International Journal of Soft Computing 8 (4): 305-312, 2013
ISSN: 1816-9503
© Medwell Journals, 2013

Forecasting Criteria Air Pollutants Using Data Driven Approaches: An Indian Case Study

[1]S. Tikhe Shruti, [1]K.C. Khare and [2]S.N. Londhe
[1]Department of Civil Engineering, Sinhgad College of Engineering,
Pune-411041, Maharashtra State, India
[2]Department of Civil Engineering, Vishwakarma Institute of Information Technology,
Pune-411038, Maharashtra State, India

Abstract: Forecasting air pollutant trends especially in metropolitan cities of India has become a vital issue as air pollution has immediate and severe impacts on human health. Criteria pollutants like Oxides of Sulphur (SOx), Oxides of Nitrogen (NOx) and Respirable Suspended Particulate Matter (RSPM) have either reached or exceeded the acceptable limits specified by Central Pollution Control Board of India for Pune city which is at the second position as far as pollution levels of India are concerned. In the present research, two soft computing approaches namely Artificial Neural Networks (ANN) and Genetic Programming (GP) are used to predict the air quality parameters (SOx, NOx, RSPM) a few time steps in advance for Pune city. Six models have been developed based on daily average data values of pollutant concentrations spanning >7 years. ANN, one of the proven tools in estimation and prediction of air quality has been used and the results of the models are compared with GP which is rarely used tool in the field of air quality modelling and forecasting. The performance of the models has been compared using r, RMSE and d. Considering the complexity of the air pollution phenomenon, it was found that GP Models are robust and could work well as compared to ANN.

Key words: Air quality, criteria pollutants, ANN, GP, India

INTRODUCTION

Air pollution is a complex issue, fuelled by multiple sources ranging from vehicular exhaust, industrial emissions, emissions from fossil fuels, construction activities to domestic activities. Air pollution may cause pernicious effects on human health, especially in areas with high population density. Forecasting air quality is one of the most sought after topic of research today for urban air pollution studies and specifically for prediction of pollution episodes, i.e., high pollutant concentrations causing adverse health effects (Niska *et al.*, 2004). Air quality models play a vital role in all aspects of air pollution control and air quality planning where prediction is a major component (Nagendra and Khare, 2006). Air quality forecasts provide the public with air quality information which allows people to take precautionary measures to avoid or limit their exposure to unhealthy levels of air pollution. Hence, it is quite essential to predict criteria pollutants.

Urban air pollution involves physical and chemical process ranging over a wide scale of time and space. In order to model the urban systems, extensive data such as emissions from various sources (stationary and mobile) influence of buildings and other obstacles, meteorology of the area, information about turbulence profile, heat flux previous values of the pollutants, etc. is required. It is practically very difficult to collect the above-mentioned data (except pollutant concentrations), hence temporal models are handy in such situations. They can be used easily for forecasting purpose because historical sequence of the pollutant concentrations is readily available from pollution control authorities of the country. Air pollution is a time dependent phenomenon which further justifies the use of time series approach for forecasting of criteria air pollutants.

Several techniques are available to predict future pollutant concentrations including fixed box methods, linear regression methods, Computational Fluid Dynamics (CFD) simulation, artificial intelligence, etc. Conventional technique like numerical method require detailed source information and consume a lot of time and effort to forecast and also found to be weak particularly when used to model nonlinear systems (Barai *et al.*, 2007). This leaves a scope for another approach like data driven techniques which are found to be suitable to model non-linear systems.

Corresponding Author: S. Tikhe Shruti, Department of Civil Engineering, Sinhgad College of Engineering, Pune-411041, Maharashtra State, India

Artificial Neural Networks (ANN) are already been regarded as a cost effective method to achieve the prediction of air pollutants in time series and have become popular since, last decade (Lu *et al*., 2004). The literature reveals that GP a relatively new approach has been applied successfully to solve complex Civil engineering problems such as structural optimisation, soil classification, prediction of scour depth of circular piles, algal bloom prediction and also for prediction of climate change. Better predictive capabilities of GP have also been reported, especially for the peak values for wave forecasting (Londhe, 2008). Air pollution happens to be the complex civil engineering problem hence an attempt is made to assess the success of GP for air quality prediction.

The present research aims at forecasting criteria pollutants such as Oxides of Sulphur (SOx), Oxides of Nitrogen (NOx) and Respirable Suspended Particulate Matter (RSPM) concentration 1 day in advance for one of the polluted metropolitan city (Pune) of India using ANN as well as GP and comparing them with respect to their accuracy of forecast.

