AJAE

Open Access

Vol. 16, No. 1, 2021087, March 2022
doi: https://doi.org/10.5572/ajae.2021.087
ISSN (Online) 2287-1160, ISSN (Print) 1976-6912

Research Article

A Methodological Comparison on Spatiotemporal Prediction of Criteria Air Pollutants

Pankaj Singh*, Rakesh Chandra Vaishya, Pramod Soni, Hemanta Medhi[1)]

Department of Civil Engineering, Motilal Nehru National Institute of Technology Prayagraj Uttar Pradesh 211004, India
[1)]Department of Civil Engineering, Tezpur University, Tezpur, Assam 784028, India

***Corresponding author.**
Tel: +91-7800094305
E-mail: pankaj.singh@mnnit.ac.in; st.pnkj@gmail.com

Received: 26 July 2021
Revised: 16 January 2022
Accepted: 4 February 2022

ABSTRACT Air pollution monitoring devices are widely used to quantify at-site air pollution. However, such monitoring sites represent pollution of a limited area, and installing multiple devices for a vast area is costly. This limitation of unavailability of data at non-monitoring sites has necessitated the Spatio-temporal analysis of air pollution and its prediction. Few commonly used methods for Spatio-temporal prediction of pollutants include - 'Averaging'; 'Best correlation coefficient method'; 'Inverse distance weighting method' and 'Grid interpolation method.' Apart from these conventional methods, a new methodology, 'Weighted average method,' is proposed and compared for air pollution prediction at non-monitoring sites. The weights in this method are calculated based on both on the distance and directional basis. To compare the proposed method with the existing ones, the air pollution levels of NO_2 (Nitrogen dioxide), O_3 (Ozone), PM_{10} (Particulate matter of 10 microns or smaller), $PM_{2.5}$ (Particulate matter of 2.5 microns or smaller), and SO_2 (Sulphur dioxide) were predicted at the non-monitoring site (test stations) by utilizing the available data at monitoring sites in Delhi, India. Preliminary correlation analysis showed that NO_2, $PM_{2.5}$, and SO_2 have a directional dependency between different stations. The 'average' method performed best with the mode RMSE of 18.85 μg/m^3 and R^2 value 0.7454 when compared with all the methods. The RMSE value of the new proposed method 'weighted average method' was 21.25 μg/m^3, resulting in the second-best prediction for the study area. The inverse distance weighting method and the Grid interpolation method were third and fourth, respectively, while the 'best correlation coefficient' was the worst with an RMSE value of 41.60 μg/m^3. Results also showed that the methods that used dependent stations had performed better when compared to methods that used all station data.

KEY WORDS Correlation, Prediction, Methodology comparison, Air pollutants, Correlation analysis, Delhi, Interpolation, Dependency, Spatiotemporal, Non monitoring station, Weighted average

1. INTRODUCTION

Pollutants, the undesired substances, degrade the ecosystem with their adverse effects and cause pollution. When concerned with the imbalance in the atmosphere, this pollution is called air pollution. It kills an estimated seven million every year, with low- and middle-income countries suffering the most (Osseiron and

Lindmeier, 2018; WHO, 2018). A significant relationship between atmospheric pollutants and health hazards is reported, and their exposure increases the risk of cardiovascular diseases, fertility issues, and mental health in particular (Chen *et al.*, 2018; Manan *et al.*, 2018; Merklinger-Gruchala *et al.*, 2017; Curtis *et al.*, 2006; Mortimer *et al.*, 2002; Wong *et al.*, 2001). Particulate matter variation studies in India have shown Delhi as the most polluted city when compared to Kolkata, Mumbai, Hyderabad (Singh *et al.*, 2021), with values of $PM_{2.5}$ (Particulate matter of 2.5 microns or smaller) and PM_{10} (Particulate matter of 10 microns or smaller) exceeds National Ambient Air Quality Standards (NAAQs) 60 mg/m^3 (Guo *et al.*, 2019). Studies have attributed agricultural fires as one of the reasons for this (Cusworth *et al.*, 2018). Furthermore, the existence of the relationship between the outbreak of COVID-19 and air pollution is also reported (Roy, 2021). Moreover, an increase in levels of NO_2 (Nitrogen dioxide), O_3 (Ozone), and SO_2 (Sulphur dioxide) have potential health concerns (WHO, 2018).

Considering the harmful nature of these pollutants, it is important to monitor these pollutants. However, due to the high cost of placing monitoring instruments everywhere, sometimes the pollutants can be predicted at a non-monitoring site using various available methods. There have been a number of studies around the globe that have predicted these criterion air pollutants using machine learning (Qi *et al.*, 2019; Wen *et al.*, 2019; Yeganeh *et al.*, 2018; Fan *et al.*, 2017; Zou *et al.*, 2015; Papaleonidas and Iliadis, 2013; Rigol *et al.*, 2001) and Regression models (Kerckhoffs *et al.*, 2021; Boaz *et al.*, 2019; Wang and Song, 2018; Alam and McNabola, 2015; Russo and Soares, 2014; Dominick *et al.*, 2012; Crouse *et al.*, 2009).

There have also been some studies, wherin a comparison of various methods to predict these pollutants have been done. The spatial interpolation methods such as geostatistical methods (various kriging methods), local interpolators (Thiessen polygons, IDW, splines), global interpolators (trend surfaces or regression models), and mixed methods were analyzed by Vicente-Serrano *et al.* (2003). Spatial interpolation methodologies were summarized for urban air pollution modeling based on the application for the greater area of metropolitan Athens, Greece (Deligiorgi and Philippopoulos, 2011). The proposed methodologies include Nearest neighbor method, Triangulated irregular network method, Natural neighbor method, Inverse distance weighting (IDW) method (with linear and squared IDW), Radial basis function (RBF) method, Thin plates splines method, Kriging method, and Artificial neural network (ANN) method. The application of spatial interpolation methods was given to many disciplines (Li and Heap, 2011). A total of 72 spatial interpolation methods were analyzed, and comparative performances were provided in their study by Li and Heap (2014). They included the most frequently used methods as Inverse distance weighting (IDW), ordinary kriging (OK), and ordinary co-kriging (OCK) also. The spatiotemporal prediction at the non-monitoring site was presented by Alimissis *et al.* (2018) using interpolation methods to determine the air pollution at a new location.

Some previous studies in India focused on the missing data prediction using previous data of pollutants or meteorological data or their combinations as predictors. Singh *et al.* (2012) provides a comparison of the linear (PLSR) and nonlinear (MPR, MLPN, RBFN, and GRNN) models for urban air quality prediction (RSPM, SO_2, NO_2) using data of temperature, relative humidity, wind speed, SPM, NO_2, and SO_2 data and shown more remarkable performances of GRNN models than the MLPN and RBFN. Nagendra and Khare (2006, 2005) compared among three choices of input data sets for the prediction of NO_2 firstly, using meteorological and traffic data, secondly, using only meteorological data, and lastly using only traffic data. MLR and PCA + ANN models were evaluated with statistical analysis by Mishra and Goyal (2015) for NO_2 forecasting models at Taj Mahal, Agra using NO_2, SO_2, temperature, CO, O_3, RH, WS, and WDI (wind direction index) as predictors. The performances of PCA approaches were found better than the MLR in their analysis. The performance for the MLP model was found better as compared to RBF and GRNN models for all seasons (Kumar *et al.*, 2017).

Thus, the prediction studies of air pollution have been done for a particular location to evaluate the missing data using the pollution data and meteorological data of that location or a combination of both. The studies have also been done to predict air pollution at a location using pollution data of other sites and other predictors. However, the studies predicting the air quality at a new location (other than monitoring site) using spatial interpolation methods have not been done in India.

In this study, the prediction of air quality parameters at a location of a non-monitoring site is presented over Delhi, India. The air pollution data of the neighboring monitoring sites have been used for the predictions of

NO_2, O_3, PM_{10}, $PM_{2.5}$, and SO_2. The specific objectives of the present study are:

(i) Prediction of criteria air pollutants (NO_2, O_3, PM_{10}, $PM_{2.5}$, and SO_2) at non-monitoring sites in Delhi using five different methods (average method, best correlation coefficient method, weighted average, inverse distance weighting method, and grid interpolation method).
(ii) To compare the performance of these five methods on the basis of prediction efficiency.

2. STUDY AREA AND DATA USED

The present study is conducted for Delhi, India; located between 28.42°N to 29.00°N latitude and 76.86°E to 77.36°E longitude (Fig. 1). Delhi is one of the major cities in India, with rapid growth in population, traffic, industrialization and construction activities. This has led to higher energy consumption. Moreover, the availability of alternate energy sources is limited, thus further increasing the air pollution in Delhi. The air quality observation data for the study area is obtained from CPCB (Central Pollution Control Board of India) at 'https://app.cpcbccr.com/AQI_India/', and the same is also provided at 'https://openaq.org/'. Table 1 shows the statistical summary of fifteen minutes scaled dataset of the five pollutants ($PM_{2.5}$, PM_{10}, NO_2, SO_2, and O_3) for the 46 stations over Delhi in the period of January 1, 2018, to October 30, 2020.

Analysis showed that, only 25 of these stations had missing data less than 75%, which were used for prediction and validation. The top six monitoring stations (Table 2) having maximum missing data out of the 25-monitoring station was reserved for validation, while the remaining 19 monitoring station with the lowest missing per-

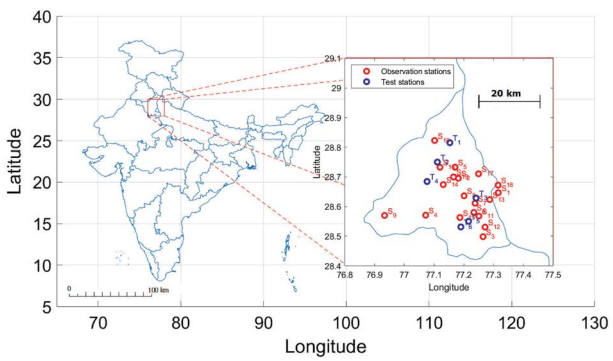

Fig. 1. The location of Delhi (the capital) in India showing 19 observation stations and 6 test stations.

