Contents lists available at SciVerse ScienceDirect

Science of the Total Environment

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

Short Communication

Forecasting of daily air quality index in Delhi

Anikender Kumar, P. Goyal *

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

ARTICLE INFO

Article history:
Received 2 February 2010
Received in revised form 25 August 2011
Accepted 29 August 2011
Available online 1 October 2011

Keywords:
Auto regressive integrated moving average (ARIMA)
Principal component regression (PCR)
Air quality index (AQI)
Meteorological variables
Pollutants

ABSTRACT

As the impact of air pollutants on human health through ambient air address much attention in recent years, the air quality forecasting in terms of air pollution parameters becomes an important topic in environmental science. The Air Quality Index (AQI) can be estimated through a formula, based on comprehensive assessment of concentration of air pollutants, which can be used by government agencies to characterize the status of air quality at a given location. The present study aims to develop forecasting model for predicting daily AQI, which can be used as a basis of decision making processes. Firstly, the AQI has been estimated through a method used by US Environmental Protection Agency (USEPA) for different criteria pollutants as Respirable Suspended Particulate Matter (RSPM), Sulfur dioxide (SO_2), Nitrogen dioxide (NO_2) and Suspended Particulate Matter (SPM). However, the sub-index and breakpoint concentrations in the formula are made according to Indian National Ambient Air Quality Standard. Secondly, the daily AQI for each season is forecasted through three statistical models namely time series auto regressive integrated moving average (ARIMA) (model 1), principal component regression (PCR) (model 2) and combination of both (model 3) in Delhi. The performance of all three models are evaluated with the help of observed concentrations of pollutants, which reflects that model 3 agrees well with observed values, as compared to the values of model 1 and model 2. The same is supported by the statistical parameters also. The significance of meteorological parameters of model 3 has been assessed through principal component analysis (PCA), which indicates that daily rainfall, station level pressure, daily mean temperature, wind direction index are maximum explained in summer, monsoon, post-monsoon and winter respectively. Further, the variation of AQI during the weekends (holidays) and weekdays are found negligible. Therefore all the days of week are accounted same in the models.

1. Introduction

Air pollution related problems have resulted in an increased public awareness of the air quality in both developing and developed countries (Kurt and Oktay, 2010). There are many air pollutants adversely affecting human health in the polluted air such as carbon monoxide (CO), RSPM, SO_2, NO_2, SPM, ozone (O_3), etc. The high concentration of these pollutants can be life threatening, causing breathing difficulty, headache and dizziness. They may even result in heart attacks (Kunzli et al., 2000). Thus, the authorities advise the monitoring and forecasting of criteria pollutants in the air. The forecasting of air pollutants can be made through models. The Gaussian dispersion models are, generally, used for air quality prediction in most of the air pollution studies. Although, the dispersion models have some physical basis, detailed information about the source of the pollutants and other parameters are not generally known (Chelani et al., 2002). In order to overcome these limitations, the statistical models are used, which facilitate the forecasting of pollutant concentrations (Finzi and Tebaldi, 1982; Ziomass et al., 1995; Polydoras et al., 1998).

AQI is an important task for general public to understand easily how bad or good the air quality is for their health and to assist in data interpretation for decision making processes related to pollution mitigation measures and air quality management. Basically, the AQI is defined as an index or rating scale for reporting daily combined effect of ambient air pollutants recorded in the monitoring sites. Recently, Van den Elshout et al. (2008) gave the review of existing air quality indices. A regression model was also used by Cogliani (2001) for air pollution forecast in cities by an air pollution index highly correlated with meteorological variables. A study of Goyal et al. (2006) is made for daily air quality forecasting of air pollutants in Delhi through ARIMA and multiple linear regression (MLR) models. The most of the air quality forecasting studies present in the literature have been made for individual air pollutants.

The main objective of the present study is to develop forecasting models for predicting the daily air quality indices, which can provide the timely information to the public to take precautionary measures to protect their health.

2. Study area

Delhi is the capital city of India spread over 1483 km^2 with 16.9 million inhabitants in 2007–08. Due to the presence of large number

* Corresponding author. Tel.: +91 11 26591309; fax: +91 11 26591386.
E-mail address: pramila@cas.iitd.ernet.in (P. Goyal).

0048-9697/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.scitotenv.2011.08.069

of industries and migration of people from neighboring states, nearly 56.27 lakh vehicles are plying on Delhi roads (Economic Survey of Delhi, 2008–2009). Delhi has one of the highest road densities as 1749 km of road length per 100 km^2 in India. Its high population growth rate, coupled with high economic growth rate, has resulted in ever-increasing demand for transportation creating excessive pressure on the city's existent transport infrastructure. The city faces acute transport management problems leading to air pollution, congestion and resultant loss of productivity.

Delhi has a semi-arid climate with high variation between summer and winter temperatures. Because of its proximity to the Himalayas, cold waves from the Himalayan region dip temperatures across it. The average annual rainfall is approximately 714 mm, most of which falls during the monsoons, in July and August (Economic Survey of Delhi, 2008–2009). In view of air pollution, the most important season in Delhi is winter (December–February). This period is dominated by cold, dry air and ground-based inversion with low wind conditions ($u \leq 1$ m s^{-1}), which occur frequently and increases the concentration of pollutants (Anfossi et al., 1990). The summer season (March, April, May) is governed by high temperature and high winds, the monsoon season (June, July, August) is dominated by rains and post-monsoon season is influenced (September, October, November) by moderate temperature and wind conditions.

