



# Applicability of machine learning in modeling of atmospheric particle pollution in Bangladesh

Shihab Ahmad Shahriar<sup>1</sup> · Imrul Kayes<sup>1</sup> · Kamrul Hasan<sup>1</sup> · Mohammed Abdus Salam<sup>1</sup> · Shawan Chowdhury<sup>2</sup>

Received: 16 April 2020 / Accepted: 9 July 2020  
© Springer Nature B.V. 2020

## Abstract

Atmospheric particle pollution causes acute and chronic health effects. Predicting the concentrations of PM<sub>2.5</sub> and PM<sub>10</sub>, therefore, is a prerequisite to avoid the consequences and mitigate the complications. This research utilized the machine learning (ML) models such as linear-support vector machine (L-SVM), medium Gaussian-support vector machine (M-SVM), Gaussian process regression (GPR), artificial neural network (ANN), random forest regression (RFR), and a time series model namely PROPHET. Atmospheric NO<sub>x</sub>, SO<sub>2</sub>, CO, and O<sub>3</sub>, along with meteorological variables from Dhaka, Chattogram, Rajshahi, and Sylhet for the period of 2013 to 2019, were utilized as exploratory variables. Results showed that the overall performance of GPR performed better particularly for Dhaka in predicting the concentration of both PM<sub>2.5</sub> and PM<sub>10</sub> while ANN performed best in case of Chattogram and Sylhet for predicting PM<sub>2.5</sub>. However, in terms of predicting PM<sub>10</sub>, M-SVM and RFR were selected respectively. Therefore, this study recommends utilizing “ensemble learning” models by combining several best models to advance application of ML in predicting pollutants’ concentration in Bangladesh.

**Keywords** Machine learning · SVM · ANN · RFR · GPR · PROPHET · Particulate matter · Bangladesh

## Introduction

Atmospheric particulate matter (PM) pollution, particularly PM<sub>2.5</sub> and PM<sub>10</sub>, poses a severe and growing threat to global public health (Orioli et al. 2018). Exposure to the high concentration of PM has a strong association with different health hazards such as respiratory diseases, cancer, and cardiovascular disease. (Kim et al. 2015). In a clinical meta-analysis, Kim et al. (2015) revealed that about 3% of cardiopulmonary and 5% of lung cancer deaths are attributable to PM exposure globally. The study also argued that the existence of PM in the atmosphere poses more threat to public health than that of other ambient air pollutants. Moreover, a new study revealed

that an increase of 1 g m<sup>-3</sup> in PM<sub>2.5</sub> could accelerate the death rate of the coronavirus disease 2019 (COVID-19) by 15% (Wu et al. 2020). Thus, numerous scientific studies illustrated the strong evidence of the association between health hazards and PM concentration. It occurs, mostly, for the size and composition of the particles. Both particles are constituted by other subclasses of pollutants with the major ones being water-soluble ions, i.e., sulfates, nitrates, ammonium, and minor constituents such as metal ions, organic and elemental carbon, and volatile organics. They can be emitted into the air from natural or anthropogenic sources, and secondary formation in the atmosphere (Lu et al. 2016).

Since air quality is vital for health and the environment, it is essential to regulate proper controlling mechanisms. Pollution modeling can act as a preliminary step of controlling mechanisms (Salnikov and Karatayev 2011). Generally, atmospheric pollution modeling (APM) is defined as the numerical tool that illustrates the casual relationships among emissions, meteorology, atmospheric concentrations, depositions, and other factors (Daly and Zannetti 2007). The APM techniques mainly categorized into three types, i.e., physical model, dispersion model, and machine learning model (Sportisse 2007). However, some other models are broadly implemented in atmospheric sciences, i.e., Gaussian models, Lagrangian models, and Eulerian Models. Commonly, these models work

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s11869-020-00878-8>) contains supplementary material, which is available to authorized users.

