Abstract: The purpose of the present study is to develop a model to forecast the concentrations of some important atmospheric pollutants over Kolkata (22° 32′N; 88° 20′E), India during the period from 1st April 2009 to 30th November 2010 with considerable accuracy and adequate lead time. The pollutants considered in this study are respiratory suspended particulate matter (RSPM), nitrogen oxide (NOx), and sulphur oxide (SOx). The auto regressive (AR) models with different orders and radial basis function network (RBFN) model are developed to attain the objective. The skill of both the models is compared. The results of the study reveal that the 3rd order Auto-Regressive Model, AR (3) represents the best statistical model for the prediction of concentrations of all the three different pollutants over Kolkata. The study thus, depicts that the pollutants can be predicted with considerable accuracy and 3 days or 72 hours lead time using AR (3) model. The skill of the AR (3) model is compared with RBFN model. The result further reveals that the percentage error in forecast with 72 hours lead time is much less with RBFN model than AR model.

Keywords: Concentration of pollutants; prediction; AR model; RBFN model

I. INTRODUCTION

Air pollution is a broad term applied to all chemical and biological agents that modify the natural characteristics of the atmosphere and it may be defined as a situation in which substances that result from anthropogenic activities are present at concentrations sufficiently high above their normal ambient levels to produce a measurable effect on humans, animals, vegetations, materials, environment and weather.

The atmosphere is a complex, dynamic natural gaseous system that is essential to support life on planet earth. Worldwide air pollution is responsible for large numbers of casualties and cause of respiratory diseases. Urban air pollution is one of the most important environmental problems and is the result of human activities. It has diverse causes and sources, such as industrial, commercial, agricultural and domestic activities. In the recent times, there has been growing concern about the consequence of air pollution on environment and atmospheric weather. Clean dry air consists nitrogen (78.09%), oxygen (20.94%) by volume and the remaining 0.97% is composed of a gaseous mixture of carbon dioxide (CO₂), helium (He), argon (Ar), krypton (Kr), ozone (O₃), nitrous oxide (N₂O) and xenon (Xe), as well as very small amount of some other organic and inorganic substances whose amount in the atmosphere vary with time and place. Air pollutants may exist in gaseous or particulate form. Concentrations are commonly expressed either in mass per unit volume (μg/m³ of air) or as a volume mixing ratio (1 ppm = 10⁻⁶ v/v; 1 ppb = 10⁻⁹ v/v). Gaseous and particulate air pollutants may be separated operationally by use of a filter.

Air pollutants emitted directly into the atmosphere from a source is termed primary. Thus, carbonaceous particles from diesel engine exhaust and sulphur dioxide from power stations are examples of primary pollutants. In contrast, secondary pollutants are not emitted as such, but are formed within the atmosphere. The most commonly considered secondary pollutant is ozone, formed as a result of molecular oxygen in the stratosphere, and of nitrogen dioxide in the troposphere.

The present study is aimed at forecasting the concentrations of different pollutants, respiratory suspended particulate matter (RSPM), nitrogen oxide (NOx), and sulphur oxide (SOx) over Kolkata. The study adopts the method of statistical lagged time series analysis. Different orders of auto – regressive (AR) models are attempted. The AR models with minimum prediction error (%), minimum white noise variance, minimum BIC (Bayesian Information Criteria) and AIC (Akaike Information Criteria) values are identified as the best predictive models. Radial basis function network (RBFN) model is developed to forecast the concentration of the said pollutants. The forecast with Auto regressive (AR) model and radial basis function network (RBFN) model are compared to assess the skill of AR and RBFN models in forecasting the concentration of the said pollutants.

II. MATERIAL AND METHOD

A. Data Analysis

There prevails a variety of air pollutants in the Earth’s atmosphere and diversion from an allowable concentration of the pollutants lead to environmental as well as health hazards [1, 2]. The data sets explored in this study consist of the daily average (24 - hrs average) concentrations of Respiratory Suspended Particulate Matter (RSPM), Nitrogen Oxides (NOx) and Sulphur Oxides (SOx).

Among the three, RSPM is the most damaging component of air-borne particulate contamination and cannot be blocked effectively through cloth or paper filters. Exposure to RSPM can affect breathing and aggravate respiratory and cardiovascular diseases.
NOx include various nitrogen compounds like nitrogen dioxide (NO2) and nitric oxide (NO) which form acid rain, contribute to global warming, hamper the growth of plants and individually, may affect ecosystems, both on land and within water.