2.1. Air quality data

In this study, daily air quality data of RSPM, SO_2, NO_2 and SPM over a period of 2000–2006 at ITO, obtained from Central Pollution Control Board (CPCB), Delhi has been used. ITO is chosen in the present study as it has a continuous air quality monitoring station of CPCB and is a busiest traffic intersection, located in central Delhi.

2.2. Meteorological data

The 24 hourly averaged surface meteorological variables as daily maximum temperature (t_{max}), minimum temperature (t_{min}), daily temperature range (difference between daily maximum and minimum temperature, t_{range}), mean temperature (t_{mean}), wind speed (wsp), wind direction index (wdi), relative humidity (rh), vapor pressure (vp), station level pressure (slp), rainfall (rf), sunshine hours (ssh), cloud cover (cc), visibility (v) and radiation (rd), observed at Safderjung Airport, about 5.7 km to ITO in Delhi, have been acquired from the Indian Meteorological Department (IMD), Pune for the 7 years period of 2000–2006.

3. Methodology

There are primarily two steps involved for forecasting of daily AQI.

i) The estimation of AQI through USEPA method using daily observed concentration of air pollutants. In this formulation, the formation of sub-indices of each pollutant and the breakpoints aggregation of sub indices are made according to the Indian National Ambient Air Quality Standard. The results of epidemiological studies are indicating the risk of adverse health effects of specific pollutants. In order to assess the status of air quality and its effects on human health, the range of index values, applicable to Indian cities have been taken from earlier study by Nagendra et al. (2007), which reflects the different ranges as “[a]Good (0–100)”, “[b]Moderate (101–200)”, “[c]Poor (201–300)”, “[d]Very Poor (301–400)” and “[e]Severe (401–500)” (Table 1), where all the values of SO_2, NO_2, RSPM and SPM are in μg/m^3.

[a] Good: air quality is acceptable; however, for some pollutants, there may be a moderate health concern for a very small number of people.

[b] Moderate: members of sensitive groups may experience health effects.

[c] Poor: members of sensitive groups may experience more serious health effects.

[d] Very poor: triggers health alter, everyone may experience more serious health effects.

[e] Severe: triggers health warnings of emergency conditions.

Table 1
Propose sub-index and breakpoint pollutant concentration for Indian-AQI.

SI. no.	Index values	Descriptor	SO_2 (24-h avg.)	NO_2 (24-h avg.)	RSPM (24-h avg.)	SPM (24-h avg.)
1	0–100	Good[a]	0–80	0–80	0–100	0–200
2	101–200	Moderate[b]	81–367	81–180	101–150	201–260
3	201–300	Poor[c]	368–786	181–564	151–350	261–400
4	301–400	Very poor[d]	787–1572	565–1272	351–420	401–800
5	401–500	Severe[e]	>1572	>1272	>420	>800

Superscripted letters has been taken directly from the reference Nagendra et al. (2007).

The AQI formula (EPA, 1999) for four criteria pollutants RSPM, SO_2, NO_2 and SPM is given as:

$$I_P = \left[\frac{(I_{Hi}-I_{Lo})}{(BP_{Hi}-BP_{Lo})}\right](C_P-BP_{Lo}) + I_{Lo}, \tag{1}$$

where I_P = the AQI for pollutant ‘p’,

C_P actual ambient concentration of the pollutant ‘p’,
BP_{Hi} the breakpoint given in Table 1 that is greater than or equal to C_p,
BP_{Lo} the breakpoint given in Table 1 that is less than or equal to C_p,
I_{Hi} the sub index value corresponding to BP_{Hi},
I_{Lo} the sub index value corresponding to BP_{Lo}.

The overall AQI is now determined on the basis of the AQI for above pollutant ‘p’ and highest among them is declared as the overall AQI for that day. In the present study, AQI is determined at ITO for a period of seven years (2000–2006), which has also been analyzed with respect to weekdays (Mon–Fri) and weekends (Sat and Sun) on seasonal basis.

ii) The three statistical models are used for forecasting the daily AQI. These models are namely time series auto regressive integrated moving average (ARIMA) (model 1), principal component regression (PCR) (model 2) and combination of both (model 3) and their brief formulations are explained in Appendix A. The AQI, as estimated in step 1, and daily averaged meteorological variables, listed in previous section, for a period of 2000–2005 have been used as input parameters to the models for training purposes. The same process has been followed for all four seasons.

Once the training of all three models is completed, the above models are used to predict daily air quality of the future year 2006. It is noticeable that data of the year 2006 has not been used in building the models. Further, the forecasted values of daily AQI, resulted from models, are evaluated by comparing them with daily observed concentrations of air pollutants of the year 2006. A quantitative assessment of accuracy of model's output has been made through the statistical parameters.

4. Results and discussion

The daily AQI has been estimated using monitored concentrations of criteria pollutants in all the four seasons over the period from 2000 to 2006. The percentage of very poor and severe descriptors of AQI is found to be 81.52%, 81.51%, 69.54% and 38.04% during summer, winter, post-monsoon and monsoon seasons respectively. As one can see, summer and winter seasons have very poor and severe descriptors of

AQI. It can be expected in winter due to worst meteorological scenario. However, the same status of AQI is observed in summer, which may be due to the accumulation of dust particles, originating from neighboring areas in Delhi. They are increasing the RSPM and SPM concentrations. The percentage of very poor and severe descriptors in post-monsoon and monsoon seasons is less in comparison to those found in summer and winter seasons, may be due to the washing out of the pollutants by precipitation.