✉ Imrul Kayes  
ikayes1@lakeheadu.ca

<sup>1</sup> Department of Environmental Science and Disaster Management, Noakhali Science and Technology University, Noakhali 3814, Bangladesh

<sup>2</sup> School of Biological Sciences, The University of Queensland, Brisbane, Australia

based on continuous emission records and conservation of mass (Gaussian models), the trajectory of air parcel, and wind data (Lagrangian models), and gridded atmospheric properties (Eulerian models). Apart from those, a prognostic model, namely, chemical transport model (CTM) (e.g., GEOS-Chem, CMAQ, WRF-Chem) processes emission, transport, and chemical conversion of trace gases and aerosols simultaneously with meteorological parameters. These models incorporate atmospheric science and multi-processing computational approaches, including the real-time updated emission inventory inputs, and meteorological records (Daly and Zannetti 2007). Unfortunately, the application of these models is further limited by some complexities in terms of geophysical characteristics, i.e., land use and terrain (Jiménez and Dudhia 2013). However, several recent studies argued that the traditional deterministic models struggle to capture the non-linearity among pollutants' concentration, meteorology, land use, and emission and dispersion sources (Shimadera et al. 2016; Chen et al. 2017). To tackle and minimize the limitations of the models, however, machine learning algorithms seem promising (Rybarczyk and Zalakeviciute 2018).

The traditional statistical approaches are limited by describing the variables based on probability and statistical average. In contrast, machine learning models such as artificial neural network (ANN), support vector machine (SVM), and random forest regression (RFR) have been performed as the most popular classifiers to efficiently overcome the non-linear uncertainties and trends to accomplish better forecasting accuracy (Joharestani et al. 2019). Though the models do not unambiguously simulate the environmental process, in general, they exhibit better prognostic performance than the CTMs on spatiotemporal scale in the existence of extensive monitoring records (Marshall et al. 2008). Several studies have been conducted in different countries to evaluate the performance of machine learning models in the field of air quality modeling and forecasting (Kang et al. 2018). However, based on relevant literature, the study of machine learning in air pollution modeling was limited in Bangladesh, though multiple studies were performed to investigate the particulate pollution (Begum et al. 2011; Begum and Hopke 2018). The most used statistical technique to forecast air quality in Bangladesh was Seasonal ARIMA (Islam et al. 2020). Therefore, considering these observations, the study aims to investigate the application of machine learning models, i.e., ANN, L-SVM, M-SVM, Gaussian process regression (GPR), RFR, and a time series model namely PROPHET, on particle pollution modeling in four metropolitan cities in Bangladesh. Among them, ANN, L-SVM, and RFR were used in many air pollution research across the world. However, a limited study found in terms of investigating GPR, and M-SVM machine learning algorithms for pollution modeling (Rybarczyk and Zalakeviciute 2018). A new additive time series model PROPHET, developed by Facebook's Core Data Science

Team, was also implemented in this study to compare with the results of machine learning models. The rationale using this model was its specialty to forecast non-linear trends with yearly, weekly, seasonality, and holiday effects. Besides, the study will demonstrate the regional relationships between  $PM_{2.5}$  and  $PM_{10}$  concentration with meteorological parameters and the other air pollutants, i.e., nitrogen oxides ( $NO_x$ ), sulfur-di-oxide ( $SO_2$ ), carbon-mono-oxide (CO), and ozone ( $O_3$ ), which will be later considered as the exploratory variables to investigate the models.