Artificial Neural Networks (ANN): Artificial Neural Networks (ANN) are intelligent systems that have the capacity to learn, memorize and create relationships among the data. ANN is made up by simple processing units, the neurons which are connected in a network by a large number of weighted links where the acquired knowledge is stored and over which signals or information can pass.

These interconnected neurons combine the input parameters, the strength of such combination is determined by comparing with bias and executing a result in proportion to such a strength. ANNs learn by example hence it is trained first with examples by using various algorithms which converge the solution by reducing the error between the network output and the target by distributing the performance error between the weights and biases associated with each neuron. Then, the network is tested for unseen inputs (ASCE Task Committee on Application of Artificial Neural Networks in Hydrology, 2000).

Artificial Neural Networks map any random input with random output by self learning, without any fixed mathematical form assumed beforehand and without necessarily having the knowledge of the underlying physical process. The ANN Model is given in Fig. 1. The input values are summed up, a bias is added to this sum and then the result is passed through a nonlinear transfer function, like the sigmoidal function. Mathematically this is equivalent to:

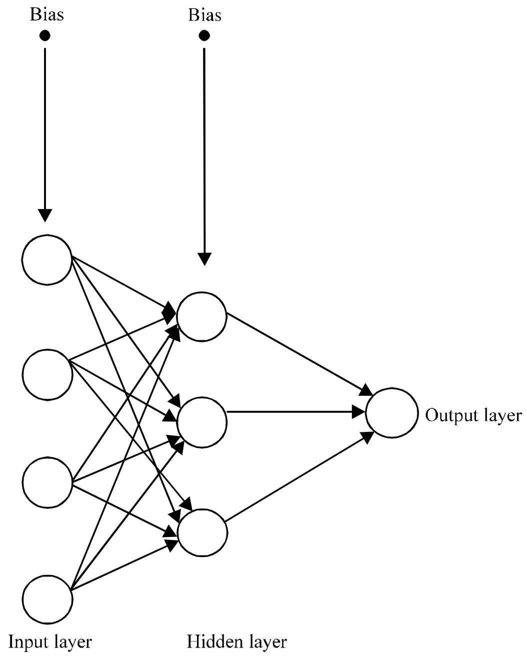

Fig. 1: The ANN Model

$$\text{Out} = \frac{1}{1+e^{-\text{sum}}} \quad (1)$$

Where:

$$\text{sum} = (X_1W_1+X_2W_2+ \ldots)+\beta \quad (2)$$

Where:

$X_1, X_2,...$ = Inputs
$W_1, W_2,...$ = Weights
β = Bias

Before its application, the network is required to be trained and this is done by using a variety of training algorithms, like standard Backpropagation, Conjugate Gradient, Quasi-Newton and Levenberg-Marquardt, etc. For more information about ANN, readers are referred to Bose and Liang (2000). All training algorithms are basically aimed at reducing the global error, E, between the network output and the actual observation, as defined:

$$E = \sum(O_n-O_t)^2 \quad (3)$$

Where:

O_n = The network output at a given output node
O_t = The target output at the same node

The summation is carried out over all output nodes for a given training pattern and then for all patterns. For general applications of ANN in atmospheric sciences, readers are referred to Gardner and Dorling (1999). The present study uses three layered Feed Forward Back Propagation network to predict SOx, NOx and RSPM levels in Pune (State: Maharashtra of India) 1 day in advance using commercial software MATLAB.

Genetic programming: Genetic Programming (GP) is a search algorithm based on principle of Darwin's theory of evolution. It is a generalization of genetic algorithms which starts with an initial population composed by a set of individuals randomly created. The fitness of the individuals is evaluated and then the parents are selected out of these individuals. The parents are then made to yield offsprings through the process of crossover mutation and reproduction. Creation of offsprings is continued in an iterative manner till a specified number of offsprings are produced in a generation and further till another specified numbers of generations are created. Resulting offsprings (equation or program) at the end of the process is the required solution of the problem. Detailed explanation of concepts related to GP can be found by Koza (1992). Figure 2 shows the typical process of Genetic programming. The present study uses GP as a data driven approach to predict SOx, NOx and RSPM levels in Pune (State: Maharashtra of India) 1 day in advance using commercial software discipulus.

