RESEARCH ARTICLE OPEN ACCESS

Short Term Air Quality Forecast Using Data Driven Approaches

Shruti S. Tikhe [#1], Dr. Mrs.K.C. Khare [*2], Dr. S. N. Londhe [#3]
[# 1]Department of Civil Engineering, Sinhgad College of Engineering Pune, Maharashtra, India-411041
[*2]Department of Civil Engineering, Symbiosis Institute of Technology, Pune, Maharashtra, India- 412115
[#3]Department of Civil Engineering, Vishwakarma Institute of Information Technology, Pune, Maharashtra, India- 411048

Abstract
Predicting air quality is a challenge, when there are uncertainties involved in the availability as well as accuracy of the desired data. Most of the developing countries including India often have no formalized forecasting approach. Little data (which may be of suspect quality) and inadequate institutional structure to support data collection are the main concerns. Present study suggests a forecasting tool that is seldom used in the field of air quality forecasting; but is a proven robust technique.
Thirty six models have been developed with Genetic Programming (GP) and Artificial Neural Networks (ANN) considering daily average concentrations of meteorological parameters as well as pollutant concentrations spanning from 2005-2008 for one of the most polluted metropolitan city of India. The models are specific to cases when all the significant input parameters (data) are not available because of various reasons.
ANN is used as a benchmarking tool for estimation and prediction of air quality and the results are compared with GP .Performance of all the models has been assessed using r (correlation coefficient), RMSE (root mean square error) & d (d statistics).Compared to ANN, GP models seem to work well in all cases considered because of unavailability of data and have an advantage of pollution forecasting equation generated by the model. These equations can be of help for real time forecasting.
Keywords- Air quality, ANN, GP, Western Maharashtra

I. INTRODUCTION

Air pollution is a growing problem arising from domestic heating, high density vehicular traffic, electricity production and expanding commercial and industrial activities parallel with urban population. Monitoring and forecasting of air quality parameters in the urban area is important due to health impacts. Air quality is the result of various complex processes which include meteorology, emissions, chemical reactions amongst the pollutants and transport of the pollutants. Artificial Intelligence techniques are successfully used in modelling of highly complex and nonlinear phenomenon of air pollution. Air pollution models play an important role in science because of their capability to assess the relative importance of the relevant process.

Approaches for air quality modelling can be stated as deterministic (analytical and numerical), Stochastic (statistical), Physical and Soft computing. Deterministic models integrate the equations of fluid motion to predict the air quality. Stochastic models are based on the fact that diffusion has certain statistical nature (eg. Gaussian plume model). Physical models are scaled models of stack & terrain features. Soft computing models resemble biological processes and on the basis of the data availability they result in to the solution with acceptable tolerance. Data constraints and specific purpose for which prediction is needed lead to the selection of the particular model [1].

Most of the urban air pollution models require information about source inventory, their emissions, types of pollutants, their rate of release, climate of the region and other meteorological parameters. Data collection plays a vital role in air quality modelling and forecasting. Many a times it is difficult to obtain the required data on a continuous basis. Hence model should be robust enough to accommodate such fluctuations in data collection. Traditional forecasting techniques are found to be weak particularly when used to model nonlinear systems. This leaves a scope for data driven approaches which are found to be suitable to model the nonlinear systems.

Artificial Neural Network (ANN) models are regarded as the benchmarking tools and they usually presented better performance than the linear ones due to the nonlinear behaviour associated with pollutant formation. However, they are included in a group called black box models, having limited interpretation. Moreover, the selection of the optimal network architecture and the computation time are the main disadvantages of these models.

As many factors could influence the performance of models, their development should have more degrees of freedom. In stochastic processes, such as the prediction of pollutant concentrations, the

structure of the models should be more flexible. In this context, Genetic Programming (GP) could be a successful methodology, as it does not assume in advance any structure for the model whereas it can optimize both the structure of the model and its parameters, simultaneously [2].

The present work aims at development of pollution forecasting models for criteria pollutants such as Oxides of Sulphur (SO_x), Oxides of Nitrogen (NO_x) and Respirable suspended Particulate Matter (RSPM)) which can work well even in the situation of fluctuations in data availability with ANN as well as GP and comparing the results with respect to accuracy of forecast.

II. LITERATURE REVIEW

ANN is one of the proven tools in the field of air quality modelling and forecasting whereas GP is relatively new approach which is evident from the literature references.

Bonzar et al. (1993)[3] constructed a multilayer perceptron to predict atmospheric sulphur dioxide concentrations in a highly polluted industrialised area of Slovenia. Yi and Prybutok (1996) [4] described a multilayer perceptron that predicts surface ozone concentrations in an industrialised area of North America. Comrie (1997) [5] compared ozone forecasts made by multilayer perceptron and regression models. Gardner and Dorling (1999) [6] used neural network for hourly prediction of NO_x and NO_2 concentrations in London. Kolehmainen et al. (2001) [7] used neural network & periodic component for air quality forecasting. Dahe Jiang et al. (2004) [8] developed ANN model to forecast air pollution index for Shanghai. Shiva Nagendra and Mukesh Khare (2004) [9] developed ANN based line source model for vehicular exhaust emission predictions of urban roadways of India. G. Grivas and A. Chaloulakou (2006) [10] used ANN for prediction of PM_{10} hourly concentration in Greece. Saleh M. Al-Alawi et al. (2008) [11] used ANN for prediction of ground level concentration of ozone. Atakankurt and Ayse Betul Oktay (2010) [12] used ANN to forecast air pollutant indicator levels three days in advance. Jose C. M. Pires et al. (2010) [13] have used Multigene Gnetic Programming for one day ahead prediction of PM_{10} in Portugal. Pires et al. (2011) [2] 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. Tikhe et al. (2013) [14] carried out a comparative study to forecast one day ahead, criteria air pollutants for Pune using ANN & GP and found that, GP proved to be better compared to ANN.