The daily AQI, as obtained above, and meteorological parameters, as listed in previous section, are used as input to the models for forecasting the daily AQI. The AQIs of weekends and weekdays have also been analyzed quantitatively in order to the see the variation of concentration of air pollutants in different days of the week, which are found as 336.29, 267.48, 313.15 and 333.96 (weekdays) and 334.55, 259.77, 309.43 and 327.95 (weekends) in summer, monsoon, post monsoon and winter seasons respectively. All the seven days of the week are accounted same in the model, since the variation of AQI in different days in week are likely negligible.

In previous studies of air pollution forecasting (e.g., Cogliani, 2001; Sousa et al., 2007 etc.), the previous day's values of explanatory variables (air quality and meteorological variables) are considered as input to the forecasting models. In the present study, the same practice has been followed. The logic behind using one previous day's values of explanatory variables is explained clearly through an analysis of correlation coefficient between given day's AQI and input variables up to three previous day's as shown in Table 2, which reflects that the correlation coefficients between given day's AQI and 1st previous day, 2nd previous day and 3rd previous day are found to be positive and raging from 0.78 to 0.35, in all the four seasons. However, only one previous day's AQI has been considered because of the better correlation compared to two previous days or three previous days. Finally, one previous day's data is decided to be considered as input to the models.

First of all, the model 1 of order (1, 1, 1) has been trained using the previous 6 years 2000 to 2005 of AQIs for forecasting the daily AQI of the year 2006. The time series are developed for forecasting the AQI as given below:

$$w_t = 0.64 + \alpha_t - 0.99\ \alpha_{t-1} \tag{2}$$

$$w_t = 0.71 + \alpha_t - 0.98\ \alpha_{t-1} \tag{3}$$

$$w_t = 0.32 + \alpha_t - 0.75\ \alpha_{t-1} \tag{4}$$

$$w_t = 0.64 + \alpha_t - 0.99\ \alpha_{t-1,} \tag{5}$$

where $\alpha_t \approx \mathrm{NID}\ (0, \sigma^2)$, and σ^2 is the variance of white noise.

A same set of equations as given below is developed and used for forecasting the daily AQI in summer, monsoon, post-monsoon and winter seasons respectively in the year 2006. The evaluation of model's forecasted and observed AQIs has been made through statistical parameters and is given in Table 3, which shows that the model is performing satisfactory in all the seasons and has better results in winter season with respect to the Normalized Mean Square Error (NMSE), Root Mean Square Error (RMSE) and correlation coefficient. However, the model is under-predicting in the summer, monsoon and winter seasons and is over-predicting only in post-monsoon season with respect to fractional bias.

Table 2
Correlation between AQI with previous three day's AQI for summer, monsoon, post monsoon and winter seasons.

S.N	Season	Correlation between AQI with previous day's AQI	Correlation between AQI with previous two day's AQI	Correlation between AQI with previous three day's AQI
1	Summer	0.6654	0.4789	0.3460
2	Monsoon	0.7750	0.6350	0.5290
3	Post Monsoon	0.7786	0.7062	0.6317
4	Winter	0.6368	0.4576	0.3727

The model 2 as already discussed in above Section 3, is used in different seasons, based on the daily input data of explanatory variables of the years 2000–2005, the model is resulted 7 principal components (PC's) in all seasons except post-monsoon and cumulative variance 87.54, 84.26 and 79.51 in summer, winter and monsoon respectively. Similarly the 6 PC's, and cumulative variance 81.29 in post-monsoon season, are used as input to model.

Further, the equations of the model 2, as given below, have been developed for different seasons based on above data:

$$\begin{aligned}[\mathrm{AQI}] = {} & 0.1150 - 0.0980\times[\mathrm{PC1}] - 0.1660\times[\mathrm{PC2}] + 0.0530\times[\mathrm{PC4}] \\ & + 0.1070\times[\mathrm{PC5}] + 0.0550\times[\mathrm{PC6}] - 0.1580\times[\mathrm{PC7}]\end{aligned} \tag{6}$$

$$\begin{aligned}[\mathrm{AQI}] = {} & -0.0520 - 0.2960\times[\mathrm{PC1}] - 0.0410\times[\mathrm{PC2}] - 0.2790\times[\mathrm{PC3}] \\ & + 0.1860\times[\mathrm{PC5}]\end{aligned} \tag{7}$$

$$\begin{aligned}[\mathrm{AQI}] = {} & 0.0980 - 0.1410\times[\mathrm{PC1}] + 0.2240\times[\mathrm{PC2}]\times 0.0730\times[\mathrm{PC3}] \\ & - 0.2170\times[\mathrm{PC5}]\end{aligned} \tag{8}$$

$$\begin{aligned}[\mathrm{AQI}] = {} & 0.0310 - 0.2000\times[\mathrm{PC1}] + 0.0660\times[\mathrm{PC3}] + 0.0800\times[\mathrm{PC4}] \\ & - 0.1530\times[\mathrm{PC5}] - 0.2190\times[\mathrm{PC6}],\end{aligned} \tag{9}$$

where PC1, PC2, …, PC7 are defined as principal components 1, 2, …, 7. The daily AQI of the year 2006 as resulted from model 2 are compared with those of model 1 and observed values in order to evaluate the models' performance (Table 3).

Table 3 indicates that the model 2 is over-predicting in the summer and post-monsoon and is under-predicting in monsoon and winter seasons, whereas model 1 is over predicting only in post-monsoon season.

The above results reflect that the combination of model 1 and model 2 may achieve better results compare to those of individual models. However, the model 2 itself deals with the issue of collinearity and model 1 is exploiting the autocorrelation in the explanatory variables.