Table 1. Statistical Summary of the five pollutants used in the study area.

Pollutant	Mean ($\mu g/m^3$)	Median ($\mu g/m^3$)	Standard deviation ($\mu g/m^3$)	Missing data (%)
NO_2	42.54	32.90	35.77	72.58
O_3	29.40	15.30	36.29	73.10
PM_{10}	195.42	151.00	153.96	72.09
$PM_{2.5}$	104.39	67.00	106.59	72.15
SO_2	12.63	10.70	11.78	73.68

Table 2. Test stations' names and locations.

S. no.	Station name	Latitude	Longitude
1	Alipur, Delhi - DPCC	28.8153	77.1530
2	DTU, Delhi - CPCB	28.7500	77.1113
3	ITO, Delhi - CPCB	28.6286	77.2411
4	Mundka, Delhi - DPCC	28.6847	77.0766
5	Sirifort, Delhi - CPCB	28.5504	77.2159
6	Sri Aurobindo Marg, Delhi - DPCC	28.5313	77.1901

Table 3. Observation Stations' names and locations.

S. no.	Station name	Latitude	Longitude
1	Anand Vihar, Delhi - DPCC	28.6468	77.3160
2	Ashok Vihar, Delhi - DPCC	28.6954	77.1817
3	Dr. Karni Singh Shooting Range, Delhi - DPCC	28.4986	77.2648
4	Dwarka-Sector 8, Delhi - DPCC	28.5710	77.0719
5	Jahangirpuri, Delhi - DPCC	28.7328	77.1706
6	Jawaharlal Nehru Stadium, Delhi - DPCC	28.5803	77.2338
7	Major Dhyan Chand National Stadium, Delhi - DPCC	28.6113	77.2377
8	Mandir Marg, Delhi - DPCC	28.6364	77.2011
9	Najafgarh, Delhi - DPCC	28.5702	76.9338
10	Narela, Delhi - DPCC	28.8228	77.1020
11	Nehru Nagar, Delhi - DPCC	28.5679	77.2505
12	Okhla Phase-2, Delhi - DPCC	28.5308	77.2713
13	Patparganj, Delhi - DPCC	28.6237	77.2872
14	Punjabi Bagh, Delhi - DPCC	28.6740	77.1310
15	R K Puram, Delhi - DPCC	28.5632	77.1869
16	Rohini, Delhi - DPCC	28.7325	77.1199
17	Sonia Vihar, Delhi - DPCC	28.7105	77.2494
18	Vivek Vihar, Delhi - DPCC	28.6723	77.3152
19	Wazirpur, Delhi - DPCC	28.6997	77.1654

centage was used for prediction model development purposes (Table 3).

3. METHODOLOGY

In this study five different methods (Simple Averaging, Best Correlation Coefficient Method, Weighted Average Method, Inverse Distance Weighting method, and Grid Interpolation Method) have been evaluated in terms of prediction efficiency. Theses are the methods which are commonly used in the previous studies (Jin *et al.*, 2011). The correlation analysis was performed to check the dependency among the chosen monitoring stations. The correlation analysis was performed between 19 monitoring stations using the entire data at 15 min intervals at seasonal (Winter: January and February), Summer: March, April, and May, Monsoon: June to September, and Post-monsoon period: October to December) and annual scales.

3.1 Correlation Analyses

The correlation coefficients between all possible pairs of 19 observation monitoring stations were evaluated for each season and pollutant. The value of the correlation coefficient to be satisfactory depends on the purpose for which it is used and the nature of raw data. The broad classification of correlation coefficients for their correlation strength is given in Asuero *et al.* (2006). The pair of stations showing average or high correlation was treated to have a positive dependency. A linear model was established among the correlation coefficients between each pair as output variable, distances between two stations, and direction of the line joining the two stations as input variables.

$$r(O_i, O_j) = C_1 \times d(S_{O_i}, S_{O_j}) + C_2 \times \theta(S_{O_i}, S_{O_j}) + C_3 \quad (1)$$

where, $r(O_i, O_j)$ is the correlation coefficient between the observation of the i^{th} and j^{th} observation monitoring stations.

If the spatial coordinates of the i^{th} and j^{th} observation monitoring stations are given as

$$S_{O_i} = (\text{lat}_i, \text{long}_i)$$

$$S_{O_j} = (\text{lat}_j, \text{long}_j)$$

then, $d(S_{O_i}, S_{O_j})$ is the linear distance between the i^{th} and j^{th} observation monitoring stations whose mathematical expression given by

$$d(S_{O_i}, S_{O_j}) = \sqrt{(\text{lat}_i - \text{lat}_j)^2 + (\text{long}_i - \text{long}_j)^2} \quad (2)$$

$\theta(S_{O_i}, S_{O_j})$ is the direction of the line joining i^{th} and j^{th} observation monitoring stations, whose mathematical expression is given as

$$\theta(S_{O_i}, S_{O_j}) = \left| \frac{(\text{lat}_i - \text{lat}_j)}{(\text{long}_i - \text{long}_j)} \right| \quad (3)$$

The constants C_1, C_2, and C_3 are determined by multiple linear regression among correlation coefficient, linear distance, and direction of the line joining two observation stations for each season and pollutants. Thus, the possible correlation coefficient between any test station and observation station for any pollutant can be evaluated in any season using the established linear model.

$$r(T_i, O_j) = C_1 \times d(S_{T_i}, S_{O_j}) + C_2 \times \theta(S_{T_i}, S_{O_j}) + C_3 \quad (4)$$

where, $r(T_i, O_j)$ is the correlation coefficient, $d(S_{T_i}, S_{O_j})$ is the linear distance, and $\theta(S_{T_i}, S_{O_j})$ is the direction of the line joining between i^{th} test station and j^{th} observation monitoring station.

3.2 Prediction Methods

3.2.1 Averaging Method

The model established in the correlation analysis was used to evaluate the dependency of pollutants concentration of observation stations (monitoring stations) on the test station pollutants concentration based on the spatial characteristic of the test station with respect to observation stations. The correlation coefficient was evaluated between test stations and observation stations for each pollutant using the established linear model (Equation 4).

This method gives the average value of the pollutant of nearby monitoring stations, which shows positive dependency ($R > 0.5$) with the site to be predicted. It can be mathematically expressed as

$$C(S_{T_i}, t_0) = \frac{\sum_{j=1}^{n} C(S_{O_j}, t_0)}{n} \quad (5)$$

where, $C(S_{T_i}, t_0)$ is the pollutant concentration at i^{th} test station on time t_0, and $C(S_{O_j}, t_0)$ is the pollutant concentration at the j^{th} dependent observation station on time t_0.

3.2.2 Best Correlation Coefficient Method

The linear model was established among the correlation coefficients as output variable, distances between

two stations, and direction of the line joining the two stations as input variables (Equation 4). Using this linear model, the correlation coefficients between the desired point location and available data point locations are predicted from the distances and directions.

This method states that the data point of the best correlation station with the desired location should be considered the predicted values for the desired location.

$$r_{\text{best}}(T_i, O_k) = \max(r(T_i, O_1), r(T_i, O_2), \ldots\ldots, r(T_i, O_n)) \quad (6)$$

$$C(S_{T_i}, t_0) = C(S_{O_k}, t_0) \quad (7)$$

where, $C(S_{T_i}, t_0)$ is the pollutant concentration at i^{th} test station on time t_0, $C(S_{O_k}, t_0)$ and is the pollutant concentration at k^{th} dependent observation station having maximum correlation coefficient on time t_0.

3.2.3 Weighted Average Method

The correlation coefficient was evaluated between test stations and observation stations for each pollutant in each season using the established linear model (Equation 4). The pollutants concentrations were assumed to be linearly dependent on the nearby dependent observation stations.

$$C_f = M_D \times B \quad (8)$$

$$C_f = \begin{bmatrix} C(S_{O_i}, t_0) \\ C(S_{O_i}, t_{-1}) \\ . \\ . \\ C(S_{O_i}, t_{-k}) \end{bmatrix} \quad (9)$$

$$M_D = \begin{bmatrix} C(S_{O_1}, t_0) & C(S_{O_2}, t_0) & . & . & C(S_{O_n}, t_0) \\ C(S_{O_1}, t_{-1}) & C(S_{O_2}, t_{-1}) & . & . & C(S_{O_n}, t_{-1}) \\ . & . & . & . & . \\ . & . & . & . & . \\ C(S_{O_1}, t_{-k}) & C(S_{O_2}, t_{-k}) & . & . & C(S_{O_n}, t_{-k}) \end{bmatrix} \quad (10)$$

$$B = \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ b_n \end{bmatrix} \quad (11)$$

where $C(S_{T_i}, t_0)$ is the pollutant concentration at i^{th} test station on time t_0, and $C(S_{O_j}, t_0)$ is the pollutant concentration at the j^{th} dependent observation station on time t_0. C_f is the observation station data, and M_D is its dependent station data matrix. B is the coefficients of proportionality between C_f and M_D.

The coefficients of linearity (weights for each dependent observation station) for any observation station are assumed to be linearly dependent on the distance of the observation station and the direction of the line joining that observation station.

$$B(S_{O_i}, S_{O_j}) = K_1 \times d(S_{O_i}, S_{O_j}) + K_2 \times \theta(S_{O_i}, S_{O_j}) \quad (12)$$

Now the coefficients B' between test station and observation stations are back-calculated using K_1 and K_2.

$$B'(S_{T_i}, S_{O_j}) = K_1 \times d(S_{T_i}, S_{O_j}) + K_2 \times \theta(S_{T_i}, S_{O_j}) \quad (13)$$

Thus, in this method, mass concentration pollutants at test are given as a sum of multiplication of pollutants concentration data of dependent observation stations with their weights (proportionality coefficients determined earlier for each observation station).

$$C(S_{T_i}, t_0) = \sum_{j=1}^{m} C(S_{O_j}, t_0) \times B'(S_{T_i}, S_{O_j}) \quad (14)$$

3.2.4 Inverse Distance Weighting (IDW) Method

The principle of this method is that the observation station having more distance from the test station will affect the least to test station and vice versa. In this method, the weights are distributed among the dependent observation stations according to their inverse distance from the test station. The dependency of the observation stations is determined from the correlation analysis (Equation 4). The generic equation for the Inverse Distance Weighting (IDW) method (Bartier and Keller, 1996) was given as

$$C(S_{T_i}, t_0) = \frac{\sum_{k=1}^{m} C(S_{O_k}, t_0) \times \frac{1}{d(S_{T_i}, S_{O_k})}}{\sum_{j=1}^{n} \frac{1}{d(S_{T_i}, S_{O_j})}} \quad (15)$$

where $C(S_{T_i}, t_0)$ is the pollutant concentration at i^{th} test station on time t_0, and $C(S_{O_j}, t_0)$ is the pollutant concentration at j^{th} dependent observation station on time t_0.