## Methodology

### Air monitoring stations and data

This study used 24-h air quality ( $PM_{2.5}$ ,  $PM_{10}$ ,  $NO_x$ ,  $SO_2$ , CO,  $O_3$ ) and meteorological data (mean temperature, rainfall, relative humidity, barometric pressure, wind speed, wind direction, and solar radiation) of four air monitoring stations, i.e., Dhaka, Chattogram, Rajshahi, and Sylhet, which were provided by the Department of Environment (DoE), Government of Bangladesh. Among them, Dhaka and Chattogram are ranked 19th and 76th having poor air quality, respectively, in the world (WHO 2018). The first air monitoring station (MS-1) in Dhaka was placed at 23.78° N and 90.36° E and characterized by heavy traffic and transportation. The MS-1 was positioned about 100 m away from the main road, and the height of the roof was approximately 7 m above the ground. The second air monitoring site (MS-2), Chattogram, was located at 22.32° N and 91.80° E. The sampling inlets of MS-2 were positioned on the flat roof of the monitoring site shelter, about 7 m above the ground. Unlike MS-1, the site was a residential area and, therefore, not much influenced by local sources. The location of the third monitoring site (MS-3), Rajshahi, was at 24.38° N and 88.61° E, which was approximately 3 km north from the downtown and 10 m away from a moderate traffic source. The sampler inlet was placed on a flat roof which was 5 m above the ground. The fourth and final monitoring site (MS-4) of this study was situated at Sylhet (24.89° N and 91.87° E). The location of MS-4 was 20 m far from the Kin Bridge of Sylhet and characterized with moderate traffic. The roof height was about 12 m above the ground. For every station, the intake nozzle of the sampler was placed 1.8 m above the roof with proper ventilation. To measure  $PM_{2.5}$  and  $PM_{10}$  concentrations, an automatic and real-time suspended particulate monitor (Beta Gauge 101 M; ENVIRONMENT SA, France) was used. The UV-fluorescence AF22M (TELEDYNE/API, USA), chemiluminescence gas analyzer AC32M, UV photometric ozone analyzer-42M, and dispersive infra-red carbon monoxide analyzer-12M (ENVIRONMENT SA, France) were utilized to measure the

concentration of SO<sub>2</sub>, NO<sub>x</sub>, O<sub>3</sub>, and CO, respectively. To maintain quality assurance and control of the data, calibration was routinely performed. While processing, the data were checked for outliers and if 75% of the data in a day were not available for any parameter due to power failure or equipment's nonoperational, values were considered as non-representative and excluded from the analysis. The amount of total captured data for MS-1, MS-2, MS-3, and MS-4 was 90.4%, 90%, 87.6%, and 86.5%, respectively, from January 2013 to June 2019.

## Data pre-processing

The study performed data pre-processing to maintain the consistency of the raw dataset containing 2372 observations. To process the checking and removal of spatiotemporal outliers from raw data, the *Z* scores method was used before the calculation of statistical parameters, in consistency with previous studies (Barzeghar et al. 2020). Firstly, the series data were transformed into *Z* scores. The observations in the transformed series were rejected from the original series meeting the following conditions: (i) having absolute *Z* score is greater than 4 ( $|Z_i| > 4$ ), (ii) the increment from the previous value of the series is larger than 9 ( $Z_t - Z_{t-1} > 9$ ), and (iii) the ratio of the *Z* score value to its centered mean of order 3 (MA3) being greater than 2 ( $Z_t/MA3(Z_t) > 2$ ). On the other hand, to correct the missing values, the study used the nearest neighbor method (NN) which aims to provide unbiased and valid estimates of associations based on information from the available data. It is also widely known as the standard method to handle missing values in many areas of research (Li et al. 2019). The algorithm is a similarity-based concept that relies on distance metrics. In this work, we used the Minkowski norm (*D*) given by Eq. (1) as a metric to evaluate distance in form of the Euclidean, when  $p = 2$ :

$$D = (\sum_{i=1}^n |x_i - y_i|^p)^{1/p} \quad (1)$$

where  $x_i$  and  $y_i$  are the test sample and training data, respectively.

## Feature selection

A model-free method, i.e., Boruta algorithm (BA), was used in this study to identify the features for the models' prediction. The overall procedure was conducted in six steps, i.e., (a) creating duplicate copies of predictors, i.e., meteorological parameters and pollutants; (b) performing the random shuffle original predictors and duplicate copies of predictors to remove their correlation with the outcomes; (c) applying RFA to find out the most important predictors based on the higher mean values; (d) estimating the *Z* score by using mean and standard deviation; (e) finding the  $Z_{\max}$  score on duplicates

predictors; (f) repeating the above steps till iteration completes for all the air monitoring stations separately.

## PROPHET

PROPHET implements the decomposition of the time series with three main components which are seasonality, overall trends, and holidays (Papacharalampous and Tyralis 2018). The Eq. (2) is a result of those three components:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \quad (2)$$

where  $g(t)$  = the trend function which models non-periodic changes in the value of the time series,  $s(t)$  = periodic changes (e.g., weekly and yearly seasonality), and  $h(t)$  = the effects of holidays which occur on potentially irregular schedules over one or more days;  $\epsilon_t$  = any idiosyncratic changes which are not accommodated by the model.

