training them in a supervised manner with a highly popular algorithm. In this present study, ANN model has been used to forecast, hourly concentrations of $PM_{2.5}$ in the Delhi megacity. Like other studies, it has been observed that the concentration of pollutants present continuously throughout the day, but hourly variations of concentrations affecting much more to develop statistical models. Therefore, the hourly data has been chosen for training and validation.

The neuron is the basic information processing unit of an ANN. It consists a set of links, describing the neuron inputs, with weights W_1, W_2, …, W_m and an adder function (linear combiner) for computing the weighted sum of the inputs i.e. $u = \sum_{j=1}^{m} W_j X_j$. Finally, activation function φ for limiting the amplitude of the neuron output, i.e. $y = \varphi(u + b)$, where 'b' denotes bias. All the processes are depicted in Fig. 4.

The combination of neurons into multilayer structures gives the power of pattern recognition and prediction. The multi layer feed-forward network comprises of an input layer, hidden layer and output layer. Specifically, the input layer is a layer that is directly connected to outside information. All data in the input layer will be feed-forwarded to the hidden layer as the next layer. Meanwhile, the hidden layer functions as feature detectors of input signals and releases them to the output layer. Finally, the output layer is considered as a collector of the features detected and as a producer of the response. In the networks, the output from output layer is the function of the linear combination of hidden unit's activation; whereas the hidden unit's activation function is in the form of a non-linear function of the weighted sum of inputs. Under the assumption that concentration pattern does not change significantly from one day to the next, the proposed model can be used to forecast concentrations for the consecutive hour by providing values of new predictor variables.

2.5. Neuro-fuzzy (NF) modeling: application of artificial intelligence techniques

Forecasting of high concentration of pollutants, e.g., haze episode is a complex phenomenon. Despite many methods usually used for forecasting of pollutants, none of them is commonly accepted and does not give the satisfactory results. In fact, most of the results come from physical methods, which may not be reliable because of problems with obtaining credible data of pollutants, especially those coming from the communal sources and traffic, it is hard to use them in operational modeling. For this reason, the combination of neural network and fuzzy logic methods, are used in the present study, which do not affect the physical part of the phenomenon, but allow to prepare properly and fast forecasting by means of discovering the unknown correlation between collected data. Fuzzy logic and neural networks are natural complementary tools in building intelligent systems. The neural networks are low-level computational structures that perform well when dealing with raw data and fuzzy logic deals with reasoning on a higher level, using linguistic information, acquired from domain experts. Integrated Neuro-Fuzzy systems can combine the parallel computation and learning abilities of neural networks with the human-like knowledge representation and explanation abilities of fuzzy systems. As a result, neural networks become more transparent, while fuzzy systems become capable of learning. A fuzzy system is prepared through IF-THEN rule on the basis of membership functions defined for input and output variables of the system. This fuzzy system is trained on neural network on the basis of the input data. The structure of a Neuro-Fuzzy system is similar to a multi-layer neural network. In general, a Neuro-Fuzzy system has: (i) input and output layers, (ii) three hidden layers that represent membership functions and fuzzy rules. The selection of membership function (type and number) depends on characteristics of input and output variables that can be decided by experts on the basis of experiment, observation and experience. Fig. 5 shows the architecture of Neuro-Fuzzy structure. The layer wise descriptions are given in Appendix.

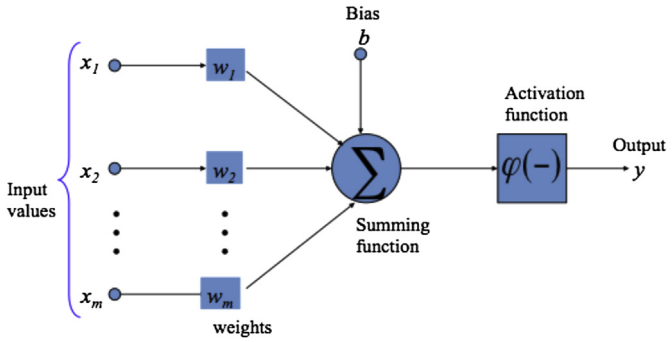

Fig. 4. Schematic structure of artificial neural network.