3.2.5 Grid Interpolation Method

In this method, the grid interpolation (using the nearest neighbor method) has been done for query points

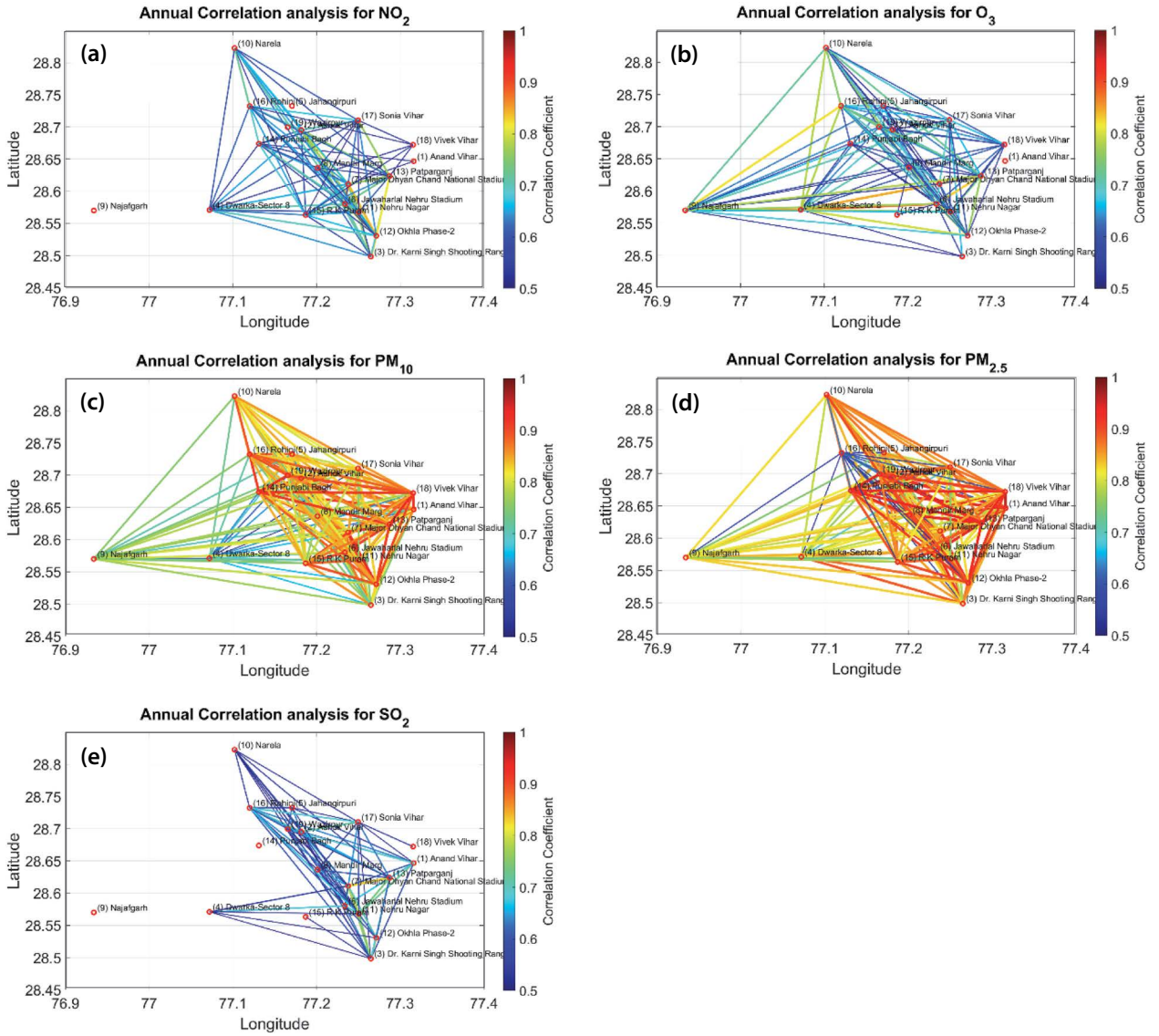

Fig. 2. Spatial plot of observation stations showing the strength of correlation coefficient calculated on annual data over three years using colour and weight of the line for NO_2, O_3, PM_{10}, $PM_{2.5}$ and SO_2.

(entire grid) over the scattered pollutants data by the fitting surface over the area. The mass concentration data obtained by interpolation at the query point (location at which the pollution concentrations are to be predicted) is considered the predicted data.

a. Grid interpolation with dependent station data

The observation stations dependent on the particular test station ($R > 0.5$) were used as input data for interpolation in a specific season for a specific pollutant.

b. Grid interpolation with all station data

The data of all observation stations were used as input for the interpolation in a particular season for a specific pollutant.

4. RESULTS AND DISCUSSION

4.1 Correlation Analyses

The performance analyses in this study were carried using the Root Mean Square Error (RMSE) and Correlation Coefficient (R^2). Considering O and P as the observed and predicted concentrations, the formula for RMSE and R^2 are as shown below:

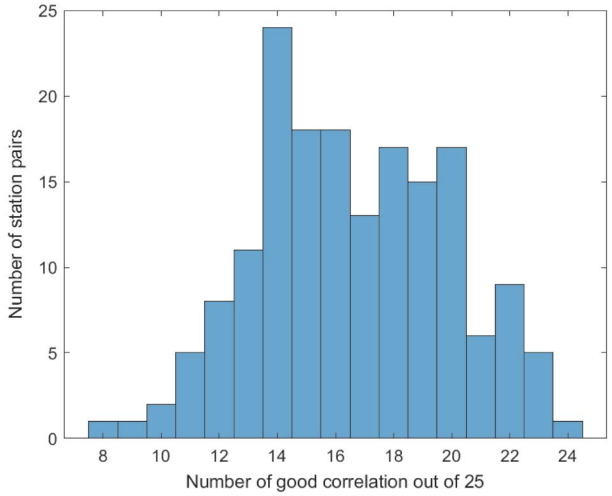

Fig. 3. Histogram showing the frequency of the combination of observation station pairs with the number of times having good strength of correlation out of 25 cases (for five pollutants in 5 seasons).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2} \quad (16)$$

$$R^2 = \left(\frac{\sum_{i=1}^{N}(O_i-\bar{O})(P_i-\bar{P})}{\sqrt{\sum_{i=1}^{N}(O_i-\bar{O})^2}\sqrt{\sum_{i}^{N}(P_i-\bar{P})^2}}\right)^2 \quad (17)$$

The correlation plots between 19 observation stations for different pollutants are shown in Figs. 2(a) to (e). Lower value of correlation coefficient does not provide sufficient relation, hence, in this study a threshold value of 0.5 was considered. All values having correlation coefficient less than 0.5 were removed in this study. The color and width of the line represent the magnitude of the correlation between the two observation stations. The larger width and red color signifies higher correlation whereas, the smaller width and blue color represent a lower correlation.

It can be seen that the correlations are strongest for PM_{10} (Fig. 2(c)) and $PM_{2.5}$ (Fig. 2(d)), whereas weakest for SO_2 (Fig. 2(e)). However, a weak directional dependency was also observed for SO_2 along the NW-SE direction, as the strength of the correlation is always lower in this direction.

The number of pairs of stations showing a good correlation with their frequency of occurrence along with five pollutants and five seasons is shown in the histogram (Fig. 3). Dwarka-Sector 8, Delhi - DPCC shows a good correlation in 24 out of 25 cases (for five pollutants along five seasons) with Jawaharlal Nehru Stadium, Delhi - DPCC always a good correlation with the Jawaharlal Nehru Stadium monitoring station. The monitoring station pairs having good correlation in 23 out of 25 cases are given in Table 4.

Table 4. Pair of observation stations showing good strength of correlation 23 times out of 25 cases.

S. no.	First element of pair	Second element of pair
1	Nehru Nagar, Delhi - DPCC	Jawaharlal Nehru Stadium, Delh - DPCC
2	Patparganj, Delhi - DPCC	Ashok Vihar, Delhi - DPCC
3	Patparganj, Delhi - DPCC	Major Dhyan Chand National Stadium, Delhi - DPCC
4	Sonia Vihar, Delhi - DPCC	Ashok Vihar, Delhi - DPCC
5	Sonia Vihar, Delhi - DPCC	Patparganj, Delhi - DPCC

4.2 Prediction of Pollutant's Concentration

The predictions of the pollutant's concentration at 15 minutes intervals have been made for three years by five various methods in all seasons for five pollutants at six validation stations. For the purpose of inter-comparison between different pollutants, instead of RMSE, the PRMSE (Percentage RMSE) was used to prepare the box plots. These box plots were prepared both for 15-minute interval and daily averages.

4.2.1 Nitrogen dioxide (NO_2)

The box plots showing PRMSE for NO_2 in different seasons and by different methods are shown in Fig. 4. Overall, the weighted-average prediction method shows the best performance having the lowest median PRMSE and least interquartile range for three seasons. However, on the annual scale, even the Average method shows better consistency and accuracy.

Due to the conversion of the 15 minutes interval prediction data into daily data, the interquartile range and median RMSE have been reduced to 12.92 $\mu g/m^3$ on average (Fig. 5). It can still be seen that the pattern of performance among the methods of prediction is similar to 15-minute interval case.