SOx is released primarily from burning fuels that contains sulfur (like coal, oil and diesel fuel) and high concentrations of which can affect breathing and may aggravate existing respiratory and cardiovascular diseases.

The data is collected from 1st April 2009 to 30th November 2010 over Kolkata, India. The concentrations of the pollutants (μg/m³) are collected from Central Pollution Control Board (CPCB), Kolkata. The total number of observations is without any gap, so that the pollutant concentrations can generate temporally equispaced data point to create three different time series for each of the pollutants.

B. Methodology

The methodologies adopted in this study are:

1. Computation of Autocorrelation Coefficients (ACC) for lag1, 2, 3
2. Computation of Percentage Error of Prediction (PE) for different AR models.
4. Application of Bayesian and Akaike Information Criteria (BIC and AIC).
5. Identification & Selection of best Predictive Model.
6. Development of Radial Basis Function Network (RBFN) Model
7. Computation of Percentage Error of Prediction (PE) with RBFN model
8. Skill comparison of AR model and RBFN model in predicting the concentration of atmospheric Pollutants

C. Statistical Approach

One of the very important applications of statistical method in meteorology is in making sense of new set of data. In most of the cases, the data are paired. Often an abbreviated, single - valued measure of association between two variables, x and y, is needed. In such situations, the data analysts almost automatically and sometimes fairly uncritically compute a correlation coefficient [3]. Usually, the term “correlation coefficient” is used to mean the “Pearson product - moment coefficient of linear correlation” between two variables x and y.

Pearson correlation coefficient (PCC) is the ratio of the sample covariance of the two variables to the product of the two standard deviations. It measures the degree of association between the variables x & y;

$$\rho_{xy} = \frac{\text{cov}(x,y)}{s_x s_y}$$

Mathematically, PCC is expressed as,

\[
\rho_{xy} = \left[ \frac{1}{n-1} \sum_{i=1}^{n} [(x_i - \bar{x})(y_i - \bar{y})] \right] \left[ \left( \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{\frac{1}{2}} \left( \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \right)^{\frac{1}{2}} \right]^{\frac{1}{2}}
\]

\[
= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y}) \left[ \left( \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{\frac{1}{2}} \left( \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \right)^{\frac{1}{2}} \right]^{\frac{1}{2}}
\]

The primes denote the anomalies or subtraction of the mean values from the original values. The Pearson product -moment correlation coefficient is neither robust nor resistant. It has two important properties; firstly, the correlation coefficients lie between (–) 1 and 1, that is $-1 \leq \rho_{xy} \leq 1$.

If $\rho_{xy}$ is equal to 1, there is perfect positive linear association between the two variables, x and y. The second property is that the correlation coefficient is not affected either by a change in scale or by a change in origin. The said correlation coefficient is relevant for the cases where two parameters are involved. In meteorology, there are cases where the dependence of a time series upon its own past history is to be discerned. In such cases the concept of Pearson correlation coefficient is mapped into autocorrelation coefficient (ACC). To view the concept, the following time series is considered.

The random variable $X_t$ realizes the value $x_t$ at time $t$. Thus the series is;

$$X_1, X_2, X_3, \ldots, \ldots, X_n$$

To construct the autocorrelations of different lags, the same time series is written beside the original series with one or more values according to the lag. For example, for Lag-1 ACC the same series is written in two different ways:

$$X_1, X_2, X_3, \ldots, \ldots, X_n$$

$$X_1, X_2, X_3, \ldots, \ldots, X_{n-1}, X_n$$

The Lag-1 ACC is the PCC for the pairs

$$\left( X_{2}, X_{1} \right), \left( X_{3}, X_{2} \right), \ldots, \left( X_{n}, X_{n-1} \right)$$

Similarly, higher order autocorrelations can be computed by shifting the series to further time steps. The concept of ACC discussed so far is important for analysis of continuous time series that requires auto regression approach [4]. Since the last seventies, statistical approach has become a popular tool to predict meteorological phenomena over the globe. Autoregressive models [4] are suitable for prediction of time series. Such models can represent the correlation structure of a time series very successfully. The most widely used time series model for prediction through autoregressive analysis is Box-Jenkins model [4].
The simplest Box-Jenkins model is the first order autoregressive [AR(1)] model, where the predictand is the value, \( x_{t+1} \), of the time series at \( t+1 \), and the predictor is the current value of the time series, \( x_t \). The AR (1) model can be presented as,

\[ x_{t+1} - \mu = \phi (x_t - \mu) + \epsilon_{t+1} \] (2)