## Artificial neural network

ANN models are based on the interactions of neurons by transferring signals to another one along with weighted connections (Feng et al. 2015). Besides, in ANN model systems, each neuron is coupled with all preceding neurons and following the layers by links (Bai et al. 2016). In the input layer, every input value is regarded as a neuron. For the success of ANN, all input values should be weighted firstly, and then, the weighted values are processed into the hidden layers. In that layer, every neuron produces output values. The Eq. (3) calculates the outputs:

$$o = f\left(\sum_{j=0}^n w_{ji}x_j\right) \quad (3)$$

where  $f$  = non-linear function,  $x_j$  =  $j$ th input to the neuron,  $w_j$  =  $j$ th synaptic weight, and  $n$  = the number of inputs (Gomez-Sanchis et al. 2006). All data sets need to be normalized before the experimentation of the ANN model. This is the basic method of artificial studies. The equation of normalization (Eq. 4) is given below:

$$NI_{ij} = \frac{I_{(i,j)} - \min(j)}{\max(j) - \min(j)} \quad (4)$$

where  $I$  = input value,  $NI$  = standardized value,  $i$  = number of patterns, and  $j$  = value of variables. In this study, multilayer perceptrons (MLP) was used as it is the most classical type of ANN. After experimenting on several MLP structures, the study decided to utilize two hidden layers.

## Random forest regression

As the RF model is a supervised learning algorithm, three user-defined properties should be determined in the application of this model. They are the number of predictors that are

used to make each tree (mtry), the number of trees in the forest (ntree), and the minimum number of terminal nodes (nodesize). Three user-defined parameters should be determined in RF modeling, which are the number of variables used to grow each tree (mtry) that creates the strength of the trees in the forest and establishes the correlations among them, the number of trees in the forest (ntree), and the minimum number of terminal nodes (nodesize) which should be set to fit an RF. The predictive performance of the RF model, however, is enhanced by increasing the tree strength and decreasing the number of correlations among trees (Brokamp et al. 2017). Firstly, the  $n$  number of training sample subset is expressed as  $D_1, D_2, \dots, D_n$  from total training data set  $D$  using BS. Secondly, based on the subsets,  $n$  number of regression trees is created. Accordingly,  $n$  number of regression result is obtained. Finally, the optimal output is established by aggregating the results of each regression trees. In this study, we selected mtry = 5 and ntree = 500 using BS from the input data. Therefore, regression trees grew based on training data for each of one-fourth the total samples.

### Gaussian process regression

Gaussian process models (GPM) are probabilistic and non-parametric in nature which generally works on the basic principles of Bayesian probability. In this study, exponential GPR was experimented. The task of GPR was to infer a mapping from a set of “ $D$ ” dimensional regression vectors denoted by the regression matrix  $X = [x_1, x_2, \dots, x_n]^T$  to a vector of output data  $y$ . This denotes:

$$D = \{x_{ij}, y_i\}_{i=1}^N; x_i \in X; y_i \in \mathbb{R} \quad (5)$$

The outputs ( $y_1, y_2, \dots, y_n$ ) are usually assumed to be noisy realizations of the underlying function  $f(x_i)$ . A GP model assumes that the output is a realization of a GP with a joint probability density function:

$$f = N(m, k); m : X \rightarrow \mathbb{R}; k : C(x_i, x_j) \quad (6)$$

where  $m$  = mean function and  $k$  = covariant function. Generally, the GP model assumes that the output is a realization of a GP (here, noted as  $N$  in Eq. 6) with a joint probability density function with the mean covariance being functions of the inputs. Usually, the  $m$  is defined as 0 and  $k$  defines the characteristics of the process to be modeled. To make the predictions, the study used the posterior and the marginal likelihood for selecting hyperparameters. The posterior predictive distribution is expressed as the following equation:

$$p(y|x) = \iint p(y|f, X, \theta), p(f|\theta), p(\theta), df, d\theta \quad (7)$$