4.2.2 Ozone (O_3)

The presence of outliers shows the variations of predictions at different test stations are more in the summer sea-

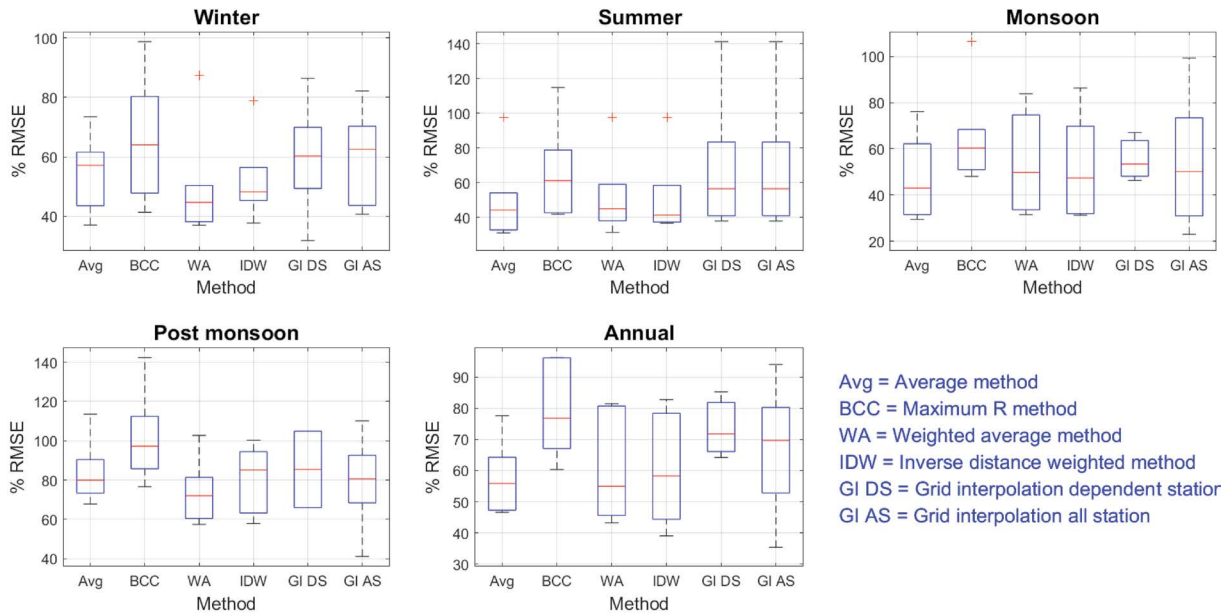

Fig. 4. Box plot of PRMSE of predictions by various methods at six validation stations in various seasons for NO_2.

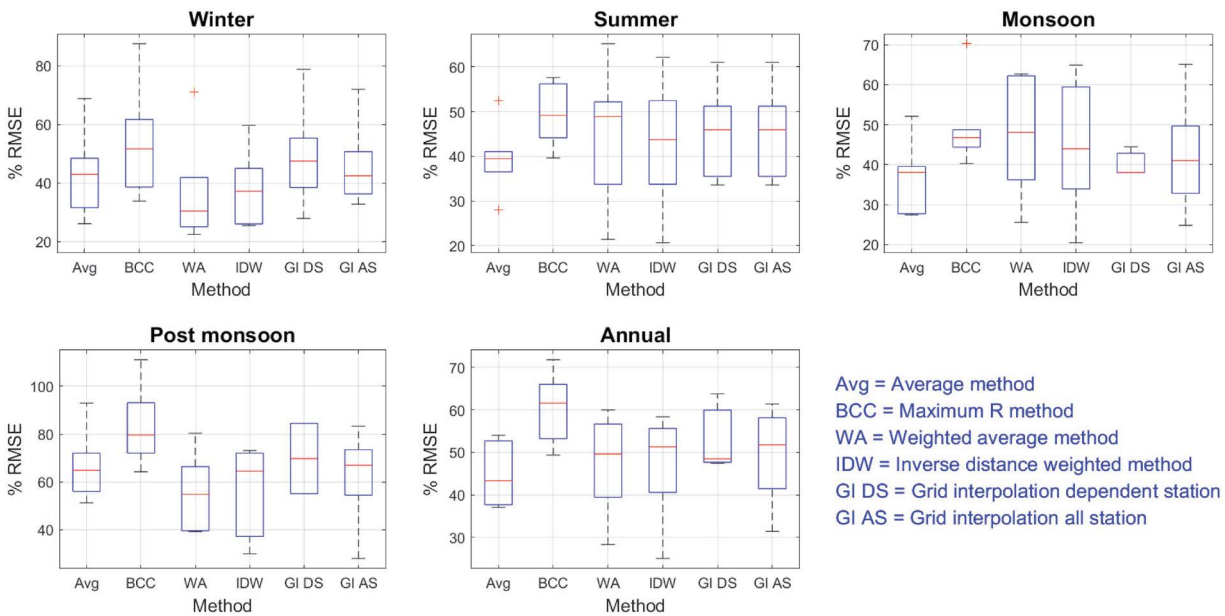

Fig. 5. Box plot of PRMSE of predictions by various methods at six validation stations in various seasons for NO_2 data (15 minutes data) converted to daily data.

son for the prediction of O_3 (Figs. 6 and 7). All the methods of prediction have almost similar behavior for the prediction of O_3. The missing box plot for any method shows the unpredictability of that method due to the lack of dependent observation station data ($R > 0.5$) in the respective season.

4.2.3 Particulate Matter (PM_{10})

The RMSE of predictions is greater for PM_{10} than NO_2 and O_3. The weighted average method has worse prediction (highest median PRMSE and interquartile range in all seasons) than other methods (Figs. 8 and 9) in all seasons for the prediction of PM_{10}.

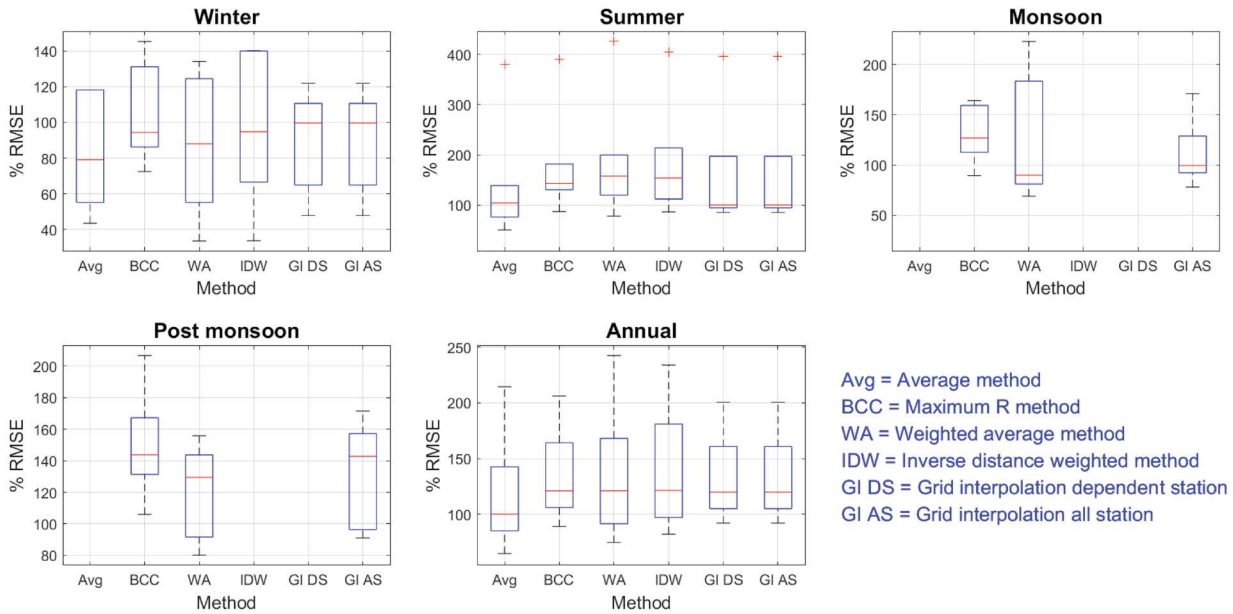

Fig. 6. Box plot of PRMSE of predictions by various methods at six validation stations in various seasons for O_3.

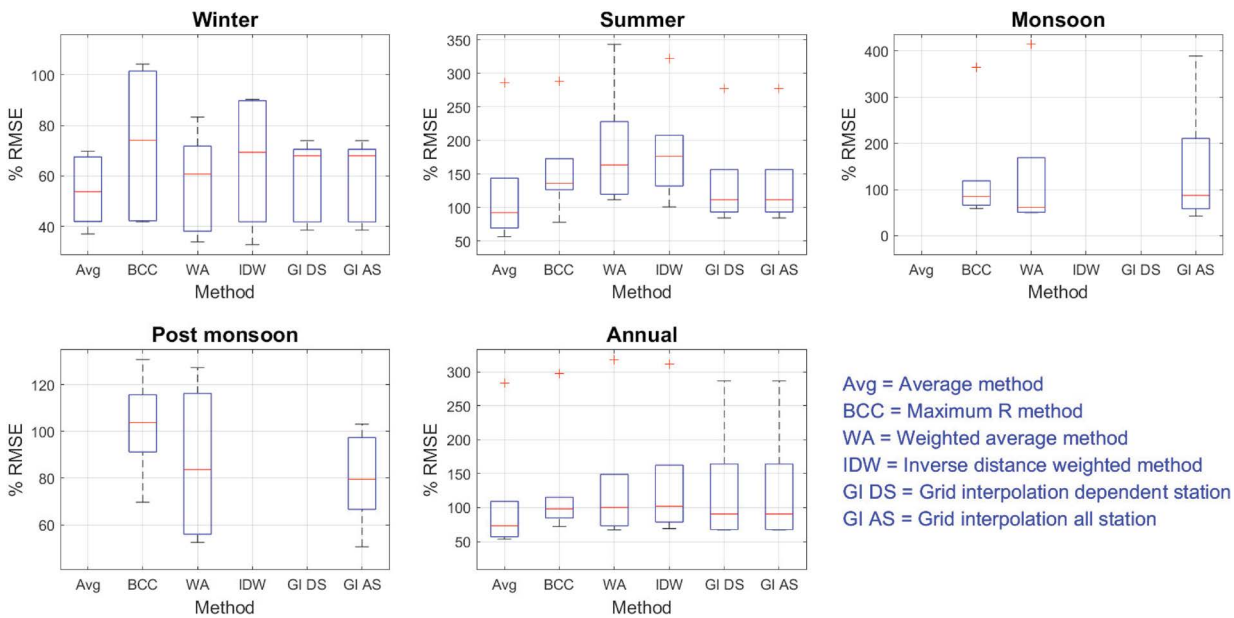

Fig. 7. Box plot of PRMSE of predictions by various methods at six validation stations in various seasons for O_3 data (15 minutes data) converted to daily data.