Working of the model 3, which is combination of model 1 and model 2 of the years 2000 to 2005, is as follows:

First of all, the output of model 1 with previous day's explanatory variables is used as input to the model 3, which is formed in the correlation matrix for each season separately. These predictor variables, which include explanatory variables and output of model 1, are transformed into principal components through eigenvalue matrix of these variables. Those PC's, who have the cumulative amounts of variance approximately 80%, are retained and rests of the components are ignored and are shown in Table 4. Only 7 PC's in each season with cumulative variance 84.72, 83.27, and 79.99 instead of 16 variables in winter, summer and monsoon seasons respectively are given as input to this model. Similarly 6 PC's with cumulative variance 81.73 in post-monsoon season are used as input to the model. The Communalities of each original variable using the first 7 PC's in summer, winter and monsoon seasons and 6 PC's in post-monsoon season have also shown in Table 5, which reveals that these PC's “explain” 99.14% of the rainfall, 99.29% of the station level pressure, 96.10% of mean temperature and 98.71% of wind direction index in summer, monsoon, post-monsoon and winter seasons respectively.

The equations of the model 3 are developed on the basis of above data for all four seasons separately as shown below. In order to find out the significance of the variables, the *t*-test is applied to the above equations. The variables, those are found insignificant/statistically invalid, have been removed from the model equations. Finally these

Table 3
Comparison of ARIMA and PCR models predicted and observed values for year 2006.

S.N.	Season	ARIMA				PCR			
		RMSE	NMSE	Correlation coefficient	Fractional Bias	RMSE	NMSE	Correlation coefficient	Fractional Bias
1	Summer	35.30	0.0106	0.6641	0.0022	35.70	0.0113	0.7286	−0.0002
2	Monsoon	62.55	0.0537	0.5664	0.0132	68.93	0.0694	0.6557	0.0878
3	Post Monsoon	57.70	0.0354	0.5381	−0.0083	62.93	0.0416	0.6188	−0.0212
4	Winter	26.74	0.0054	0.6475	0.0011	58.39	0.0301	0.6748	0.1492

equations are used to forecast daily AQI in different seasons of the year 2006.

$$[\mathrm{AQI}]\mathrm{S} = -0.0440 - 0.1940 \times [\mathrm{PC1}] - 0.1570 \times [\mathrm{PC2}] + 0.0994 \times [\mathrm{PC5}] + 0.0417 \times [\mathrm{PC6}] - 0.1540 \times [\mathrm{PC7}] \tag{10}$$

$$[\mathrm{AQI}]_{\mathrm{M}} = -0.0490 - 0.3110 \times [\mathrm{PC2}] - 0.0560 \times [\mathrm{PC3}] - 0.2450 \times [\mathrm{PC4}] + 0.1130 \times [\mathrm{PC5}] + 0.0870 \times [\mathrm{PC6}] \tag{11}$$

$$[\mathrm{AQI}]_{\mathrm{PM}} = 0.0203 - 0.1860 \times [\mathrm{PC1}] + 0.2040 \times [\mathrm{PC2}] + 0.0589 \times [\mathrm{PC3}] - 0.212 \times [\mathrm{PC5}] \tag{12}$$

$$[\mathrm{AQI}]_{\mathrm{W}} = -0.071 - 0.2570 \times [\mathrm{PC1}] + 0.0584 \times [\mathrm{PC3}] + 0.1450 \times [\mathrm{PC4}] - 0.0944 \times [\mathrm{PC5}] - 0.1920 \times [\mathrm{PC6}] - 0.0888 \times [\mathrm{PC7}], \tag{13}$$

where $[\mathrm{AQI}]_{\mathrm{S}}$, $[\mathrm{AQI}]_{\mathrm{M}}$, $[\mathrm{AQI}]_{\mathrm{PM}}$ and $[\mathrm{AQI}]_{\mathrm{W}}$ are forecasted AQI for summer, monsoon, post-monsoon and winter seasons respectively. The statistical evaluation of model's forecasted AQI with observed AQI has been made and is given in Table 6. It is observed in Table 3 and Table 6 that model 3 is performing better than model 1 and model 2 in all the seasons with respect to most of the statistical parameter. Fractional bias shows that model 3 is under-predicting in all the seasons except post-monsoon season for the year of 2006.

In order to limit the presentation to manageable size, graphical presentation of model 3 is given instead of model 1 and model 2 for all four seasons. It is justifiable because it is the combination of model 1 and model 2. The graphical presentation of model 3 is shown in Fig. 1 (a), (b), (c) and (d) of all four seasons for the period 2000–2005. The correlation coefficients (R) between observed and model 3's forecasted values of years 2000–2005 have found as 0.6721, 0.7801, 0.7896 and 0.6397 in summer, monsoon, post-monsoon and winter seasons respectively. The Eqs. (10)–(13) are used for forecasting the daily AQI of the year 2006. The comparison of model's forecasted and observed AQI of the year 2006 has been shown graphically in Fig. 2 (a), (b), (c) and (d) for summer, monsoon, post-monsoon and winter seasons respectively. It is noticeable that observed and predicted AQIs are found to be the maximum and minimum in the same season i.e., monsoon season, which is also supporting the good performance of model 3.