Where, \( \mu \) is the mean of the time series; \( \phi \) is the autoregressive parameter; and \( \epsilon_{t+1} \) is the random quantity corresponding to the residual in the ordinary regression. The time series of \( x \) is assumed stationary, so that the mean is the same for each interval of time. The autoregressive model, as presented in equation (2) can represent the serial correlation of a time series. Parameter estimation in the first order autoregressive model is straightforward. The mean of the time series, \( \mu \), is simply the sample average of the dataset, under the assumption that the time series is stationary. The estimated autoregressive parameter is equal to the sample lag-1 autocorrelation coefficient. This estimation can be presented as,

\[ \hat{\phi} = r_1 \] (3)

Here, \( r_1 \) is the sample autocorrelation coefficient for lag-1. For the resulting probability model to be stationary, it is required that \(-1 < \phi < 1\). The full autocorrelation function for a time series governed by a first-order autoregressive process is given by [1],

\[ r_k = \phi^k \] (4)

The first-order autoregressive model, as described the equation (2) predicts \( x_{t+1} \) using \( x_t \) as predictor. The first order autoregressive process can readily be extended up to any order \( k \) to obtain autoregressive predictive models of higher order, i.e. to predict \( x_{t+1} \) using predictors further back in time. The general autoregressive model to order \( K \) or AR (K) model is,

\[ x_{t+1} - \mu = \sum_{k=1}^{K} \phi_k (x_{t-k+1} - \mu) + \epsilon_{t+1} \] (5)

Obviously, the AR (K) model requires \( K \) autoregressive parameters \( \phi_k \) to be estimated. This estimation is most easily done by applying Yule-Walker equations [5],

\[ r_1 = \hat{\phi} + \hat{\phi} r_1 + \hat{\phi}^2 r_2 + \cdots + \hat{\phi}^K r_{K-1} \]

\[ r_2 = \hat{\phi} r_1 + \hat{\phi}^2 r_2 + \cdots + \hat{\phi}^{K-1} r_{K-2} \]

\[ \vdots \]

\[ r_K = \hat{\phi}^{K-1} r_1 + \hat{\phi}^{K-2} r_2 + \cdots + \hat{\phi}^0 \]

Here, \( \hat{\phi}_k \) = 0 for \( k > K \).

The theoretical autocorrelation function corresponding to a particular set of the \( \phi_k \) values can be determined by solving (6) and then applying

\[ \rho_m = \sum_{k=1}^{K} \phi_k \rho_{m-k} \] (7)

Equation (7) holds for lags \( m \) greater than or equal to \( k \), with \( \rho_0 = 1 \).

Autoregressive model AR (2) and AR (3) are two special cases of equation (5).

For AR (2), Yule-Walker equations will be,

\[ r_1 = \hat{\phi}_1 + \hat{\phi}_1 r_1 + \hat{\phi}_2 r_2 \]

\[ r_2 = \hat{\phi}_1 r_1 + \hat{\phi}_2 r_2 + \hat{\phi}_3 r_3 \] (8)

For AR (3), Yule-Walker equations will be,

\[ r_1 = \hat{\phi}_1 + \hat{\phi}_1 r_1 + \hat{\phi}_2 r_2 + \hat{\phi}_3 r_3 \]

\[ r_2 = \hat{\phi}_1 r_1 + \hat{\phi}_2 r_2 + \hat{\phi}_3 r_3 + \hat{\phi}_4 r_4 \]

\[ r_3 = \hat{\phi}_1 r_2 + \hat{\phi}_2 r_3 + \hat{\phi}_3 r_4 \] (9)

The sample residual variance \( s^2_e(m) \), also known as white noise variance, is computed [3],

\[ s^2_e(m) = \left[ 1 - \hat{\phi}_m^2(m) \right] s^2(m-1) \] (10)

Here, \( s^2_e(m) \) is the estimated white-noise variance of the \( m \)th auto regression and \( s^2_e(m-1) \) is the estimated white-noise variance for the previously fitted (one order smaller) model. For autoregressive model of lag - 0, the white noise variance \( s^2_e(0) \) is simply the variance for the time series itself.