4.2.4 Particulate Matter ($PM_{2.5}$)

The presence of outliers shows the more significant variations of predictions at different test stations in every season for the prediction of $PM_{2.5}$ (Figs. 10 and 11). All the predictions methods have almost similar behaviour for the prediction of $PM_{2.5}$ except the Weighted average method in some seasons (winter and post-monsoon). The prediction by the Weighted average method is the worst among all the methods for $PM_{2.5}$ prediction.

4.2.5 Sulphur dioxide (SO_2)

The missing box plot for any method shows the unpre-

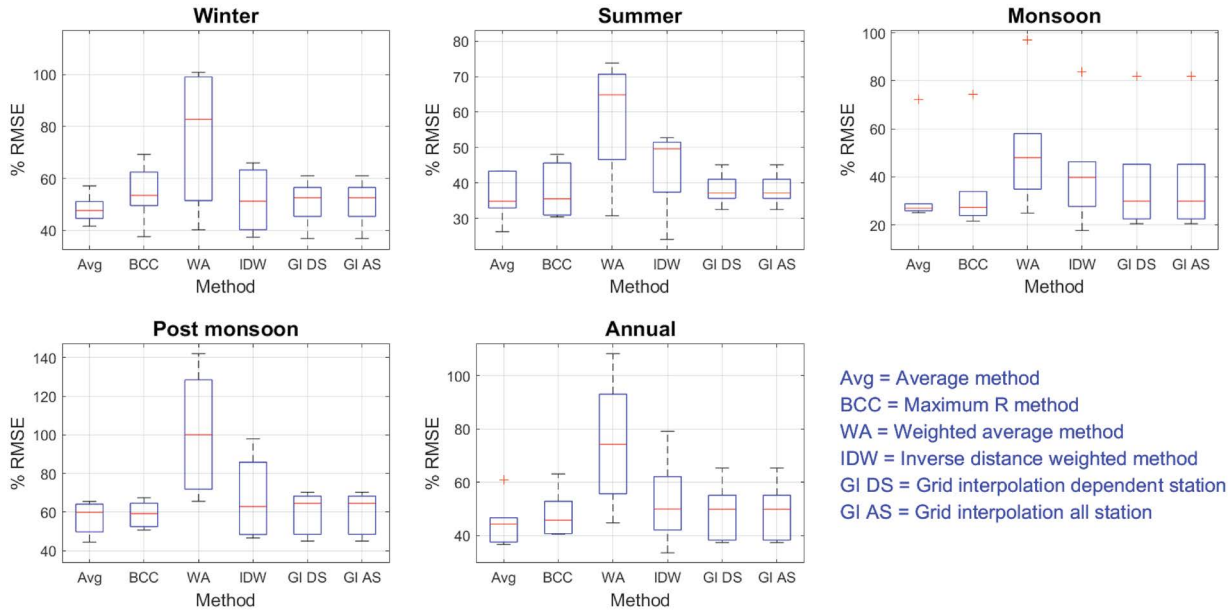

Fig. 8. Box plot of PRMSE of predictions by various methods at six validation stations in various seasons for PM_{10}.

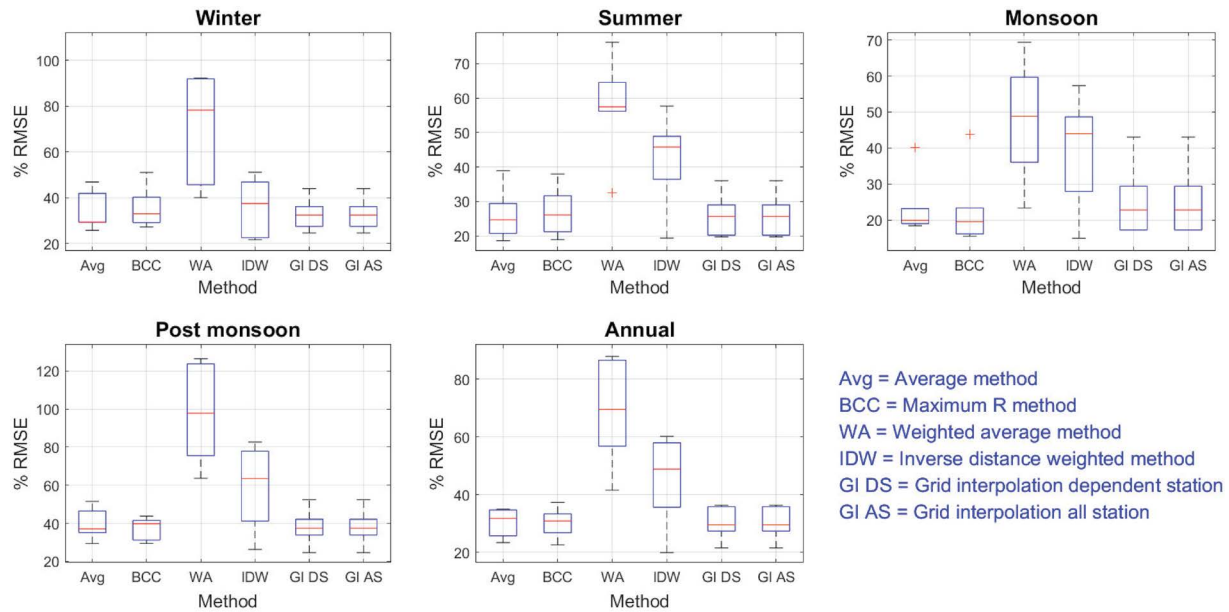

Fig. 9. Box plot of PRMSE of predictions by various methods at six validation stations in various seasons for PM_{10} data (15 minutes data) converted to daily data.

dictability of that method due to the lack of dependent observation station data in the respective season. All the methods of prediction have almost similar behaviour for the prediction of SO_2 (Figs. 12 and 13).

4.3 Performance of Methodologies

The RMSE and R^2 of predictions were calculated for each prediction method in each season for each pollutant and at each test station. These RMSE and R^2 were grouped for each methodology, and histogram plots

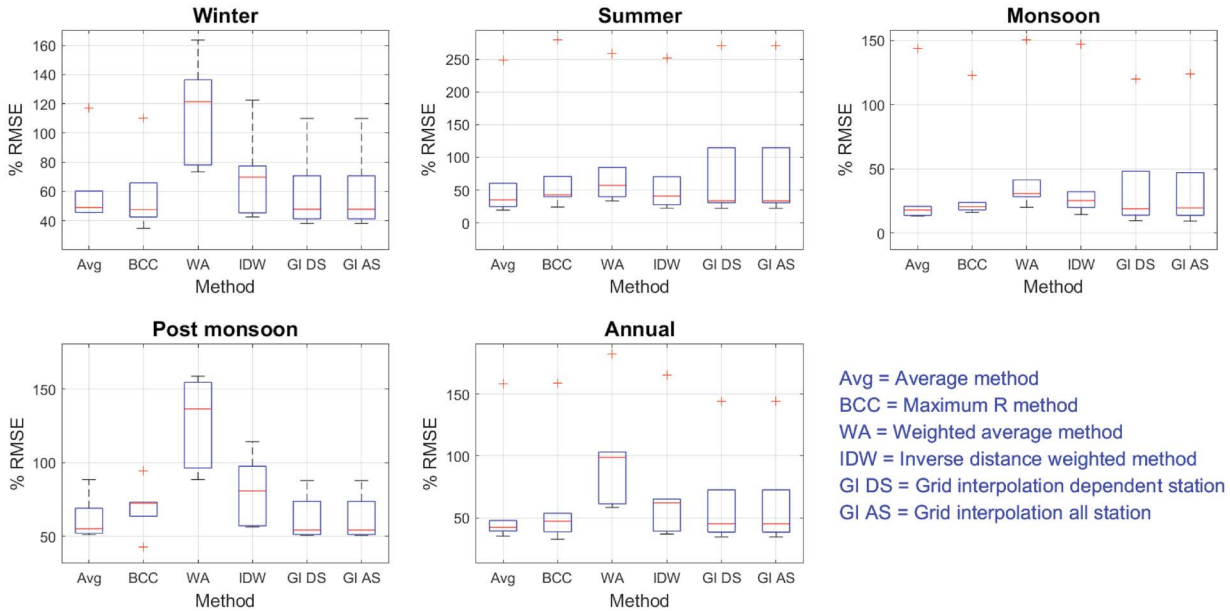

Fig. 10. Box plot of PRMSE of predictions by various methods at six validation stations in various seasons for $PM_{2.5}$.