The Statistical evaluation between observed and model's predicted values for a period of 2000–2005 and the year 2006 has been made for different seasons as given in Table 6. The NMSE and correlation coefficient (R) are found as (0.0086, 0.7539) in summer season which are followed by (0.0341, 0.6237) in post monsoon; (0.0390, 0.6804) in winter and (0.0531, 0.6710) in monsoon seasons during 2006, which indicates that the overall performance of model 3 has been found to be better than those of model 1 and model 2. Seasonal performance of model 3 is also discussed and found that model 3 is performing better in summer compared to other seasons. In the end the results for forecasting of daily AQI using three statistical techniques are summarized as: ARIMA (model 1), based on time series approach but PCR (model 2) involves the effect of meteorological variables. Model 2 is performing better in comparison to model 1 with respect to correlation coefficients, while model 1 is showing better results with respect to RMSE and NMSE. So the unique features of model 1 and model 2 are combined as model 3, which gives a more accurate result than the individual models (model 1 and model 2). The statistically error analysis of model 3 evaluation for all four

Table 4
Eigenvalues and explained variance of the computed PC's for summer, monsoon, post monsoon and winter seasons.

Seasons	Principal component	Eigenvalue	% of variance	Cumulative variance (%)
Summer	1	5.3635	33.5218	33.5218
	2	2.2412	14.0075	47.5293
	3	1.5842	9.9012	57.4306
	4	1.1967	7.47937	64.9100
	5	1.0084	6.3025	71.2125
	6	0.9758	6.0990	77.3115
	7	0.9547	5.9670	83.2785
Monsoon	1	5.6927	35.5790	35.5793
	2	1.7054	10.6590	46.2381
	3	1.3753	8.5956	54.8337
	4	1.0740	6.7125	61.5462
	5	0.9794	6.1216	67.6677
	6	0.9449	5.9057	73.5734
	7	0.9068	5.6676	79.9910
Post-monsoon	1	6.6429	41.5181	41.5181
	2	2.3063	14.4143	55.9325
	3	1.1971	7.48187	63.4143
	4	1.0804	6.7525	70.1668
	5	0.9851	6.1570	76.3238
	6	0.8668	5.4136	81.7375
Winter	1	4.1145	25.7156	25.7156
	2	3.2229	20.1431	45.8587
	3	2.2611	14.1318	59.9906
	4	1.2649	7.9056	67.8962
	5	0.9672	6.0455	73.9418
	6	0.9062	5.6640	79.6058
	7	0.8191	5.1199	84.7258

Table 5
Communalities of each original variable for summer, monsoon, post monsoon and winter seasons.

Variable	Summer	Monsoon	Post-Monsoon	Winter
$\mathrm{AQI}_{\mathrm{d}-1}$	0.90831	0.85492	0.8124	0.93691
$\mathrm{t}_{\mathrm{mean}}$	0.95964	0.8856	0.96105	0.95137
rh	0.8269	0.89244	0.80939	0.88801
vp	0.73326	0.58686	0.89258	0.80423
rf	0.99149	0.77907	0.92314	0.88761
wsp	0.8423	0.80994	0.71017	0.66405
wdi	0.9812	0.95418	0.94777	0.98711
rd	0.9742	0.87154	0.92174	0.86224
$\mathrm{t}_{\mathrm{max}}$	0.94216	0.84522	0.8914	0.90553
$\mathrm{t}_{\mathrm{min}}$	0.95447	0.79279	0.92923	0.91641
ssh	0.6536	0.66905	0.74477	0.76996
slp	0.81894	0.99292	0.50117	0.50791
v	0.65626	0.65957	0.66994	0.7643
cc	0.96922	0.43376	0.62538	0.85809
$\mathrm{t}_{\mathrm{range}}$	0.83283	0.81332	0.80254	0.87576
ARIMA	0.90346	0.83733	0.85591	0.93494

Table 6
Comparison of model3s' predicted and observed values in years 2000–2005 and year 2006.

S.N.	Season	2000–2005				2006			
		RMSE	NMSE	Correlation coefficient	Fractional bias	RMSE	NMSE	Correlation coefficient	Fractional bias
1	Summer	40.47	0.01406	0.6721	−4.63E-02	31.99	0.0086	0.7539	8.74E-04
2	Monsoon	55.16	0.0436	0.7801	2.06E-06	61.94	0.0531	0.6710	0.0445
3	Post Monsoon	41.76	0.0177	0.7896	−2.12E-04	56.55	0.0341	0.6237	−0.0056
4	Winter	41.83	0.1282	0.6397	−3.10E-05	65.76	0.0390	0.6804	0.1707

seasons shows that model is performing satisfactory in all the seasons but is performing better in summer than the other seasons. The seasonal significance of model 3's parameters has been assessed through PCA, which indicates that daily rainfall, station level pressure, daily mean temperature and wind direction index are maximum explained in summer, monsoon, post-monsoon and winter respectively. The NMSE and correlation coefficient of model 3 are found as 0.0086 and 0.7539, while the NMSE and correlation coefficient are (0.0106, 0.6641) and (0.0113, 0.7286) for model 1 and model 2 respectively during the summer season. The comparison between observed and model's predicted values suggest that the model 3 can be used for future prediction/forecasting of daily AQI in other urban cities.