3. Results and discussion

3.1. Variations of $PM_{2.5}$ concentrations with local meteorological conditions

The occurrences of severe haze episodes are found to be associated with high concentrations of $PM_{2.5}$ in the urban area of Delhi with favorable weather conditions. The episode was occurring in several days in certain months of dry periods as discussed in the previous section. A correlation matrix of $PM_{2.5}$ with other air pollutants and meteorological variables during haze days are shown in Table 1. The negative correlation is visible between the $PM_{2.5}$ concentration with relative humidity, temperature and dew point temperature. Similarly, the positive correlations among the investigated variables, is the positive correlation between CO and pressure. This demonstrates the importance to stabilize the air pollutants in the atmosphere and the continuum of the hazy conditions and the atmosphere tended to be very stable, preventing deep convection and dispersion of pollutant of atmosphere. Haze can also closely relate with the atmospheric visibility, the most effective indicator. However, the fog could reduce visibility in addition to haze (Goyal et al., 2014). It is observed from the whole data that when the relative humidity is greater than 90%, fog is the major factor that reduces visibility, whereas, when relative humidity is less than 90%, haze is formed in the ambient atmosphere. In order to distinguish haze from fog, relative humidity is also analyzed in the present study.

The graphical presentation of $PM_{2.5}$ along with relative humidity, temperature during the continue 46 h and 39 h have been shown in Fig. 6(a) and (b) during April and May respectively. It again shows that there are negatively correlation of $PM_{2.5}$with-temperature, dew point temperature and the relative humidity.

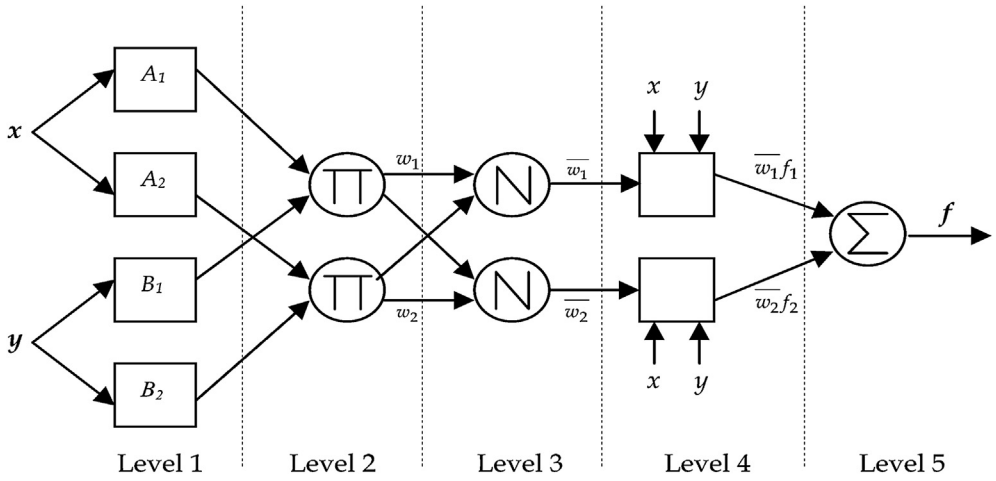

Fig. 5. Artificial Intelligence based the structure of Neuro fuzzy (NF) model.

Table 1
Correlation matrix of air pollutant concentrations and meteorological variables for urban area Delhi, India.

	$PM_{2.5}$	$PM_{2.5}$(p)	Pres	CO	NO_2	O_3	SO_2	T	WDI	DT	RH	WS	Vis	NH_3
$PM_{2.5}$	1.00													
$PM_{2.5}$(p)	0.61	1.00												
Pres	0.27	0.32	1.00											
CO	0.26	0.34	0.01	1.00										
NO_2	0.06	0.11	0.01	0.24	1.00									
O_3	−0.28	−0.29	−0.01	−0.46	−0.29	1.00								
SO_2	0.14	0.21	−0.03	0.29	0.16	−0.18	1.00							
T	−0.62	−0.60	−0.42	−0.38	−0.09	0.43	−0.09	1.00						
WDI	0.00	0.02	0.18	−0.08	−0.11	0.20	−0.19	−0.13	1.00					
DT	−0.31	−0.37	−0.32	−0.16	0.03	−0.08	0.08	0.59	−0.36	1.00				
RH	0.04	−0.01	0.07	0.04	0.05	−0.31	−0.15	−0.17	−0.08	0.42	1.00			
WS	0.28	0.18	0.18	0.08	−0.03	0.04	0.19	−0.28	0.13	−0.29	−0.33	1.00		
Vis	−0.07	0.01	0.05	−0.03	−0.08	0.28	−0.01	−0.04	0.15	−0.44	−0.58	0.38	1.00	
NH_3	−0.08	−0.03	−0.07	0.09	0.16	−0.34	0.14	0.08	−0.28	0.39	0.27	−0.14	−0.23	1.00

While during few hours, the variables are not following negative correlations, which may be due to the influence of other factors. The wind rose diagram for continues 46 haze hours (Fig. 6(c)) is showing the prevailing winds from WNW and the significant frequencies are also observed from the northwest and southwest. This supports the fact of winds that the dust from western side placed Sahara desert may also cause a haze episode in Delhi. Again, also the wind rose diagram for continues 39 haze hours (Fig. 6(d)) showing winds from east and the significant frequencies also from the southeast. The highest concentrations are observed in light to moderate wind conditions, mainly when winds are parallel to the highway. Therefore, the highway seems to play major role in the elevated $PM_{2.5}$ concentrations, with a clear relationship between $PM_{2.5}$ concentrations and wind speed.