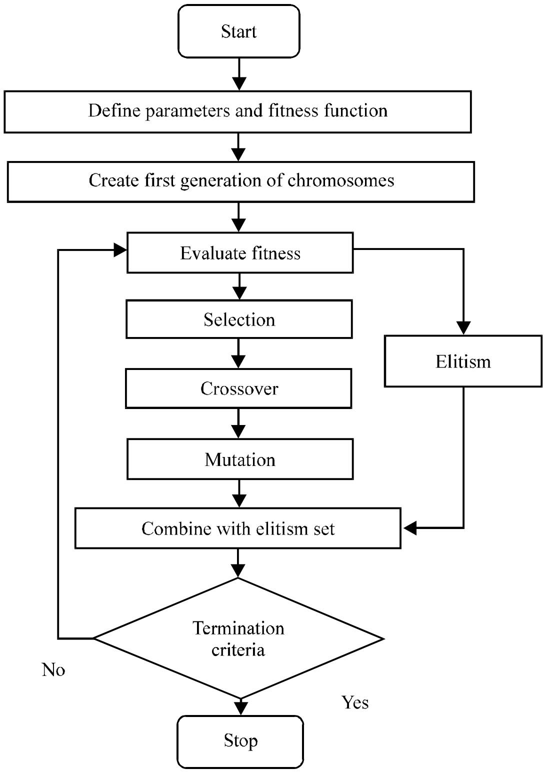

Fig. 2: A typical GP flowchart

Motivation: Air pollution is a non-linear problem which consists of interaction of various atmospheric elements. The process is complex and difficult to be mapped with linear models. ANN Models usually present better performance than the linear ones but they are included in a group called Black Box Models due to limited interpretation.

In the air quality forecasting, especially, the selection of optimal input subset becomes a tedious task due to high number of measurements from heterogeneous sources and their non-linear interactions. Moreover, due to a complex interconnection between the input patterns of ANN and the architecture of ANN (related to the complexity of the input and output mapping, the amount of noise and the amount of training data), the selection of ANN architecture must be done simultaneously. These aspects requires the formulation of search problem and the investigation of search techniques which are capable of facilitating model development research and resulting more reliable and robust ANN Models. In this context, the GP have proven to be powerful techniques due to its ability to solve linear and non-linear problems by exploring all regions of the state space and exploiting promising areas through genetic operations.

Considering the characteristics of air pollution problem and the advantages of GP, an attempt is made in this study to test the feasibility of GP for air quality modelling.

Literature review: ANN is one of the proven tools in the field of air pollution forecasting where as GP is relatively a new approach which is evident from following literature references.

Three ANN Models were tried for hourly predictions of PM_{10} at Athens on the basis of hourly data measured for 2 years (Raimondo *et al*., 2007). First model consisted of MLP with ANN which used full set of input parameters (temperature, relative humidity, wind speed and wind direction), second MLP ANN Model used the only input parameters as are suggested by Genetic algorithm optimisation procedure and the third model without meteorological input variables (time series). Researchers have got better results with GA MLP Model. PM_{10} daily average prediction for earlier and below threshold values at Sweden using hourly data of principal air pollutants and meteorological parameters was carried out by using three layered ANN and also by SVM (Grivas and Chaloulakou, 2006). According to their study, the accuracy of ANN Model is about 80% and increases with input parameters. Integrated ANN Model was developed to forecast the maxima of 24 h average of PM_{10} concentration 1 day in advance for five monitoring stations in Chile using meteorological parameters as input and ultimately researchers found the ANN tool as satisfactory (Perez and Reyes, 2006). Barai *et al.* (2007) have developed various types of neural networks such as SOFM, CPDM, RNM and SNCM in order to forecast CO, PM, NOx and SO_2 for long (annual) and short (daily) term horizon and found that SOFM are the best amongst the alternatives available.

Pires *et al.* (2010) have used multigene genetic programming for 1 day ahead prediction of PM_{10} in Portugal. Pires *et al.* (2011) have tried GP to predict next day hourly average concentration of O_3 for Portugal and have found that GP could identify the significant inputs for O_3 prediction.

There are no evidences of application of GP for air pollution forecasting for any metropolitan city of India. The present research attempts to use GP approach for 1 day ahead prediction of indicator pollutant forecasting.