5. Conclusions

The present study focuses on the forecasting of daily air quality in terms of AQI in urban city Delhi. The AQI, based on air quality of air pollutants, is a simple number, easily understandable by general public to know how bad or good air quality is? The methodology of forecasting the AQI through different statistical models has been discussed in details. The three statistical models namely ARIMA, PCR and combination of ARIMA and PCR have been used for forecasting. The performance of all the three models have been tested against observed AQI and discussed in all the four seasons over a period of 7 years (2000–2006). On the basis of previous section of results and discussion, the following conclusions are drawn:

(i) All the seven days in the week have been considered same in the model as the concentration of air pollutants do not have any noticeable variation in weekdays and weekends.
(ii) Model 3, which is a combination of two models namely ARIMA and PCR is dealing with the issue of exploiting the autocorrelation and collinearity in the variables, is performing satisfactory. Hence, this model 3 can be used for air quality forecasting in other urban cities of India. Although, the model is performing satisfactory, there are many uncertainties present in the model. These uncertainties may be involved at the time of development of model equations or may be due to the quality of the input

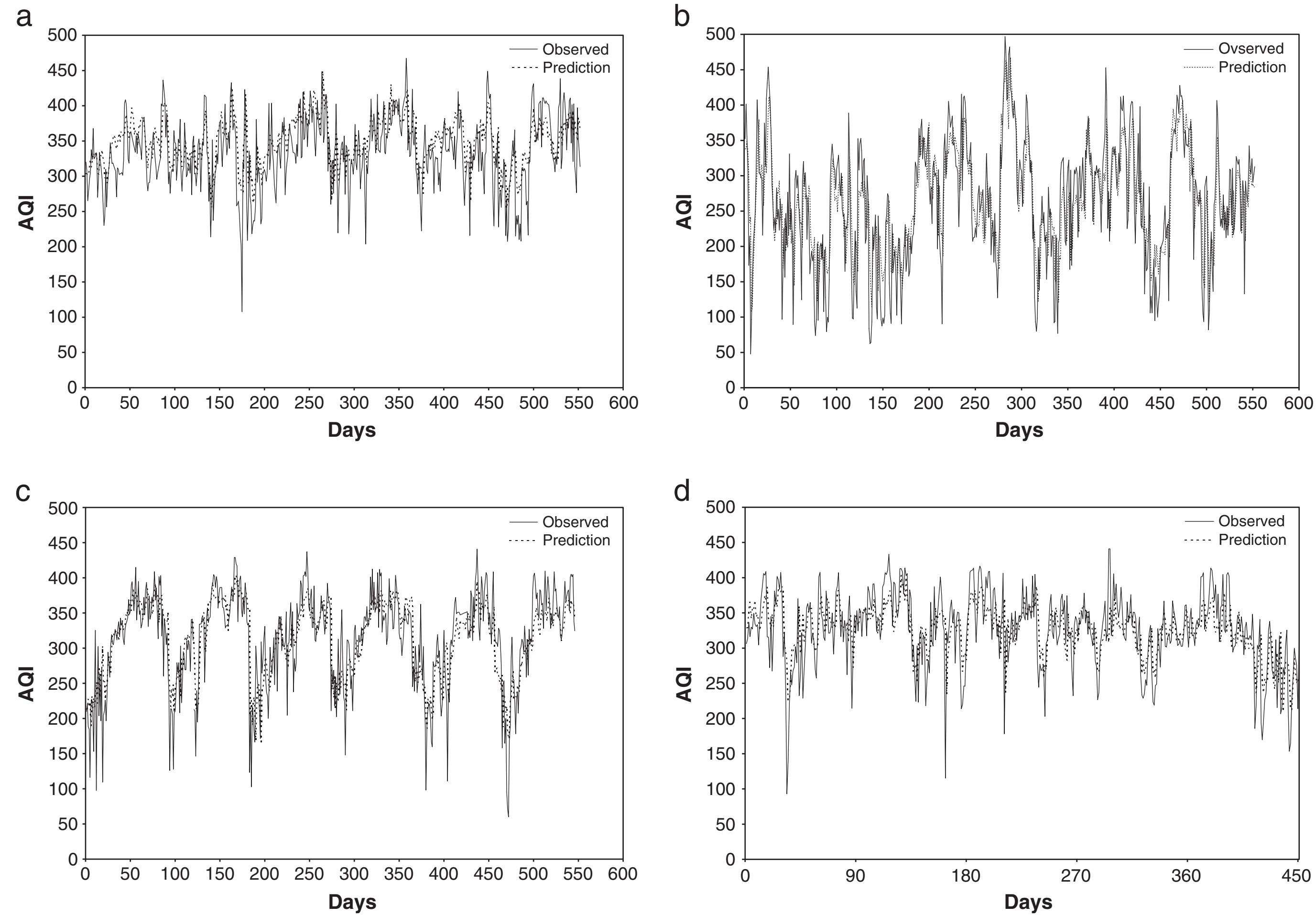

Fig. 1. Comparison of observed and model 3 predicted values of daily AQI in (a) summer, (b) monsoon, (c) post-monsoon and (d) winter seasons during the years 2000–2005.

