Chemosphere 272 (2021) 129611

Contents lists available at ScienceDirect

Chemosphere

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

Regression-based flexible models for photochemical air pollutants in the national capital territory of megacity Delhi

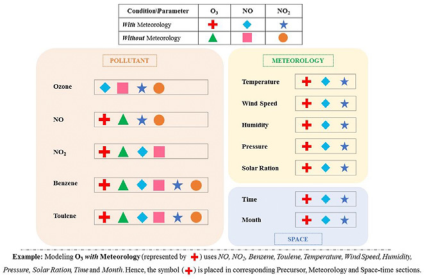
Komal Shukla [a, 1], Nikhil Dadheech [a], Prashant Kumar [b, c], Mukesh Khare [a, d, *]

[a] Department of Civil Engineering, Indian Institute of Technology, Delhi, New Delhi, India
[b] Global Centre for Clean Air Research (GCARE), Department of Civil and Environmental Engineering, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, GU2 7XH, United Kingdom
[c] Department of Civil, Structural & Environmental Engineering, Trinity College Dublin, Dublin, Ireland
[d] Centre of Excellence for Research on Clean Air, Indian Institute of Technology, Delhi, New Delhi, India

HIGHLIGHTS

- Site-specific models with meteorology have better performance over the indicative city level model.
- Lower degrees of polynomial transformation on pollutant prediction show smaller error.
- Hourly averaged observations are well-suited for the prediction of NO, NO_2 and O_3.
- Random forest regression approach produces better models than linear regression for O_3.
- O_3 prediction shows the highest dependence on solar radiation, NO_2, wind speed and NO respectively.

GRAPHICAL ABSTRACT


Condition\Parameter
O₃
NO
NO₂
With Meteorology
Without Meteorology
POLLUTANT
Ozone
NO
NO₂
Benzene
Toulene
METEOROLOGY
Temperature
Wind Speed
Humidity
Pressure
Solar Ration
Time
Month
SPACE
Example: Modeling O₃ with Meteorology (represented by ✚) uses NO, NO₂, Benzene, Toulene, Temperature, Wind Speed, Humidity, Pressure, Solar Ration, Time and Month. Hence, the symbol (✚) is placed in corresponding Precursor, Meteorology and Space-time sections.

ARTICLE INFO

Article history:
Received 29 November 2019
Received in revised form
31 December 2020
Accepted 6 January 2021
Available online 11 January 2021

Handling Editor: Derek Muir

Keywords:
Pollutant forecasting
Polynomial transformation
Random forest regression
Linear regression
O_3
NO and NO_2

ABSTRACT

Modelling photochemical pollutants, such as ground level ozone (O_3), nitric oxide (NO) and nitrogen dioxide (NO_2), in urban terrain was proven to be cardinal, chronophagous and complex. We built linear regression and random forest regression models using 4-years (2015–2018; hourly-averaged) observations for forecasting O_3, NO and NO_2 levels for two scenarios (1-month prediction (for January 2019) and 1-year prediction (for 2019)) — *with* and *without* the impact of meteorology. These flexible models have been developed for, both, localised (*site-specific models*) and combined (*indicative of city-level*) cases. Both models were aided with machine learning, to reduce their time-intensity compared to models built over high-performance computing. O_3 prediction performance of linear regression model at the city level, under both cases of meteorological consideration, was found to be significantly poor. However, the site-specific model with meteorology performed satisfactorily ($r = 0.87$; RK Puram site). Further, during testing, linear regression models (site-specific and combined) for NO and NO_2 with meteorology, show a slight improvement in their prediction accuracies when compared to the corresponding equivalent linear models without meteorology. Random forest regression with meteorology performed satisfactorily for indicative city-level NO ($r = 0.90$), NO_2 ($r = 0.89$) and O_3 ($r = 0.85$). In both regression techniques, increased uncertainty in modelling O_3 is attributed to it being a secondary pollutant, non-linear

[*] Corresponding author. Department of Civil Engineering, IIT Delhi, India.
E-mail addresses: mukeshk@civil.iitd.ac.in, kharemukesh@yahoo.co.in (M. Khare).

[1] Current affiliation: Postdoctoral researcher, University of North Carolina, Institute of Environment, USA.

https://doi.org/10.1016/j.chemosphere.2021.129611

dependency on NO_x, VOCs, CO, radicals, and micro-climatic meteorological parameters. *Analysis of importance* among various precursors and meteorology have also been computed. The study holistically concludes that site-specific models with meteorology perform satisfactorily for both linear regression and random forest regression.

1. Introduction

Photochemical air pollution is a challenge to most of the developing countries of the 21st century (Ghude et al., 2016). It affects human well-being, ecology, infrastructure as well as agricultural systems (Gurjar et al., 2016). Ground level ozone (O_3) is a photochemical pollutant of paramount importance and has been ranked 33rd in the global ranking of health risk factors for total deaths from all causes (Faridi et al., 2018). Among the criterion pollutants, experienced across the nation, O_3 is being considered a 'new-age pollutant' for tropical countries such as India (Sharma and Khare, 2017). By constantly violating the prescribed standards (especially in summers and post-monsoon), O_3 has been growing into a critical urban air quality concern in highly polluted environments of Delhi, India (Hazarika et al., 2019; Kumar et al., 2020). Its chemical precursors: oxides of nitrogen (NO_x) and volatile organic compounds (VOCs) have been in constant dominance across the region; and are an extensive problem for the last decade (Jenkin et al., 2017; Lu et al., 2018).

Human mortality of about 0.25 million in 2015 was attributed to O_3 exposure, especially due to causing chronic lung diseases. Additionally, about 0.10 million premature deaths every year in India are linked with O_3 exposure; of which, 42% are exclusively over the Indo-Gangetic plain (Health Effect Institute, State of Global Air, 2017). Typically, adverse effects of pollution show a considerable effect on the economy, owing to its impact on human, animal and plant health; and these losses, as estimated to be about 7.7% of the national GDP (Amann et al., 2017; Lin et al., 2012; WHO, 2016). Continuous and belligerent degradation in the photochemical air quality specifically over Indian Capital, New Delhi, has raised significant attention; and in many instances tagged, Delhi, as one of the most polluted cities in the world (Mukherjee and Agrawal, 2016).

Long-term exposure to high levels of O_3 may cause a serious decrement in the lung function of children, increase possibilities of asthma and other breathing issues such as chest pain and coughing (Ghude et al., 2008). Moreover, regular contact with its precursors i.e. VOCs can be harmful as it may lead to conjunctival irritations and other health-related issues (Paoletti et al., 2014). The uptake of O_3 in plants may alter the leaf physiology and reduce growth by altering phenology, i.e. number and timing of flowers (Ainsworth et al., 2012). While on the material, direct corrosive effect on plastics, natural rubber, textiles, paints and surface coating is observed (Screpanti and De Marco, 2009). Despite increasing attentiveness on O_3 pollution, in both the science and policy communities, the severity of the pollutant can be well-traced by the findings that, "India has been consistently reported to have one of the highest numbers of premature deaths due to O_3 pollution, which also adversely affects wheat and soybean crop yields" (Pozzer et al., 2015; Zheng et al., 2009).

The sources responsible for O_3 generation (thermal power stations, transport and industrial emissions, domestic use of coal and fossil fuels etc.) have been researched extensively (Ojha et al., 2016). O_3 and its precursors share a complex relationship, owing to interactions between meteorology and chemical processes over a large spatial scale along with an extended timeline. Several investigations have been carried out for Delhi (Ghude et al., 2008; Kumar and Foster, 2009; Sharma et al., 2016; Tiwari et al., 2015) on O_3 and other photochemical pollutants. The studies cumulatively examined the spatial and temporal distribution of pollutants by capturing the trends, studied diurnal cycles, analysed for the response of living species using various models (such as the exposure-plant response of ambient ozone using Ethylenediurea), modelled photochemical pollutants and performed sensitivity analysis etc.

The severity of photochemical air pollution invokes the development of observation-based photochemical kinetic models (PKMs) for O_3 that includes its chemical precursors and meteorological parameters. Formation of such models involve capturing variation, thus, it is easier to forecast the trends for primary pollutants (NO and NO_2) due to their linear nature (i.e. statistical persistence in NO_x), whereas O_3 shows anti-persistent behaviour due to its composite secondary nature (Chelani, 2013). Some regression studies target NO_x ($= NO + NO_2$) (de Foy et al., 2018) and O_x ($= O_3 + NO_2$) (Notario et al., 2012; Clapp et al., 2001) to understand the intermittent consumption of O_3 and NO_2. O_3 is a secondary pollutant and therefore formed through the photochemical reactions between NO_x and VOCs along with CO, through a series of free radical reactions in the presence of sun light (Shukla et al., 2018a; Tiwari et al., 2015). Involvement of complex precursors, radicals and dynamic diurnal pattern makes it difficult to develop an observation-based model for O_3 prediction over any city. These above-mentioned challenges along with rising photochemical pollution strengthen the imperative need to develop an observation-based model linking O_3, NO_x, VOCs and meteorology for highly polluted environments like Delhi. The developed model can be used, either to forecast the pollutant levels or in case of absence/missing of observations.

This paper follows an approach of a polynomial transformation of the data followed by linear regression and random forest regression. Polynomial transformation is usually executed to introduce non-linearity in the dataset. Both models were aided with machine learning to reduce their time-intensity when compared to models built over high-performance computing. The pre-requisite for such techniques is the presence of continuous ground observation of precursors and meteorological parameters for the particular day for which the O_3 prediction is to be made. For this purpose, hourly-averaged pollutant observations and meteorological parameters for 2015–2018 were taken for Delhi, India.

Historically, several studies have been conducted to understand *ground level O_3–precursor* relationship with the help of regression-based analysis (Abdul-Wahab, 2003; Al-Alawi et al., 2008; Khedairia and Khadir, 2012; Özbay et al., 2011). In a similar vein, many studies have analysed O_3 characteristics over Delhi (Ali et al., 2012; Beig and Ali, 2006; Ganguly, 2009; Jain et al., 2005; Mahapatra, 2010; Mishra and Goyal, 2016; Pallavi and Chirashree, 2011). These studies explained the seasonal, annual and diurnal trends of pollutants on *supersites* of Delhi, India and elsewhere. However, none of the studies to-date have worked on forecasting of photochemical pollutants based on their relationship with precursors and meteorological variables. Even the state-of-the-art published research on the development of regression models on ground level pollutant observations has significant gaps and has not been carried out for photochemical pollutants such as O_3, NO_x and VOCs. Therefore, there arises a need for the application of regression models for the above-mentioned scenario. This work is one of the first studies to investigate the regression models for explaining the relationship between emerging photochemical pollutants and meteorological parameters.

2. Methodology

2.1. Study area

The National Capital Territory (NCT) of India covers an area of approximately 1482 km^2, with more than 11 million people in it (Census of India, 2011) and its current population is estimated to be 19.5 million (Populationu, 2020). Past few decades have seen rapid industrial, transportation and real-estate sector growth in the city which led to accelerated degradation of air quality (Coe et al., 2015). Delhi has emissions generating from within (vehicular) and transported from outside (west to east; crop and fossil fuel burning) (Dumka et al., 2018). This study undertakes 3 supersites* in Delhi (Fig. 1): (i) Ramakrishna Puram (RKP; 28°33′46.1"N, 77°11′10.2"E), situated in South West Delhi, is mainly a residential and institutional colony. It has been referred to as an institutional colony because there exist various academic and non-academic institutions (such as embassies, universities, hospitals etc.) nearby. This site has been selected due to heavy traffic on the adjacent major roads such as ring road (Shukla et al., 2020; Guttikunda and Gurjar, 2012); (ii) Punjabi Bagh (PB; 28°39′47.1"N, 77°07′25.2"E), located in West Delhi, is primarily a residential area along with few industries in Mayapuri and Mangolpuri. The increased vehicular emissions mainly on the neighbouring Rohtak Road and Ring Road are the major reasons for the site performing poor in the air quality index (Shukla et al., 2020); and (iii) Mandir Marg (MM; 28°38′02.3"N, 77°12′00.8"E), based in Central Delhi, is also primarily a residential location along with few industries situated nearby. Mandir Marg has one of the busiest routes with high vehicular emissions (Shukla et al., 2020).

**Supersites include a major area of a city such as colonies, sectors or wards with a significant proportion of the population living in it (Solomon and Sioutas, 2008). A supersite can be residential, institutional or industrial, such as Anand Vihar, Punjabi Bagh etc. in case of Delhi.*

Among the three aforementioned sites, two have residential and industrial land uses while the third has a residential and institutional land use. The ratiocination against this is that, as per the Delhi Development Authority (DDA, 2020), a major proportion of Delhi-NCT is residential.

