Forecasting of Ozone Episodes through statistical and artificial intelligence based models over Delhi metropolitan area

P. GOYAL*, DHIRENDRA MISHRA AND ABHISHEK UPADHYAY
Centre for Atmospheric Sciences
Indian Institute of Technology Delhi
Hauz Khas, New Delhi
INDIA
*E-mail- pramila@cas.iitd.ernet.in web- http://web.iitd.ac.in/~pramila

Abstract:- Ozone is the one of the most phytotoxic air pollutants, and causes considerable damage to ecological system throughout the world. Urban regions tend to have maximum ozone values in the late afternoon and minimum values in the early morning hours. The 8-hourly analysis of past data pattern in each season at different monitoring stations in Delhi suggests that the ozone episodes have occurred approximately 31% times in the year 2012 and its occurrence in the summer season is observed to be the worst in Delhi. In the present study, the air pollutants and meteorological parameters have been used to model the ozone episodes in Delhi through different modelling techniques, e.g., multiple linear regressions (MLR), artificial neural network (ANN) and artificial intelligence based Neuro-Fuzzy (NF). An analysis of the meteorological variables during ozone episodes revealed a good correlation between the ozone (O_3) concentrations and variables like pressure, dew point temperature and nitrogen dioxide (NO_2). The dry weather e.g., summer season is also favored for ozone formation. Further, the different statistical measures e.g., correlation coefficients (R), root mean square error (RMSE), fractional bias (FB) and index of agreement (IOA), factor of two (FAC2) have been used to assess the performance of models.

Key-Words: Ozone Episode, ANN, MLR, Statistical Measure, Neuro-Fuzzy, Delhi.

1 Introduction

Ozone is one of the most important air pollutants. It is formed in photochemical reactions, with concentrations affected by weather and the supply of chemical precursors, including other criteria pollutants, e.g., nitrogen oxides (NO_X), carbon monoxide (CO). Local weather conditions can also affect ozone concentrations and thus influence respiratory health [1] through a number of processes, including chemical production, and dilution and deposition of ozone. The ozone episodes are known for the period of usually a few days up to weeks with high ozone concentrations in the ambient atmosphere, i.e., the 8-hourly concentrations of ozone exceeds National Ambient Air Quality Standard (NAAQS). It is characterized by daily thresholds value to protect human health. Ozone episodes occur under specific meteorological conditions characterized by large stagnant areas of high pressure [2]. Since the formation of ozone requires sunlight, it mainly occurs during summer.

Ground-level ozone is a highly reactive oxidant and is unique among pollutants because it is not emitted directly into the air. It is a secondary pollutant that results from complex chemical reactions in the atmosphere. In the presence of the sun's ultraviolet radiation (hv), oxygen (O_2), nitrogen dioxide (NO_2), and volatile organic compounds (VOCs) react in the atmosphere to form ozone and nitric oxide (NO) through the reactions given in Equations (1) and (2).

ISBN: 978-1-61804-308-5

$$NO_2 + hv \rightarrow NO + O \quad (1)$$

$$O + O_2 \rightarrow O_3 \quad (2)$$

$$O_3 + NO \rightarrow NO_2 + O_2 \quad (3)$$

Resultant ozone, however, is quickly reacted away to form nitrogen dioxide by the process given in Equation (3) [3]. This conversion of ozone by NO is referred to as titration.

Epidemiological studies have shown a broad range of effects of ground-level ozone on health, leading to excess daily mortality and morbidity. The majority of epidemiological studies are concentrated on acute health consequences (due to short-term high concentrations of ozone). For instance, the association between acute exposures of ozone and short-term mortality has been shown in several studies [4] [5] [6] [7]. There are several large multi-city studies relating the numbers of hospital admissions for respiratory diseases [8] and chronic obstructive pulmonary diseases [9] to ambient ozone levels. The other main effects include emergency department visits for asthma, respiratory tract infections and exacerbations of existing airway diseases [10], as well as the decline of lung function [11].

Analysis and forecasting of air quality parameters are important topics of atmospheric and environmental research. In many of our applications, data are generated in the form of a time series. Therefore, time series analysis is a major task in forecasting or prediction studies, where one tests and predicts known or estimated observations for past times using them as input into the model to see how well the output matches the known observations. This methodology of forecasting, known as hind cast, is widely used in research studies. To keep in line with the common practice, the term forecast has been used for the models developed in the present paper.