MATERIALS AND METHODS

Pune is one of the fastest developing metropolitan cities in India which generates about 181.957 tonne of toxic waste daily. It is located on the Deccan Plateau at the confluence of Mula Mutha Rivers and at an elevation of about 560 m above mean sea level at Karachi. Location sketch of study area can be found in Fig. 3. Accelerating growth in the transport sector, a booming construction industry and a growing industrial sector are responsible for deteriorating air quality of the city which has resulted into bad health impacts.

Under National Ambient air Monitoring Program (NAMP) and also State Ambient Air Monitoring Program (SAMP), the upper limits set for daily average concentration of SOx, NOx and RSPM by CPCB (Central Pollution Control Board, India) and MPCB (Maharashtra Pollution Control Board) are 80, 80 and 100 μg/m^3, respectively. Since, last few years it has been observed that the upper limits have been crossed for SOx, NOx and RSPM with a maximum value recorded as high as 195, 138 and 370 μg/m^3, respectively. The present study aims at predicting values of SOx, NOx and RSPM, 1 day in advance which can provide an indication about the prevailing air quality on the next day.

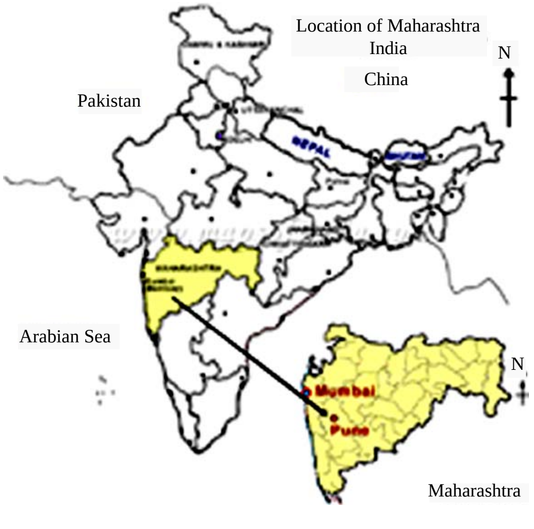

Fig. 3: The study area (Pune, Maharashtra, India)

Data used for the study consists of daily average values of SOx, NOx and RSPM concentrations recorded for the period of January 2005 to December 2011 by MPCB under SAMP for Pune (Maharashtra State, India). The values are recorded using High Volume Sampler by Improved West and Gaeke Method for SOx, Sodium Arsenite Method for NOx and by Filter Paper Method for RSPM. As the earlier measured data of the above mentioned criteria pollutants is only used for this research the data driven approaches of artificial neural networks and genetic programming are employed to develop the SOx, NOx and RSPM Forecast Models 1 day in advance and the results are compared for forecasting accuracy.

Model development: Input data selection for Air Quality Models can be done in a variety of ways such as correlations, sensitivity analysis, principal component analysis and also by trial and error. As air pollution is a time dependent phenomenon, a record of seven antecedent values of the pollutant concentrations were considered which represent a picture of pollution concentration for the preceding week which will have impact on the future trends of pollutants. Out of the seven earlier values, the most influential inputs were identified using correlation analysis. A correlation coefficient of each antecedent value in a cumulative way is calculated with the targeted output and it was found that three, four and three antecedent values prove to be the optimum inputs for SOx, NOx and RSPM Models, respectively. Table 1 highlights the optimum number of input parameters for SOx, NOx and RSPM Models.

ANN Model: The available data was of total 2548 values of daily average concentration of SOx, NOx and RSPM

Table 1: Optimum number of input parameters

No. of input parameters	R (SOx)	R (NOx)	R (RSPM)
1	0.764	0.789	0.892
2	0.557	0.773	0.899
3	0.767*	0.879	0.900*
4	0.759	0.897*	0.900
5	0.755	0.884	0.897
6	0.223	0.892	0.897
7	0.734	0.891	0.897

*Highlighted figures indicate correlation coefficients corresponding to optimum number of input parameters