Several studies in the past have considered between 1 and 6 sites to build a city-level pollutant forecast model, with varying levels of prediction accuracies. Wang et al. (2013) have used a novel technique, Single Point Areal Estimation (SPA), to extend the pollutant mass concentrations obtained at a single station to a citywide scale for Beijing metropolis. Kheirbek et al. (2013) had adopted hourly O_3 data from seven regulatory monitors in/around the counties of New York City to assess for air quality health impacts and disparities at City level. García et al. (2011) have developed an O_3 prediction model using Artificial Neural Network (ANN)

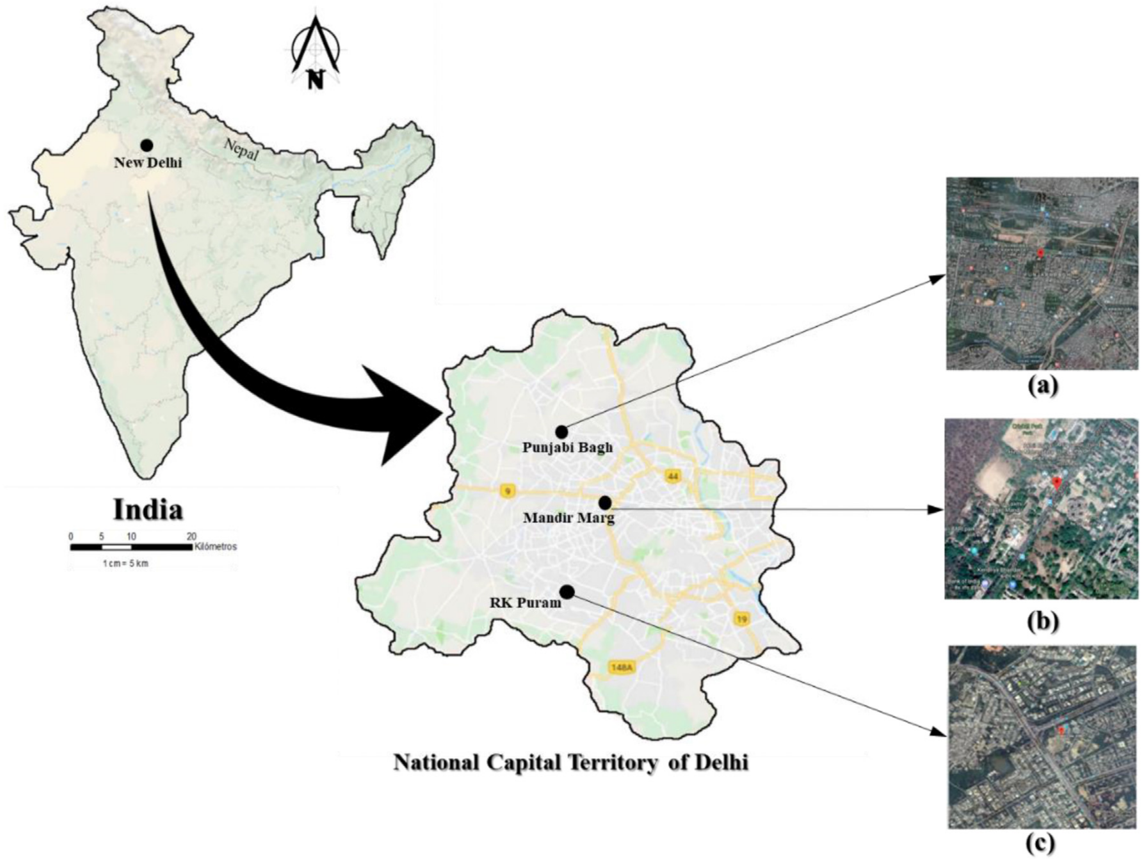


Fig. 1. Sites with varying land-use patterns under observation in Delhi, India a) Punjabi Bagh b) Mandir Marg c) RK Puram.

as the base framework. They have adopted the data pertaining to chemical variables from a single station (Miravalle Station; South of the City) and meteorological parameters from another station (Chapala station; City Centre) to model O_3 in the entire city of Guadalajara, Mexico. Ruiz-Suarez et al. (1994) and Ruiz-Suarez et al. (1995) have developed and employed neural network paradigms (Bidirectional Associative Memory (BAM) and Holographic Associative Memory (HAM)) towards short-term forecasting of ozone for Mexico City, using data from five stations of RAMA (Mexico City's automatic air quality monitoring network).

For the combined (*indicative of city-level*) model, we aggregate emissions and meteorology of all three Delhi-NCT supersites in one matrix. In current study, 'Combined model' or 'indicative of city-level model' refer to a model that has been developed using machine learning algorithm while using the collated parametric data from all hotspots as input (i.e. input to training the algorithm/model). This collation combines the characteristics of the hotspot sites and inculcates them into the final developed city-level model. Since these three stations carry different innate (land-use, topography, terrain etc.) and incidental (meteorology and pollutant emissions) characteristics, it is assumed that combining the data obtained from the above-mentioned stations would inherently represent the data corresponding to the entire city, in general. The applicability of a combined (indicative of city-level) model can be extremely important in case of unavailability of measurement stations at any location point in the city. The combined model in this study gives uses three stations to provide an indicative ambient ozone concentration which may be assumed as representing Delhi city. The model would strengthen as we increase the number of sites.

During the hot summer afternoons, O_3 concentrations in many parts of Delhi are often found to exceed even 200 $\mu g/m^3$, against the 8-h average (100 $\mu g/m^3$) standard provided by the NAAQS (Fig. 2). Data appertaining to hourly-averaged observations for 4 years (2015–2018), has been adopted for analysis in polynomial transformation. Statistical analysis of ground monitored ozone and its precursors (VOCs: Benzene, Toluene; NO_x: NO, NO_2) at key hotspot sites in Delhi aforementioned, elucidates the complex photochemistry (Table 1).

This study is based on the generalization of the data, therefore an indicative city-wide model can be formed since the model needs concentration of pollutants and meteorological parameters as input to predict the concentration of label class. Hence, after the inclusion of meteorological parameters, the model becomes *generalized at* city-level. However, the performance may be affected due to the significant difference between the spread of the data in multiple sites.

Ground observations of pollutant concentrations and data pertinent to meteorological parameters are taken from an inventory of the Central Pollution Control Board (CPCB), India, continuous monitoring system. At all the monitoring sites (shown in Fig. 1), O_3 is measured using online ozone analyser (model O342 M, Environment SA, France), which works on UV absorption technology (CPCB, 2016); NO_2 is measured using Jacob and Hochheiser modified (NaOH–$NaAsO_2$) method and Gas-Phase Chemiluminescence; and VOCs are measured using Gas Chromatography (GC) based continuous analyser, adsorption and desorption followed by GC-MS analysis. The observations in the adopted data are missing for some days due to maintenance work at the monitoring station or any defect in the measuring instruments.

2.2. Methods for analysis and model building framework

The study uses regression analysis, which has been used in various areas of research such as boundary integrals (Sladek et al., 2000), time-series auto-regression and evaluation of other existing transformations like logarithmic and square root projections (Kumar and Foster, 2009; Pearce et al., 2011; Tao et al., 2012). Linear and random forest regression technique combined with machine learning have been used in this paper to perform meteorological and precursor adjustment, for prediction of O_3, NO and NO_2 (method chart in Fig. 3).

The model is trained after pre-processing the observations. Proper quality assurance has been adopted for the dataset through pre-processing which includes only those data points which contain the entire information i.e. concentrations or measurements corresponding to all features. The approach is basically to delete the entire data point if any of the pollutant concentration is null i.e. not recorded by the station. Also, it has been observed in the data that for a period of significant weeks the concentration of major pollutants was recorded zero which could not be the case. Consequently, to assure the quality those data points have also been deleted. The pre-requisite for the data to train a machine learning-based regression model is that it should not have missing data. However, one can perform data imputation i.e. filling missing data based on an average or a distribution but it cannot be completely accurate. Therefore, to maintain the complete accuracy of the data, the deletion operation was performed. Also, the high concentration data points were not removed because ozone itself contains several spikes in its distribution and if these values were removed then models would not have learnt enough during training as the distribution function may not be differentiable at every point.

Polynomial transformation provides more flexibility over individual distributions of different emissions and meteorology as it can be safely assumed that most of the distributions can be expressed through polynomial expressions. It is done through the introduction of new columns in the data matrix by raising the polynomial order of the entire data point (row of the data matrix). This is done to morph the data points into a polynomial curved shape in a graph of nth dimension (range for this study is from 1st to 10th) space. It increases the number of features significantly to establish the relationship between features (*precursors*) and labelled data (*to be forecasted*).

The regression technique (can be linear regression or random forest regression), that follows polynomial transformation, can establish a relationship between the labelled data Y_i (O_3, NO, NO_2 in this case) and features X_1, X_2, X_3……. X_n (= known concentrations of other pollutants including meteorological parameters such as toluene, benzene, temperature etc.).

Linear regression performed in this study can be well understood using linear optimization theory. Given a dataset D = $\{(x^{(i)}, y^{(i)})\}m_{i=1}$, linear regression optimization condition can be written as below. Equation (1) (Neal, 2009) provides the optimal hyperplane as:

$$J(w) = \min_{w,b}\frac{1}{2m}\sum_{i=1}^{m}\left|\left|w^{T}x^{(i)} + b - t^{(i)}\right|\right|^{2} \tag{1}$$

Here, (w,b) represents the hyper-plane fitting the data. When J(w) is the error function, m is the total number of reading of the data used in training, w is the coefficients of the equation, $x^{(i)}$ is the feature data of ith reading, b is the constant to adjust the noise, $y^{(i)}$ is the concentration of pollutant to be predicted for ith reading, $w^{T}x^{(i)} + b$ is $Y_{predicted}$ and $t^{(i)}$is Y_{actual}. $J(w)$ has been used to minimise the error and calculate the coefficients of the regression equation using the training data. Efforts were put to reduce the error, which was calculated by comparing $Y_{predicted}$ and Y_{actual}. The above condition in case of linear regression gives the following equation:

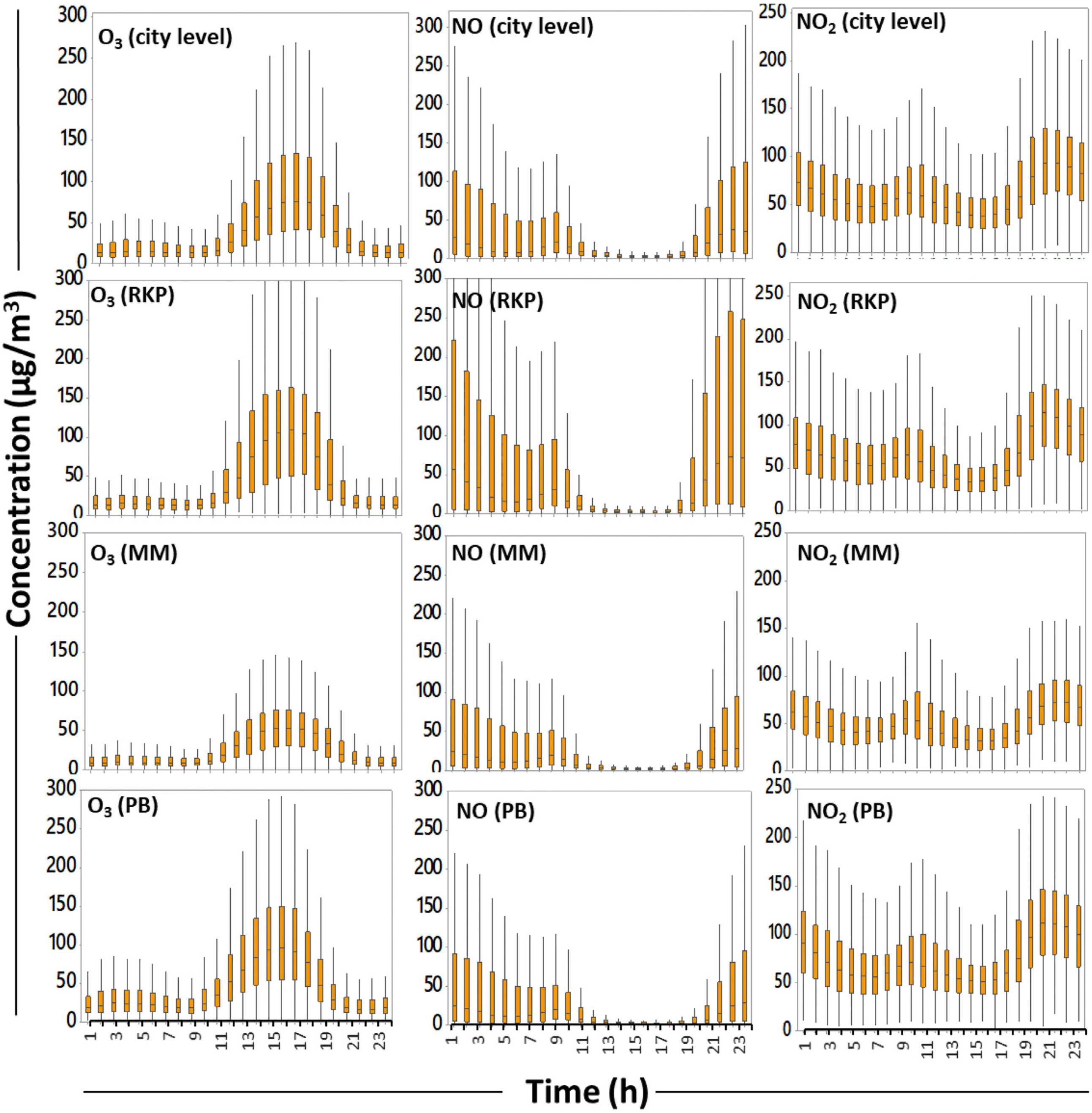


Fig. 2. Average diurnal variation of photochemical pollutants at city level and sites level.

$$W= (X^TX)^{-1}(X^TY) \quad (2)$$

Where W is the matrix of coefficients of the equation, X is the feature data which will be known in future and Y is the known labelled data which we have to predict in the future, but during training – initial data will be provided to train the model.

Linear regression has been used – in Athens and Helsinki, for predicting NO_x and PM_{10} (Vlachogianni et al., 2011); in Morocco, for evaluating various O_3 prediction models (Oufdou et al., 2018); in Portugal, daily average O_3 coupled with principal component analysis (Sousa et al., 2007); next-day PM_{10} concentration in Malaysia (Ul-Saufie et al., 2013). Furthermore, Random forest regression is used on hourly photochemical pollutants to improve the predictions. Random forest regressor is a machine learning method for classification and regression, and has multiple decision trees (Hu et al., 2017). These decision trees are divided based on each feature after setting up a particular threshold value and this way, data is divided in different branches of the trees. All the component trees use a random sample subset from the dataset. For every individual tree, equal probability induced predictors are selected. The output is calculated by taking the mean and aggregation of every individual component tree.