The chemistry of ozone formation in large areas around the world is well understood, yet there are significant uncertainties about its distribution, behavior, and associated trends. These questions are important for large cities in the developing world, particularly in the urban area of India. Though ozone prediction models exist, there is still a need for more accurate models. Development of these models is difficult because the variations of meteorological variables and photochemical reactions involved in ozone formation are complex. Therefore, models, depending on statistical as well as artificial intelligence techniques, are used as an alternative approach with a wide spectrum of methodologies such as multiple linear regression (MLR), artificial neural networks (ANN), and Neuro-Fuzzy (NF) models. Time series forecasting has been dominated by linear methods (e.g., MLR) for decades as they are easy to develop, implement, and relatively simple to understand and interpret. However, these models are not able to capture any nonlinear relationships in the data. Further, ANN provides a promising alternative tool for forecasters due to its inherently nonlinear structure, which is particularly useful for capturing the complex underlying relationship in many real world problems.

2 Methodology

2.1 Study Area

Delhi,in the list of most populated metropolis in the world, is one of the most highly polluted cities in India and it ranks at second position in reference to

the annual mean of concentrations of criteria pollutants while Beijing is at the tenth position. Air quality forecast is useful to the public to protect their health. Short term forecasting of air quality is required to take preventive action during episodes of airborne pollution. Therefore, the main objective of this study is to forecast hourly ozone concentrations for Delhi using statistical as well as artificial intelligence techniques. The study has been made for available data of ozone episode in March, April and May 2012 (summer season) at IGI Airport, Delhi.

Fig.1: Study area of Delhi with CPCB monitoring stations

Central Pollution Control Board (CPCB), New Delhi, continuously monitors the air pollution data at different locations in Delhi. These locations include Indira Gandhi International (IGI) Airport, Income Tax Office (ITO), Shadipur and Dwarka in Delhi, India. Fig.1 shows these monitoring stations over the study area of Delhi. The hourly observed concentrations of ozone are obtained from Central Pollution Control Board (CPCB)' website and the hourly meteorological data at Safdarjung Airport monitoring station has been obtained from Indian Meteorological Department (IMD), New Delhi.

2.2 Assessments of Ozone episode

In the present study, the observed concentrations of O_3 pollutant have been investigated at all of the above monitoring stations, with the aim to develop the reliable ozone episodes forecasting models. Firstly, the available yearly average concentrations have been analyzed for the years 2010 to 2013 (Fig.2). The annual trend is found to be mixed and the annual average values are showing an increasing trend at IGI, Airport followed by ITO, Dwarka, Shadipur. The ITO monitoring station represents a heavy vehicular traffic area in the northeast part of Delhi [12] [13], while IGI, Airport, located at National highway, is also surrounded by construction of newly upcoming buildings in the southwest part of Delhi. It has recorded the most frequent occurrence of ozone episodes. However, the overall yearly levels of ozone are approximately stable, but are above the National Ambient Air Quality Standards at IGIAirport.

ISBN: 978-1-61804-308-5

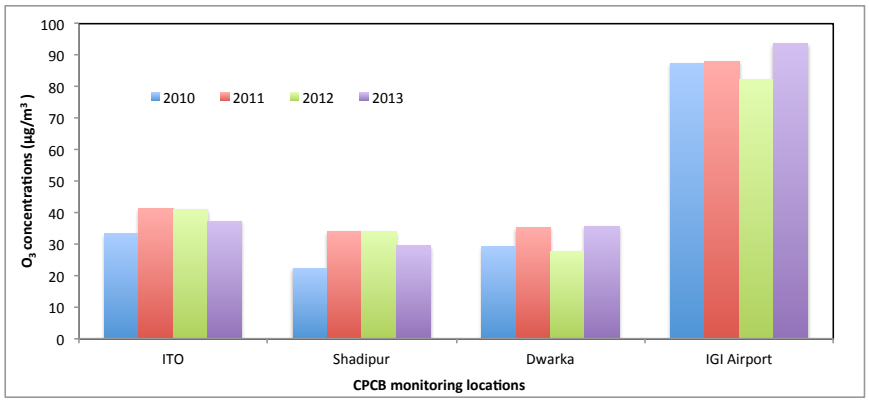


Fig.2: Annual average of Ozone concentrations at CPCB monitoring locations

Fig.3 shows the summer seasons' variation of the observed 8-hourly averaged values of O_3 concentrations at both stations. The variations of concentrations are observed to be almost the same in throughout the season. It is noticeable that all the values of O_3 concentrations are always higher at IGI, Airport than ITO. They have been observed to be high and frequently occurring at daytime. The maximum O_3 concentrations can be explained on the basis of large photochemical ozone production [14]. One of the reasons for high O_3 episodes may be the transportation.