3.2. Modeling: training and validation

The haze formation criteria have also been investigated using the hourly average data for the year 2012 to find the linear regression amongst various concerned variables. The MLR model is used to determine the statistically significant regression parameters with the help of windows software SPSS (version 17.0), which achieves the least sum of squared errors with the training set. It is represented by:

$$PM_{2.5} = 50.302 + 0.888 \times PM_{2.5}(p) + 0.182 \times SO_2 - 0.259 \times O_3 - 0.933 \times DT \quad (2)$$

where, DT-dew point temperature, $PM_{2.5}(p)$ − previous hour concentrations of $PM_{2.5}$, O_3 − ozone concentrations, and SO_2 − sulfur dioxide. The above model has been performing well during training and validation periods. The scatter plot between the observed and predicted concentrations of $PM_{2.5}$ pollutants through Equation (2) is shown in Fig. 7(a). The statistical performances of the model are shown in Table 2.

The ANN model is developed to forecast the haze hours in Delhi using the MATLAB8.1 (licensed, IIT Delhi). The model is built with the selected variables to forecast the haze episode in terms of $PM_{2.5}$ concentrations and multilayer perceptron (MLP) architecture is with on hidden layers, eleven neurons, and with a square activation function. The first layer is the input layer, which consist the air pollutants, i.e., CO, NO_2, SO_2, NH_3, O_3 and previous hour's concentration of $PM_{2.5}$ including meteorological variables viz. temperature, relative humidity, wind speed, wind direction index and dew point temperature as the input in the proposed ANN model. The next layers are hidden layers, where the values of neurons in each layer are chosen. Here two hidden layers and different value of neurons are chosen to optimize the ANN performance. The last layer is the output layer, which consists of the target of the