As far as air quality management of Pune is concerned, Central Pollution Control Board (CPCB) efforts are mainly directed towards hot spot area monitoring and control strategy management. Indian Institute of Tropical Meteorology is actively involved in development of latest emission inventories using GIS methodology.

There are no evidences of application of GP for air pollution forecasting of Pune. The present work attempts to use GP approach for one day ahead prediction of indicator pollutants considering data fluctuations.

III. ARTIFICIAL NEURAL NETWORKS

Artificial Neural Networks (ANNs) 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 strength. ANN 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 [15].

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 underlying physical process. The ANN model is given in Figure 1.

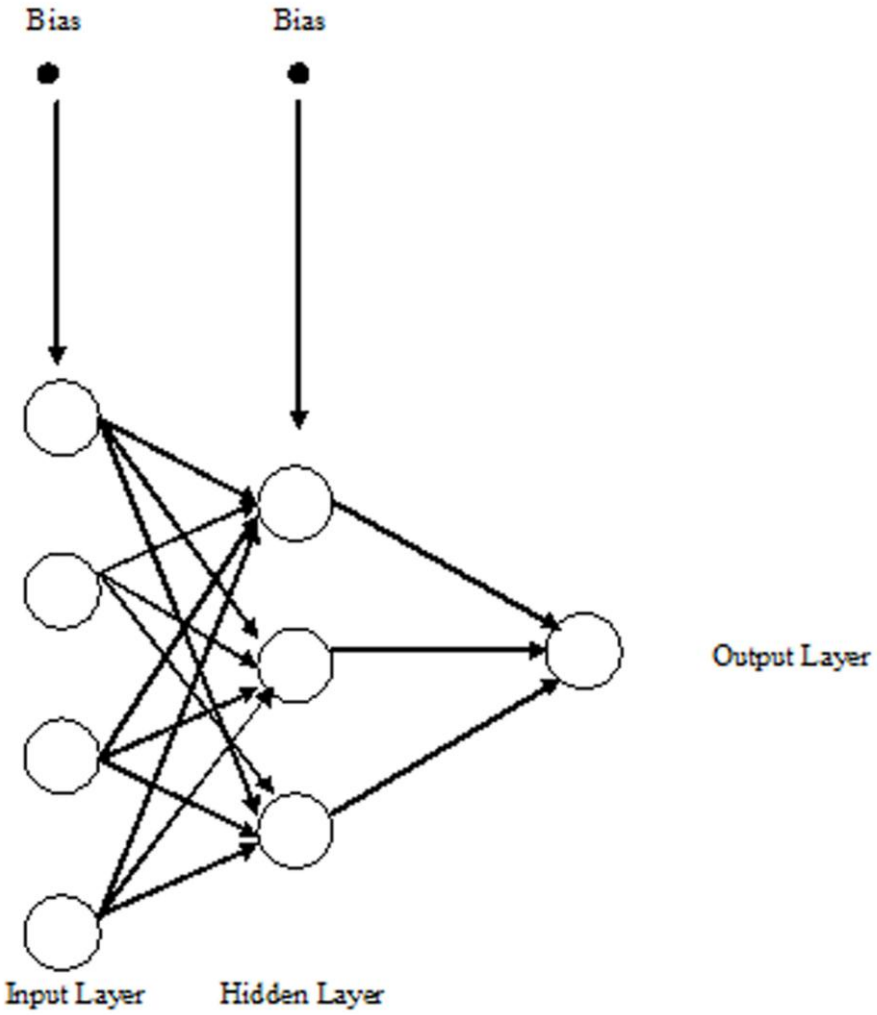

Fig. 1. The ANN Model

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,

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

where sum = ($X_1W_1+X_2W_2+$ …..) +β .....(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 [16].

All training algorithms are basically aimed at reducing the global error, E, between the network output and the actual observation, as defined below:

$$E = \sum (O_n - O_t)^2 \quad \text{.....(3)}$$

Where O_n is the network output at a given output node and O_t is 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 (1998) [7].

The present paper uses three layered Feed Forward Back Propagation neural network to predict SO_x, NO_x and RSPM levels one day in advance for Pune (State:-Maharashtra of India) using commercial software MATLAB 07.

IV. GENETIC PROGRAMMING (GP)

Genetic programming (GP) is an artificial intelligence methodology that uses principles of the Darwin's Theory of Evolution. Its search strategy is based on Genetic Algorithms (GA) introduced by John Holland in 1960s [17]. GA use bit strings as chromosomes and are commonly applied in function

optimization. This algorithm has several disadvantages, for example, the length of the strings is static [18] . Additionally, the size and the shape of the model, solution of a given problem are generally not known in advance. Similar to GA, the GP introduced by Koza in 1990s [18] , is based on simple rules that imitate biological evolution. It is a good alternative to GA due to its valuable characteristic such as the flexible variable-length solution representation. Moreover, GP enables the automatic generation of mathematical expressions. The expressions are represented as tree structures which contain functions as nodes and terminals as leafs. Terminals are the input variables and constants and functions are all operators that are available to solve the problem .GP uses the genetic operations (selection, crossover and mutation). In selection, part of population (the fittest individuals) is retained and the remainder new generation is the result of genetic operations on the individuals of the actual population. In crossover, two individuals are selected, their tree structures are divided at a randomly selected crossover point and the resulting sub-trees are recombined to form two new individuals. In mutation, a random change is performed on a selected individual by substitution. Offspring are produced in a generation and further till another specified numbers of generations are created through the process of crossover and mutation. Detailed explanation of concepts related to GP can be found in [18] . Figure 2 shows the typical process of Genetic Programming.