The analysis of 8-hourly averaged surface ozone data illustrates that on a large number of days, the surface ozone values at Delhi exceed the National Ambient Air Quality Standard (NAAQS) (8-hourly average- 100 μ g/m^3). The concentrations of O_3 are observed more than that of the prescribed standards at IGI, Airport, which has the capacity of around 1,300 passengers per hour. Therefore, the detailed analysis of "ozone episode" is required at IGI, Airport. Thus, in present study, season-wise ozone forecasting models have been developed. The models, performing best among all the considered forecast models, have been recommended for practical use in any urban area.

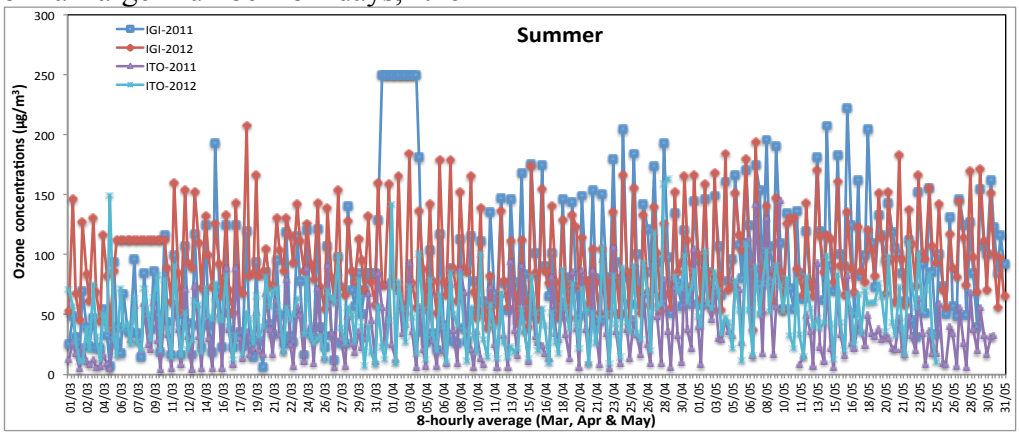


Fig.3: 8-hourly Ozone concentrations at ITO and IGI Airport, Delhi during summer.

ISBN: 978-1-61804-308-5

2.3 Multiple Linear Regression (MLR)

In general, MLR techniques have been used for forecasting and can be expressed as a function of a certain number of factors. It includes one dependent variable to be predicted and two or more independent variables. The MLR can be expressed as in Equation (4):

$$Y= b_1 +b_2 X_2+.....+b_k X_k+ e \quad (4)$$

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, O_3 is the dependent variable, concentrations of other 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 [15].

2.4 Artificial Neural Network

ANN has been applied in order to forecast, hourly mean concentrations of O_3 pollutants in the Delhi during ozone episode periods. 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, has been used for required output [16]. All the processes are depicted in Fig. 4.

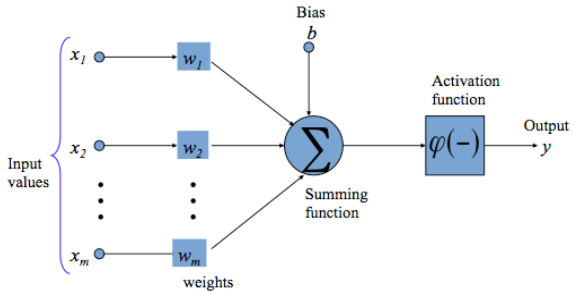


Fig. 4: Schematic structure of artificial neural network

Basically, the multilayer network incorporates three layers consists an input layer, hidden layer and output layer. The first layer, i.e., input layer directly collects the information from outside. All data in the input layer are feed-forwarded to the next layer, i.e., to the hidden layer, which functions as feature detectors of input signals and sent the information to the output layer. Lastly, the features has been collected by output layer and works as a producer of the response. In ANN, the output from last layer is the function of the linear combination of hidden unit's activation function and collected 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) model

Neural network and fuzzy logics are natural complementary tools in NF model. The neural networks are low-level computational structures that perform well when dealing with raw data while 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 neurons as the human-like knowledge representation and explanation abilities of fuzzy logic. Thus, the Neuro-Fuzzy system is more powerful as the neural networks become more transparent and fuzzy logic becomes capable of learning [17].