Table 2: Criteria used for ANN Based Models

Items	Criteria used in the present study
Network architecture	Input neurons = Number of input variables (as in Table 1). Output neurons = Number of output variables (one variable for each model). Hidden neuron = Smallest number of neuron which yield a minimum prediction error on the validation dataset (ASCE Task Committee on Application of Artificial Neural Networks in Hydrology, 2000). ANN SOx = 3:5:1, ANN NOx = 4:2:1, ANN RSPM = 3:2:1
Neuron activation function	Input neuron = Identity function. Output neuron = Identity function. Hidden Neuron = Hyperbolic tangent function (Khare and Nagendra, 2007). 'logsig' and 'purelin' for all the models
Learning parameters	The learning parameters converge to the network configuration and give best performance on the validation data with least epochs (Khare and Nagendra, 2007). ANN SOx = 400 epochs, ANN NOx = 300 epochs, ANN RSPM = 300 epochs
Criteria for initialisation of the network weights	Network weights are uniformly distributed in the range of -1 to 1
Training algorithm	Levenberg Marquardt
Stopping criteria	Performance goal/epochs
Performance indicator	r, RMSE, d

spanning over the year 2005 to 2011. Three ANN Models were developed namely ANN SOx, ANN NOx and ANN RSPM for 1 day ahead prediction of the criteria pollutants. Symbolically models can be written as:

$$\text{ANN SOx }(t+1) = f\left(\text{SOx }(t),\ \text{SOx }(t-1),\ \text{SOx }(t-2)\right) \tag{4}$$

$$\text{ANN NOx }(t+1) = f\begin{pmatrix}\text{NOx }(t),\ \text{NOx }(t-1), \\ \text{NOx }(t-2),\ \text{NOx}(t-3)\end{pmatrix} \tag{5}$$

$$\text{ANN RSPM }(t+1) = f\left(\text{RSPM }(t),\ \text{RSPM }(t-1),\ \text{RSPM }(t-2)\right) \tag{6}$$

Criteria used for ANN Based Pollutant Forecast Models: For the present study, couple of trials for deciding data division were taken. Training and testing dataset, varying from 40-85% (for training and remaining data for testing) were taken and found that 75% data for training and 25% of data for testing (for NOx and RSPM Models) and 80% data for training and 20% for testing (for SOx Model) yield better results. Hence, the same data division is used for the study. Readers are requested to refer (Khare and Nagendra, 2007) for more details of the training and testing data division.

Table 2 indicates the criteria used for ANN Models in the present study. The MATLAB neural network toolbox is used to develop models based on above criteria.

GP Model: Three GP Models were developed namely GP SOx, GP NOx and GP RSPM for the same data and the data division for which the best results were obtained for respective ANN Models so that they can be compared. The GP Models were developed on selection of major control parameters such as fitness function in terms of mean square error, initial population size, mutation frequency and the crossover frequency. Table 3 indicates the GP parameters used for the present study. The Commercial Software discipulus was used to develop the GP Models.

Table 3: GP parameters

Parameters	GP SOx	GP NOx	GP RSPM
Initial population size	1025	242	282
Mutation frequency	83.76%	100%	87.21%
Crossover frequency	70.46%	48.86%	34.00%
Performance indicator	r, RMSE, d	r, RMSE, d	r, RMSE, d

Model assessment: The testing performance of all six models was assessed by plotting time series plots as well as by statistical parameters like correlation coefficient (r), Root Mean Square Error (RMSE) and descriptive statistics (d).

Correlation coefficient (r) is a measure of the trends of predicted values as compared to the observed (measured) values. It is independent of the scale of the data. Higher value of r indicates better result and r = 1.00 indicates a perfect correlation. The Root Mean Square Error (RMSE) is a measure of the differences between values predicted by a model or an estimator and the values actually observed. RMSE is a good measure of accuracy. These individual differences are also called residuals and the RMSE serves to aggregate them into a single measure of predictive power. Lesser value of RMSE is preferred.

The d is a descriptive statistics. It reflects the degree to which the observed variate is accurately estimated by the simulated variate. The d is not a measure of correlation or association in the formal sense but rather a measure of the degree (based on ensemble average) to which the model predictions are error free. At the same time, d is a standardized measure in order that it may be easily interpreted and cross-comparisons of its magnitudes for a variety of models, regardless of units can readily be made. It varies between 0 and 1. A computed value of 1 indicates perfect agreement between the observed and predicted observations while 0 connotes complete disagreement.

Out of the three statistical measures, r and d are the measures of goodness of fit whereas RMSE is an absolute error measure. The model evaluation based on r mostly fails due to the presence of lag between source emission quantity and the ambient pollutant concentration. The lag is due to adverse meteorological conditions (inversion) which implies the accumulation of pollutants in the ambient environment during odd hours of the day when there are no source emissions (ASCE Task Committee on Application of Artificial Neural Networks in Hydrology, 2000). In such a situation, for air quality models d statistics appears to be the most relevant measure as it would signify whether the models are error free.