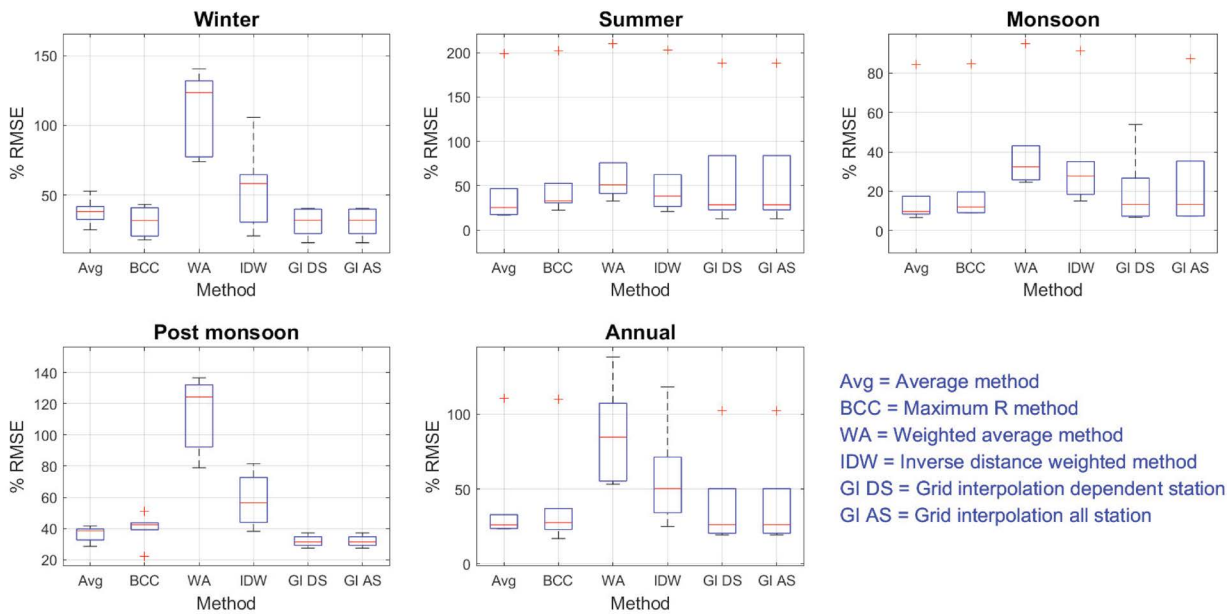

Fig. 11. Box plot of PRMSE of predictions by various methods at six validation stations in various seasons for $PM_{2.5}$ data (15 minutes data) converted to daily data.

were shown representing the frequencies of occurrence of RMSE with mode, median and average values.

The performance of the Average method was found to be having the best predictions most of the time with mode RMSE of 18.85 (Fig. 14), R^2 of 0.74 (Fig. 15), and the performance of the Best correlation coefficient (BCC) method was found to be having the worst predictions most of the times with mode RMSE of 41.60 (Fig. 14). These results are similar to some of the previous studies (Alam *et al.*, 2015: max $R^2 = 0.66$; Crouse *et al.*, 2009: max $R^2 = 0.8$; Qi *et al.*, 2019: max $R^2 = 0.72$).

The scatter plot of the prediction having least and

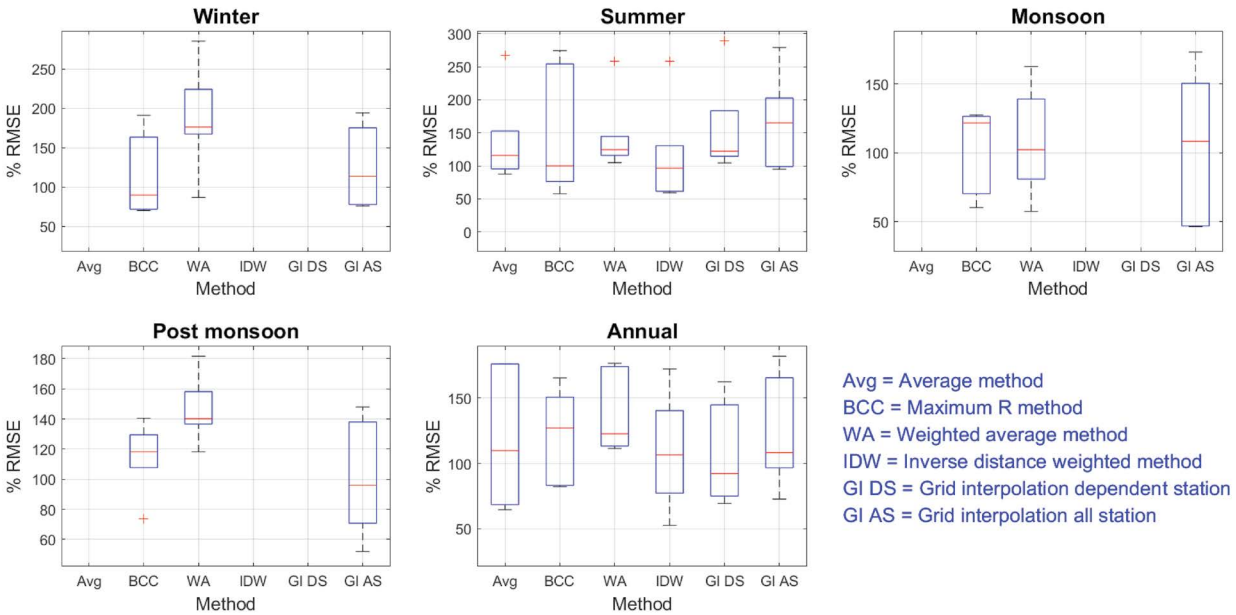

Fig. 12. Box plot of PRMSE of predictions by various methods at six validation stations in various seasons for SO_2.

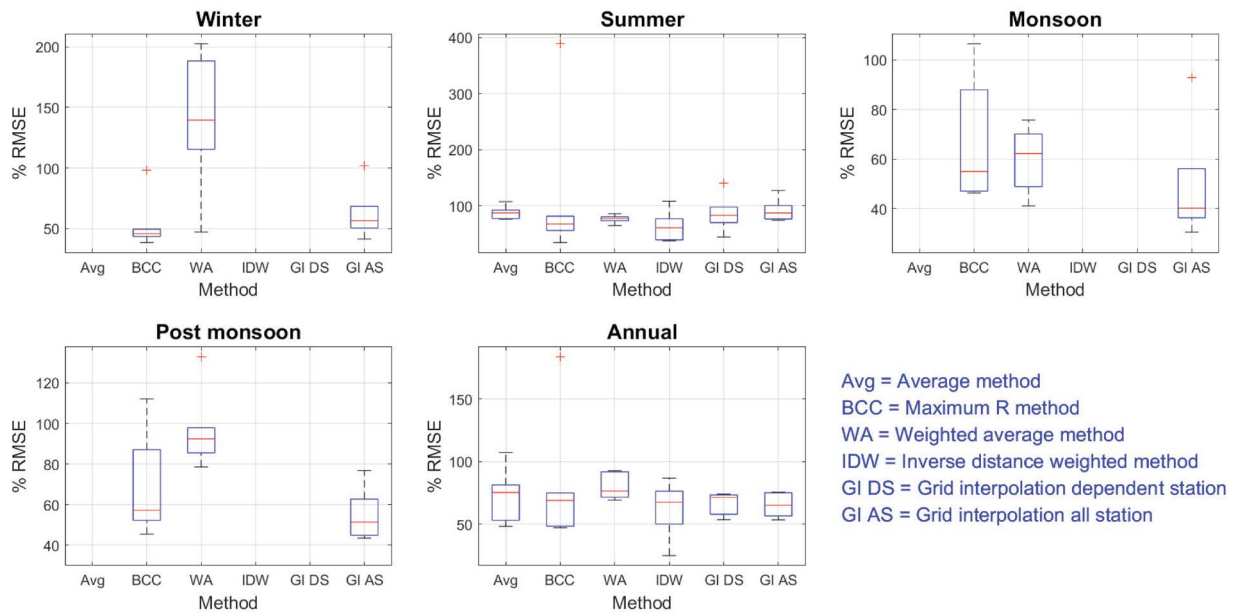

Fig. 13. Box plot of PRMSE of predictions by various methods at six validation stations in various seasons for SO_2 data (15 minutes data) converted to daily data.

maximum RMSE among all the combinations of prediction methods, seasons, and test stations have also been shown for each pollutant. The method showing the best performance for any of the stations/seasons is chosen, and the corresponding scatter plot is shown in Fig. 16(a)–(e).

The least RMSE has been achieved to predict $PM_{2.5}$ by Grid interpolation using all observation station data in monsoon season (Fig. 16(d)) with a high correlation coefficient of 0.88. In contrast, the least RMSE has been achieved to predict SO_2 by Grid interpolation using all observation station data in monsoon season (Fig. 16(e))

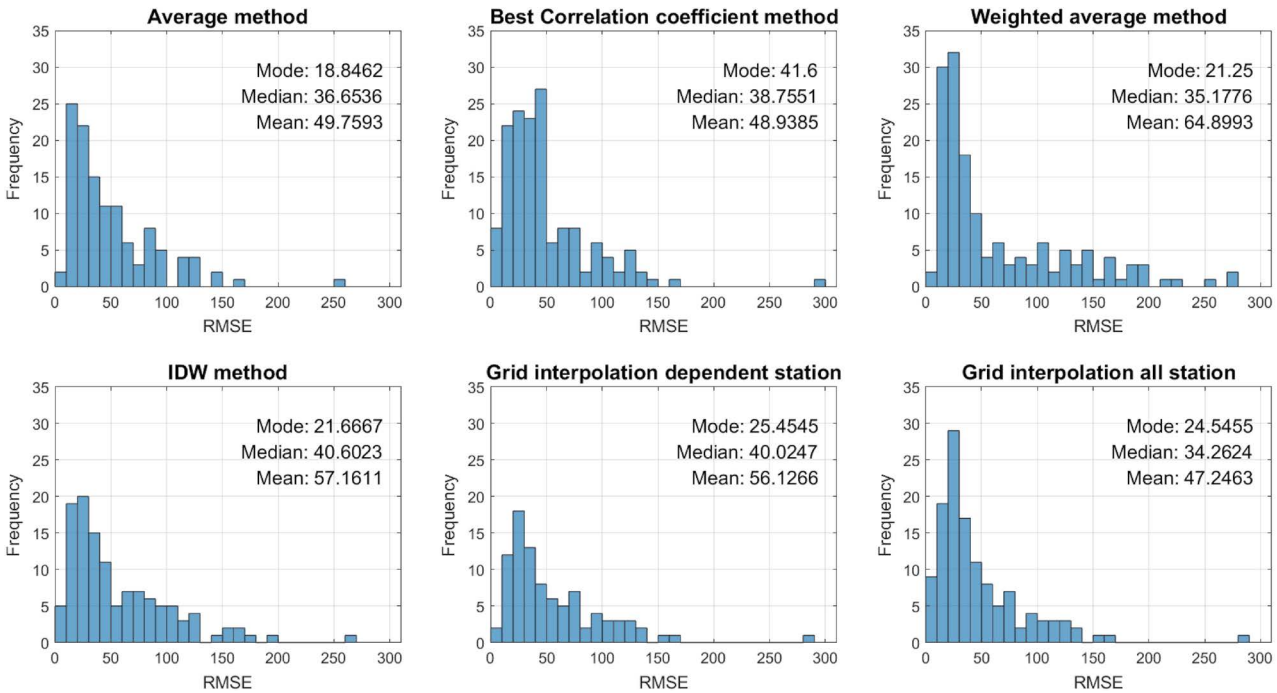

Fig. 14. Histogram plot showing frequencies of RMSE over various pollutants, seasons, and test stations for each prediction method.