A fuzzy system is prepared through *if-then* rule on the basis of membership functions, which is 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 same methodology has been adopted as our previous study [17].

3 Results and Discussion

3.1 Correlation Matrix

The ozone episode hours have been investigated at IGI, Airport over the summer months in the years 2012. The hourly averaged data have been enlisted to find the linear regression amongst various concerned variables. Only significant variables have been chosen as the input in the model to save the computation time [18]. There are only nine significant parameters containing air pollutants and meteorological variables as shown in correlation matrix (Table 1).

Table 1: Correlation matrix of air pollutants and meteorological variables in Delhi for the year 2012.

	O_3	CO	DewT	NH_3	NO_2	P	$PM_{2.5}$	RH	SO_2	WS	T	Vis
O_3	1.00											
CO	-0.47	1.00										
DewT	-0.35	0.14	1.00									
NH_3	-0.31	0.24	0.35	1.00								
NO_2	-0.73	0.65	0.29	0.43	1.00							
P	0.05	-0.03	0.02	0.02	-0.03	1.00						
$PM_{2.5}$	-0.17	0.17	-0.19	-0.01	0.13	-0.01	1.00					
RH	-0.44	0.26	0.63	0.20	0.39	0.24	-0.03	1.00				
SO_2	-0.27	0.45	0.13	0.33	0.43	0.13	0.15	0.15	1.00			
WS	0.21	-0.05	-0.20	-0.04	-0.21	0.00	0.11	-0.31	0.02	1.00		
T	0.48	-0.33	-0.11	-0.15	-0.34	-0.42	-0.21	-0.55	-0.38	0.09	1.00	
Vis	0.29	-0.21	-0.29	-0.13	-0.29	0.03	-0.20	-0.42	-0.05	0.26	0.09	1.00

where, T- temperature, P-pressure, WS- wind speed, DewT- dew point temperature, Vis-visibility, and RH- relative humidity.

ISBN: 978-1-61804-308-5

3.2 Ozone episode modelling: Training and Validation

In order to limit the presentation, a case study of the year 2012 has been made.

The ozone episodes data have been collected for March, April and May 2012 (summer season), at IGI Airport, Delhi. It has been observed that the daily concentration of pollutant is not varying significantly. However, the hourly data shows a noticeable variation. Therefore, hourly data has been chosen for training and validation of models, which are collected and processed into an hourly format, and used to construct data series. There are some errors and/or incomplete information for some particular hours. Therefore, rows containing any incomplete information are removed from the data series. Finally, the entire table consists of 2002 rows. The columns represent variables consist of CO, NO_x, SO_2, NH_3, O_3, temperature, wind speed, relative humidity, visibility and dew point temperature. The first 75% dataset is chosen for training and last 25% data for validation. Therefore, the training data set containing 1500 hours data (data for March, April and May) and 502 hours data (data from 10^{th} May month) for validation. All the data are normalized into the range of [0, 1] by linear scaling, i.e., the input and output data are converted into values between zero and one.

The MLR model is determined by the combination of statistically significant regression parameters with the help of windows software SPSS (version 17.0), which achieves the minimum sum of squared errors with the training set. The best optimism equation is:

$$O_3 = 3.238 \times T + 1.237 \times P + 0.377 \times SO_2 + 0.355 \times RH + 2.457 \times Vis - 0.638 \times NO_2 - 2.671 \times DewT - 1171.461 \quad (5),$$

where, the aberrations have their usual meaning.

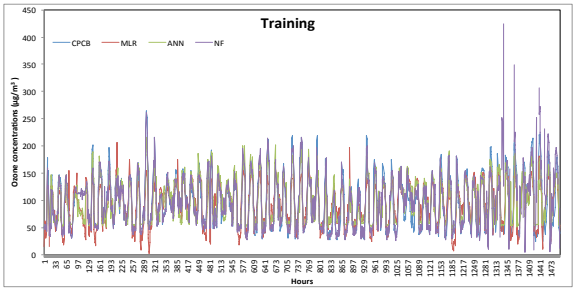
The ANN model is developed to forecast the ozone hours in Delhi using the MATLAB 8.1 (licensed, IIT Delhi). The model is built with the selected variables to forecast the ozone episode and the architecture is with two hidden layers, ten neurons, and with a square activation function. The first layer is the input layer, which consist the air pollutants, i.e., CO, NO_2, SO_2, NH_3, O_3 and meteorological variables viz. temperature, pressure, relative humidity, wind speed, and dew point temperature as the input in the proposed ANN model. 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 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.