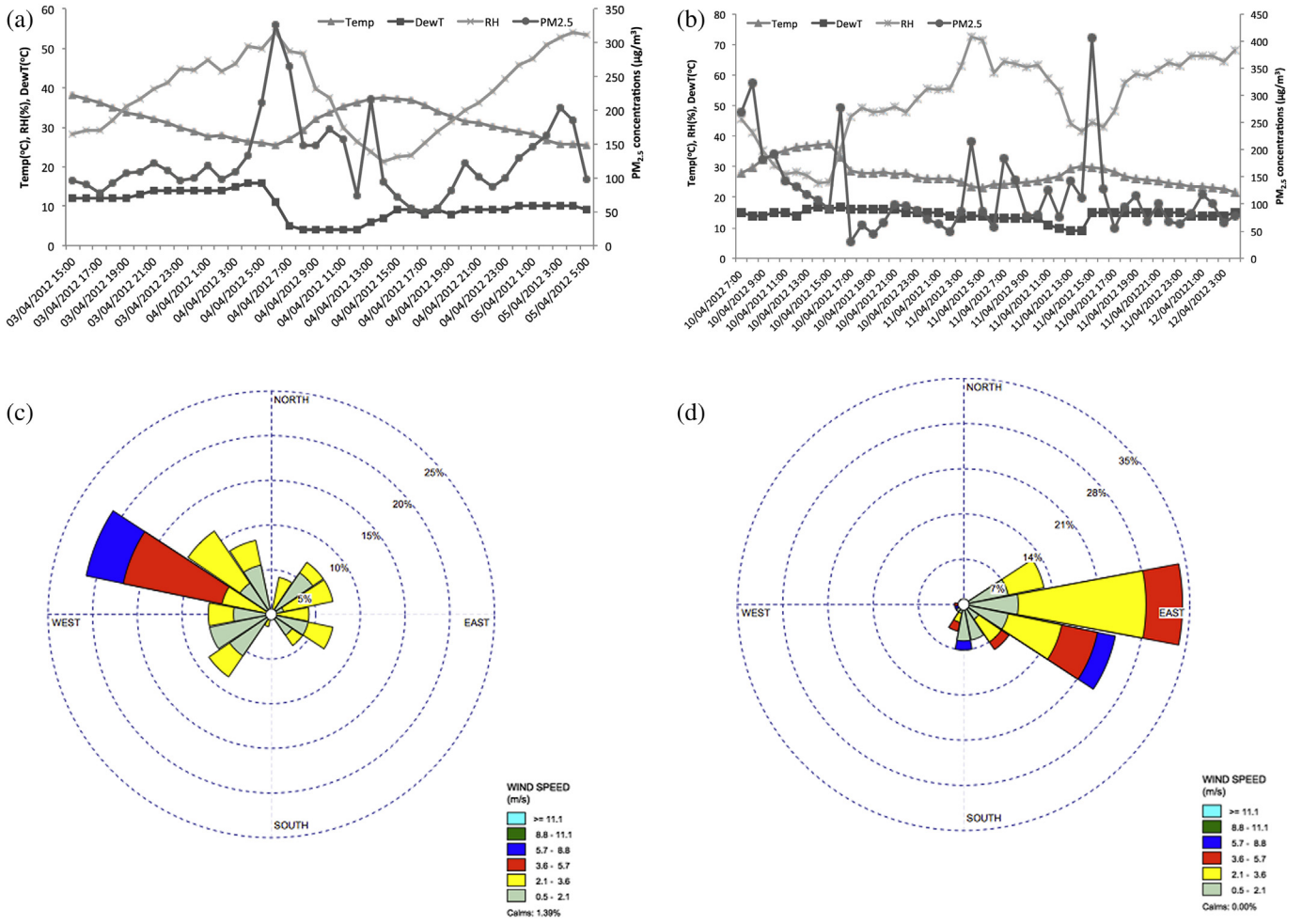

Fig. 6. Variations of $PM_{2.5}$ concentrations with temperature, relative humidity, and dew point temperature during continue (a) 39-h; (b) 46-h haze episode and Wind rose diagram during continue (c) 39-h; (d) 46-h haze episode.

forecasting model. Concerning the ANN model specifications and the way that their results are evaluated, the training process of the ANN models is based on the back-propagation algorithm. The ANN models have been trained with the hourly data of air pollutants and meteorological variables. The hyperbolic tangent sigmoid function uses the transfer function. The scatter plot between the observed and predicted concentrations of $PM_{2.5}$ pollutants during the training phase from the chosen above described ANN model is shown in Fig. 7(b).

Artificial intelligence based NF model has been developed to forecast the haze episode over an urbanized region of Delhi. The training and validation are performed by MATLAB8.1 (licensed IIT Delhi). The selection of 'training data set' is important in the development of model, which is taking the haze hour's data for nine months (January to September 2012). The input layer consists nine parameters containing air pollutants, i.e., CO, NO_2, SO_2, O_3 and previous hour's concentration of $PM_{2.5}$ and meteorological variables viz. temperature, relative humidity, wind speed, and dew point temperature. Once the input data have been loaded, Sugeno FIS has been generated showing input as well as output variables. The FIS has been trained by hybrid algorithm, i.e., backpropagation method and fuzzy logic. The input with "4" categories of Gaussian membership functions have been used for the model development, which are categorized as moderate, bad, very bad, and hazardous. It is trained for 30 epochs and observed that there was a minimum error of 0.94856 after 30 epochs i.e. becomes almost constant. The FIS has generated 262144 rules after training of the artificial intelligence model using 'and' operator. For the training period, the scatter plot between the observed and predicted concentrations of $PM_{2.5}$ pollutants from NF model is shown in Fig. 7(c).

All the above three models have been trained and validated separately with the hourly data of air pollutants and meteorological variables. The performances of models for forecasting of $PM_{2.5}$ pollutants for three months haze hour data (649 h) have been shown in Fig. 8. The models performances are evaluated by the calculation of several performance indexes that are described in Mishra and Goyal (2015). The performances computed between observed and predicted $PM_{2.5}$ concentrations from MLR, ANN and NF models in training and validation phases are given in Table 2. It can be concluded that without the application of fuzzy logic method, ANN presented slightly better performance than MLR.

The value of correlation coefficient (R) between observed and predicted $PM_{2.5}$ concentrations are 0.25, 0.53, and 0.72 for the trained MLR, ANN and NF respectively. The artificial intelligence based NF model is showing a better agreement with the observed values than the other two models. Also, the indexes of agreement (IOA) between have been found to close on its ideal value, i.e., 0.62, 0.78, and 0.80 for MLR, ANN and NF respectively. The values of