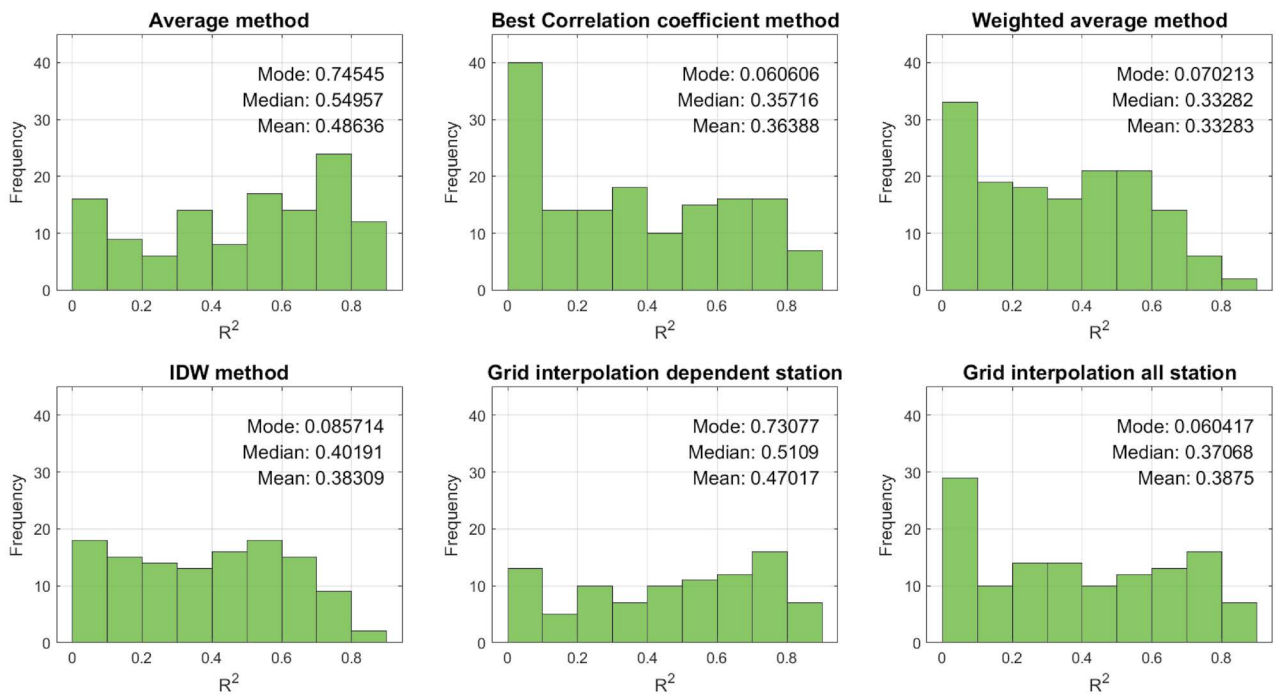

Fig. 15. Histogram plot showing frequencies of R^2 over various pollutants, seasons, and test stations for each prediction method.

with a very low correlation coefficient of 0.15.

Most of the time, the best predictions have been achieved in monsoon season and with the grid interpolation method at all stations.

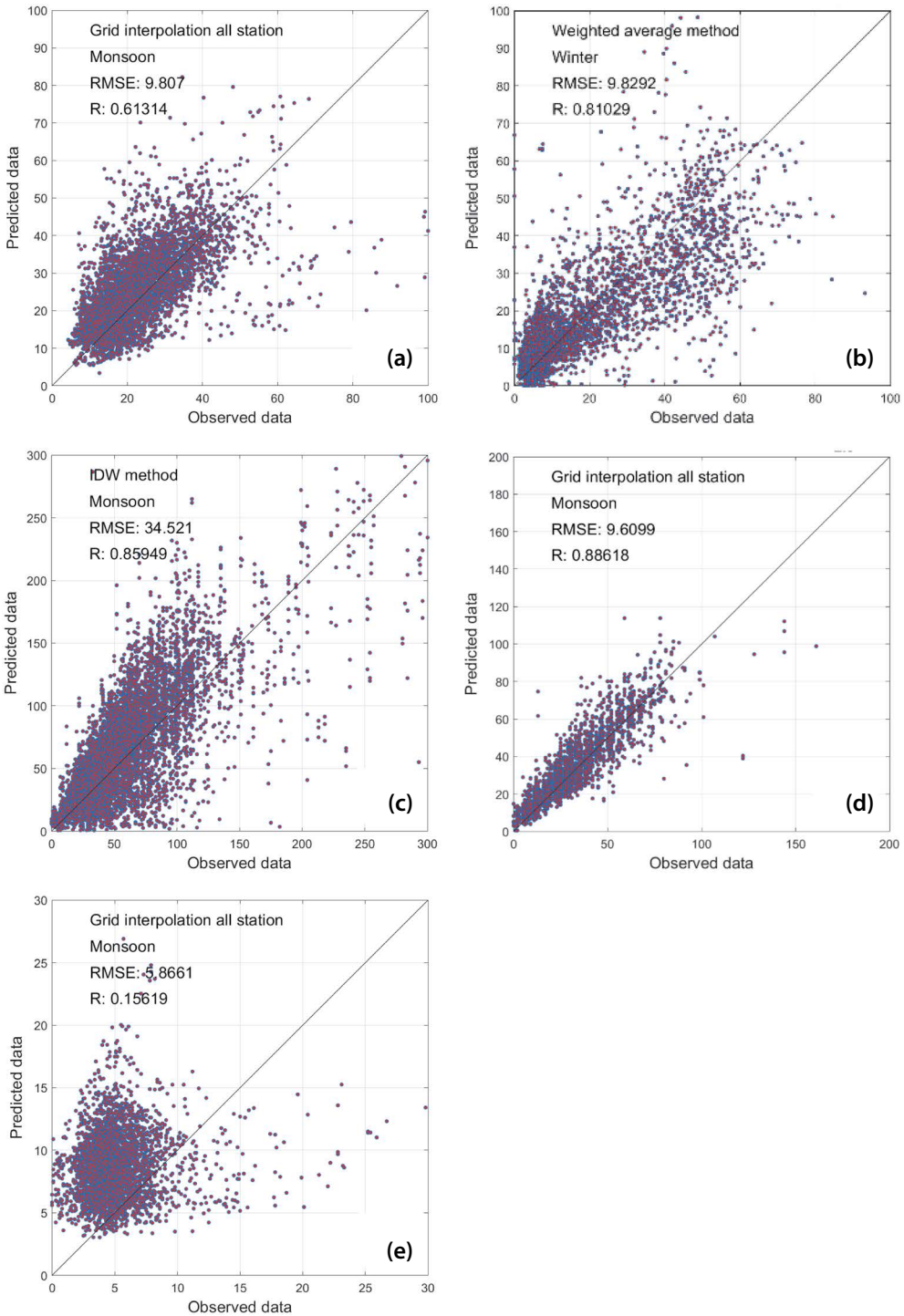

Fig. 16. Scatter plot of the best prediction (having least RMSE) among all seasons, method of predictions, and test stations for (a) NO_2, (b) O_3, (c) PM_{10}, (d) $PM_{2.5}$, (e) SO_2.

5. SUMMARY

The increasing availability of historical data, the number of monitoring stations, and computing resources have facilitated us to develop more advanced models for air pollution prediction. In this study, a methodological

comparison to predict the mass concentration of NO_2, O_3, PM_{10}, $PM_{2.5}$, and SO_2 at any non-monitoring site was carried out.

A correlation analysis was done to check the interdependency of pollutants concentration data among the different observation stations. The correlation matrix was calculated based on the pollutant's concentration data of three years. Further, a model is established based on the correlation coefficients between any pair, their distance, and the direction of the line joining among the 19 monitoring stations of Delhi. This model predicts the strength of correlation between non-monitoring sites and 19 monitoring stations of Delhi.

Five different methods were adopted to predict mass concentrations at non-monitoring sites using the mass concentration data of those monitoring sites, showing a good correlation with non-monitoring sites. Six additional monitoring stations' mass concentrations data were used to validate predicted mass concentrations at those sites. The percentage errors and RMSE were calculated, and comparison was carried out among the methodologies in different seasons for each pollutant.

The correlation between the two monitoring stations has no relation with the direction of their spatial position (direction of the line joining them). The results showed that a simple average on the dependent station model performed best over other methods. The performance of the proposed method is also consistent both for optimal prediction at all stations and has the second minimum mode RMSE among different prediction methods. These methods can be used in future studies and other regions for air pollutant prediction.

6. CONCLUSIONS

The conclusions of the present study can be enumerated as:

1) Based on the correlation analysis, we found that the stations showed higher correlations for the Particulate matter ($PM_{2.5}$ and PM_{10}) compared to other pollutants.
2) Overall, the Grid interpolation method with dependent station was found the best (lowest median RMSE = 5,107.5 μg/m^3) whereas the Weighted Average was the worst (maximum RMSE = 9,734.9 μg/m^3).
3) Averaging is found to be the best prediction method based on mode RMSE(= 18.85 μg/m^3), whereas the Best correlation coefficient method was found to be the worst prediction method (mode RMSE 41.60 μg/m^3).
4) Overall, the prediction for SO_2 was the best among all the pollutants whereas for O_3, it was the worst. Considering different seasons, it was easier to predict the pollutants in the monsoon season, whereas it was most difficult in the post monsoon season. This can be attributed to corresponding level of pollution in these seasons.
5) The methods that used dependent station data (Average, IDW, and GI DS) always have greater mode R^2, whereas the methods that used all station data (BCC, WA, and GI AS) always have mode R^2 smaller as 0.1 (Fig. 15).