RESULTS AND DISCUSSION

Owing to numerous advantages of ANNs such as adaptive learning, self organisation, real time operation, capability of handling nonlinear systems, etc., they are used in this study as the benchmarking tool for 1 day ahead prediction of criteria pollutant in Pune city. The prediction is also carried out by a relatively new approach of GP by testing for unseen inputs and the qualitative and quantitative performance is judged by means of correlation coefficient, root mean square error and d statistics between the observed and forecasted values and also by plotting the time series plots between the same.

The ANN as well as GP Model exhibited a reasonable performance in testing between the observed and forecasted pollutant concentrations for all the models which is evident from the mentioned performance indicators (Table 4). From the times series plot (Fig. 4) for SOx, it can be clearly seen that ANN Models work well as compared to GP for prediction of 1 day ahead concentration of SOx but in ANN researchers define the structure initially and weights are found by learning algorithm whereas in GP the functions are defined and it would result into the optimal solution. Thus, the solution given by GP is always optimal which is not the case with ANN. This fact is verified by increased value of r and d and decreased RMSE as compared to ANN Model. Hence, it can be concluded that GP works better compared to ANN for SOx.

Considering the time series plot for NOx (Fig. 5) it can be seen that both (ANN as well as GP) models have worked well but GP Models are even better for prediction of peak values where ANN was found to be insufficient.

Table 4: Performance indicator of the developed models

	ANN			GP		
Pollutant	R	RMSE ($\mu g/m^3$)	D	r	RMSE ($\mu g/m^3$)	d
SOx	0.822	7.574	0.898	0.8635	6.422	0.925
NOx	0.897	12.657	0.924	0.9280	10.259	0.961
RSPM	0.928	18.402	0.958	0.9300	18.031	0.961

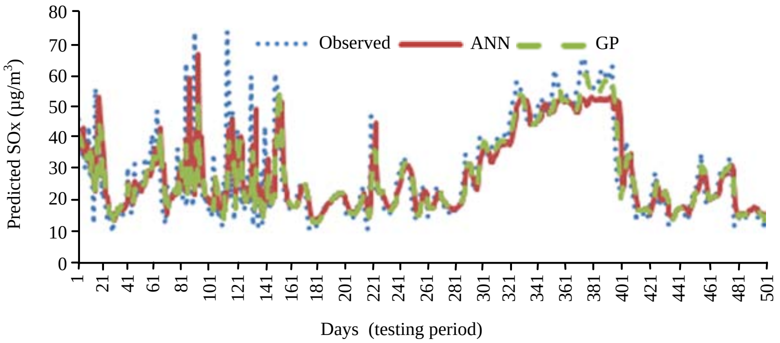

Fig. 4: Time series plot for SOx

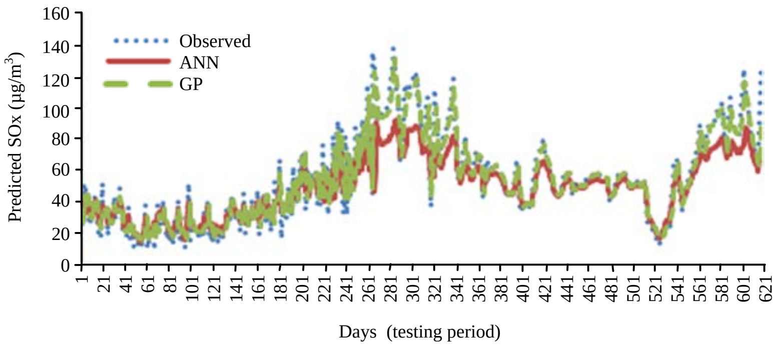

Fig. 5: Time series plot for NOx

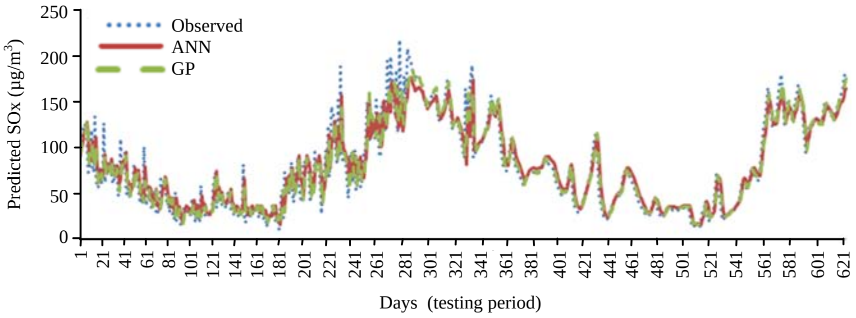

Fig. 6: Time series plot for RSPM

The r and d values for GP Model are higher than those for ANN Model where as RMSE is lesser as compared to ANN Model.