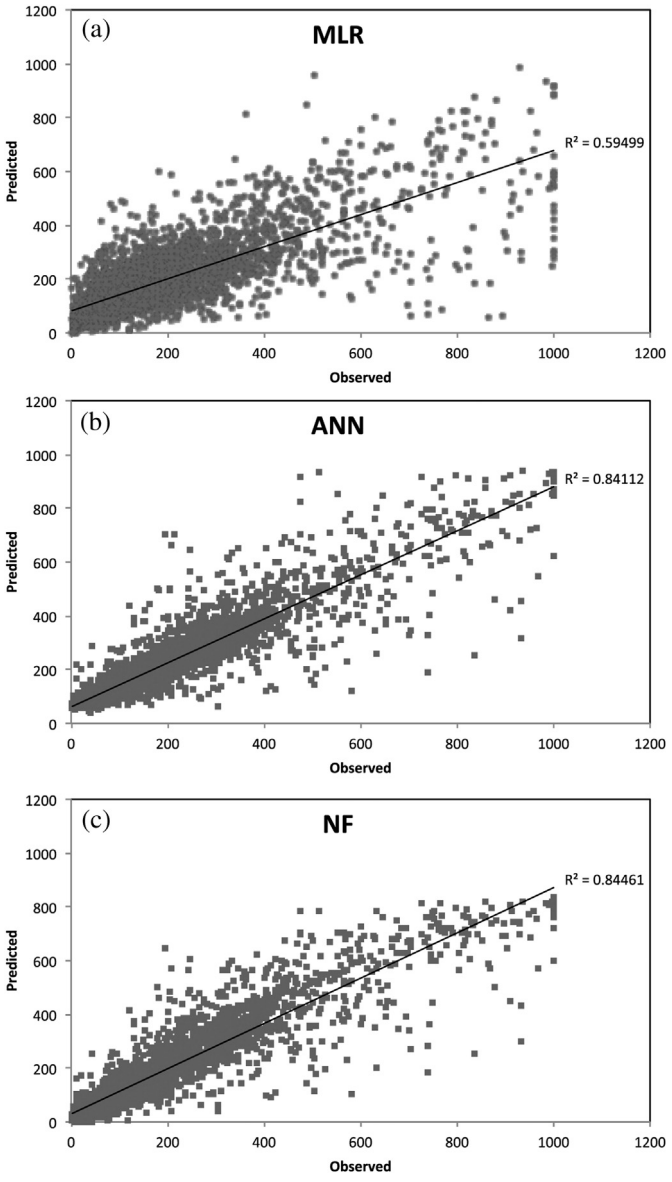

Fig. 7. Scatter plot between the observed and predicted $PM_{2.5}$ concentration ($\mu g/m^3$) for (a) MLR model; (b) ANN model; and (c) NF model.

Table 2
Statistical performance of MLR, ANN and NF models.

	MLR		ANN		NF	
	Training	Validation	Training	Validation	Training	Validation
R	0.77	0.51	0.92	0.53	0.92	0.72
IOA	0.95	0.72	0.98	0.78	0.98	0.80
NMSE	0.165	0.653	0.062	0.641	0.065	0.526
FB	0.058	0.146	0.071	−0.008	0.034	−0.206
FAC2	0.73	0.77	0.81	0.81	0.90	0.84

NMSE have been found as 0.676 for MLR, 0.640 for ANN, and 0.525 for NF. The values of fractional bias (FB) for all models are observed close to 0.0 as well as an under-prediction for ANN and NF models. The values of more than 84% are lying between the factors of two (FAC2) for NF model. Thus, the models are showing the acceptable results and the values of statistical measures for artificial intelligence based NF model is found close to be the corresponding ideal values and it can be concluded that NF model is performing better than other ANN and MLR models. It can also be observed that the high concentration values are not predictable by the proposed model, which may be due to the local anthropogenic emission activities. Furthermore, the improvement of the predictive ability of the constructed NF model could be reached by the use of several other parameters such as solar radiation, sunlight duration, based on the disability of the data over space and time. Overall, the performance of the artificial intelligence based NF model is observed satisfactory. However, the employed input meteorological variables are generally available from routine weather prediction models. Thus, the developed model can be used for operational forecasts of haze episodes over urban areas like Delhi.