Random forest has been researched to perform better than other linear regression techniques aided with machine learning (Archer and Kimes, 2008; Hengl et al., 2015; Nicolas et al., 2016). Random forest regression is proven to produce good predictions for air pollutants such as $PM_{2.5}$, NO and NO_2 in Poland (Kamińska, 2018), monthly $PM_{2.5}$ in China (Huang et al., 2018), excellent in advancing

Table 1
Average (2015–2018) photochemical pollutant concentrations and meteorological variables.

Pollutant and meteorological – hourly average (2015–2018)			
Pollutant Concentration (μg/m^3)	RKP	PB	MM
	Average ± σ	Average ± σ	Average ± σ
Ozone	50 ± 58	41 ± 53	32 ± 38
NO	56 ± 122	26 ± 66	28 ± 57
NO_2	66 ± 45	66 ± 60	55 ± 38
NO_x	140 ± 190	95 ± 134	88 ± 95
Benzene	7 ± 10	3 ± 6	3 ± 3
Toluene	16 ± 13	N/A	11 ± 15
Meteorological Parameter	RKP	PB	MM
	Average ± σ	Average ± σ	Average ± σ
Temperature (°C)	25 ± 8	18 ± 12	24 ± 9
Humidity (% Rh)	53 ± 21	44 ± 26	54 ± 21
Wind Speed (m/s)	1 ± 0	0.9 ± 0.76	1 ± 3
Pressure (hPa)	688 ± 262	768 ± 125	702 ± 76
Solar Radiation (w/m^2)	117 ± 144	76 ± 109	114 ± 141

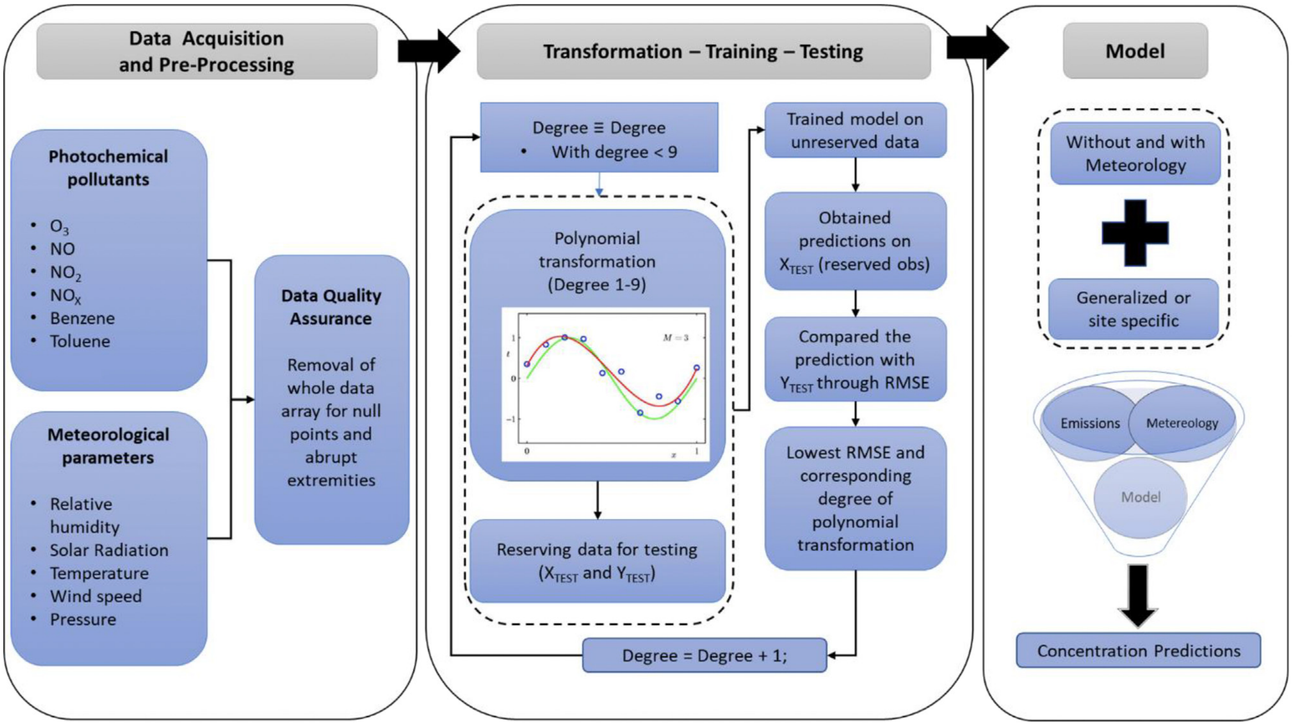

Data Acquisition and Pre-Processing
Photochemical pollutants
• O3
• NO
• NO2
• NOX
• Benzene
• Toluene
Meteorological parameters
• Relative humidity
• Solar Radiation
• Temperature
• Wind speed
• Pressure
Data Quality Assurance
Removal of whole data array for null points and abrupt extremities
Transformation – Training – Testing
Degree ≡ Degree
• With degree < 9
Polynomial transformation (Degree 1-9)
M = 3
t
0
−1
0
x
1
Reserving data for testing (XTEST and YTEST)
Trained model on unreserved data
Obtained predictions on XTEST (reserved obs)
Compared the prediction with YTEST through RMSE
Lowest RMSE and corresponding degree of polynomial transformation
Degree = Degree + 1;
Model
Without and with Meteorology
Generalized or site specific
Emissions
Metereology
Model
Concentration Predictions

Fig. 3. Model development framework for photochemical pollutants.

$PM_{2.5}$ in USA (Liu et al., 2018) and O_3 in China (Zhan et al., 2018). The approach is also beneficial for downscaling meteorological parameters such as wind (Davy et al., 2010) and temperature (Hutengs and Vohland, 2016). The random forest model performs better than linear regression because of its structural algorithms. Unlike a linear regression model, it can exploit more context from the feature and increase the training data through its decision trees and branches (Li et al., 2014). Kamińska (2018) also concluded that a random forest model is better than linear regression for mapping the mathematical reality when predicting dynamically varying features. Their study also concurs that meteorology is an important factor for predicting NO and NO_x.

The random forest in this study consisted of 10 trees, which is a hyperparameter given to the model and we have used the default value of the random-forest python library. The default number comes after much detailed analysis of varying the number of trees and evaluating the models. Therefore, it has been decided to go with the default value. Random forest regression also yielded 'attributed importance' of different variables in making the prediction. The matrices include variables with their percentage. Higher percentage denotes that more importance is given to that variable in determining the prediction (Gregorutti et al., 2017). The study also formed different combinations of variables for predicting ground level O_3 to understand the role of various variables.

The models were built separately for the data from all three sites, allowing different aspects of pollutants, and comparison between '*with or without meteorology data*' for a *specific site* to include local variations. The model is trained after pre-processing the observations adopted from CPCB pollutant inventory. In this study, while executing both the models, 90% of the total data adopted from the field were randomly selected for model training while the remaining 10% has been selected randomly and was used for testing (validating) the trained model. This ratio was adopted to arrive at finer prediction and hard training of the model. Upon training and validating the model for 4-year data (2015−18), the third phase of model development, i.e. model performance with meteorology, is assessed for two scenarios — 1-year prediction (against 2018−19) and 1-month prediction (against January 2019).

The forecasting only depends on the features used i.e. pollutants concentration and meteorological parameters; and hence, it has no dependency on the period. It can be used for any duration providing the features are contained inside its spread. If they cross the variance then it will be a situation which was not taught to the model while training.

2.3. Performance indices

The coefficient of correlation (R) and Root Mean Square Error (RMSE) have been considered to evaluate the performance of the developed models.

$$R = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \underline{Y_i})^2 - \sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2}{\sum_{i=1}^{n}(Y_i - \underline{Y_i})^2}} \tag{3}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2}{n}} \tag{4}$$

R is a numerical evaluation measure, meaning a statistical relationship between the actual and predicted values, with scale starting from 0 (no correlation) to 1 (perfect correlation). *RMSE* is used as an evaluation factor in several other air pollutant studies (Chaloulakou, 2003). After training and testing, *RMSE* is calculated for the models corresponding to all the degrees of transformation to find out the best model (lowest RMSE). The degree which had the lowest RMSE corresponding to it, was noted as the optimum polynomial degree.

3. Results

Regression models of hourly O_3, NO and NO_2 were produced using linear regression and random forest technique under varying cases of meteorology. Random forest regression improves prediction for photochemical pollutants over linear regression.

3.1. Regression models for ozone

Ground level O_3 in Delhi city does not have direct point sources. Its formation is highly NO_x sensitive and less VOC sensitive (Shukla and Khare, 2019). The overall trend of NO_x emissions has been increasing (rapid increase in NO_2), leading to more favourable conditions for O_3 generation in the city (Shukla et al., 2018b). However, the emerging policies and control on NO_x emissions might lead this zone due to being VOC sensitive in future. It has also been observed that micro-meteorology has become most important in forming/destructing O_3 at any point (Jing et al., 2016). To understand O_3 generation, the performance of regression-based models of all the sites for predicting O_3 (using hourly average) *with* and *without meteorology* has been evaluated. In the majority of the cases, the transformed degree of the original dataset into a 2nd to 4th polynomial degree. The observed order of the performance is also similar for all the sites i.e. smaller error observed for *site-specific model considering meteorological parameters* compared to the *site-specific model without meteorology*. O_3 in observations exhibits a sharp daytime increase in concentration due to photo-oxidation of precursor gases (NO_x, CO, CH_4, NMHC and VOCs). While after sunset, the loss of O_3 is due to its titration by NO and surface deposition produces low mixing ratios (Coyle et al., 2002).

Preliminarily, a multiple linear regression (MLR) model for daily averaged data was formulated and tested. This daily averaged model was not able to capture the diurnal profile changes of photochemical pollutants for Delhi city. Performance of the models drastically improved when hourly averaged concentrations of O_3 and precursors were used instead of daily averaged data or maximum daily 8-h average (MDA8) to build a model. The analysis results into the observation that initially the optimised degree of transformed training dataset was found to be either 1st or 2nd, but in the hourly dataset, it shifted to 2nd to 4th. It indicates that the relationship between ozone and other pollutants is corresponding to degree 2 or 4 in nature when the observations are carried out more vertically i.e. hourly. A similar approach of the linear regression model for O_3 has been discussed by Jing et al. (2016) and they trained their basic linear regression model on pollutants and seasonality as features, which are similar to the feature identification step of this study. The results obtained from this study are consistent with findings from their study and that meteorology plays a crucial role in calculating the concentration of ozone. To enhance the prediction accuracy and assert the effect of seasonal variation, solar radiation, time and month have also been added to the features list. The nature of O_3, due to being secondary pollutant and diurnal, is well captured in a model designed using hourly average observations. Further, random forest regression has been applied on hourly O_3 observations (Zhong et al., 2017 and Lei et al., 2018). Markedly, when the random forest is used on hourly O_3 concentrations for predictions, the impact of NO_x emissions and meteorology i.e. hourly changing solar radiation and relative humidity are embedded better (Tiwari et al., 2015; Gioda et al., 2018). The study formed 10 trees under random forest regression.

After training, Ground level O_3 modelling was tested with meteorology for city-level through hourly average observations using linear regression technique showed $R^2 = 0.45$ ($r = 0.67$, $RMSE = 37.07$), and using random forest technique $R^2 = 0.74$ ($r = 0.85$, $RMSE = 25.65$) (Table 2). A comparative assessment involving produced correlation from linear regression and random forest with a meteorology case is represented by a Taylor diagram (Fig. 4). A Taylor diagram can show model performance changes between any 2 modelling approaches e.g. 2 different model's, their versions or setups (Taylor, 2001). Clearly, during testing phase, the random forest *with* meteorology has achieved excellent correlations (varying from 0.80 to 0.94) and relative standard deviations for site-specific and indicative city-level model against linear regression *with* meteorology (Fig. 5). For instance, during testing, RK Puram site has achieved ground level O_3 predictions with highest correlation (r) i.e. 0.92 (random forest), against 0.86 (linear regression).

The features that have been taken to predict and build a model for ground level O_3 are below:

1. Features (with meteorology case): NO, NO_2, benzene, toluene, temperature, humidity, wind speed, pressure, solar ration, time, month
2. Features (without meteorology case): NO, NO_2, benzene, toluene, time, month

3.2. Importance of different variables in ozone formation

Importance of different predictive variables in forecasting ground level O_3, was determined as a summation of the increment

Table 2
Regression model indices after model testing for Ozone using hourly averages (2015–2018).