The interest of the study was the log marginal likelihood, as the quality of its approximation and the posterior

approximation of the study was linked:

$$l(\theta) = -\frac{1}{2} \log(|k|) - \frac{1}{2} y^T k^{-1} y - \frac{N}{2} \log(2\pi) \quad (8)$$

### Support vector machine

In this study, the predicted concentration of particulate matter was followed by following SVM operated formula:

$$\begin{aligned} \hat{Y}_0 &= \sum_{i=1}^n K(X_i, X_0) (\alpha_i - \alpha_i^*); K(X_i, X_0) \\ &= \exp(-|X_i - X_j|)^2 / Y \end{aligned} \quad (9)$$

where  $\alpha_i$  and  $\alpha_i^*$  = support vectors and  $K(X_i, X_0)$  = radial basis kernel function. The practice of an appropriate kernel function (KF) is one of the main features in SVM applications since SVMs are characterized by the usage of KF. It provides the capability of representing non-linear data in the input spaces that in essence are linear; then, an optimization procedure can be applied as in the linear case. This delivers a means to dimension the problem properly; however, the results still depend on the good selection of a set of training datasets. The Gaussian kernel function (GKF) is defined as-

$$K(X_i, X_0) = \exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (10)$$

The GKF provides an estimate for the consistency of the forecast in the form of the variance of the predictive distribution and the analysis can be used to estimate the evidence in favor of a particular choice of the covariance function. The covariance or kernel function can be seen as a model of the data, thus providing a principled method for model selection (Singh et al. 2013). In this study, SVM in linear SVM (l-SVM) and medium Gaussian SVM (mG-SVM) was used.

### Evaluation metrics

The performance of the models was evaluated based on the coefficient of determination ( $R^2$ ) (Eq. 12), root mean square error (RMSE) (Eq. 13), and mean absolute error (MAE) (Eq. 14)

$$R^2 = \frac{\sum_{i=1}^n (y_i - y_{\text{mean}})^2}{\sum_{i=1}^n (x_i - x_{\text{mean}})^2} \quad (12)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n ((x_i - y_i)^2)}{n}} \quad (13)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |x_i - y_i| \quad (14)$$



where  $x_i$  =  $i$ th observed value, ( $i = 1, 2, 3, \dots, n$ ),  $x_{\text{mean}}$  = mean of observed value,  $y_i$  =  $i$ th simulated value, and  $y_{\text{mean}}$  = mean of simulated value.

## Results and discussions

### Data overview

The descriptive statistics for air pollutants and meteorological variables are presented in Table 1. Among the four stations, the highest annual mean concentrations of  $\text{PM}_{10}$  were observed in Dhaka ( $160.4 \mu\text{g m}^{-3}$ ) and the lowest was observed in Rajshahi ( $109.4 \mu\text{g m}^{-3}$ ). Like  $\text{PM}_{10}$ , the highest concentration of  $\text{PM}_{2.5}$  belonged to Dhaka ( $90.5 \mu\text{g m}^{-3}$ ) and the lowest concentration was in Rajshahi ( $67.1 \mu\text{g m}^{-3}$ ). The annual mean concentrations of  $\text{PM}_{2.5}$  and  $\text{PM}_{10}$  in Chattogram were  $70.6 \mu\text{g m}^{-3}$  and  $133.2 \mu\text{g m}^{-3}$  which was the 2nd highest among the stations. The seasonal and annual patterns of  $\text{PM}_{2.5}$  and  $\text{PM}_{10}$  were illustrated with the comparison of WHO air quality standards in Fig. S6. It demonstrated that annual averages of the  $\text{PM}_{2.5}$  and  $\text{PM}_{10}$  concentration in the air of the Dhaka, Chattogram, Rajshahi, and Sylhet are greater than the standards of WHO. In Dhaka, it is about six times greater than the standard. Moreover, the annual PM concentration of stations surpassed the value of Bangladesh Air Quality Standard (BNAQS). The standard value of annual  $\text{PM}_{2.5}$  and  $\text{PM}_{10}$  according to BNAQS is  $15 \mu\text{g m}^{-3}$  and  $50 \mu\text{g m}^{-3}$  respectively (Table S5).