4. Conclusions

The present study reveals that the increasing concentrations of fine particulate matter and the restrictive nature of the atmosphere to disperse or transport of pollutants are the reason behind the haze phenomenon. Meteorological conditions would trap the pollutants and contribute to the build-up of pollutants observed during the haze. Several extreme haze episodes were occurring in Delhi urban area during the year 2012. In the present study, the favorable conditions for haze formation for Delhi urban area has been investigated and concluded that either less than 90% of relative humidity or more than 50 $\mu g/m^3$ of the concentrations of $PM_{2.5}$ with calm winds or both simultaneously are responsible for haze formation. An investigation of the meteorological variables during haze episodes revealed a good correlation between the $PM_{2.5}$ concentrations and some variables, e.g., temperature and relative humidity. The April and May months of the year 2012 were recorded the most persistent haze episodes and the least affected by the meteorological variables. The dry weather is also favored for haze formation as the maximum frequency of haze formation was occurring in the summer season of Delhi. An artificial intelligence based NF model has been conducted to develop a prognostic model that could make a reliable forecasting of the high level of $PM_{2.5}$ concentrations in an urban area with high traffic and industrial influences. The most convincing advantage of artificial intelligence based NF model is that the capability of generalization over the test data is higher than ANN and MLR methods. The proposed prognostic model shown precise and very effective predictions compared with the conventional MLR and ANN methods. Finally, it should be mentioned that NF models forecasting ability does present a limited precedence against ANN and MLR models. Clearly, this study has indicated that the artificial intelligence based NF model provided a well suited method and gave promising results for modeling of highly non-linear air pollution problem at urban area like Delhi.

Acknowledgments

The authors acknowledge the financial and infrastructural support provided by parent institute, Indian Institute of Technology Delhi for the present study. The authors also would like to thank CPCB, New Delhi and Larry D. Oolman, University of Wyoming for proving the data required for this study.

Appendix

The membership function for neuro-fuzzy modellingis chosen arbitrary for each input variable and on the basis of trial and error method best combination of membership value and type is assigned for each input variable. For a given model total number of fuzzy rules will be:

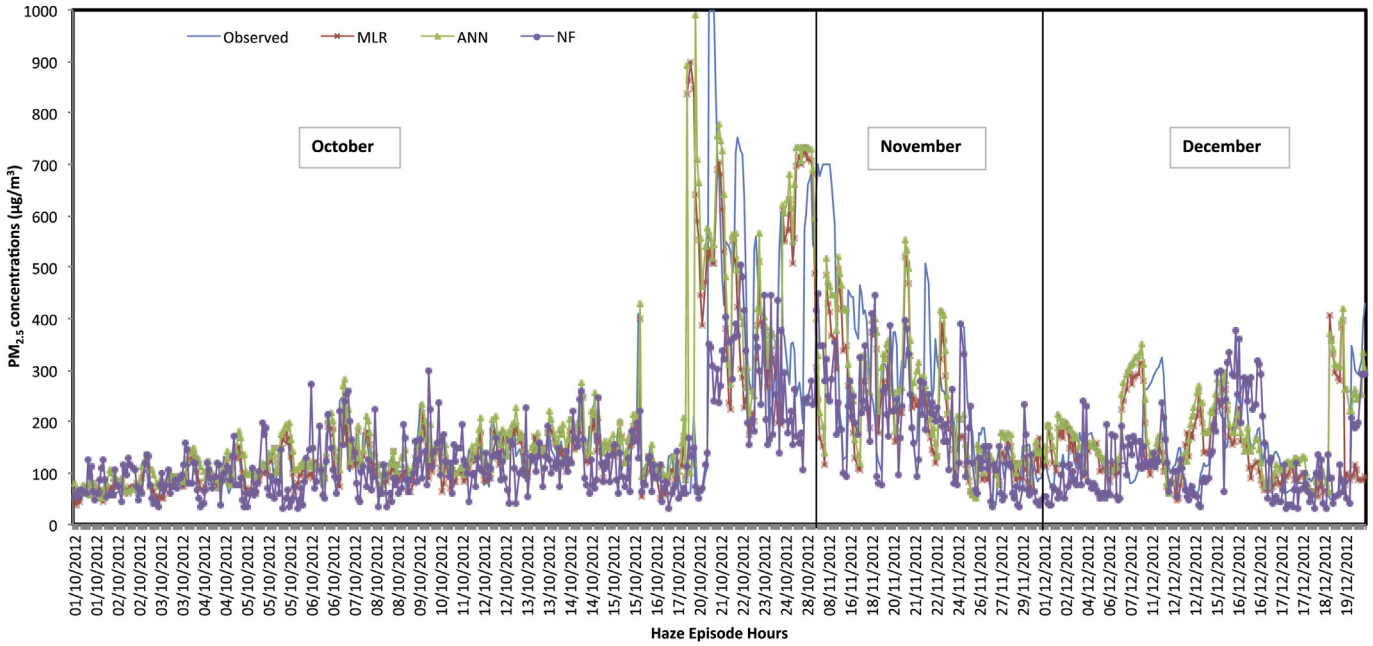

Fig. 8. Comparisons of the developed MLR, ANN and NF models for haze episodes over three months.

$$N = M^n \tag{3}$$

Where '*N*' is number of fuzzy rules, '*M*' is membership value and '*n*' is the number of input variables.