Technique	Linear Regression			Random forest		
Area	R^2	RMSE after polynomial fit	Degree min RMSE	R^2	RMSE after polynomial fit	Degree min RMSE
with meteorology						
Combined (city level Delhi)	0.45	37.07	2	0.74	25.65	1
R.K. Puram	0.75	27.58	3	0.85	21.69	1
Punjabi Bagh	0.51	38.41	3	0.66	31.42	1
Mandir Marg	0.46	25.24	2	0.74	17.54	1
without meteorology						
Combined (city level Delhi)	0.29	41.88	3	0.65	29.19	1
R.K. Puram	0.45	41.90	3	0.78	26.47	1
Punjabi Bagh	0.24	47.03	3	0.56	35.59	1
Mandir Marg	0.23	30.22	3	0.68	19.34	1

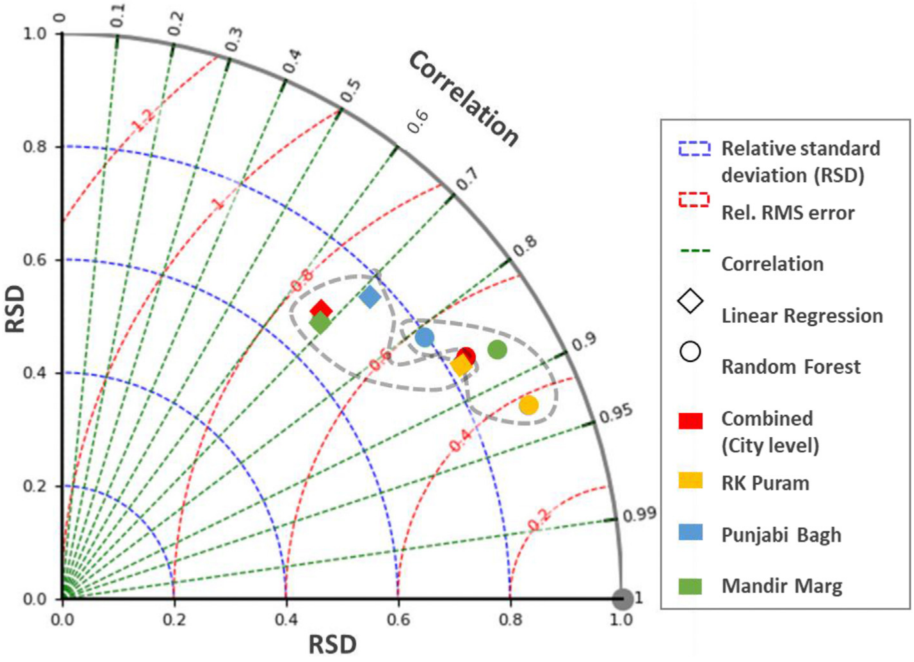


Fig. 4. Taylor diagram representing hourly ozone predictions using linear regression and random forest regression.

in re-substituting estimates across all the individual tree nodes (Breiman, 2001). Importance value is the percentage of the maximum sum, and its maximum value is 100 for the most important predictive variable. The importance or say dependence of O_3 calculation on other variables through random forest regression has been calculated for all 3 sites and indicative city level (Fig. 6). Among all the variables and for all cases (city-level or site-specific), the highest dependence on the prediction of ozone concentration was observed to be on relative humidity (28–44%). The next variables that were in the order of dependence are solar radiation, NO_2, NO and benzene. It should be noted that NO_2 plays a dominant role while regressing at certain sites in comparison to solar radiation. These results were in line with the findings of the model formulated by Abdullah et al. (2019), where relative humidity was found to be one of the significant predictors of O_3.

3.3. Regression models for NO and NO_2

The major source of NO in an urban atmosphere for that of Delhi is the burning of fossil fuels, biomass, lightning and microbiological emission from soil (Singh et al., 2011). A clear rise in the NO (from 2015 to 2018) along with the presence of other VOCs is observed, which can be an attributed cause of increasing ground-level O_3 in the Delhi region (Shukla and Khare, 2019). Historically, NO_x emission has been found to be increasing, and have reached 1,84,000 tons in 2012 from 1,20,500 tons in 2001 (approx. 52.6% rise), due to fuel and technology conversion from *petrol and diesel with 2-stroke engines* to *CNG with 4-stroke engines* (Goel et al., 2015). The current status invokes the development of flexible regression based models to understand formation of NO and NO_2.

For NO, testing performance of the models from linear regression is better in terms of obtained R^2 and RMSE than ground level O_3 (Table 3) and it also follows almost the same trend with respect to the degree of polynomial transformation corresponding to least RMSE. In case of testing regression on NO (Fig. 7a and c), site specific models are showing better correlation that indicative city-level regression models, while there is a significant difference between *with* and *without* the meteorological case. Peculiarly, NO regression models improve (produces lower RMSEs, while R^2 is slightly improved as well) when developed with meteorology. This infers that NO (label) regression models tested to be the best only when precursors (like NO_2 and various VOCs (features)) are taken (Bisht

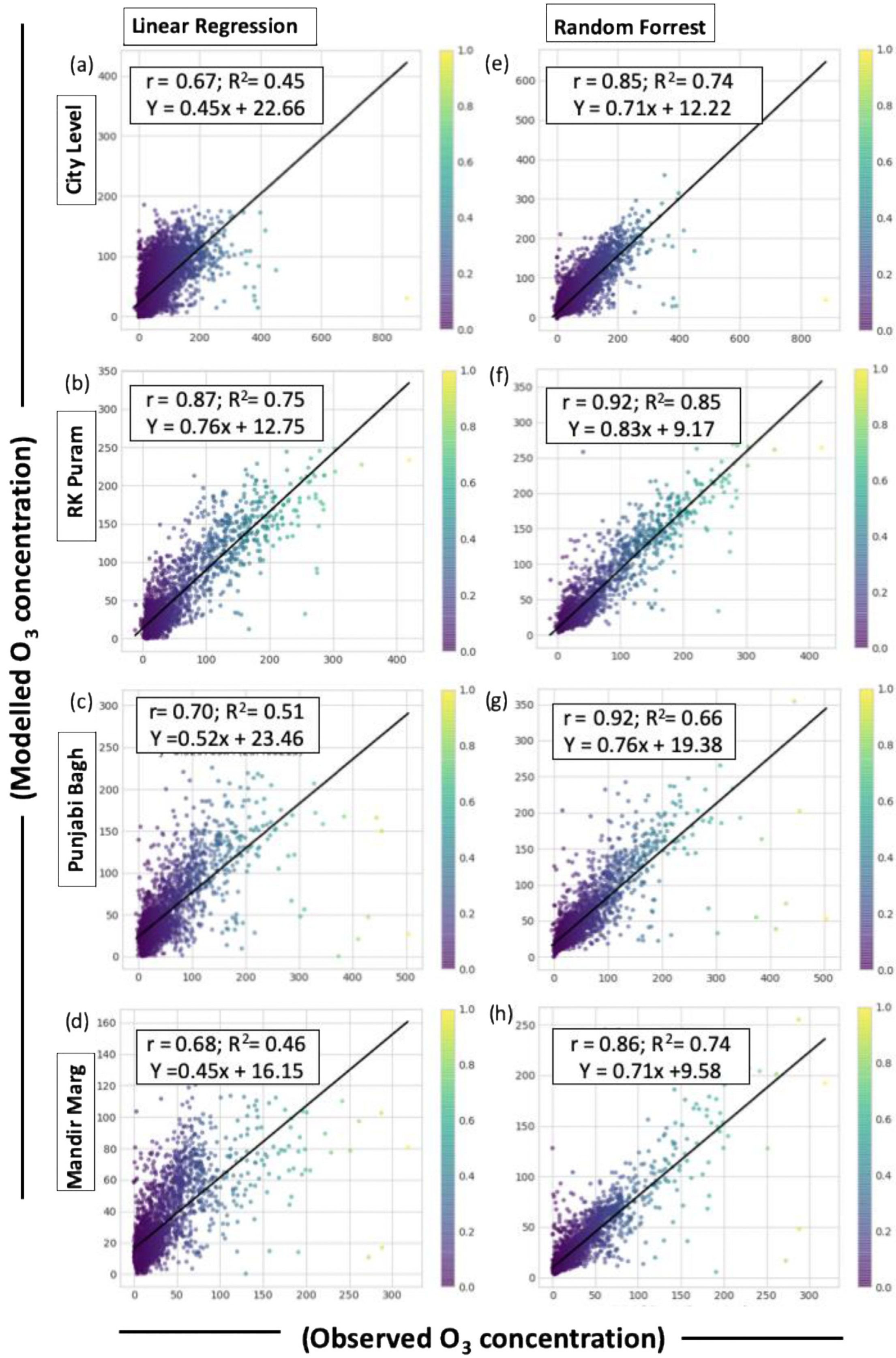


Fig. 5. City-level ozone testing results obtained using: linear regression (a) and random forest (e); Site-specific ozone prediction using: linear regression (b–d) and random forest (f–h).

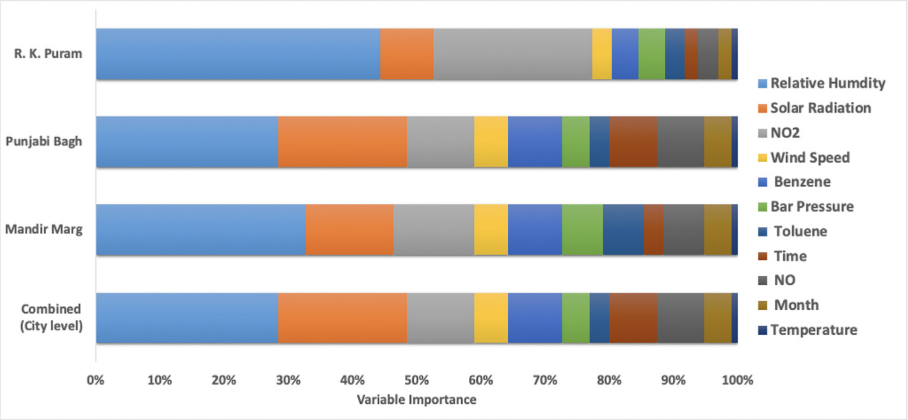


Fig. 6. Importance of various variables (precursors) against ground level O_3 using random forest regression.

Table 3
Testing results of linear regression and random forest model for NO and NO_2 using hourly averages (2015–2018).

NO	Linear			Random Forest		
Area	R^2	Optimal RMSE after polynomial fitting	Degree with minimum RMSE	R^2	Optimal RMSE after polynomial fitting	Degree with minimum RMSE
with meteorology						
Combined (City-level)	0.55	62.6	3	0.81	39.9	1
R.K. Puram	0.56	89.6	3	0.83	55.2	1
Punjabi Bagh	0.61	48.2	4	0.80	39.9	1
Mandir Marg	0.63	36.4	2	0.77	28.5	1
without meteorology						
Combined (City-level)	0.40	72.0	3	0.75	46.20	1
R.K. Puram	0.49	98.0	3	0.76	66.55	1
Punjabi Bagh	0.53	53.2	4	0.68	48.52	1
Mandir Marg	0.49	40.9	3	0.74	30.46	1
NO_2	Linear			Random Forest		
Area	R^2	Optimal RMSE after polynomial fitting	Degree with minimum RMSE	R^2	Optimal RMSE after polynomial fitting	Degree with minimum RMSE
with meteorology						
Combined (City-level)	0.50	35.8	2	0.79	23.0	1
R.K. Puram	0.72	25.1	3	0.81	19.8	1
Punjabi Bagh	0.67	36.7	3	0.84	33.1	1
Mandir Marg	0.53	23.4	2	0.77	16.8	1
without meteorology						
Combined (City-level)	0.42	38.6	3	0.75	25.49	1
R.K. Puram	0.52	32.9	3	0.75	22.98	1
Punjabi Bagh	0.58	41.3	4	0.65	39.82	1
Mandir Marg	0.37	27.2	2	0.74	18.09	1

et al., 2015). In contrast to regression on O_3, which yields high dependence on meteorology, both NO and NO_2 are not much affected with meteorological parameters (Tables 2 and 3). The observed value of R^2 is conspicuously high as compared to O_3. This is contributed by the stable nature of NO concentrations in comparison with O_3. NO does not have any effective seasonal variations which result in its persistence (Xu et al., 2018). The stable and accurate test results of NO can be credited to the more or less equal rate of formation and consumption of NO in the atmosphere.

The testing performance of models for all the sites to predict NO_2 (Table 3) is evaluated with a similar methodology. It is clearly observed that for most of the cases, degree 1 is producing the best test performance among all other degrees of the original observations which are transformed into the polynomial data. After analysing testing results, NO_2 models are also found to be best for *with meteorology cases*, where not much difference is observed between site-specific and indicative city-level. It is concluded that NO_2 has a very similar pattern as NO, as they are found as a mixture of gas-phase organic molecules which is represented as NO_x (Pusede et al., 2015). It exhibits similar chemical interference with O_3. It constructs O_3 and acts as a vital component of the photochemical cycle. The photolysis of NO_2 ($\lambda \leq 424$ nm) leads to formation of atomic oxygen (O^3 P) and NO (5). Here, (O^3 P) reacts with atmospheric O_2 and forms O_3 (6). This leads to a null photochemical cycle, where O_3 now combines with NO to form NO_2 (7). In further, complex sets of reactions, oxidants like HO and HO_2 are also involved in conversion of NO to NO_2 (Wang et al., 2017). Photochemical formation of O_3, NO and NO_2 is interdependent and thus regression techniques produce a good insight (Figs. 7 and 8).