Further, the artificial intelligence based NF model has been developed to forecast the ozone episode in Delhi urban area. In the present study, the training and validation are performed by MATLAB8.1 (licensed IIT Delhi). 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., back propagation method and fuzzy

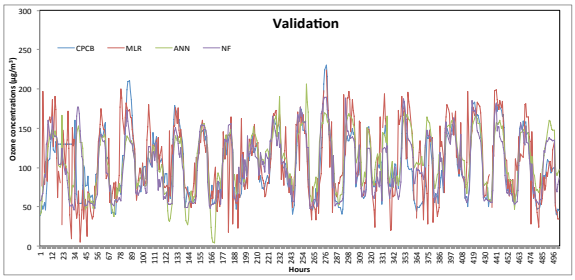
ISBN: 978-1-61804-308-5

logic. The six meteorological parameters with "4" categories of Gaussian membership functions have been used as input for the model development, which are categorized as good, moderate, bad, and hazardous. There are 30 epochs and 4096 rules in the model.

The evaluation of all three models have been made by providing comparison plots between the observed and models forecasted concentrations of ozone for training and validation phases at IGIAirport as shown in Figs.5(a) and5(b),which show a similar trend for almost all the models. Whereas the values of Neuro- Fuzzy (NF) model are found to be more close to observed CPCB values compared to that of other two models.



(a)



(b)

Fig.5: Plot between observed and model's forecasted values of ozone at IGI Airport for training and validation of all three models

Table 1: Correlation matrix of air pollutants and meteorological variables in Delhi for the year 2012.

	O_3	CO	DewT	NH_3	NO_2	P	$PM_{2.5}$	RH	SO_2	WS	T	Vis
O_3	1.00											
CO	-0.47	1.00										
DewT	-0.35	0.14	1.00									
NH_3	-0.31	0.24	0.35	1.00								
NO_2	-0.73	0.65	0.29	0.43	1.00							
P	0.05	-0.03	0.02	0.02	-0.03	1.00						
$PM_{2.5}$	-0.17	0.17	-0.19	-0.01	0.13	-0.01	1.00					
RH	-0.44	0.26	0.63	0.20	0.39	0.24	-0.03	1.00				
SO_2	-0.27	0.45	0.13	0.33	0.43	0.13	0.15	0.15	1.00			
WS	0.21	-0.05	-0.20	-0.04	-0.21	0.00	0.11	-0.31	0.02	1.00		
T	0.48	-0.33	-0.11	-0.15	-0.34	-0.42	-0.21	-0.55	-0.38	0.09	1.00	
Vis	0.29	-0.21	-0.29	-0.13	-0.29	0.03	-0.20	-0.42	-0.05	0.26	0.09	1.00

where, T- temperature, P-pressure, WS- wind speed, DewT- dew point temperature, Vis-visibility, and RH- relative humidity.

ISBN: 978-1-61804-308-5

The models performances in training and validation are quantified by statistical error analysis using several performance indexes, e.g., correlation coefficient (R), Index of agreement (IOA), normalized mean square error (NMSE), fractional bias (FB) and the factor of two (FAC2) [19], which are given in Table 2. On the basis of values of R, one can conclude that artificial intelligence based NF model is performing best among all the models. Where as, IOA and NMSE for all the models are found to be close to their ideal values. The RMSE values of all the models are neutralized by the values of NMSE. The values of FB for all models in training phase are under-predicting and are over predicting in validation. The NF model has 92% values within FAC2, which is considered to be very good for model performance. Finally, it can be concluded that the NF model is performing best among all the models though all of them are performing satisfactory.

Thus, the developed model can be used for daily forecasting of ozone episodes in any urban city like Delhi.

4 Conclusions

The meteorological variables like Relative Humidity, Temperature and dew point temperature have good correlation with ozone, whereas NO_2 and CO is showing the similar behavior. Summer season is observed to have most persistent episodes of ozone.