Time series plot for RSPM (Fig. 6) indicates that both the models (ANN as well as GP) could work well for 1 day ahead RSPM prediction (except for few peaks). statistical measures also indicate that r and d values are higher for GP Model and RMSE is lower as compared to ANN Model.

Out of the three statistical performance indicators, d value is the most relevant criteria as it accounts for the lag due to adverse meteorological conditions. Air pollution is a complex phenomenon where in lag effect is required to be considered for forecasting models. All the GP Models show higher value of d as compared to respective ANN Models which justifies that GP can be used for Air pollution forecasting.

CONCLUSION

The criteria pollutant forecasting temporal models are developed for Pune city (Maharashtra State, India) using data driven approaches of Artificial Neural Networks and Genetic Programming with a lead time of 1 day. Both the tools worked reasonably well in terms of prediction accuracy for the dataset of 2005 to 2011. The GP technique seems to research better than ANN. GP being a relatively new approach, needs to be explored further for short term as well as long term forecast of criteria pollutants with certain considerations such as climatic conditions, seasonal variations.

REFERENCES

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology, 2000. Artificial neural networks in hydrology I: Preliminary concepts. J. Hydrol. Eng., 5: 115-123.

Barai, S.V., A.K. Dikshit and S. Sharma, 2007. ANN for air quality prediction: A comparative study. J. Soft comput. Ind. Appl., 39: 290-305.

Bose, N.K. and P. Liang, 2000. Neural Network Fundamentals with Graphs, Algorithms and Applications. Tata McGraw-Hill Publishing Company Limited, New Delhi.

Gardner, M.W. and S.R. Dorling, 1999. Artificial neural networks: The multilayer perceptron: A review of applications in the atmospheric sciences. J. Atmos. Environ., 32: 2627-2636.

Grivas, G. and A. Chaloulakou, 2006. Artificial neural network models for prediction of PM_{10} hourly concentrations in the Greater Area of Athens, Greece. Atmos. Environ., 40: 1216-1229.

Khare, M. and S.A. Nagendra, 2007. Artificial neural networks in vehicular pollution modelling. J.Studies Comput. Intell., 41: 41-45.

Koza, J.R., 1992. Genetic Programming on the Programming of Computers by Means of Natural Selection. MIT Press A Bradford Book, USA.

Londhe, S.N., 2008. Soft computing approach for real-time estimation of missing wave heights. J. Ocean Eng., 35: 1080-1089.

Lu, W.Z., W.J. Wang, X.K. Wang, S.H. Yan and J.C. Lam, 2004. Potential assessment of a neural network model with PCA/RBF approach for forecasting pollutant trends in Mong Kok urban air, HongKong. Environ. Res., 96: 79-87.

Nagendra, S.M. and M. Khare, 2006. Artificial neural network approach for modeling nitrogen dioxide dispersion from vehicular exhaust emissions. Ecol. Modell., 190: 99-115.

Niska, H., T. Hiltunen, A. Karppinen, J. Ruuskanen and M. Kolehmainena, 2004. Evolving the neural network model for forecasting air pollution time series. Eng. Applied Artif. Intel., 17: 159-167.

Perez, P. and J. Reyes, 2006. Integrated neural network model for PM10 forecasting. Atmosp. Environ., 40: 2845-2851.

Pires, J.C.M., Alvim-Ferraz, M.C.M., M.C. Pariera and F.G. Martins, 2010. Prediction of PM10 concentration through Multigene genetic programming. Atmos. Pollut. Res., 1: 305-310.

Pires, J.C.M., M.C.M. Alvim-Ferraz, M.C. Pariera and F.G. Martins, 2011. Prediction of troposphere ozone concentration: Application of a methodology based on Darwin's Theory of Evolution. Expert Syst. Appl., 38: 1903-1908.

Raimondo, G., M. Alfonso and M.A. Walter, 2007. Machine learning tool to forecast PM10 level. Proceedings of the 11th International Conference on knowledge based and Intelligent Information and Engineering Systems, September 12-14, 2007, Vietri Sul Mare, Italy.