$$NO_2 \ \rightarrow NO + O(3_P)\ (\leq 424\ nm) \qquad (5)$$

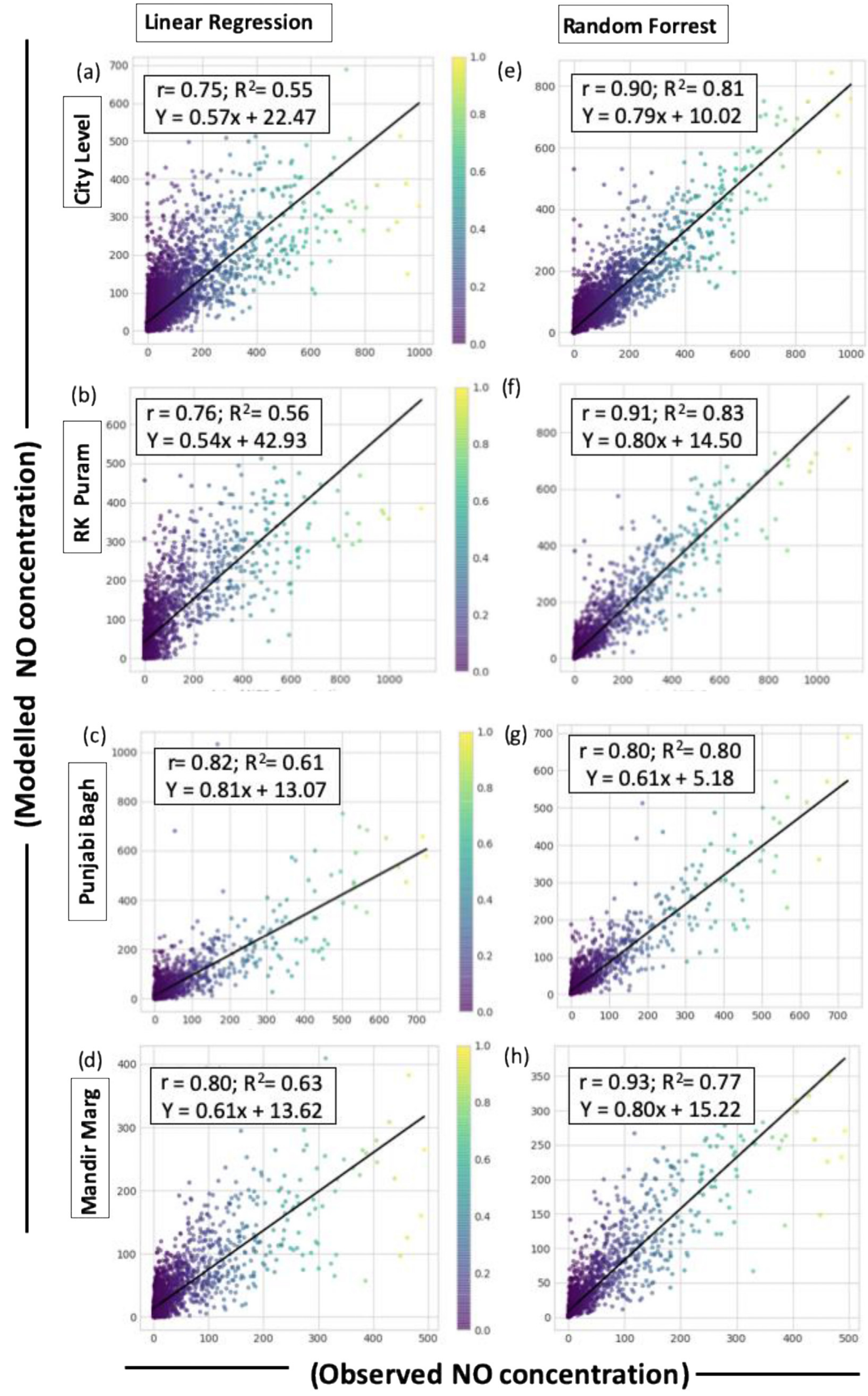


Fig. 7. City-level NO testing results obtained using: linear regression (a) and random regression (e); Site specific NO predictions using: linear regression (b–d) and random forest (f–h).

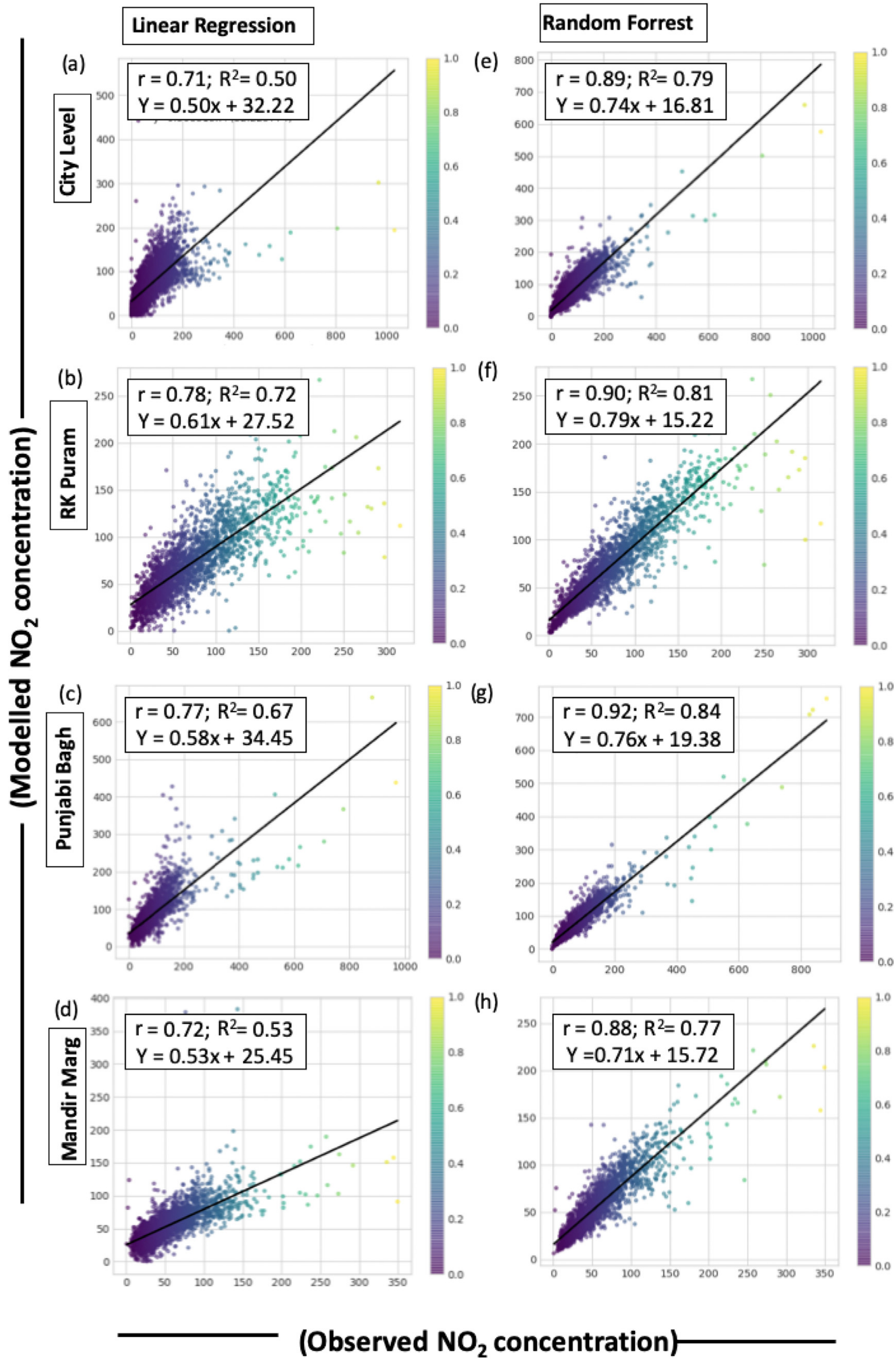


Fig. 8. City-level NO_2 testing results obtained using: linear regression (a) and random regression (e); Site-specific NO_2 predictions using: linear regression (b–d) and random forest (f–h).

$$O\,(3_P)\;+\;O_2 \rightarrow O_3\;(M\;body\;as\;reac\;\tan\;t\;and\;product) \tag{6}$$

$$O_3 + NO = NO_2\;+O_2 \tag{7}$$

However, model testing shows that NO_2 has slightly less accuracy as compared to the NO, and it may be associated with the formation of nitrate and reactions to form aerosols. The source of the NO_2 emission lies close to the ground such as fossil fuel combustion and biomass burning. NO_2 (lifetime < 1 day) varies with meteorological parameters, photolysis rate and concentration of hydroxide radicals (Sheel et al., 2010). This variation can also be a reason for NO performing slightly better in prediction compared to NO_2 (Fig. 8 and Table 3). Also, NO_2 is one of the major sources of O_3 and hence justifies its association with O_3 concentration (Shukla et al., 2017; Shukla and Khare, 2019; Zheng et al., 2009). It also exhibits enhanced seasonal variations such as its concentration decreases during monsoon and increases during summer.

The features that have been taken to predict and build a model for NO and NO_2 are below:

1. Features for NO (with meteorology case): O_3, NO_2, benzene, toluene, temperature, humidity, wind speed, pressure, solar ration, time and month
2. Features for NO (without meteorology case): O_3, NO_2, benzene, toluene, time and month
3. Features for NO_2 (with meteorology case): O_3, NO, benzene, toluene, temperature, humidity, wind speed, pressure, solar ration, time and month
4. Features for NO_2 (without meteorology case): O_3, NO, benzene, toluene, time and month

The efficient test performance of random forest as compared to linear regression is associated with consideration of only linear dependency between labels and features in linear regression. However, Random forest classifies the data in different branches of the trees and, hence, brings non-linearity into consideration. It also works on the concept of decision trees which is related to partitioning the data based on Gini impurity function, ball trees and KD trees (Bogdan and Mozgovoy, 2019; Notario et al., 2012; Laber et al., 2019).

3.4. Coefficients of the regression equation: O_3, NO and NO_2

The coefficients for regression equations for prediction of O_3, NO and NO_2 along *with* and *without* meteorology have been produced through the linear regression (Equations (1) and (2)). The hourly average concentration of O_3, NO and NO_2 for Delhi city can be computed from the (polynomial degree 1) equations below with considering the effect of meteorological parameters.

$$[\text{Hourly } O_3]_{\text{combined(indicative city-level)=}}0.013[NO] + 0.003[NO_2] - 0.081[B] - 0.079[Tol] + 0.201[Temp] - 0.870\,[RH] + 0.096\,[SR] + 0.526\,[WS] + 0.074\,[P] + 0.090[T] - 0.189\,[M] + 19.907 \tag{8}$$

$$[\text{Hourly NO}]_{\text{combined(indicative city-level)}} = 0.046[O_3] + 0.598\,[NO_2] + 0.534[B] + 1.251\,[Tol] - 0.784\,[Temp] + 0.948[RH] - 1.539\,[WS] - 0.063[P] - 0.008[SR] - 0.162[T] - 0.332\,[M] - 1.099 \tag{9}$$

$$[\text{Hourly } NO_2]_{\text{combined(indicative city-level)}} = 0.003[O_3] + 0.173\,[NO] + 0.254[B] + 0.176\,[Tol] - 0.103\,[Temp] - 0.260[RH] - 2.901\,[WS] + 0.065[P] - 0.059[SR] + 0.963[T] - 0.095[M] + 20.935 \tag{10}$$

*Units and abbreviation: NO ($\mu g/m^3$), NO_2 ($\mu g/m^3$), NO_x ($\mu g/m^3$), B: Benzene ($\mu g/m^3$), Tol: Toluene ($\mu g/m^3$), Temp: Temperature (°C), RH: Relative humidity (%),SR: Solar radiation (W/m^2), WS: Wind speed (m/s), P: Pressure (bar), T: Time (hour) and M: Month (month unity).

**All the regression equation coefficients with various polynomial degrees for individual sites and indicative city level model for *with* and *without meteorology* have been provided separately with this paper.

Additionally, analysing the observation versus modelled concentrations after testing, it is observed that random forest based O_3 models predict best for a range 0–200 $\mu g/m^3$, while for extremely higher concentrations (~250–300 $\mu g/m^3$) the prediction is not accurate and fidelity of the model diminished (Fig. 9a); for NO (Fig. 9b), most of observations lie between 0 and 400 $\mu g/m^3$; and for NO_2 (Fig. 9c), most of observations lie between (0–200 $\mu g/m^3$). Random forest models satisfactorily even for extremely high concentrations of NO and NO_2. This ill-performance corresponding to elevated concentrations of O_3 might be because “If the data peak goes outside the standard deviation then it is basically beyond the range of the data on which the model was trained, in other words, these peaks are also known as outliers”. The model cannot predict well for the outliers because it does not see much of outliers during the training and does not generalize the parameters (coefficients and bias) accordingly. As already specified in section 2.2, if the features cross the variance then it will be a situation which was not taught to the model while training.

The obtained results for O_3, NO and NO_2 (in the testing phase; shown in Figs. 5, 7 and 8 and Tables 2 and 3) veritably establishs that random forest regression has accomplished admirable correlations for site-specific and indicative city-level relative to linear regression — *with* and *without* meteorology. This shows that the photochemical kinetic model developed using random forest regression has trained better; and hence, best-suited for predicting future concentrations.

3.5. Model performance: forecasting O_3, NO and NO_2

In order to access model's accuracy, the third phase of model development i.e. model performance (or forecast) has been performed for two scenarios. Upon training (development) and testing (validating) the model for 4-year data (2015–18), 1-year prediction (against 2018–19) and 1-month prediction (against January 2019) were executed. The predictions were performed for 4 degrees of polynomial transformation and the degree which had the highest R^2 value (when compared against 2019 observed field data) has been reported as the best forecast.

At combined city level, ozone and NO predictions were found to be better forecasted for one future month (i.e. January 2019; $R^2 = 0.91$ for O_3; $R^2 = 0.70$ for NO) than for full year (i.e. 2019; $R^2 = 0.65$ for O_3; $R^2 = 0.34$ for NO). Contrasting pattern were observed for NO_2 prediction, as it primarly emitted from fossil fuel combustion in Delhi and is less likely dependent on seasonality. Ideally, when indicative city-level model is considered, the January 2019 results (1-month predictions) should be better compared to the annual results (2019) but it is not the case with NO_2 and it could be because of high standard deviation of the moving average of NO_2 concentration in the area. Fig. 10 presents a time series comparision (predicted vs observed) for O_3, NO and NO_2 at the combined city-level for two forecast scenarios (January 2019 and Annual 2019). At site-level, RK Puram had better O_3 prediction for January 2019 while Punjabi bagh showed best forecast for 2019 full-year (Table 4). This observation has been reversed in case of NO prediction. For NO_2 predictions, among site-level forecasts, RK Puram showed the highest R^2 for both the scenarios (1-month and 1-year).