Yule-Walker equations can potentially be used to autoregressive models of any order. However, in order to have a predictive model of the best fit for the time series, it is essential to decide which model is the best fit for the time series. The Bayesian Information Criteria popularly known as BIC statistic [6] and Akaike information Criteria popularly known as AIC statistic [7] are used to decide among alternative orders of autoregressive models. For any order \( m \), the BIC and AIC statistics are expressed as,

\[ BIC(m) = n \ln \left( \frac{n \sum_{m=1}^{n} s^2_e(m)}{n - m - 1} \right) + (m+1) \ln n \] (11)

\[ AIC(m) = n \ln \left( \frac{n \sum_{m=1}^{n} s^2_e(m)}{n - m - 1} \right) + 2(m+1) \] (12)

While experimenting with autoregressive models of different orders, the order having minimum AIC and minimum BIC is identified to provide the best autoregressive predictive model for the time series.

D. Radial Basis Function Network (RBFN)

The Radial Basis Function Network (RBFN) is embedded in a three layer neural network with one input layer, one hidden layer and one output layer, where each hidden unit implements a radial activation function. In the present study, Gaussian activation function is used for RBFN;
TABLE 1
TABULAR PRESENTATION OF THE ORDER SELECTION OF AUTOREGRESSIVE MODELS TO PREDICT THE RSPM CONCENTRATION (APRIL 2009 TO NOVEMBER 2010) OVER KOLKATA.

<table>
<thead>
<tr>
<th>Lag , m</th>
<th>rm</th>
<th>Sr2 (m)</th>
<th>BIC (m)</th>
<th>AIC(m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.74813</td>
<td>3607.3099</td>
<td>2388.6548</td>
<td>2381.3150</td>
</tr>
<tr>
<td>2</td>
<td>0.66395</td>
<td>3405.0738</td>
<td>2378.6016</td>
<td>2367.5920</td>
</tr>
<tr>
<td>3</td>
<td>0.64321</td>
<td>3256.0007</td>
<td>2372.3013</td>
<td>2357.6218</td>
</tr>
</tbody>
</table>

TABLE 2
TABULAR PRESENTATION OF THE ORDER SELECTION OF AUTOREGRESSIVE MODELS TO PREDICT THE NOx CONCENTRATION (APRIL 2009 TO NOVEMBER 2010) OVER KOLKATA.

<table>
<thead>
<tr>
<th>Lag , m</th>
<th>rm</th>
<th>Sr2 (m)</th>
<th>BIC (m)</th>
<th>AIC(m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.74813</td>
<td>3607.3099</td>
<td>2388.6548</td>
<td>2381.3150</td>
</tr>
<tr>
<td>2</td>
<td>0.66395</td>
<td>3405.0738</td>
<td>2378.6016</td>
<td>2367.5920</td>
</tr>
<tr>
<td>3</td>
<td>0.64321</td>
<td>3256.0007</td>
<td>2372.3013</td>
<td>2357.6218</td>
</tr>
</tbody>
</table>

TABLE 3
TABULAR PRESENTATION OF THE ORDER SELECTION OF AUTOREGRESSIVE MODELS TO PREDICT THE SOx CONCENTRATION (APRIL 2009 TO NOVEMBER 2010) OVER KOLKATA.

<table>
<thead>
<tr>
<th>Lag , m</th>
<th>rm</th>
<th>Sr2 (m)</th>
<th>BIC (m)</th>
<th>AIC(m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.32794</td>
<td>6.9657</td>
<td>576.2377</td>
<td>568.8979</td>
</tr>
<tr>
<td>2</td>
<td>0.36006</td>
<td>6.4081</td>
<td>558.7185</td>
<td>547.7088</td>
</tr>
<tr>
<td>3</td>
<td>0.34858</td>
<td>6.1555</td>
<td>553.7377</td>
<td>539.0582</td>
</tr>
</tbody>
</table>

\[
\Phi_j(X) = \left[ (X - \mu_j)^T \sum_j (X - \mu_j)^{-1} \right] \tag{13}
\]

For \( j = 1, \ldots, L \); where \( L \) is the number of hidden unit and \( X \) in the input matrix. \( \mu_j \) and \( \sum_j \) are the mean and covariance matrix of the \( j \)th Gaussian function. The output layer implements a weighted sum of the hidden units output as;

\[
\Psi_k(X) = \sum_{j=1}^{L} \lambda_{jk} \Phi_j(X) \tag{14}
\]

For \( k = 1, \ldots, M \); where \( M \) is the number of output unit and \( \lambda_{jk} \) are the output weights.