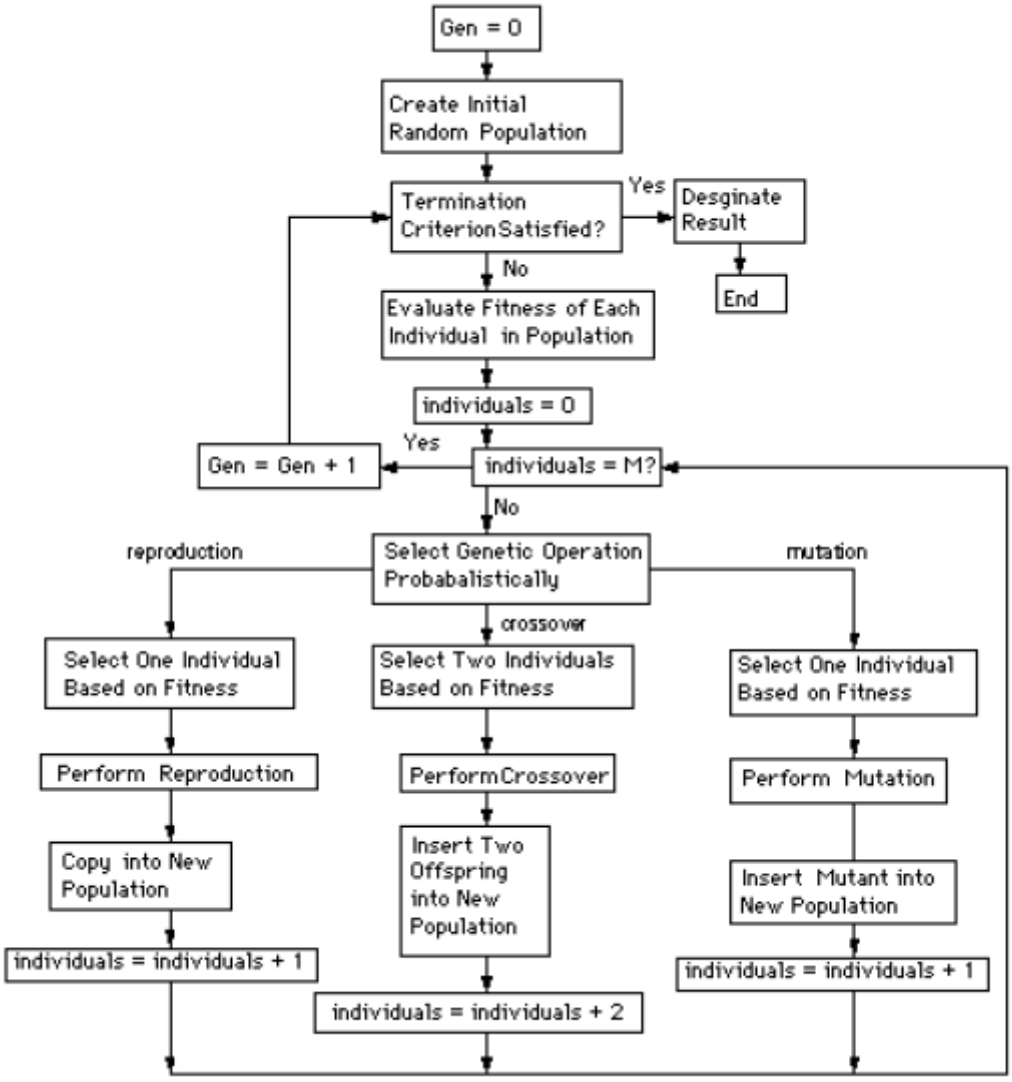

Fig.2. Typical GP flowchart

V. STUDY AREA AND DATA

Pune is one of the fastest developing metropolitan cities in India which generates about 181.957 tonnes of toxic waste daily [19] . It is located in Western Maharashtra on the Deccan Plateau at the confluence of Mula Mutha Rivers and at an elevation of about 560m above mean sea level at Karachi. Location sketch of Study area can be found in Figure 3.

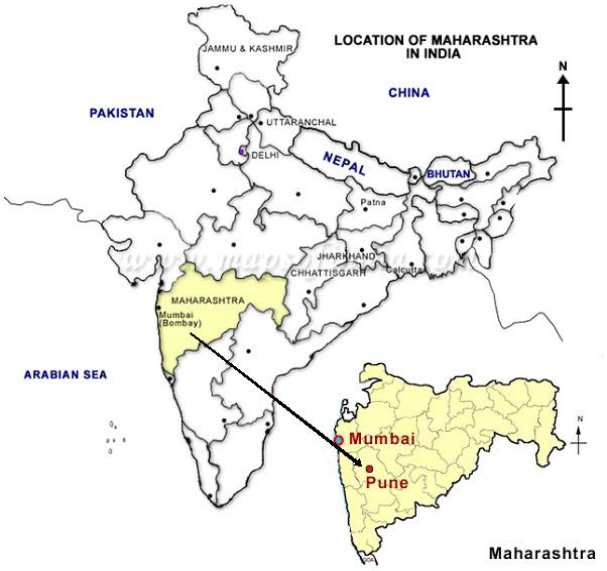

Fig. 3. The Study Area (PUNE, MAHARASHTRA, INDIA)

Accelerating growth in the transport sector, booming construction industry and growing industrial sector are responsible for deteriorating air quality of the city, which has resulted into bad health impacts. There has been a continuous rise in the level of criteria (indicator) air pollutants namely SO_x, NO_x and RSPM. Under National Ambient Air Monitoring Program (NAMP) and also State Ambient Air Monitoring Program(SAMP), the upper limits set for daily average concentrations of SO_x, NO_x and RSPM by CPCB (Central Pollution Control Board, India) and MPCB (Maharashtra Pollution Control Board) are 80, 80 and 100 $\mu g/m^3$ respectively. Since last few years it has been observed that the upper limits have been crossed for SO_x , NO_x and RSPM with a maximum value recorded as high as 195 , 138 and 370 $\mu g/m^3$ respectively . The present study aims at predicting values of SO_x, NO_x and RSPM, one day in advance which can provide an indication about the prevailing air quality on the next day.