Ozone predictions at Mandir marg were observed to the worst possible (for January 2019; $R^2 = 0.02$). The major reason for the

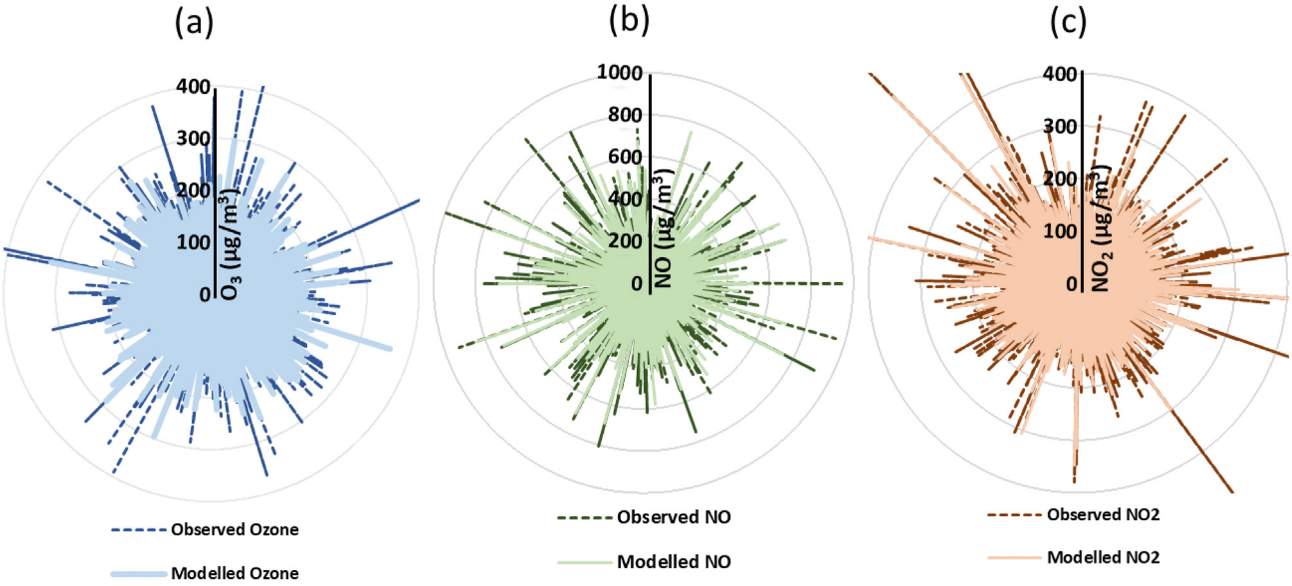


Fig. 9. Polar diagram representing testing performance for random forest *modelled* versus *observed* concentrations of (a) O_3, (b) NO, and (c) NO_2.

poor performance on the January month compared to the entire year might be because "most of O_3 measurements during January month for the selected input duration (2015–18) was missing at Mandir Marg site". This, accompanied with the usage of 'month' as a feature for model training, led to a case where the machine learning algorithm did not learn much for the January month. Also, the January field measurements were missed significantly for 2015 in the training and testing data at Punjabi bagh and RK Puram but there existed consistent data for the next input years (2016, 2017 and 2018). Therefore, site-specific and combined models could learn effectively in these cases.

4. Summary and discussions

We developed and trained two observation-based flexible photochemical kinetic models (linear and random forest regression models) and compared it with testing data, to assess model accuracy. Among the developed models, random forest regression was used to project pollutant concentrations for two scenarios. The training was carried using data adopted from three supersites of Delhi-NCT, where O_3 levels constantly violate the prescribed standards. These algorithms follow an approach of a polynomial transformation of the data (which introduces non-linearity into the dataset) before the application of regression techniques. Both the models were aided with machine learning (to reduce their time-intensity) and were applied to two scales (Site- and indicative city-level).

- In most pollutant-modelling scenarios, site-specific models with meteorology generally perform better compared to a indicative city-level combined model with or without meteorology. Nevertheless, there exist some cases where indicative city-level models were found to better suit and perform than site-specific models, which can be ascribed to high variation in the observed pollutant concentrations from those sites
- While forecasting O_3, R^2 values were observed to be relatively less because of seasonal variations, on the other hand NO and NO_2 models are found to be quite stable with better results which accredit to their stability in the atmosphere. Formation of nitrate and aerosols is hinted to be the reason for the poor performance of NO_2 compared to NO.
- Based on testing results, it can deduced that random forest regression improved O_3 modelling over linear regression with greater acceptability, i.e. correlation of 0.92 for site-specific (RKP) and 0.85 for indicative city-level; reinforcing the census that random forest regression is best suited to evolve models for secondary pollutants.
- NO does not have any effective seasonal variations which result in its persistence (Xu et al., 2018). The stable and accurate prediction of NO can be credited to the more or less equal rate of formation and consumption of NO in the atmosphere. NO_2 may be associated with the formation of nitrate and reactions to form aerosols. NO_2 (lifetime < 1 day) varies with meteorological parameters, photolysis rate and concentration of hydroxide radicals (Sheel et al., 2010). Also, It exhibits enhanced seasonal variations (such as concentration decrease in monsoon and increase in summer). These could be the probable reasons for NO forecast performing slightly better than NO_2.
- Collated parametric data from all selected hotspots was used as input for training the algorithm to develop city-level combined model. This collation combines the characteristics of the hotspot sites and inculcates them into the final developed indicative city-level model. And hence, this particular version of models will only have the features of the 3 chosen sites i.e. predominantly residential characteristics with a fraction of other land-use attributes.
- As the combined model is based on machine learning (on a similar note to site-specific models), it is capable of receiving inputs from any number of sites; making this model extremely flexible, adaptable and pliant. A large (longer temporal datasets) and diverse (larger spatial dataset or data from disparate land-uses) input for training this model results in increased model performance. It has to be admitted that for a production-level combined model, which stakeholders generally use for making judicious decisions, to predict pollutants across the entire city, one should consider data (pollutant concentrations and meteorological

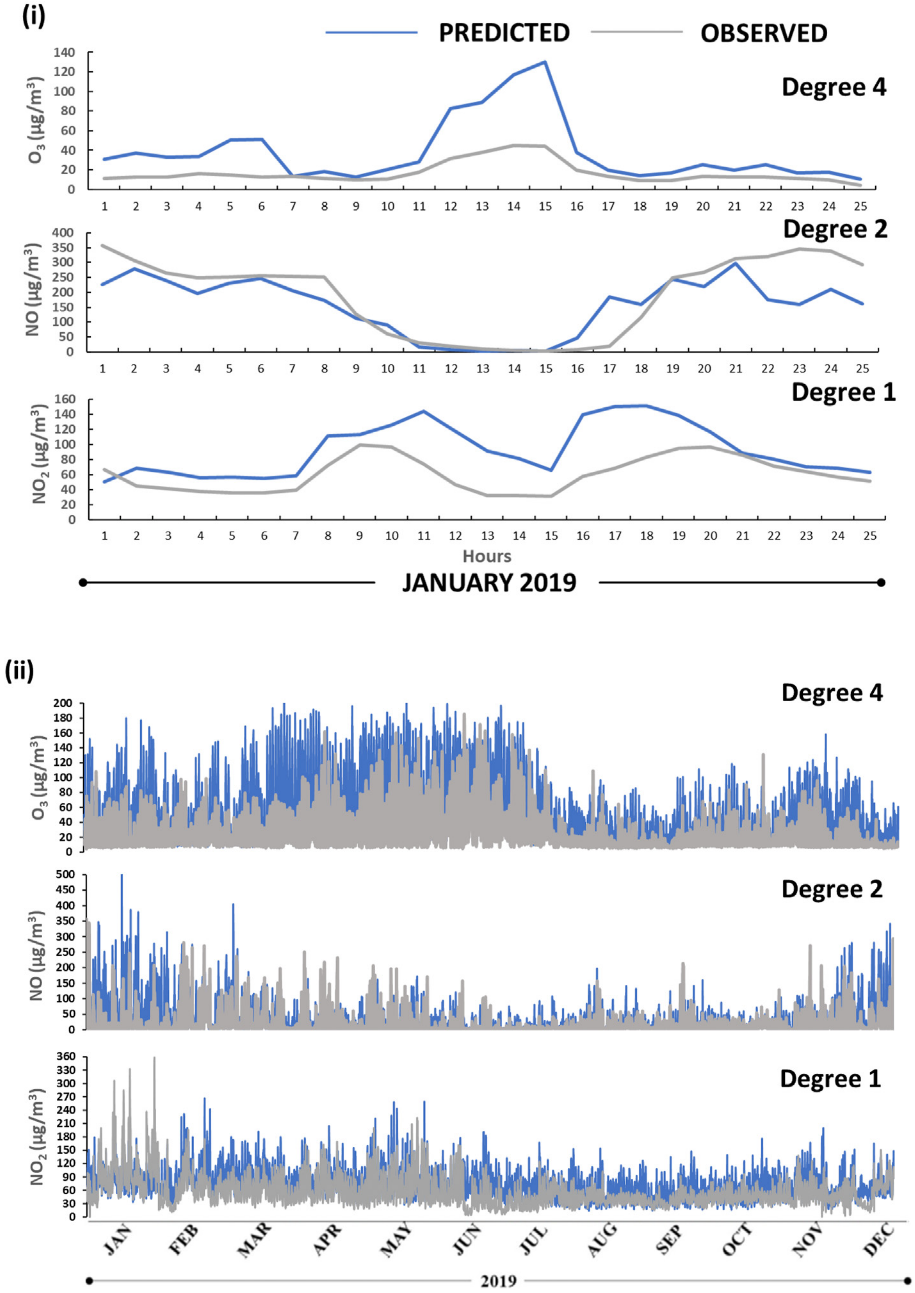


Fig. 10. Random forest regression model performance (predicted versus observed): (i) for one month (January 2019); (ii) for one full year (2019).

parameters) from all possible sites and monitoring stations from diverse land-uses.

- In general, As 1-month prediction of January 2019 was observed to better suited forecast relative to full-year prediction for 2019, it can be inferred that random forest regression should be used for monthly (or shorter) forecasts; while incorporating incremental learning in due course (to inculcate seasonality and better train the algorithm).
- As these models are built with focus on flexibility, they can be replicated for other cities and pollution sites; making them extremely utilitarian and remarkably effective for wide-range implementations. Also, these models are simple and easily understood structures, giving them minimal operational costs and in comparatively less time.
- While performing the polynomial transformation, it has been observed that only lower degree transformation could furnish best results. In most cases for Ozone (site-specific models), 3rd degree polynomial transformation was observed to be optimal. However, for NO and NO_2, both 3rd and 4th degree polynomial transformation have given the most accurate results. In general, polynomial transformation exhibits a comparably smaller error till 3rd or 4th degree, and then the error increases abruptly as the degree increases. The minimum error which can be observed below 5th-degree polynomial transformation is comparatively insignificant compared to the error corresponding to 10th-degree polynomial transformation. As explained by Bishop (2006), this sharp increase in error is because of the overfitting of the model.

Table 4
Performance of random forest regression based model to predict O_3, NO, NO_2 for two scenarios (1-month prediction and 1-year prediction).

Area	Prediction for January 2019			Prediction for 2019 (Full year)		
	R^2	Optimal RMSE after polynomial fitting	Degree with maximum R^2	R^2	Optimal RMSE after polynomial fitting	Degree with maximum R^2
Ozone						
Combined (City-level)	0.9126	30.5931	4	0.6547	30.7189	3
R.K. Puram	0.8648	34.8630	1	0.4832	32.6531	3
Punjabi Bagh	0.5339	31.0502	3	0.6067	27.9837	2
Mandir Marg	0.0212	23.3120	3	0.3059	13.8197	1
NO						
Combined (City-level)	0.7049	78.4754	2	0.3417	41.9877	2
R.K. Puram	0.3991	116.3283	1	0.4530	111.9550	1
Punjabi Bagh	0.5275	84.0075	1	0.2855	38.8895	1
Mandir Marg	0.6219	58.9812	1	0.5682	33.8293	2
NO_2						
Combined (City-level)	0.4028	41.5856	1	0.4404	30.6276	1
R.K. Puram	0.7216	30.1515	1	0.7220	31.6242	1
Punjabi Bagh	0.5657	37.3623	1	0.3818	34.5473	1
Mandir Marg	0.3363	43.8510	3	0.4731	21.9390	1

- The multiple linear regression-based (MLR) PKM used against O_3 prediction in the current study has obtained relatively lower R^2 compared to the MLR formulated by Abdullah et al. (2019). Although the coefficients of determination against O_3 forecast were somewhat low, MLR of the current study is considered reliable and accurate, since the present model uses 11 input features compared to 8 used by Abdullah et al. (2019). Even though the random forest regression employed the present study has performed marginally better than 'the MLR model developed by Abdullah et al. (2019)' and 'random forest regression by Rekha et al. (2018)' in estimating O_3 levels, it performed poorly in comparison to 'the Multivariate Adaptive Regression Splines (MARS) model applied by Rekha et al. (2018)'. Despite using deep convolutional neural networks (CNN) for O_3 prediction, Eslami et al. (2019) have reported similar Pearson correlation coefficient (r) when evaluated against the MLR and random forest regression in the current study. The performance of the random forest regression, in quantifying O_3, was found to be similar against 'the improved auto-regressive (AR) method employed by Zhang et al. (2011)'. Forecasting NO and NO_2 with meteorology using MLR and random forest regression was better performed compared to the random forest model used by Kamińska et al. (2018).