The quality of prediction is assessed from the performance of the test set of data. Percent errors of prediction (PE) are calculated as [8];

\[
PE = \left( \frac{Y_{dp} - Y_{da}}{Y_{da}} \right) \tag{16}
\]

Where, \( \left\{ \right\} \) denotes the average of test cases. The predicted and actual values of the parameters are denoted by \( Y_{dp} \) and \( Y_{da} \) respectively. The root mean square error (RMSE) and mean absolute error (MAE) are computed for testing (validation) the output data;

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_{da} - Y_{dp})^2} \tag{17}
\]

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |Y_{dp} - Y_{da}| \tag{18}
\]

The predicted and actual values of the parameters are denoted by \( Y_{dp} \) and \( Y_{da} \) respectively.

III. RESULTS & DISCUSSIONS

The autoregressive models with different orders and radial basis function network (RBFN) are developed to forecast the concentration of different pollutants over Kolkata during the period from 1st April, 2009 to 30th November 2010 covering all the seasons of the year. The result show that the temporal variability of the pollutants possesses a definite pattern (Figures 1-3).

The result further depicts that the accuracy in the forecast of pollutants is more with AR (3) model (fig4).

The percentage error in prediction (PE) with AR (3) model is observed to be minimum among the different orders of autoregressive models for all the three pollutants (Figure 4). The white noise variance is observed to be minimum with AR (3) model than other orders of autoregressive models for all the three pollutants (Tables 1 - 3).

![Fig. 1 Diagram showing the actual and predicted RSPM concentration (in µg/m³) using (a) AR 1, (b) AR 2 and (c) AR 3 models.](image-url)
Fig. 2. Diagram showing the actual and predicted NOx concentration (in µg/m³) using (a) AR 1, (b) AR 2 and (c) AR 3 models.

Fig. 3. Diagram showing the actual and predicted SOx concentration (in µg/m³) using (a) AR 1, (b) AR 2 and (c) AR 3 models.

Fig. 4. Diagram showing the prediction error of different Auto Regressive (AR) models for (a) RSPM, (b) NOx and (c) SOx concentration forecast.

The BIC & AIC statistics also observed to be minimum with AR (3) model than the other different orders of autoregressive models for all the three pollutants (Tables 1 - 3).

The statistical approach makes it evident that the forecast of the concentration of selected pollutants over Kolkata can be done with 3rd order auto regressive model AR (3) with maximum accuracy and 3 day (72 hours) lead time. The forecast quality of AR (3) model is observed to be better than AR (1) and AR (2) models (Figure 4).

The radial basis function network (RBFN) developed to forecast the concentration of RSPM, NOx and SOx with 3 days lead time shows better forecast potential than the AR models (Figure 5).
The quality of forecast with different models is assessed by computing error matrices. The error matrices are formed with three types of performance errors like percentage error (PE), mean absolute error (MAE) and root mean square error (RMSE) (equations - 16, 17 and 18). The best statistical auto regressive model AR (3) shows the values of PE, MAE and RMSE to be 31%, 26% and 22.5% for the said pollutants while RBFN forecast depicts the values of PE, MAE and RMSE to be 16%, 13% and 13.5% respectively (Figure 6). The superiority of RBFN forecast for NOx and SOx has also been revealed (Figure 4 and 6).

Figure 5. Diagram showing the actual and 3 days advance predicted values of (a) RSPM, (b) NOx and (c) SOx concentrations using Radial Basis Function Network (RBFN) model.

IV. CONCLUSION

All the analyses lead to conclude that neural network method (RBFN) has got supremacy over statistical methods in medium range forecast (72 hours advance) of the concentrations of selected pollutant over Kolkata. However the 3rd order auto-regressive model, AR (3) is observed to be the best statistical model than the other two auto regressive model (AR (2) and AR (1)) for the prediction of concentrations of all the three different pollutants, namely, RSPM, NOx and SOx over Kolkata.

The present study is devoted to observe the potential of auto – regressive and radial basis function network models to forecast the concentration of different atmospheric pollutants.
individually. Future perspective is to incorporate the meteorological parameters with the environment parameters to view the influence of one and / or the other on the variability of atmospheric weather.

ACKNOWLEDGEMENT

The author acknowledges Central Pollution Control Board (CPCB) for sharing the pollution data and Council of Scientific and Industrial Research (CSIR) for financial assistance. The authors acknowledge the anonymous reviewers for the constructive comments, which aided in improving the clarity of the paper.

REFERENCES