Data used for the study consists of daily average values of metrological parameters and pollutant concentrations recorded for the period of January 2005 to December 2008 by India Meteorological Department (IMD) and MPCB respectively for Pune city.

Meteorological parameters such as rainfall (RF) is measured by rainguages , temperature difference (TD) is recorded by thermometer, relative humidity (RH) is measured by hygrographs, station level pressure & vapour pressure (SLP &VP) is measured by barometer , solar radiation (SR) is recorded using solarimeter and wind speed (WS) is measured by anemometer. Pollutant concentrations are recorded using High Volume Sampler by Improved West and Gaeke Method for SO_x , Sodium Arsenite Method for NO_x and by Filter paper method for RSPM. As the previously measured data of the above mentioned criteria pollutants is used for this work, the data driven approaches of Artificial Neural Networks and Genetic Programming are employed to develop one day ahead continuous forecasting models of SO_x, NO_x and RSPM and the results are compared for forecasting accuracy.

VI. MOTIVATION

Air pollution is a nonlinear problem which consists of interaction of various elements. Those elements can be grouped as meteorological parameters, emission inventory, terrain characteristics, traffic characteristics and previous pollutant concentrations. Air quality modelling considering all the above elements is a realistic approach towards forecasting but it is expensive .It is very difficult to acquire data of all above referred parameters. Instrumental errors, adverse meteorological conditions etc. are also the reasons for non availability of continuous and accurate data of all the meteorological parameters as well as pollutant concentrations.

In such situations, the models developed should be robust enough to be useful for prediction of short term concentration of the criteria air pollutants.

The process of air pollution 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 parameters 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 require the formulation of search problem and the investigation of search techniques which are capable of facilitating model development work and

resulting in more reliable and robust ANN models. In this context, GP has proven to be a powerful technique due to its ability to solve linear and non-linear problems as it can capture the underlying trend better than ANN by exploring all regions of the state space and utilizing promising areas through genetic operations [20]. Further GP can also result into an equation which can be used for real time forecasting.

It is challenge to develop a pollution forecasting model which can run on a continuous basis especially when data availability is a major constraint. In this paper an attempt has been made to develop a continuous forecasting model using ANN and GP for each of the three criteria pollutants.

VII. MODEL DEVELOPMENT

One day ahead pollution forecasting model has been developed taking into account all the conditions of availability of the data.

When all the meteorological parameters responsible for the phenomenon and previous pollutant concentrations are available the model can be written as

Model A-

SO_x (t+1) = f(RF, TD, SLP, VP, WS, SR, RH,SO_x) t .....(4)

NO_x (t+1) = f(RF, TD, SLP, VP, WS, SR, RH,NO_x) t .....(5)

RSPM (t+1) = f(RF, TD, SLP, VP, WS, SR, RH,RSPM) t .....(6)

This could be the most sensible model as it includes almost all the major meteorological parameters which are measured daily , but it would only work when the data pertaining to all the inputs is available, which is not always the situation. Hence correlation of all the inputs (the meteorological parameters as well as previous pollutant concentrations) with the output (the pollutant concentration) is calculated and the inputs are arranged in the order of their correlation with the output (Table I).

TABLE I
Correlation of inputs with the output

Sr. No.	Input parameter	Correlation (r) with		
		SO_x	**NO_x**	**RSPM**
1	Relative Humidity (RH)	0.268	0.131	0.220
2	Wind Speed(WS)	0.024	0.260	0.375
3	Solar Radiation(SR)	0.200	0.076	0.207
4	Station Level Pressure (SLP)	0.277	0.062	0.344
5	Vapour Pressure(VP)	0.077	0.072	0.596
6	Temperature Difference (TD)	0.364	0.133	0.517
7	Rainfall (RF)	0.201	0.030	0.153
8	Previous concentration of the pollutant (t-1)	0.725	0.675	0.812

The three most influential inputs as identified by correlation analysis have been considered in the second model which is designated as forecasting model with top three causes (Table II).

TABLE II
The most and the least influential parameters responsible for the phenomenon

Sr. No.	Pollutant	Top 3 causes	Bottom 3 causes
1	SO_x	SO_x(t-1),TD,SLP	WS, VP, SR
2	NO_x	NO_x (t-1), WS, TD	RF, SLP, VP
3	RSPM	RSPM(t-1),VP, TD	RF, SR, RH

Model B-

SO_x (t+1) = f(top 3 inputs) t .....(4)

NO_x (t+1) = f(top 3 inputs) t .....(5)

RSPM (t+1) = f(top 3 inputs) t .....(6)

This model can be adopted when top three meteorological parameters affecting the process are available.

In the worst case scenario, if these top most inputs are not available , then the effect of three bottom most inputs (as identified by correlation analysis and depicted in Table 2) is studied on pollution prediction by a forecasting model which considers only bottom three causes.