### Seasonal variation of PM

The overall statistics of seasonal meteorological patterns across the stations are represented in supplementary Tables S1, S2, S3, and S4. It was cleared that the concentration of  $\text{PM}_{10}$  and  $\text{PM}_{2.5}$  across the stations was highest in winter whereas it was lowest in the monsoon season. The fluctuation pattern with the seasonal variation throughout the stations was almost the same. In winter, the highest mean concentration of  $\text{PM}_{2.5}$  and  $\text{PM}_{10}$  was observed in Dhaka ( $186.6 \mu\text{g m}^{-3}$  and  $284.9 \mu\text{g m}^{-3}$  respectively) and the lowest in Sylhet ( $146.0 \mu\text{g m}^{-3}$ ) and Rajshahi ( $207.7 \mu\text{g m}^{-3}$ ) respectively. However, in monsoon, it was found the highest  $\text{PM}_{2.5}$  concentrations in Rajshahi ( $31.4 \mu\text{g m}^{-3}$ ) and lowest in Sylhet ( $26.4 \mu\text{g m}^{-3}$ ). From the above statistics, it is clear that there is a relation among the particulate matters and meteorological variables of the seasons. Like this study, Manju et al. (2018) found the similar relationship among the meteorological parameters and air pollutants in India. The correlation among the meteorological parameters and the particulate matters of the study is illustrated in Fig. S4 and Fig. S5. Begum et al. (2011) revealed that brick kilns are responsible for the highest concentration of PM in Dhaka during winter as northwestern wind transports the PMs from the brick kilns located in Dhaka. However, a positive correlation was found between PMs and temperature during the monsoon season in Dhaka. It can be addressed by high summer temperatures after the rainfall in that season. Besides, the combined effect of high wind and temperature can accelerate the concentration (Kayes et al. 2019). Apart from the significant correlation with meteorological parameters, PM was also highly correlated with other gaseous air

**Table 1** Annual mean ( $\mu$ )  $\pm$  standard error (SE), and standard deviation (SD) of the air pollutants' concentration and meteorological parameters across the four air monitoring stations in Bangladesh during 2013–2019

Variables	Unit	Dhaka		Chattogram		Rajshahi		Khulna	
		$\mu \pm \text{SE}$	SD	$\mu \pm \text{SE}$	SD	$\mu \pm \text{SE}$	SD	$\mu \pm \text{SE}$	SD
$\text{PM}_{2.5}$	$\mu\text{g m}^{-3}$	$90.5 \pm 1.4$	69.2	$70.6 \pm 1.1$	51.7	$67.1 \pm 1.2$	57.8	$68.6 \pm 1.2$	56.8
$\text{PM}_{10}$	$\mu\text{g m}^{-3}$	$160.4 \pm 2.3$	110.8	$133.2 \pm 1.8$	88.4	$109.4 \pm 1.6$	79.5	$109.5 \pm 1.7$	80.0
$\text{NO}_x$	ppb	$131.2 \pm 2.1$	102.1	$127.2 \pm 2.1$	102.1	$117.5 \pm 1.7$	100.2	$120.6 \pm 1.8$	101.5
$\text{SO}_2$	ppb	$23.3 \pm 0.5$	22.3	$21.3 \pm 0.5$	19.7	$20.7 \pm 0.5$	18.9	$21.9 \pm 0.6$	19.1
CO	ppm	$2.04 \pm 0.02$	0.98	$1.54 \pm 0.02$	0.97	$1.22 \pm 0.02$	0.72	$1.42 \pm 0.02$	0.89
$\text{O}_3$	ppb	$15.8 \pm 0.2$	11.4	$13.8 \pm 0.2$	11.4	$12.7 \pm 0.2$	10.7	$13.1 \pm 0.3$	10.8
Temp	$^{\circ}\text{C}$	$26.1 \pm 0.1$	4.5	$25.8 \pm 0.1$	3.1	$27.5 \pm 0.1$	3.9	$27.5 \pm 0.1$	3.9
RH	%	$68.7 \pm 0.3$	10.9	$71.8 \pm 0.2$	11.5	$78.1 \pm 0.2$	8.6	$78.2 \pm 0.2$	8.7
R	mm	$0.7 \pm 0.00$	2.1	$0.2 \pm 0.0$	1.7	$0.4 \pm 0.0$	1.2	$0.7 \pm 0.00$	1.7
WS	$\text{ms}^{-1}$	$3.0 \pm 0.1$	3.7	$2.1 \pm 0.0$	0.8	$1.0 \pm 0.0$	0.7	$0.7 \pm 0.0$	0.5
WD	o	$177 \pm 1.0$	50.3	$159.3 \pm 0.8$	40.5	$179.2 \pm 1.3$	63.6	$177.7 \pm 1.3$	62.5
SR	$\text{Wm}^{-2}$	$210.9 \pm 2.9$	140.1	$176.4 \pm 1.3$	63.9	$169.0 \pm 1.3$	61.2	$168.8 \pm 1.3$	61.8
BP	mb	$1006.7 \pm 0.1$	6.3	$1009.1 \pm 0.1$	4.8	$1008.4 \pm 0.1$	6.0	$1008.3 \pm 0.1$	6.0