7. LIMITATIONS

One of the major limitations of this study is the lack of available information about point source pollution. If emission data about these major point sources were available, proper weights could be assigned to these stations based on distance and direction from the source of pollution. Moreover, the distances between the stations is small, therefore the outcomes of the study may be limited only up to the spatial extent of the study area.

REFERENCES

Alam, M.S., McNabola, A. (2015) Exploring the modeling of spatiotemporal variations in ambient air pollution within the land use regression framework: Estimation of PM_{10} concentrations on a daily basis. Journal of the Air & Waste Management Association, 65(5), 628–640. https://doi.org/10.1080/10962247.2015.1006377

Alimissis, A., Philippopoulos, K., Tzanis, C.G., Deligiorgi, D. (2018) Spatial estimation of urban air pollution with the use of artificial neural network models. Atmospheric Environment, 191, 205–213. https://doi.org/10.1016/j.atmosenv.2018.07.058

Asuero, A.G., Sayago, A., González, A.G. (2006) The correlation coefficient: An overview. Critical Reviews in Analytical Chemistry, 36(1), 41–59. https://doi.org/10.1080/10408340500526766

Bartier, P.M., Keller, C.P. (1996) Multivariate interpolation to incorporate thematic surface data using inverse distance weighting (IDW). Computers and Geosciences, 22(7), 795–

799. https://doi.org/10.1016/0098-3004(96)00021-0
Boaz, R.M., Lawson, A.B., Pearce, J.L. (2019) Multivariate air pollution prediction modeling with partial missingness. Environmetrics, 30(7), e2592.
Chen, S., Oliva, P., Zhang, P. (2018) Air Pollution and Mental Health: Evidence from China. https://doi.org/10.3386/W24686
Crouse, D.L., Goldberg, M.S., Ross, N.A. (2009) A prediction-based approach to modelling temporal and spatial variability of traffic-related air pollution in Montreal, Canada. Atmospheric Environment, 43(32), 5075–5084. https://doi.org/10.1016/j.atmosenv.2009.06.040
Curtis, L., Rea, W., Smith-Willis, P., Fenyves, E., Pan, Y. (2006) Adverse health effects of outdoor air pollutants. Environment International, 32(6), 815–830. https://doi.org/10.1016/j.envint.2006.03.012
Cusworth, D.H., Mickley, L.J., Sulprizio, M.P., Liu, T., Marlier, M.E., Defries, R.S., Guttikunda, S.K., Gupta, P. (2018) Quantifying the influence of agricultural fires in northwest India on urban air pollution in Delhi, India. Environmental Research Letters, 13(4), 044018. https://doi.org/10.1088/1748-9326/aab303
Deligiorgi, D., Philippopoulos, K. (2011) Spatial interpolation methodologies in urban air pollution modeling: application for the greater area of metropolitan Athens, Greece. Advanced Air Pollution, 17, 341–362.
Dominick, D., Juahir, H., Latif, M.T., Zain, S.M., Aris, A.Z. (2012) Spatial assessment of air quality patterns in Malaysia using multivariate analysis. Atmospheric Environment, 60, 172–181. https://doi.org/10.1016/j.atmosenv.2012.06.021
Fan, J., Li, Q., Hou, J., Feng, X., Karimian, H., Lin, S. (2017) A spatiotemporal prediction framework for air pollution based on deep RNN. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 4(4W2), 15–22. https://doi.org/10.5194/isprs-annals-IV-4-W2-15-2017
Guo, H., Sahu, S.K., Kota, S.H., Zhang, H. (2019) Characterization and health risks of criteria air pollutants in Delhi, 2017. Chemosphere, 225, 27–34. https://doi.org/10.1016/j.chemosphere.2019.02.154
Kerckhoffs, J., Hoek, G., Gehring, U., Vermeulen, R. (2021) Modelling nationwide spatial variation of ultrafine particles based on mobile monitoring. Environment International, 154(2), 106569. https://doi.org/10.1016/j.envint.2021.106569
Kumar, N., Middey, A., Rao, P.S. (2017) Prediction and examination of seasonal variation of Ozone with meteorological parameter through artificial neural network at NEERI, Nagpur, India. Urban Climate, 20, 148–167. https://doi.org/10.1016/j.uclim.2017.04.003
Li, J., Heap, A.D. (2011) A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors. Ecological Informatics, 6(3), 228–241. https://doi.org/10.1016/j.ecoinf.2010.12.003
Li, J., Heap, A.D. (2014) Spatial interpolation methods applied in the environmental sciences: A review. Environmental Modelling & Software, 53, 173–189. https://doi.org/10.1016/j.envsoft.2013.12.008
Manan, D.N.A., Aizuddin, A.N., Hod, R. (2018) Effect of Air Pollution and Hospital Admission: A Systematic Review. Annals of Global Health, 84(4), 670. https://doi.org/10.29024/AOGH.2376
Merklinger-Gruchala, A., Jasienska, G., Kapiszewska, M. (2017) Effect of Air Pollution on Menstrual Cycle Length - A Prognostic Factor of Women's Reproductive Health. International Journal of Environmental Research and Public Health, 14(7), 816. https://doi.org/10.3390/IJERPH14070816
Mishra, D., Goyal, P. (2015) Development of artificial intelligence based NO_2 forecasting models at Taj Mahal, Agra. Atmospheric Pollution Research, 6(1), 99–106. https://doi.org/10.5094/APR.2015.012
Mortimer, K.M., Neas, L.M., Dockery, D.W., Redline, S., Tager, I.B. (2002) The effect of air pollution on inner-city children with asthma. European Respiratory Journal, 19(4), 699–705. https://doi.org/10.1183/09031936.02.00247102
Nagendra, S.M.S., Khare, M. (2005) Modelling urban air quality using artificial neural network. Clean Technologies and Environmental Policy, 7(2), 116–126. https://doi.org/10.1007/s10098-004-0267-6
Nagendra, S.M.S., Khare, M. (2006) Artificial neural network approach for modelling nitrogen dioxide dispersion from vehicular exhaust emissions. Ecological Modelling, 190(1–2), 99–115. https://doi.org/10.1016/j.ecolmodel.2005.01.062
Osseiron, N., Lindmeier, C. (2018) 9 out of 10 people worldwide breathe polluted air. https://www.who.int/news/item/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action
Papaleonidas, A., Iliadis, L. (2013) Neurocomputing techniques to dynamically forecast spatiotemporal air pollution data. Evolving Systems, 4(4), 221–233. https://doi.org/10.1007/s12530-013-9078-5
Qi, Y., Li, Q., Karimian, H., Liu, D. (2019) A hybrid model for spatiotemporal forecasting of $PM_{2.5}$ based on graph convolutional neural network and long short-term memory. Science of the Total Environment, 664, 1–10. https://doi.org/10.1016/j.scitotenv.2019.01.333
Rigol, J.P., Jarvis, C.H., Stuart, N. (2001) Artificial neural networks as a tool for spatial interpolation. International Journal of Geographical Information Science, 15(4), 323–343. https://doi.org/10.1080/13658810110038951
Roy, M.P. (2021) Air pollution and Covid-19: experience from India. European Review for Medical and Pharmacological Sciences, 25(8), 3375–3376. https://doi.org/10.26355/eurrev_202104_25749
Russo, A., Soares, A.O. (2014) Hybrid Model for Urban Air Pollution Forecasting: A Stochastic Spatio-Temporal Approach. Mathematical Geosciences, 46(1), 75–93. https://doi.org/10.1007/s11004-013-9483-0
Singh, K.P., Gupta, S., Kumar, A., Shukla, S.P. (2012) Linear and nonlinear modeling approaches for urban air quality prediction. Science of the Total Environment, 426, 244–255. https://doi.org/10.1016/j.scitotenv.2012.03.076
Singh, V., Singh, S., Biswal, A. (2021) Exceedances and trends of particulate matter ($PM_{2.5}$) in five Indian megacities. Science of the Total Environment, 750, 141461. https://doi.org/10.1016/j.scitotenv.2020.141461

Vicente-Serrano, S.M., Saz-Sánchez, M.A., Cuadrat, J.M. (2003) Comparative analysis of interpolation methods in the middle Ebro Valley (Spain): Application to annual precipitation and temperature. Climate Research, 24(2), 161–180. https://doi.org/10.3354/cr024161

Wang, J., Song, G. (2018) A Deep Spatial-Temporal Ensemble Model for Air Quality Prediction. Neurocomputing, 314, 198–206. https://doi.org/10.1016/j.neucom.2018.06.049

Wen, C., Liu, S., Yao, X., Peng, L., Li, X., Hu, Y., Chi, T. (2019) A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. Science of The Total Environment, 654, 1091–1099. https://doi.org/10.1016/j.scitotenv.2018.11.086

Wong, C.M., Ma, S., Hedley, A.J., Lam, T.H. (2001) Effect of air pollution on daily mortality in Hong Kong. Environmental Health Perspectives, 109(4), 335–340. https://doi.org/10.1289/EHP.01109335

WHO (World Health Organization) (2018) Ambient (outdoor) air pollution [Fact sheet]. https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

Yeganeh, B., Hewson, M.G., Clifford, S., Tavassoli, A., Knibbs, L.D., Morawska, L. (2018) Estimating the spatiotemporal variation of NO_2 concentration using an adaptive neuro-fuzzy inference system. Environmental Modelling & Software, 100, 222–235. https://doi.org/10.1016/j.envsoft.2017.11.031

Zou, B., Wang, M., Wan, N., Wilson, J.G., Fang, X., Tang, Y. (2015) Spatial modeling of $PM_{2.5}$ concentrations with a multifactoral radial basis function neural network. Environmental Science and Pollution Research, 22(14), 10395–10404. https://doi.org/10.1007/s11356-015-4380-3