Model C-

SO_x (t+1) = f(bottom 3 inputs) .....(7)
NO_x (t+1) = f(bottom 3 inputs) t..... (8)
RSPM (t+1) = f(bottom 3 inputs) t.....(9)

As the bottom three inputs may have least correlation with the output, the forecast results may not be acceptable. Hence the fourth model was developed which combines the bottom three causes with the top most cause with an intention to develop a forecasting model with reasonable accuracy. This model can be stated as...

Model D-

SO_x (t+1)= f(bottom 3 inputs + topmost input) .....(10)
NO_x(t+1)= f(bottom 3 inputs + topmost input) t .....(11)
RSPM(t+1)= f(bottom 3 inputs + topmost input) .....(12)

The above two models were specifically developed with an aim to identify the success rate of prediction in absence of all top most inputs responsible for the phenomenon and also increase in the prediction accuracy when one of the topmost input is coupled with bottom three inputs.

Sometimes a situation may arise when none of the meteorological parameters is recorded .In that case there is no alternative than developing a model based on previous values of the pollutant concentrations.

Model E-

SO_x(t+1)= f(SO_x(t),SO_x(t-1),SO_x(t-2).....).....(13)
NO_x(t+1)= f(NO_x(t),NO_x(t-1),NO_x(t-2).....).....(14)
RSPM(t+1)= f(RSPM(t),RSPM(t-1),RSPM(t-2) .....).....(15)

The above mentioned temporal model relies only on the previous values of the pollutant concentration without any consideration to the cause of the pollution. In order to make the model sensible the time series of the previous values of the pollutant concentration should be coupled with at least one of the influential cause of the pollution.

Temperature difference has the significant impact on pollutant concentrations during all the seasons which are experienced in Pune. Temperature difference also stands amongst the top three causes responsible for all the three pollutants as identified by correlation study (Table 1). Here the correlation analysis exactly matches with the physics; consequently it would be reasonable to consider time series of temperature difference along with previous value of the pollutant concentration.

Model F-

SO_x (t+1)= f(SO_x (t), TD (t), TD(t-1)).....(16)
NO_x(t+1)= f(NO_x(t),TD(t),TD(t-1)).....(17)
RSPM (t+1)= f(RSPM (t), TD (t), TD(t-1)).....(18)

The above mentioned six types of models have been considered with an objective to develop a short term criteria pollutant forecasting model for each of the three pollutants. It can work reasonably well without interruption, inspite of the data fluctuations.

A. Criteria Used for ANN based Pollutant Forecast Model

For the present study, couple of trials were taken to decide data division for the models. Training and testing dataset, varying from 40 % - 85% (for training and remaining data for testing) were taken and found that 60-80% data for training and 40-20% of data for testing yield better results . Hence the same range of data division is used for all eighteen ANN models. Readers are requested to refer to [21] for more details of the training and testing data division . Table III indicates the criteria used for ANN models. The MATLAB Neural Network toolbox is used to develop models based on above criteria.

TABLE III
Criteria for ANN model

Sr. No.	Item	Criteria used in the present study
1	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
2	Neuron activation function	Input neuron= Identity function Output neuron= Identity function Hidden Neuron= Hyperbolic tangent function ‘logsig’ and ‘purelin’ for all the models
3	Learning parameters	The learning parameters converge to the network configuration and give best performance on the validation data with least epochs

4	Criteria for initialisation of the network weights	Network weights are uniformly distributed in the range of -1 to 1
5	Training algorithm	Levenberg Marquardt
6	Stopping criteria	Performance goal / epochs
7	Performance indicator	r, RMSE, d

B. Criteria Used for GP based Pollutant Forecast Model

Eighteen GP models were developed for the same data. The data divisions for these models were adopted as similar to 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 IV indicates the GP parameters used for the present study. Commercial software GP kernel was used to develop the GP models. GP has resulted into an equation which can be used for real time forecasting.

TABLE IV
GP control parameters

Max Init Size	15- 20
Max Size	45-50
Population size (mu)	50-1000
Number of children to produce (lamda)	100-3000
Function set	**+ ,-,/,*,** sqrt
Breeding method	Tournament

VIII. MODEL ASSESSMENT

The testing performance of all thirty six models was assessed 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 results and r = 1.00 signifies 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 variant is accurately estimated by the simulated variant. 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 which can be easily interpreted and cross-compared for a variety of models regardless of units. 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 only '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 [9]. In such situations for air quality models it is likely that 'd'statistics is the most relevant evaluation criteria.

IX. RESULTS AND DISCUSSION

Above mentioned six different conditions of data availability has been considered for short term prediction of each of the three pollutants. ANN and GP are used with an aim to develop continuous forecasting models for Pune. ANN has been used as a tool for air quality forecasting since last few decades as mentioned in the literature review. Authors have found several advantages of ANN such as adaptive learning, self organisation, real time operation, capability of handling nonlinear systems etc. Considering this, ANN is used in this study as the benchmarking tool for one day ahead prediction of criteria pollutants.

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 (r) , root mean square error (RMSE) and 'd' statistics between the observed and forecasted values. GP has advantage of yielding pollution forecasting equation which can be handled easily while using the model for real time

forecasting. Whereas with other proven tools, obtaining pollution forecast is time consuming and require the skill of the programmer to run a model. The ANN as well as GP models exhibited a reasonable performance in testing between the observed and forecasted pollutant concentrations for all the models. This is clearly evident from the mentioned performance indicators as depicted in the Table V (A), (B) and (C).