5. Conclusions

Photochemical air pollutants which affect animal and plant well-being are modelled using a set of the flexible site- and indicative city-wide models. These models (linear regression and random forest regression; both assisted with machine learning) were developed to forecast ground level O_3, NO and NO_2, using the data obtained from three highly-polluted supersites of Delhi-NCT. The following conclusions are drawn:

- Integrated meteorological-emission models obtain a better equation between features and labels. Hence, both the meteorological observations and emission concentrations are used in this study. These models can be used for those areas that do not have comprehensive observations to predict the photochemical pollutants while encapsulating the corresponding meteorology with precursor emissions.
- Pollutants such as O_3 are highly sensitive to the seasonality. Since the data used in training the models include meteorological parameters for all the seasons, these models can predict the concentration in any given season. Also, the evaluation was done on testing data, which was generated randomly from all the possible seasons in a year.
- Employment of random forest regressor found solar radiation, NO_2, wind speed and NO to be the most important parameter for accurate O_3 prediction in the heavily polluted environment of Delhi.
- Through this study, it can be concluded that random forest models perform reasonably better than linear regression for predicting the concentration of photochemical pollutants. Additionally, meteorology plays a very important role in the prediction of photochemical pollutants, as it is evident from the comparisons between *with* and *without* meteorological models. These models can be used until the concentration of feature pollutants and meteorological parameters are within the variance.
- The obtained results enable this pragmatic approach (machine learning-assisted regression), not only to forecast short term ozone levels but also in capturing the ozone trends, and expanding the scientific understanding of the mechanisms underlying O_3-precursor-meteorology dynamics.
- The same methodology can be used for other cities and hence these algorithms can play a primary role in building future forecasting models for the pollutants. Consequently, through these models we can even predict the concentration of photochemical pollutants over those areas where we do not have measurement instruments installed.
- The developed regression-based models, accompanied with improved spatial and temporal resolution of input data, are felicitous for the prediction of ozone levels intended either for *early warning systems (EWS)* or *event detection and decision support systems (ED-DSS)*, for maintaining public health as well as for regional authorities to contrive strategies/policies in ameliorating the air quality.
- Current study is more inclined towards the development of a workable model instead of immediate production to real world implementations. Therefore, only prominent hotspot supersites across Delhi (that are infamous for their

consistently high pollution) were chosen and majority of work/emphasis is placed on punctilious development of the model.

Credit author statement

Komal Shukla: Conceptualization, Methodology, Software, Data curation, Formal-analysis, Investigation, Writing- Original draft, preparation Nikhil Dadheech.: Validation and Review Prashant Kumar: Technical Reviewing and Editing Mukesh Khare: Supervision and Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work has been supported by the University grant commission – Junior Research Fellowship [UGC Ref No. 35186/(NET-Dec.2014)] India, and also the Natural Environmental Research Council [grant number NE/P016510/1] through the project – An Integrated Study of Air Pollutant Sources in the Delhi National Capital Region (ASAP-Delhi) – under the UK-India NERC-MOES Programme on Air Quality and Health in Megacity Delhi. We thank Mr. Shiva Kumar Gumadivalli for his help in editing the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemosphere.2021.129611.

Reference

Abdullah, S., Nasir, N.H.A., Ismail, M., Ahmed, A.N., Jarkoni, M.N.K., 2019. Development of ozone prediction model in urban area. Int. J. Innovative Technol. Explor. Eng. 8 (10), 2263–2267.

Abdul-Wahab, 2003. The need for inclusion of environmental education in undergraduate engineering curricula. Int. J. Sustain. High Educ. 4 (2), 126–137. https://doi.org/10.1108/14676370310467140.

Ainsworth, E.A., Yendrek, C.R., Sitch, S., Collins, W.J., Emberson, L.D., 2012. The effects of tropospheric ozone on net primary productivity and implications for climate change. Annu. Rev. Plant Biol. 63 (1), 637–661. https://doi.org/10.1146/annurev-arplant-042110-103829.

Al-Alawi, S.M., Abdul-Wahab, S.A., Bakheit, C.S., 2008. Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone. Environ. Model. Software 23 (4), 396–403. https://doi.org/10.1016/j.envsoft.2006.08.007.

Ali, K., Inamdar, S.R., Beig, G., Ghude, S., Peshin, S., 2012. Surface ozone scenario at Pune and Delhi during the decade of 1990s. Journal of Earth System Science 121 (2), 373–383. https://doi.org/10.1007/s12040-012-0170-1.

Amann, M., Purohit, P., Bhanarkar, A.D., Bertok, I., Borken-Kleefeld, J., Cofala, J., et al., 2017. Managing future air quality in megacities: a case study for Delhi. Atmos. Environ. 161, 99–111. https://doi.org/10.1016/j.atmosenv.2017.04.041.

Archer, K.J., Kimes, R.V., 2008. Empirical characterization of random forest variable importance measures. Comput. Stat. Data Anal. 52 (4), 2249–2260. https://doi.org/10.1016/j.csda.2007.08.015.

Beig, G., Ali, K., 2006. Behavior of boundary layer ozone and its precursors over a great alluvial plain of the world: Indo-Gangetic Plains. Geophys. Res. Lett. 33 (24), L24813. https://doi.org/10.1029/2006GL028352.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. springer. http://users.isr.ist.utl.pt/~wurmd/Livros/school/BishopPatternRecognitionAndMachineLearning-Springer2006.pdf.

Bisht, D.S., Dumka, U.C., Kaskaoutis, D.G., Pipal, A.S., Srivastava, A.K., Soni, V.K., Tiwari, S., 2015. Carbonaceous aerosols and pollutants over Delhi urban environment: temporal evolution, source apportionment and radiative forcing. Sci. Total Environ. 521, 431–445.

Bogdan, G.M., Mozgovoy, M., 2019. October). Towards case-based reasoning with kd trees for a computer game of soccer. In: 2019 IEEE International Conferences on Ubiquitous Computing & Communications (IUCC) and Data Science and Computational Intelligence (DSCI) and Smart Computing, Networking and Services (SmartCNS). IEEE, pp. 570–572.

Breiman, L., 2001. Random forests. Retrieved from Mach. Learn. 45 (1), 5–32. https://link.springer.com/article/10.1023/A:1010933404324.

Census of India, 2011. Provisional Population Totals. Retrieved from. Government of India, New Delhi. http://www.censusindia.gov.in/2011census/PCA/PCA_Highlights/pca_highlights_file/India/Chapter-1.pdf.

Central Pollution Control Board, 2016. Retrieved from. http://cpcb.nic.in/publication-details.php?pid=OA== Last.

Chaloulakou, Archontoula, 2003. Neural network and multiple regression models for PM10 prediction in Athens: a comparative assessment. J. Air Waste Manag. Assoc. 53 (10), 1183–1190.

Chelani, A.B., 2013. Study of extreme CO, NO_2 and O_3 concentrations at a traffic site in Delhi: statistical persistence analysis and source identification. Aerosol and Air Quality Research 13 (1), 377–384. https://doi.org/10.4209/aaqr.2011.10.0163.

Clapp, L.J., Jenkin, M.E., 2001. Analysis of the relationship between ambient levels of O_3, NO_2 and NO as a function of NO_x in the UK. Atmos. Environ. 35 (36), 6391–6405. https://doi.org/10.1016/S1352-2310(01)00378-8.

Coe, H., Harrison, R.M., Lewis, A.C., Kumar, P., Khare, M., Bloss, W.J., Morawska, L., 2015. New directions: air pollution challenges for developing megacities like Delhi. Atmos. Environ. 122, 657–661. https://doi.org/10.1016/j.atmosenv.2015.10.032.

Coyle, M., Smith, R.I., Stedman, J.R., Weston, K.J., Fowler, D., 2002. Quantifying the spatial distribution of surface ozone concentration in the UK. Atmos. Environ. 36 (6), 1013–1024. https://doi.org/10.1016/S1352-2310(01)00303-X.

Davy, R.J., Woods, M.J., Russell, C.J., Coppin, P.A., 2010. Statistical downscaling of wind variability from meteorological fields. Boundary-Layer Meteorol. 135 (1), 161–175. https://doi.org/10.1007/s10546-009-9462-7.

DDA, 2020. Draft Master Plan for Delhi 2021. Delhi Development Authority (DDA), New Delhi. http://www.dda.org.in/planning/draft_master_plans.htm.

De Foy, B., 2018. City-level variations in NOx emissions derived from hourly monitoring data in Chicago. Atmos. Environ. 176, 128–139. https://doi.org/10.1016/j.atmosenv.2017.12.028.

Dumka, U.C., Kaskaoutis, D.G., Tiwari, S., Safai, P.D., Attri, S.D., Soni, V.K., Mihalopoulos, N., 2018. Assessment of biomass burning and fossil fuel contribution to black carbon concentrations in Delhi during winter. Atmos. Environ. 194, 93–109. https://doi.org/10.1016/j.atmosenv.2018.09.033.

Eslami, E., Choi, Y., Lops, Y., Sayeed, A., 2019. A real-time hourly ozone prediction system using deep convolutional neural network. Neural Comput. Appl. 1–15.

Faridi, S., Shamsipour, M., Krzyzanowski, M., Künzli, N., Amini, H., Azimi, F., et al., 2018. Long-term trends and health impact of PM2. 5 and O3 in Tehran, Iran, 2006–2015. Environ. Int. 114, 37–49.

Ganguly, N.D., 2009. Surface ozone pollution during the festival of Diwali, New Delhi, India. Journal Earth Science India 2, 224–229. Retrieved from. http://www.earthscienceindia.info/pdfupload/tech_pdf-40.pdf.

García, I., Rodríquez, J.G., Tenorio, Y.M., 2011. Artificial neural network models for prediction of ozone concentrations in Guadalajara, Mexico. In: Air Quality-Models and Application. Nicolas Mazzeo, pp. 35–52. https://doi.org/10.5772/16839.

Ghude, S.D., Jain, S.L., Arya, B.C., Beig, G., Ahammed, Y.N., Kumar, A., Tyagi, B., 2008. Ozone in ambient air at a tropical megacity, Delhi: characteristics, trends and cumulative ozone exposure indices. J. Atmos. Chem. 60 (3), 237–252. https://doi.org/10.1007/s10874-009-9119-4.

Ghude, S.D., Chate, D.M., Jena, C., Beig, G., Kumar, R., Barth, M.C., Pithani, P., 2016. Premature mortality in India due to PM $_{2.5}$ and ozone exposure. Geophys. Res. Lett. 43 (9), 4650–4658. https://doi.org/10.1002/2016GL068949.

Gioda, A., Oliveira, R.C., Cunha, C.L., Corrêa, S.M., 2018. Understanding ozone formation at two islands of Rio de Janeiro, Brazil. Atmospheric Pollution Research 9 (2), 278–288.

Goel, R., Gani, S., Guttikunda, S.K., Wilson, D., Tiwari, G., 2015. On-road PM2.5 pollution exposure in multiple transport microenvironments in Delhi. Atmos. Environ. 123, 129–138. https://doi.org/10.1016/j.atmosenv.2015.10.037.

Gregorutti, B., Michel, B., Saint-Pierre, P., 2017. Correlation and Variable Importance in Random. https://doi.org/10.1007/s11222-016-9646-1.

Gurjar, B.R., Ravindra, K., Nagpure, A.S., 2016. Air pollution trends over Indian megacities and their local-to-global implications. Atmos. Environ. 142, 475–495. https://doi.org/10.1016/J.ATMOSENV.2016.06.030.

Guttikunda, S.K., Gurjar, B.R., 2012. Role of meteorology in seasonality of air pollution in megacity Delhi, India. Environ. Monit. Assess. 184 (5), 3199–3211. https://doi.org/10.1007/s10661-011-2182-8.

Hazarika, S., Borah, P., Prakash, A., 2019. The assessment of return probability of maximum ozone concentrations in an urban environment of Delhi: a Generalized Extreme Value analysis approach. Atmos. Environ. 202, 53–63. https://doi.org/10.1016/j.atmosenv.2019.01.021.

Health Effects Institute, 2017. State of Global Air 2017. Special Report. Health Effects Institute, Boston, MA.

Hengl, T., Heuvelink, G.B., Kempen, B., Leenaars, J.G., Walsh, M.G., Shepherd, K.D., Tondoh, J.E., 2015. Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions. PloS One 10 (6), e0125814. https://doi.org/10.1371/journal.pone.0125814.

Hu, X., Belle, J.H., Meng, X., Wildani, A., Waller, L.A., Strickland, M.J., Liu, Y., 2017. Estimating PM2. 5 concentrations in the conterminous United States using the random forest approach. Environ. Sci. Technol. 51 (12), 6936–6944. https://doi.org/10.1021/acs.est.7b01210.

Huang, K., Xiao, Q., Meng, X., Geng, G., Wang, Y., Lyapustin, A., et al., 2018. Predicting monthly high-resolution PM2. 5 concentrations with random forest model in the North China Plain. Environ. Pollut. 242, 675–683. https://doi.org/10.1016/j.envpol.2018.07.016.

Hutengs, C., Vohland, M., 2016. Downscaling land surface temperatures at regional scales with random forest regression. Remote Sens. Environ. 178, 127–141. https://doi.org/10.1016/j.rse.2016.03.006.

Jain, S.L., Arya, B.C., Kumar, A., Ghude, S.D., Kulkarni, P.S., 2005. Observational study of surface ozone at New Delhi, India. Int. J. Rem. Sens. 26 (16), 3515–3524. https://doi.org/10.1080/01431160500076616.

Jenkin, M.E., Derwent, R.G., Wallington, T.J., 2017. Photochemical ozone creation potentials for volatile organic compounds: rationalization and estimation. Atmos. Environ. 163, 128–137. https://doi.org/10.1016/J.ATMOSENV.2017.05.024.

Jing, P., O'Brien, T., Streets, D.G., Patel, M., 2016. Relationship of ground-level ozone with synoptic weather conditions in Chicago. Urban Climate 17, 161–175. https://doi.org/10.1016/j.uclim.2016.08.002.

Kamińska, J.A., 2018. The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: a case study in Wrocław. J. Environ. Manag. 217, 164–174. https://doi.org/10.1016/j.jenvman.2018.03.094.

Khedairia, S., Khadir, M.T., 2012. Impact of clustered meteorological parameters on air pollutants concentrations in the region of Annaba, Algeria. Atmos. Res. 113, 89–101. https://doi.org/10.1016/j.atmosres.2012.05.002.

Kheirbek, I., Wheeler, K., Walters, S., Kass, D., Matte, T., 2013. PM 2.5 and ozone health impacts and disparities in New York City: sensitivity to spatial and temporal resolution. Air Quality, Atmosphere & Health 6 (2), 473–486. https://doi.org/10.1007/s11869-012-0185-4.