For simplicity, we assumed the fuzzy inference system under consideration has two inputs *x* and *y*, and one output *f*. Suppose that the rule base contains two fuzzy if-then rules of Takagi and Sugenos' type:

$$\text{Rule 1 : If } x \text{ is } A_1 \text{ and } y \text{ is } B_1 \text{ then } f_1 = p_1.x + q_1.y + r_1 \tag{4}$$

$$\text{Rule 2 : If } x \text{ is } A_2 \text{ and } y \text{ is } B_2 \text{ then } f_2 = p2.x + q2.y + r_2 \tag{5}$$

The adoptive Neuro-Fuzzy architecture is depicted in Fig. 5. A neural network structure consists of a number of nodes connected through directional links. Each node is characterized by a node function with fixed or adjustable parameters. The training phase of a neural network is a process to determine optimum parameter values to sufficiently fit the training data. The basic learning rule is the well-known back-propagation method that seeks to minimize some measure of error, usually a sum of squared differences between a network's outputs and desired outputs. The layer wise description is described below:

First layer is the input layer. Each neuron in this layer transmits external crisp signals directly to the next layer. That is,

$$O_i^1(x) = \mu_{A_i}(x) \tag{6}$$

where x – the input to node I, A_i – the linguistic label (small, large, etc.) associated with the node function. In other words, O_i^1 is the membership function of A_i and it specifies the degree to which the given x satisfies the quantifier A_i. Considering the statistical aspect of the air pollution modeling, Gaussian type membership functions are used in this study, described by the following equation:

$$\mu_{i_i} = \exp\left\{ -\left(\frac{x - c_i}{a_i}\right)^2 \right\} \tag{7}$$

Where a_i and c_i are memberships function parameters.

Second layer is the fuzzification layer. Neurons in this layer represent fuzzy sets used in the antecedents of fuzzy rules. A fuzzification neuron receives a crisp input and determines the degree to which this input belongs to the neuron's fuzzy set. Every node in this layer is a circle node labeled π, which multiplies the incoming signal and sends the product out.

$$O_i^2 = w_i = \mu_{A_i}(x) * \mu_{B_i}(y), \quad i = 1, 2\ldots \tag{8}$$

Third layer is a fuzzy rule layer. There is a single fuzzy rule for each neuron in this layer and receives inputs from the fuzzification neurons and represents fuzzy sets in the rule antecedents. The output of this layer is the normalized firing strengths.

$$O_i^3 = \overline{w}i = \frac{w_i}{w_1 + w_2}, \quad i = 1, 2\ldots \tag{9}$$

The total number of all possible fuzzy rules with *n* clauses (each clause is represented by a single input) is determined as

$$S = M^n \times \ T \tag{10}$$

Where *n* is the number of the system inputs, *M* is the number of fuzzy sets used to represent each input, and *T* is the number of fuzzy sets used to represent the system output.

Fourth layer is the output membership layer. Neurons in this layer represent fuzzy sets used in the consequent of fuzzy rules. The neurons of membership function to receive inputs from the corresponding fuzzy rule neurons. The probabilistic OR operations have been used for the combines of both above information. Every node i in this layer is a square node with a node function.

$$O_i^4 = \overline{w}_i * f_i = \overline{w}_i(p_i * x + q_i * y + r_i) \tag{11}$$

where: w_i – the output of layer third $\{p_i, q_i, r_i\}$ – the parameter set. Parameters in this layer will be referred to as consequent parameters.

Fifth layer is the defuzzification layer; it consists of a single output neuron. The single node is fixed with an output equal to the sum of all the rules outputs. The single node in this layer is a circle node labeled Σ that computes the overall output as the summation of all incoming signals (Jang et al., 1997), i.e.

$$O_i^5(x) = \sum_i \overline{w}_i \cdot f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \tag{12}$$

When the input–output example is presented to the system for training, the back-propagation algorithm computes the system output. Further, it can compare with the desired output of the training. The hybrid-learning algorithm is used here, which works on the principle of least-square error and the gradient descent method. The error is propagated backwards through the network from the output layer to the input layer again. Thus, the neuron activation functions are modified as the error is propagated.

Model building, training and validation are performed by MATLAB8.1 (licensed IIT Delhi). When a training input–output example is presented to the system, the back-propagation algorithm computes the system output and compares it with the desired output of the training. It uses a hybrid-learning algorithm that combines the least-squares estimator and the gradient descent method. The error is propagated backwards through the network from the output layer to the input layer. The neuron activation functions are modified as the error is propagated. To determine the necessary modifications, the back-propagation algorithm differentiates the activation functions of the neurons.