TABLE V (A)
Results SO_x model

Model	**Tools**					
	ANN			**GP**		
	r	**RMSE**	**d**	**r**	**RMSE**	**d**
A (considering all input parameters)	0.656	3.570	0.786	0.660	3.521	0.786
B(Top three input parameters)	0.671	3.762	0.768	0.673	3.735	0.777
C (Bottom three input parameters)	0.126	5.896	0.437	0.303	5.810	0.442
D(Bottom three and top one input parameters)	0.672	5.819	0.769	0.678	5.819	0.806
E(Temporal)	0.670	3.610	0.798	0.706	2.873	0.815
F (temperature & previous pollutant concentration)	0.664	3.661	0.791	0.670	3.604	0.797

TABLE V (B)
Results NO_x model

Model	**Tools**					
	ANN			**GP**		
	r	**RMSE**	**d**	**r**	**RMSE**	**d**
A (considering all input parameters)	0.722	8.360	0.815	0.800	7.238	0.885
B(Top three input parameters)	0.683	8.572	0.792	0.774	7.810	0.878
C (Bottom three input parameters)	0.088	12.132	0.174	0.108	13.302	0.366
D(Bottom three and top one input parameters)	0.731	8.010	0.825	0.860	6.027	0.919
E(Temporal)	0.765	7.694	0.869	0.809	6.920	0.881
F (temperature & previous pollutant concentration)	0.743	5.741	0.822	0.750	5.614	0.837

TABLE V (C)
Results RSPM model

Model	**Tools**					
	ANN			**GP**		
	r	**RMSE**	**d**	**r**	**RMSE**	**d**
A (considering all input parameters)	0.814	28.904	0.881	0.835	26.881	0.904
B(Top three input parameters)	0.834	26.760	0.907	0.838	26.413	0.909
C (Bottom three input parameters)	0.541	43.934	0.572	0.546	42.310	0.624

D(Bottom three and top one input parameters)	0.824	27.573	0.895	0.826	27.330	0.901
E(Temporal)	0.746	32.628	0.857	0.831	26.908	0.903
F (temperature & previous pollutant concentration)	0.823	27.369	0.901	0.830	27.196	0.903

a. Analysis of forecasting models

Model A –

It is the forecasting model which is operative when the chief meteorological parameters as well as previous pollutant concentration are available. Seven meteorological parameters namely RF, TD, SLP, VP, WS, SR, RH are considered along with previous pollutant concentration for this model. Both ANN and GP models worked reasonably well as far as NO_x and RSPM models are considered. This is evident from the decreased value of RMSE and increased value of r and d for GP compared to ANN. In the case of SO_x ; RMSE is decreasing, r is increasing and d remains the same for both GP and ANN.

Model B-

This model considers top three meteorological causes responsible for the phenomenon. This model is suitable when the data pertaining to atleast top three causes is available. The results of all GP models are better than that of ANN models for all the three pollutants. If the results of model A are compared with model B, it can be clearly indicated that model B can be opted when a limited data set of only top most causes is available. This model works better with a marginal compromise on RMSE and d values for SO_x model; r, RMSE and d for NO_x model and without any compromise for RSPM model .

Model C-

This model is developed considering the worst case scenario of non availability of the data pertaining to all causes or the top most causes. The data may be available for the bottom most causes responsible for the phenomenon. Hence the bottom three causes are considered with a view to assess the deviation in the results and to decide the suitability of the model.

The results of GP model are better compared to ANN for all the three pollutants but the accuracy of the prediction drops by 81%, 85% & 54% for SO_x, NO_x and RSPM respectively compared to that of the top three models. Hence type C model should not be used as it involves significant compromise on the results.

Model D-

Considering the failure of type C model, it is necessary to couple at least one of the significant cause with the bottom most causes. Therefore model D is developed with an assumption of availability of the data of at least one of the top most cause with three bottoms most causes. This situation may arise when only a few data measuring instruments are available. This situation will be even worse if these instruments are recording the bottom most causes.

GP shed better for all the three pollutants as compared to ANN for model type D. Sharp rise in the accuracy is seen when the topmost input is coupled with bottom three inputs. The results of model type D are almost the same as that of Model A (all causes) and model B (top three causes) above . This gives clear indication that any of the above models (except type C) can be used for real time forecasting of criteria pollutants.

Model E-

Situation may arise when the data pertaining to none of the cause is available and the pollution predictions are highly essential especially in the worst climatic conditions. In this case the only alternative available is to rely on the previous pollutant concentrations.

The temporal model (Model E) exhibits better result for GP predictions compared to ANN for all the three criteria pollutants. The results of temporal model (Model E) are even better than that of the cause effect models (Model A, Model B and Model D). Pollution is a time dependent phenomenon and previous concentrations play a great role in the value of the next day concentration. This could be the reason for the better results of Model E compared to cause effect models. But the temporal model cannot be considered as a full proof model as it lacks the important aspect of the physics behind the process. Hence this model should only be used in emergencies and with care.

Model F-

A situation of non availability of the meteorological parameters may persist for few days due to bad weather conditions and we are compelled to use model E. In this case model E can be refined

by combining it with at least one of the cause which is the most significant as well as which can be measured with relative ease and accuracy with only a few instruments. Temperature difference is the parameter which can be recorded easily and also has a great impact on the pollutant concentrations. From the correlation analysis of the available data set, temperature difference is positioned amongst the top three causes responsible for the phenomenon. This is also consistent with the physics of the process. Thus it would be a practical approach to combine temperature difference with previous pollutant concentrations.