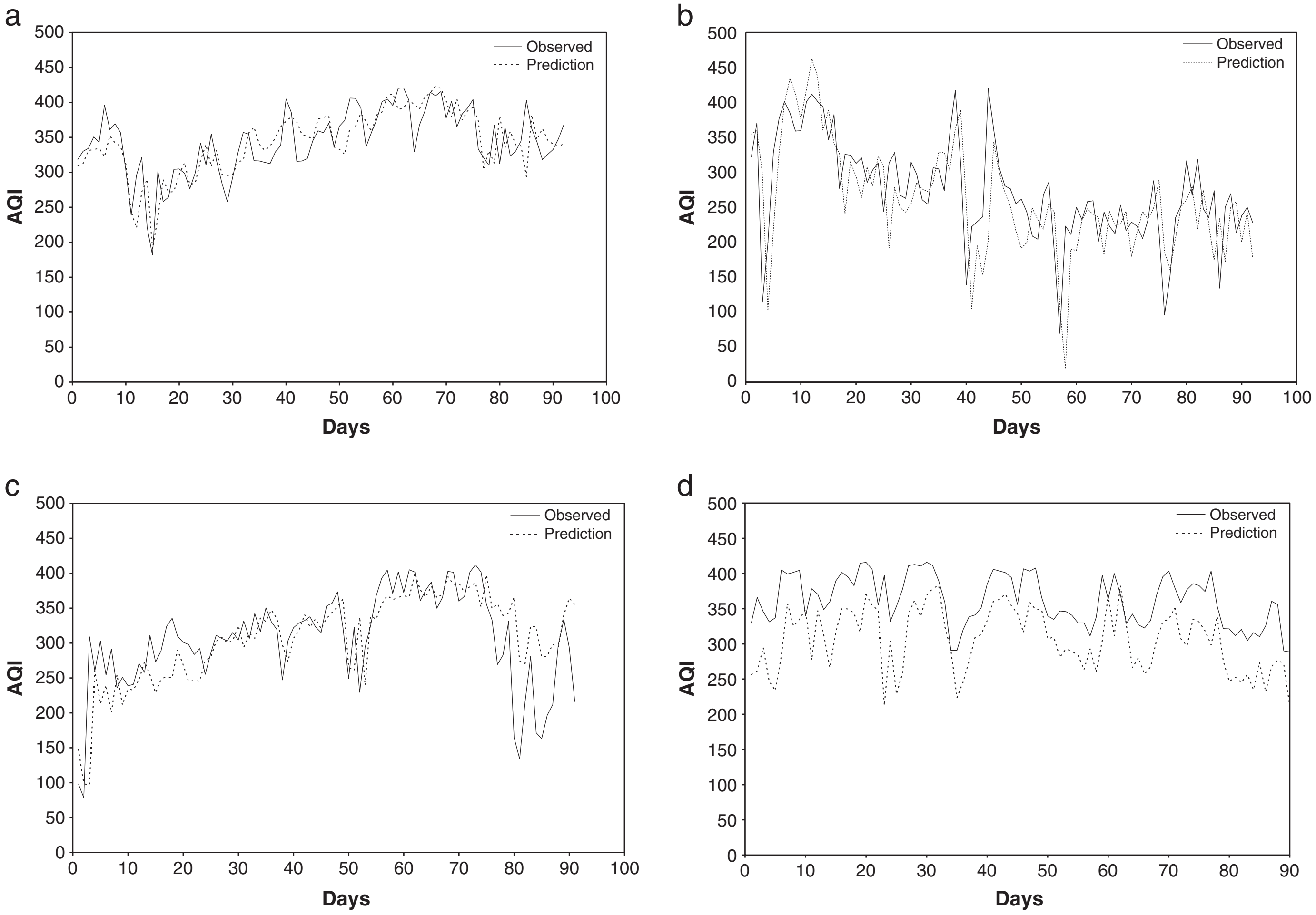

Fig. 2. Comparison of observed and model 3 predicted values of daily AQI in (a) summer, (b) monsoon, (c) post-monsoon and (d) winter seasons during the year 2006.

data like concentration of air pollutants and meteorological variables. It is also noticeable at this point that air quality and meteorological variables should be monitored at the same location, which is always not possible, e.g., in the present study, meteorological station is 5.7 km far from air pollutant monitoring station. These types of arrangements can also be used to improve the quality of input data.

(iii) The present models are statistical models, based on time series formulation and do not include any physics of problem. If these models can be combined with deterministic models like Gaussian plume or Eulerian models, the results can be improved further.

(iv) Finally, it can also be concluded that in the similar situations, the model 3 can be used for daily air quality forecasting one day in advance in any urban city. The same results can be used by Government authorities for decision making process and can also be used by general public for taking precautionary measures to protect their health.

Appendix A

Model 1: ARIMA

One of the methods used for analyzing the time series data is the Box–Jenkins ARIMA approach, which is extracting the predictable behavior of observed data (Box and Jenkins, 1984). ARIMA order (p, d, q) is defined by

$$w_t = \sum_{i=1}^{p} \phi_i w_{t-i} + \alpha_t - \sum_{j=1}^{q} \theta_j \alpha_{t-j}, \tag{A1}$$

where $w_t = \nabla^d y_t$ and d is the order of differencing, ∇ is the backward difference operator, p is order of autoregressive process and q is the order of moving average process. Φ_i and θ_j are i^{th} autoregressive parameter, j^{th} moving average parameters. y_t is the data at time t and α_t is the error term at time t.

Model 2: PCR

PCR is the combination of PCA and multiple linear regression (MLR) technique, where, MLR is based on one dependent variable to be predicted and two or more independent variables, which can be expressed in general form as:

$$Y = b_1 + b_2X_2 + \ldots + b_kX_k + e, \tag{A2}$$

where Y is dependent variable, $X_2, X_3, \ldots, X_k$ are independent variables, $b_1, b_2, \ldots, b_k$ are linear regression parameters and e is an estimated error term, which is obtained from normal distribution of independent random sampling with mean zero and constant variance. The task of regression modeling is to estimate $b_1, b_2, \ldots, b_k$, which can be done using minimum square error technique and solution can be obtained as $b = (X'X)^{-1}(X'Y)$, where X′ is transpose of X. However, when multicollinearity is present, the computation of a matrix inverse ($(X'X)^{-1}$) becomes dubious. The application of PCA with regression model aims to reduce the collinearity of the datasets, which leads the worst predictions and also determine the relevant independent variables for the prediction of air pollutant concentrations (Sousa, et al., 2007).