An artificial intelligence NF model can be considered as a reliable forecasting model for ozone episodes. The most convincing advantage of NF model is its capability of generalization the test data, which is better than ANN and MLR.

Finally, one can see that NF model provides promising results for modeling of non-linear ozone episode at urban area like Delhi and can also be used for forecasting of other air pollution events in the atmosphere.

References

[1] Ayres JG, Forsberg B, Annesi-Maesano I, et al. Climate change and respiratory disease: European Respiratory Society position statement. Eur Respir J 2009; 34: 295–302.

[2] EPA, 1996. Characterization of ambient air quality data forozone and its precursors-chapter 1. http://www.epa.gov/ttnamti1/files/ambient/pams/chap1.pdf.

[3] USEPA, 1999. Guideline for developing an ozone fore- casting program. EPA-454/R-99-009. Research Triangle Park, NC 27711: Office of Air Quality Planning and Standards.

[4] Bell ML, Dominici F, Samet JM. A meta-analysis of time-series studies of ozone and mortality with comparison to the national morbidity, mortality, and air pollution study. *Epidemiology* 2005; 16: 436–445.

[5] Gryparis A, Forsberg B, Katsouyanni K, et al. Acute effects of ozone on mortality from the ‘‘air pollution and health: a European approach’’ project. Am J Respir Crit Care Med 2004; 170: 1080–1087.

[6] Ito K, De Leon SF, Lippmann M. Associations between ozone and daily mortality: analysis and meta-analysis. *Epidemiology* 2005; 16: 446–457.

[7] Levy JI, Chemerynski SM, Sarnat JA. Ozone exposure and mortality: an empiric bayes metaregression analysis. *Epidemiology* 2005; 16: 458–468.

[8] Burnett RT, Brook JR, Yung WT, et al. Association between ozone and hospitalization for respiratory diseases in

16 Canadian cities. *Environmental Research* 1997; 72: 24–31.
[9] Anderson HR, Spix C, Medina S, et al. Air pollution and daily admissions for chronic obstructive pulmonary disease in 6 European cities: results from the APHEA project. *European Respiratory Journal* 1997; 10: 1064–1071.
[10] Amann M, Bertok I, Cofala J, et al. Baseline scenarios for the Clean Air for Europe (CAFE) programme. Final report. Laxenburg, International Institute for Applied Systems Analysis, 2005.
[11] Peters JM, Avol E, Gauderman WJ, et al. A study of twelve Southern California communities with differing levels and types of air pollution. II. Effects on pulmonary function. *American Journal of Respiratory and Critical Care Medicine* 1999; 159: 768–775.
[12] Goyal, P., Mishra, D., Kumar, A., 2013. Vehicular emission inventory of criteria pollutants in Delhi. *SpringerPlus* 2 (216).
[13] Mishra D., and Goyal, P., (2016). Quantitative Assessment of the Emitted Criteria Pollutant in Delhi Urban Area. *Aerosol and Air Quality Research*, DOI: 10.4209/aaqr.2014.05.0104 (Accepted-in press).
[14] Ghude, S. D., Jain, S. L., Arya, B. C., Beig, G., Ahammed, Y.N., Kumar, A., Tyagi, B., 2009. Ozone in ambient air at a tropical megacity, Delhi: characteristics, trends and cumulative ozone exposure indices, *Journal of Atmospheric Chemistry* 60(3), 237-252.
[15] Goyal, P., Kumar, A., 2011. Mathematical Modeling of Air Pollutants: An Application to Indian Urban City, *Air Quality-Models and Applications*, Prof. Dragana Popovic (Ed.), ISBN: 978-953-307-307-1, InTech, DOI: 10.5772/16840.
[16] Mishra, D., Goyal, P., 2015. Development of artificial intelligence based NO_2 forecasting models at Taj Mahal, Agra. *Atmospheric Pollution Research*. 6, 99-106.
[17] Mishra D., Goyal, P., Upadhyay, A., 2015. Artificial Intelligence Based Approach to Forecast PM2.5 during Haze Episodes: A Case Study of Delhi, India. *Atmospheric Environment*, 102, 239–248.
[18] Kumar, A., Goyal, P., 2013. Forecasting of air quality index in Delhi using neural network based on principal component analysis. *Pure Appllied Geophysics* 170, 711–722.
[19] Chang, J.C., Hanna, S.R., 2004. Air quality model performance evaluation. *Meteorology and Atmospheric Physics* 87, 167-196.