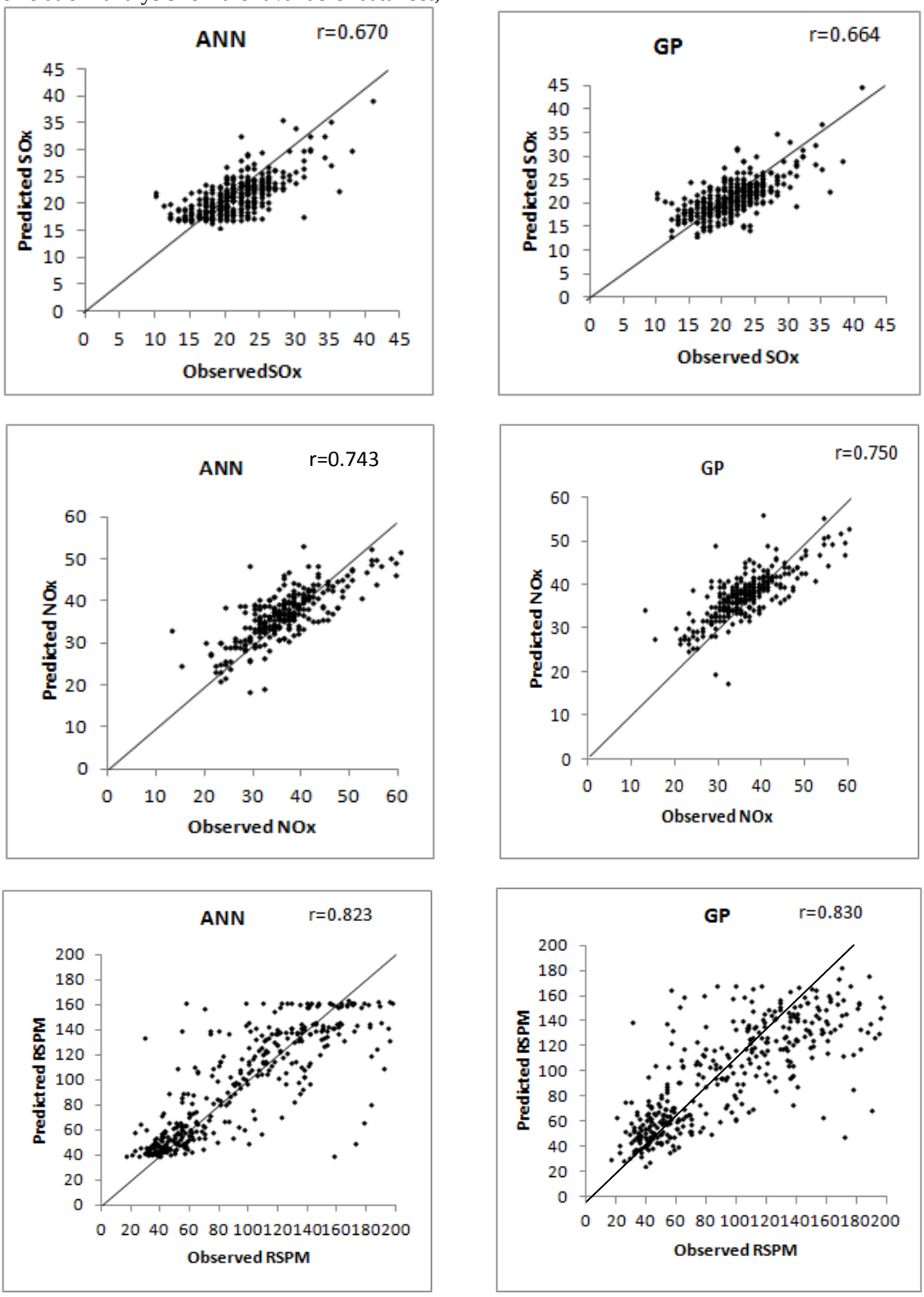

Fig. 4. Scatter Plots for model F

GP models performed better compared to ANN models for all the three pollutants which is evident from the above scatter plots. As far as r and d is concerned GP supersedes ANN with a little compromise on RMSE results for SO_x model.

The results of model E and Model F are almost the same for SO_x model and slight deviation of +/- 2-3% is observed for NO_x and RSPM models . Model F is always preferred compared to model E as it considers the significant element of physics behind the process rather than just the previous concentration of the pollutant.

Considering the results of all thirty six models it appears that GP always works better compared to ANN for all different conditions of the data. This has an added advantage of pollution forecasting equation which can be used for real time forecasting of criteria air pollutants on a continuous basis for Pune city.

X. CONCLUSION AND FUTURE SCOPE

At the outset, the study was planned in order to develop continuous air quality forecasting models irrespective of availability of the consistent data pertaining to causes affecting air quality. Thirty six models have been developed for short term prediction of criteria air pollutants namely SO_x, NO_x and RSPM for Pune city (Maharashtra State, India). These models are taking into account the fluctuations in data availability due to various reasons.

The models are developed using ANN as well as GP and the results are compared for the accuracy of forecast. It was found that GP works better than ANN with the advantage of selection of relevant inputs & development of an equation which can be used for real time forecasting of the pollutants.

This study leads to various options available for real time forecasting of criteria pollutants so that they can be predicted on a continuous basis for Pune city, which also happens to be one of the most polluted metropolitan cities of India.

GP being a relatively new approach needs to be explored further for long term forecast of criteria pollutants with certain considerations such as climatic conditions and seasonal variations.

XI. ACKNOWLEDGEMENT

Authors are grateful to India Meteorological Department (IMD) as well as Maharashtra Pollution Control Board (MPCB) for kindly providing meteorological and air quality data respectively.