References

Chameides, W.L., Yu, H., Liu, S.C., Bergin, M., Zhou, X., Mearns, L., Wang, G., Kiang, C.S., Saylor, R.D., Luo, C., Huang, Y., Steiner, A., Giorgi, F., 1999. Case study of the effects of atmospheric aerosols and regional haze on agriculture: an opportunity to enhance crop yields in China through emission controls. PNAS 96 (24), 13626–13633.

Chandra, N., 2013. Pollution in Delhi is at its most lethal level ever as study points to alarming levels of aerosol and metals in the air. Daily Mail, 13 July 2013. http://www.dailymail.co.uk/indiahome/indianews/article-2362698/Death-breath-Air-pollution-Delhi-lethal-level-ever.html#ixzz3AaBxkvzx.

Corani, G., 2005. Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning. Ecol. Model. 185, 513–529.

Department of Town and Country Planning (DTCP), 2010. Integrated Mobility Plan for Gurgaon Manesar Urban Complex. Government of Haryana. http://tcpharyana.gov.in/CIM/Doc/gurgaon mobility plan.pdf.

EPA, 2012. The National Ambient Air Quality Standards for Particle Pollution. http://www.epa.gov/air/particlepollution/2012/fshealth.pdf.

George, M.P., Jasmine, B., Sharma, A., Mishra, S., 2013. Delhi smog 2012: cause and concerns. J. Environ. Sci. Water Resour. 2 (8), 260–269.

Gardner, M.W., Dorling, S.R., 1999. Neural network modelling and prediction of hourly NO_x and NO_2 concentrations in urban air in London. Atmos. Environ. 33, 709–719.

Goyal, P., Mishra, D., Kumar, A., 2013. Vehicular emission inventory of criteria pollutants in Delhi. Springer Plus 2 (216).

Goyal, P., Kumar, A., Mishra, D., 2014. The impact of air pollutants and meteorological variables on visibility in Delhi. Environ. Model. Assess. 19 (2), 127–138.

Heo, J.S., Kim, D.S., 2004. A new method of ozone forecasting using fuzzy expert and neural network systems. Sci. Total Environ. 325, 221–237.

Hornik, K., 1991. Approximation capabilities of multilayer feed-forward networks. Neural Netw. 4, 251–257.

Jang, I.-S.R., Sun, C.T., Mimtani, E., 1997. Neuro-fuzzy and Soft Computing: a Computational Approach to Learning and Machine Intelligence. Prentise-Hall, New Jersey.

Kim, K.W., Kim, Y.J., Oh, S.J., 2001. Visibility impairment during yellow sand periods in the urban atmosphere of Kwangju, Korea. Atmos. Environ. 35, 5157–5167.

Kumar, A., Goyal, P., 2011. Forecasting of air quality in Delhi using principal component regression technique. analysis. Atmos. Pollut. Res. 2, 436–444.

Kumar, A., Goyal, P., 2013. Forecasting of air quality index in Delhi using neural network based on principal component analysis. Pure Appl. Geophys. 170, 711–722.

Mishra, D., Goyal, P., 2015. Development of artificial intelligence based NO_2 forecasting models at Taj Mahal, Agra. Atmos. Pollut. Res. http://dx.doi.org/10.5094/apr.2015.012.

Mishra, D., Goyal, P., 2014. Estimation of vehicular emission inventories using dynamic emission factors: a case study of Delhi, India. Atmos. Environ. 98, 1–7. http://dx.doi.org/10.1016/j.atmosenv.2014.08.047.

NEA (National Environment Agency), 2012, Singapore, http://blissair.com/health-effects-of-haze.htm.

Okada, K., Ikegami, M., Zaizen, Y., Makino, Y., Jensen, J.B., Gras, J.L., 2001. The mixture state of individual aerosol particles in the 1997 Indonesian haze episode. J. Aerosol Sci. 32, 1269–1279.

Pao, Y.H., 1989. Adaptive Pattern Recognition and Neural Networks. Addison Wesley, New York.

Watson, J.G., 2002. Visibility: science and regulation. J. Air Waste Manag. Assoc. 52 (6), 628–713. http://dx.doi.org/10.1080/10473289.2002.10470813.

WHO, 2010. Exposure to Air Pollution: a Major Public Health Concern. http://www.who.int/ipcs/features/air_pollution.pdf.

Zhang, F., Chen, J., Qiu, T., Yin, L., Chen, X., Yu, J., 2013. Pollution characteristics of $PM_{2.5}$ during a typical haze episode in Xiamen, China. Atmos. Clim. Sci. 3, 427–439.