Kumar, N., Foster, A.D., 2009. Air quality interventions and spatial dynamics of air pollution in Delhi and its surroundings. Int. J. Environ. Waste Manag. 4 (1/2), 85. https://doi.org/10.1504/ijewm.2009.026886.

Kumar, G.S., Sharma, A., Shukla, K., Nema, A.K., 2020. Dynamic programming-based decision-making model for selecting optimal air pollution control technologies for an urban setting. In: Smart Cities—Opportunities and Challenges. Springer, Singapore, pp. 709–729. https://doi.org/10.1007/978-981-15-2545-2_58.

Laber, E., Murtinho, L., 2019. Minimization of Gini impurity: NP-completeness and approximation algorithm via connections with the k-means problem. Electron. Notes Theor. Comput. Sci. 346, 567–576.

Lei, Y., Jeong, J.J., Wang, T., Shu, H.K., Patel, P., Tian, S., et al., 2018. MRI-based pseudo CT synthesis using anatomical signature and alternating random forest with iterative refinement model. J. Med. Imag. 5 (4), 043504.

Li, H., Leung, K.S., Wong, M.H., Ballester, P.J., 2014. Substituting random forest for multiple linear regression improves binding affinity prediction of scoring functions: cyscore as a case study. BMC Bioinf. 15 (1), 291.

Lin, M., Fiore, A.M., Horowitz, L.W., Cooper, O.R., Naik, V., Holloway, J., Wyman, B., 2012. Transport of Asian ozone pollution into surface air over the western United States in spring. J. Geophys. Res.: Atmosphere 117 (D21). https://doi.org/10.1029/2011JD016961.

Liu, Y., Cao, G., Zhao, N., Mulligan, K., Ye, X., 2018. Improve ground-level PM2. 5 concentration mapping using a random forests-based geostatistical approach. Environ. Pollut. 235, 272–282. https://doi.org/10.1016/j.envpol.2017.12.070.

Lu, X., Zhang, L., Liu, X., Gao, M., Zhao, Y., Shao, J., 2018. Lower tropospheric ozone over India and its linkage to the South Asian monsoon. Atmos. Chem. Phys. 18 (5), 3101–3118. https://doi.org/10.5194/acp-18-3101-2018.

Mahapatra, A., 2010. Prediction of daily ground-level ozone concentration maxima over New Delhi. Environ. Monit. Assess. 170 (1–4), 159–170. https://doi.org/10.1007/s10661-009-1223-z.

Mishra, D., Goyal, P., 2016. Environmental technology & innovation neuro-fuzzy approach to forecasting ozone episodes over the urban area of Delhi, India. Environmental Technology & Innovation 5, 83–94. https://doi.org/10.1016/j.eti.2016.01.001.

Mukherjee, A., Agrawal, M., 2016. Pollution response score of tree species in relation to ambient air quality in an urban area. Bull. Environ. Contam. Toxicol. 96 (2), 197–202. https://doi.org/10.1007/s00128-015-1679-1.

Neal, R.M., 2009. Pattern recognition and machine learning. Technometrics 49 (3). https://doi.org/10.1198/tech.2007.s518.

Nicolas, G., Robinson, T.P., Wint, G.W., Conchedda, G., Cinardi, G., Gilbert, M., 2016. Using random forest to improve the downscaling of global livestock census data. PloS One 11 (3), e0150424. https://doi.org/10.1371/journal.pone.0150424.

Notario, A., Bravo, I., Adame, J.A., Díaz-de-Mera, Y., Aranda, A., Rodríguez, A., Rodríguez, D., 2012. Analysis of NO, NO2, NOx, O3 and oxidant (OX= O3+ NO2) levels measured in a metropolitan area in the southwest of Iberian Peninsula. Atmos. Res. 104, 217–226. https://doi.org/10.1016/j.atmosres.2011.10.008.

Ojha, N., Pozzer, A., Rauthe-Schöch, A., Baker, A.K., Yoon, J., Brenninkmeijer, C.A.M., Lelieveld, J., 2016. Ozone and carbon monoxide over India during the summer monsoon: regional emissions and transport. Atmos. Chem. Phys. 16 (5), 3013–3032. https://doi.org/10.5194/acp-16-3013-2016.

Oufdou, h., bellanger, l., bergam, a., khomsi, k., 2018. Evaluation and Comparison of Different Daily Ozone Statistical Prediction Models for the Grand-Casablanca Area.

Özbay, B., Keskin, G.A., Doğruparmak, Ş.Ç., Ayberk, S., 2011. Predicting tropospheric ozone concentrations in different temporal scales by using multilayer perceptron models. Ecol. Inf. 6 (3–4), 242–247. https://doi.org/10.1016/j.ecoinf.2011.03.003.

Pallavi, S., Chirashree, G., 2011. Variation in the concentration of ground level ozone at selected sites in Delhi. Int. J. Environ. Sci. 1 (7), 1899–1911. ISSN 0976 – 4402.

Paoletti, E., De Marco, A., Beddows, D.C.S., Harrison, R.M., Manning, W.J., 2014. Ozone levels in European and USA cities are increasing more than at rural sites, while peak values are decreasing. Environ. Pollut. 192, 295–299. https://doi.org/10.1016/j.envpol.2014.04.040.

Pearce, J.L., Beringer, J., Nicholls, N., Hyndman, R.J., Tapper, N.J., 2011. Quantifying the influence of local meteorology on air quality using generalized additive models. Atmos. Environ. 45 (6), 1328–1336. https://doi.org/10.1016/j.atmosenv.2010.11.051.

Populationu, 2020. Delhi Population. Retrieved 09 May 2020 from http://www.populationu.com/in/delhi-population.

Pozzer, A., Giannadaki, D., Lelieveld, J., Fnais, M., Evans, J.S., 2015. The contribution of outdoor air pollution sources to premature mortality on a global scale. Nature 525 (7569), 367–371. https://doi.org/10.1038/nature15371.

Pusede, S.E., Steiner, A.L., Cohen, R.C., 2015. Temperature and recent trends in the chemistry of continental surface ozone. Chem. Rev. 115 (10), 3898–3918. https://doi.org/10.1021/cr5006815.

Rekha, G., Somula, Ramasubbareddy, Govinda, k, Ellaji, C.H., Jaya Sri, P., 2018. Ozone layer concentration prediction using machine learning techniques. In: 2018 IADS International Conference on Computing, Communications & Data Engineering (CCODE). https://doi.org/10.2139/ssrn.3167810, 7-8 February.

Ruiz-Suarez, J.C., Mayora, O., Smith-Perez, R., Ruiz-Suarez, L.G., 1994. A neural network-based prediction model of ozone for Mexico City. In: Air Pollution, vol. 94. Computational Mechanics Publications, Southampton.

Ruiz-Suarez, J.C., Mayora-Ibarra, O.A., Torres-Jimenez, J., Ruiz-Suarez, L.G., 1995. Short-term ozone forecasting by artificial neural networks. Adv. Eng. Software 23 (3), 143–149. https://doi.org/10.1016/0965-9978(95)00076-3.

Screpanti, A., De Marco, A., 2009. Corrosion on cultural heritage buildings in Italy: a role for ozone? Environ. Pollut. 157 (5), 1513–1520. https://doi.org/10.1016/j.envpol.2008.09.046.

Sharma, S., Khare, M., 2017. Simulating ozone concentrations using precursor emission inventories in Delhi – national Capital Region of India. Atmos. Environ. 151, 117–132. https://doi.org/10.1016/j.atmosenv.2016.12.009.

Sharma, S., Chatani, S., Mahtta, R., Goel, A., Kumar, A., 2016. Sensitivity analysis of ground level ozone in India using WRF-CMAQ models. Atmos. Environ. 131, 29–40. https://doi.org/10.1016/j.atmosenv.2016.01.036.

Sheel, V., Lal, S., Richter, A., Burrows, J.P., 2010. Comparison of satellite observed tropospheric NO2 over India with model simulations. Atmos. Environ. 44 (27), 3314–3321. https://doi.org/10.1016/j.atmosenv.2010.05.043.

Shukla, K., Khare, M., 2019. Behavioural Chemistry of Ground Level Ozone Formation in Heavily Polluted Environment of Delhi City. AGUFM, pp. A21G–A2645.

Shukla, K., Srivastava, P.K., Banerjee, T., Aneja, V.P., 2017. Trend and variability of atmospheric ozone over middle Indo-Gangetic Plain: impacts of seasonality and precursor gases. Environ. Sci. Pollut. Control Ser. 24 (1), 164–179. https://doi.org/10.1007/s11356-016-7738-2.

Shukla, K., Xiaoming, C., Ojha, N., Khare, M., 2018a. Air Quality Simulations over Delhi Using WRF-Chem: Effects of Local Pollution and Regional-Scale Transport, A42A-01 Presented at 2018 Fall Meeting. AGU, Washington, D.C, pp. 10–14. http://abstractsearch.agu.org/meetings/2018/FM/A42A- 01.html.

Shukla, K., Ojha, N., Khare, M., 2018b. Air Quality Simulations over Delhi Using WRF-Chem in Conference of Indian Aerosol Science and Technology Association 2018 "Aerosol Impacts: Human Health to Climate Change"2018. In: http://cas.iitd.ac.in/iasta2018/pdf/E-Proceedings_IASTA-2018.pdf.

Shukla, K., Kumar, P., Mann, G.S., Khare, M., 2020. Mapping spatial distribution of particulate matter using Kriging and Inverse Distance Weighting at supersites of megacity Delhi. Sustainable Cities and Society 54, 101997. https://doi.org/10.1016/j.scs.2019.101997.

Singh, R.P., Chauhan, A., Shaiganfar, R., Sharma, M., Beirle, S., Wagner, T., 2011. Estimation of NOx emissions from Delhi using Car MAX-DOAS observations and comparison with OMI satellite data. Atmos. Chem. Phys. 11 (21), 10871–10887. https://doi.org/10.5194/acp-11-10871-2011.

Sladek, V., Zhu, T., Atluri, S.N., Sladek, J., 2000. The local boundary integral equation (LBIE) and it's meshless implementation for linear elasticity. Comput. Mech. 25 (2–3), 180–198. https://doi.org/10.1007/s004660050468.

Solomon, P.A., Sioutas, C., 2008. Continuous and semicontinuous monitoring techniques for particulate matter mass and chemical components: a synthesis of findings from EPA's Particulate Matter Supersites Program and related studies. J. Air Waste Manag. Assoc. 58 (2), 164–195.

Sousa, S.I.V., Martins, F.G., Alvim-Ferraz, M.C.M., Pereira, M.C., 2007. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. Environ. Model. Software 22 (1), 97–103.

Tao, J., Shen, Z., Zhu, C., Yue, J., Cao, J., Liu, S., Zhang, R., 2012. Seasonal variations and chemical characteristics of sub-micrometer particles (PM 1) in Guangzhou, China. Atmos. Res. 118, 222–231. https://doi.org/10.1016/j.atmosres.2012.06.025.

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res.: Atmosphere 106 (D7), 7183–7192.

Tiwari, S., Bisht, D.S., Srivastava, A.K., Gustafsson, Ö., 2015. Simultaneous measurements of black carbon and PM2.5, CO, and NO$_x$ variability at a locally polluted urban location in India. Nat. Hazards 75 (1), 813–829. https://doi.org/10.1007/s11069-014-1351-9.

Ul-Saufie, A.Z., Yahaya, A.S., Ramli, N.A., Rosaida, N., Hamid, H.A., 2013. Future daily PM10 concentrations prediction by combining regression models and feedforward backpropagation models with principle component analysis (PCA). Atmos. Environ. 77, 621–630.

Vlachogianni, A., Kassomenos, P., Karppinen, A., Karakitsios, S., Kukkonen, J., 2011. Evaluation of a multiple regression model for the forecasting of the concentrations of NO_x and PM_{10} in Athens and Helsinki. Sci. Total Environ. 409 (8), 1559–1571. https://doi.org/10.1016/j.scitotenv.2010.12.040.

Wang, J.F., Hu, M.G., Xu, C.D., Christakos, G., Zhao, Y., 2013. Estimation of citywide air pollution in Beijing. PloS One 8 (1). https://doi.org/10.1371/journal.pone.0053400.

Wang, Y., Wang, H., Guo, H., Lyu, X.P., Cheng, H.R., Ling, Z.H., Blake, D.R., 2017. Long-term O_3-precursor relationships in Hong Kong: field observation and model simulation. Atmos. Chem. Phys.

WHO, 2016. WHO Regional Offices, World Health Organization. http://www.who.int/about/regions/en/.

Xu, W., Liu, C., Shi, K., Liu, Y., 2018. Multifractal detrended cross-correlation analysis on NO, NO2 and O3 concentrations at traffic sites. Phys. Stat. Mech. Appl. 502, 605–612.

Zhan, Y., Luo, Y., Deng, X., Grieneisen, M.L., Zhang, M., Di, B., 2018. Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment. Environ. Pollut. 233, 464–473. https://doi.org/10.1016/j.envpol.2017.10.029.

Zhang, W.Y., Han, T.T., Zhao, Z.B., Zhang, J., Wang, Y.F., 2011. The prediction of surface layer ozone concentration using an improved AR model. In: 2011 International Conference of Information Technology, Computer Engineering and Management Sciences, vol. 1. IEEE, pp. 72–75.

Zheng, J., Zhang, L., Che, W., Zheng, Z., Yin, S., 2009. A highly resolved temporal and spatial air pollutant emission inventory for the Pearl River Delta region, China and its uncertainty assessment. Atmos. Environ. 43 (32), 5112–5122. https://doi.org/10.1016/j.atmosenv.2009.04.060.

Zhong, L., Lee, C.S., Haghighat, F., 2017. Indoor ozone and climate change. Sustainable cities and society 28, 466–472.