XII. APPENDIX

Following are the equations generated by GP using software GP kernel.Out of eighteen equations generated by respective models only representative equations for model F are quoted

1. One day ahead SO_x prediction
   $SO_x(t+1)$=(sqrt(sqrt(TD(t))) + (7 + ((TD(t-1) - sqrt(TD(t-1))) - sqrt((TD(t-1) - sqrt(TD(t-1)))))))
2. One day ahead NO_x prediction
   $NO_x(t+1)$= ((((22.6347485 + 19.857563) – TD(t-1)) / (((23.0365028 + 19.857563) + (29.5049706 + (TD(t-1) + (((20.849411 + 20.076767) – TD(t-1)) / (((22.5991402 + (20.6623459 + TD(t-1)) + (TD(t-1) + (23.5913677 + 23.6871414))) / TD(t-1)))))) / TD(t-1))) + TD(t-1))
3. One day ahead RSPM Prediction
   RSPM(t+1)=(((6-sqrt(((TD(t)+sqrt(exp(sqrt((sqrt(sqrt(exp(sqrt((((sqrt(exp(sqrt((sqrt(sqrt((TD(t-1)+TD(t))))+sqrt(TD(t-1)))))) + TD(t)) + sqrt(TD(t))) + TD(t-1)))))) + TD(t-1)))))) + TD(t)))) + TD(t-1)) + TD(t))

REFERENCES

[1] Padmanabha Murthy B., *Environmental Meteorology*, I. K. International Publishing House Pvt. Ltd., New Delhi,2009.

[2] Pires, J. C. M., Alvim- Ferraz, M. C. M., Pariera, M. C. and Martins, F. G., "Prediction of troposphere ozone concentration: Application of a methodology based on Darwin's Theory of Evolution", *Expert Systems with Applications*, vol.38,issue 3, pp.1903-1908, Mar. 2011.

[3] Bonzar, M., Lesjak, M. and Mlakar, P., "A neural nwtwork-based method for the short-term predictions of ambient SO_2 concentrations in highly polluted industrial areas of complex terrain", *Atmospheric Environment. Part B. Urban Atmosphere*, vol.27, issue 2, pp.221-230, June 1993.

[4] Yi, J. And Prybutok, R., "A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialised urban area", *Environmental Pollution*, vol. 92, issue 3, pp.349-357, 1996.

[5] Comrie, A. C., "Comparing neural networks and regression models for ozone forecasting", *Journal of Air and Waste Management*. Vol. 47, issue 6, pp.653-663, 1997.

[6] Gardner, M. W. and Dorling, S. R.,"Artificial Neural Networks : The multilayer perceptron : a review of applications in the atmospheric sciences", *Atmospheric Environment*, Vol. 32, issue 14-15 , pp.2627—2636, Aug .1998.

[7] Kolehmainen M., H. Martikainen, J. Ruuskanen, "Neural networks and periodic components used in air quality forecasting". *Atmospheric Environment*, Vol.35, issue 5, pp.815-825, 2001.

[8] Dahe Jiang, Yang Zhang, Xiang Hu, Yun Zeng, Jianguo Tan, "Progress in developing an ANN model for air pollution index forecast*"*, *Atmospheric Environment*. Vol.38,issue 40, pp. 7055-7064,Dec. 2004.

[9] Shiva Nagendra and Khare, M., "Artificial neural network based line source model for vehicular exhaust emission predictions of an urban roadway", *Transportation Research Part D*,vol. 9, issue 3, pp.199-208, May 2004.

[10] G. Grivas, A. Chaloulakou, "Artificial neural network models for prediction of PM 10 hourly concentrations, in the Greater Area of Athens, Greece", *Atmospheric Environment*, vol. 40,issue 7 ,1pp. 216-1229, Mar. 2006.

[11] Saleh M. Al-Alawi ,Sabah A. Abdul-Wahab, Charles S. Bakheit, "Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone", *Environmental Modelling & Software*, Vol.23,issue 4,pp.396-403, Apr. 2008.

[12] Atakan Kurt , Ayse Betül Oktay, "Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks", *Expert Systems with Applications*., vol. 37,issue 12 , pp. 7986-8992, Dec.2010.

[13] Jose, C. M. Pires., Maria, C. M., Alvim – Ferraz, Pariera, M. C. and Martins, F. G., "Prediction of PM10 concentration through Multigene genetic programming" , *Atmospheric Pollution Research*, Vol.1,pp. 305-310, 2010.

[14] Tikhe Shruti S.,Dr. Mrs. Khare K. C., Dr. Londhe S. N., "Forecasting Criteria Air Pollutants Using Data Driven Approaches: An Indian Case Study", IOSR *Journal of Environmental Science, Toxicology and Food Technology*, vol. 3, issue 5, pp.01-08,Mar- Apr. 2013.

[15] The ASCE Task Committee on Application of Artificial Neural Networks in Hydrology "Artificial Neural Networks in Hydrology. I: preliminary concepts", *Journal of Hydrologic Engineering*, ASCE, vol. 5 , issue 2, pp.115—123,Apr. 2000.

[16] Bose, N. K., Liang, P., *Neural Network Fundamentals with Graphs, Algorithms and Applications*, Tata McGraw-Hill Publication, Delhi, 2000.

[17] Goldberg David Edward, *Genetic Algorithms in Search, Optimisation and Machine Learning.* Pearson publication, 1989.

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

[19] Central Institute of Road Transport (CIRT) Report, Inventory of emission load from vehicles in Pune city and PCMC, Revised Action Plan for Control of Air Pollution in Pune, 2000.

[20] Niska H., Teri H., Karppinenb A.,Ruuskanen J., and Kolehmainen M., "Evolving the neural network model for forecasting air pollution time series", *J .Engineering Applications of Artificial Intelligence*,vol.17, issue 2 ,pp. 159-167, Mar. 2004.

[21] Khare, M. and Nagendra, S. A., *Artificial Neural Networks in Vehicular Pollution Modelling*, Studies in Computational Intelligence, pp.41 -45, 2007.