Principal components can be computed by correlation matrix of input data matrix. The eigenvalues of the correlation matrix 'C' are obtained from its characteristic equation:

$$|C - \lambda I| = 0, \tag{A3}$$

where, λ is the eigenvalue and I is the identity matrix.

For each eigenvalue, a non zero vector e can be defined such as

$$C\ e = \lambda e, \tag{A4}$$

where the vector e is called the characteristic vector or eigenvector of the correlation matrix C associated with its corresponding eigenvalue. These eigenvectors represent the mutually orthogonal linear combination of the matrix. Their associated eigenvalues represent the amount of total variance, which is explained by each of the eigenvectors. By retaining only the first few pairs of eigenvalue–eigenvector, or principal components, a substantial amount of the total variance can be explained while explaining the higher order principal components which explain minimal amounts of the total variance and can be viewed as noise. Variance explained by ith PC is given by:

$$\text{Variance}_i = \frac{\lambda_i}{\sum_n \lambda_n}. \tag{A5}$$

The PC associated with the greatest eigenvalue, the first PC (PC1), represents the linear combination of the variables accounting for the maximum total variability in the data. The second PC explains the maximum variability that is not accounted by the PC1 and so on. In this study only the components those cumulative amounts of variance is approximately 80% should be retained. After getting the PC's, the initial data set is transformed in to the orthogonal set by multiplying the eigenvectors to the initial data set. Now this transformed data set is used as input to the MLR technique.

$$Y = \beta_0 + \beta_1(PC_1) + \beta_2(PC_2) + \ldots + \beta_n(PC_n) + e, \tag{A6}$$

where $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are the coefficients in the model equation. The coefficients of regression model have been estimated using the method of least square. Further the F test has been performed to determine whether a relationship exists between the dependent variable and the regressors. The *t*-test is performed in order to determine the potential value of each of the regressor variables in the regression model. The resulting model can be used to predict future observations.

Model 3: Combination of ARIMA and PCR

Bates and Granger (1969) were the first to introduce combining forecasts as an alternative to use one single forecast. The literature indicates that work on time series forecasting demonstrated that performance increases through combining forecasts (Makridakis et al., 1982; Clemen, 1989; Goyal et al., 2006). The idea of combining forecasts is to use each model's unique features to capture different patterns or features in the data set. In the present study, an attempt has been made to improve the PCR model by combining it with time series forecasting model i.e., ARIMA model for exploiting the autocorrelation in the variables used. Thus model 3 is used as combination of PCR and ARIMA models to improve the forecast.

Appendix B

The statistical measures, given by Chang and Hanna (2004), are used for statistical evaluation in the present study.

References

Anfossi D, Brisasca G, Tinarelli G. Simulation of atmospheric diffusion in low wind speed meandering conditions by a Monte Carlo dispersion model. Il Nuovo Climento 1990;13C:995-1006.

Bates JM, Granger CWJ. The combination of forecasts. Oper Res Q 1969;20:448–51.

Box GEP, Jenkins GM. Time series analysis forecasting and control. San Francisco: Holden-Day; 1984.

Chang JC, Hanna SR. Air quality model performance evaluation. Meteorol Atmos Phys 2004;87:167–96.

Chelani AB, Rao CVC, Phadke KM, Hasan MZ. Prediction of sulphur dioxide concentration using artificial neural networks. Environ Modell Softw 2002;17:161–8.

Clemen R. Combining forecasts: a review and annotated bibliography with discussion. Int J Forecasting 1989;5:559–608.

Cogliani E. Air pollution forecast in cities by an air pollution index highly correlated with metrological variables. Atmos Environ 2001;35:2871–7.

Economic Survey of Delhi. Planning Department, Government of NCT Delhi, June; 2009; 2008–2009.

EPA. Air quality index reporting; final rule. federal register, part III, CFR part 58; 1999.

Finzi G, Tebaldi G. A mathematical model for air pollution forecast and alarm in an urban area. Atmos Environ 1982;16(9):2055–9.

Goyal P, Chan AT, Jaiswal N. Statistical models for the prediction of respirable suspended particulate matter in urban cities. Atmos Environ 2006;40:2068–77.

Kunzli N, Kaiser R, Medina S, Studnicka M, Chanel O, Filliger P, et al. Public-health impact of outdoor and traffic-related air pollution: a European assessment. Lancet 2000;356 (2932):795–801.

Kurt A, Oktay AB. Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. Expert Syst Appl 2010;37:7986–92.

Makridakis S, Anderson A, Carbone R, Fildes R, Hibdon M, Lewandowski R, et al. The accuracy of extrapolation (time series) methods: results of a forecasting competition. J Forecasting 1982;1:111–53.

Nagendra SMS, Venugopal K, Jones SL. Assessment of air quality near traffic intersections in Bangalore city using air quality index. Transp Res D 2007;12:167–76.

Polydoras GN, Anagnostopoulos JS, Bergeles G. Air quality predictions: dispersion model vs Box–Jenkins stochastic models. An implementation and comparison for Athens, Greece. Appl Therm Eng 1998;18:1037–48.

Sousa SIV, Martins FG, Alvim-Ferraz MCM, Pereira MC. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. Environ Modell Softw 2007;22:97-103.

Van den Elshout S, Leger K, Fabio N. Comparing urban air quality in Europe in real time: a review of existing air quality indices and the proposal of a common alternative. Environ Int 2008;34(5):720–6.

Ziomass IC, Dimitrios M, Christos SZ, Alkiviadis FB. Forecasting peak pollutant levels from meteorological variables. Atmos Environ 1995;29(24):3703–11.