Temp mean temperature, RH relative humidity, R rainfall, WS wind speed, WD wind direction, SR solar radiation, BP barometric pressure

pollutants (Fig. S5). In Dhaka,  $PM_{2.5}$  was significantly correlated with  $PM_{10}$ ,  $CO$ ,  $NO_x$ , and  $SO_2$ , especially during March, April, May, and June of the year. At that time, for  $PM_{2.5}$  and  $PM_{10}$ , the highest correlation was found with  $CO$  because of the on-road traffic congestion. During the pre-monsoon and post-monsoon period, the  $SO_2$  was found highly correlated with  $NO_x$  in every air monitoring stations. During that period, for the  $NO_x$  emission, traffic was not the only significant source, but rather a substantial amount of  $NO_x$  was emitted to the atmosphere from the main source of  $SO_2$  emissions. Begum et al. (2011) revealed that the main source of  $SO_2$  emissions in Dhaka is brick kilns.

## Model execution

Before the model execution, data splitting was carried out by dividing data into two subsets, i.e., training data and testing data. The study used unsupervised 5-fold cross-validation for data splitting. In each fold, the dataset is separated into two training sets (75% of total data) and a remaining 25% as a hold-out test set utilized to evaluate the performance of the models after the training process. After dividing the data sets, the study used BA to select the most important variables before running the models. The results showed that both for  $PM_{2.5}$  and  $PM_{10}$ , temperature, RH, BP, and WD were the most important predictors among the meteorological variables across the stations (Fig. S7). For  $PM_{10}$  prediction, the relative importance scores of the most significant meteorological predictor, Temp, in Dhaka (MS-1), Chattogram (MS-2), Rajshahi (MS-3) and Sylhet (MS-4) were 13.7, 13.8, 14.27, and 14.08, respectively. The next important meteorological variable was RH, for the stations (12.8, 12.9, 13.6, and 12.11 for MS-1, MS-2, and MS-3, respectively). However, unlike  $PM_{10}$ , the most influential meteorological predictor for  $PM_{2.5}$  prediction in MS-1, MS-2, MS-3, and MS-4 was BP (15.12, 14.83, 14.7, and 14.74 for MS-1, MS-2, MS-3, and MS-4 respectively). On the other hand, among the chemical species fed into the models to predict  $PM_{2.5}$  and  $PM_{10}$ ,  $SO_2$  and  $NO_x$  were the most influential exploratory variables. In terms of  $PM_{2.5}$  prediction, the relative importance scores of  $SO_2$  were 12.58, 12.55, 12.34, and 12.35 in MS-1, MS-2, MS-3, and MS-4 respectively.

## PROPHET

The study used a time series model, i.e., PROPHET, to compare it with other machine learning models. Comparatively, PROPHET did not perform better than the machine learning models for the prediction of  $PM_{2.5}$  and  $PM_{10}$ . In terms of  $PM_{2.5}$  prediction, the  $R^2$  values of the PROPHET for MS-1, MS-2, MS-3, and MS-4 were 0.72, 0.74, 0.78, and 0.75 respectively. It showed poor results in  $PM_{10}$  prediction also. In particular, the model performs worst in MS-1 for  $PM_{10}$  prediction. Samal et al. (2019) and Ye (2019) used this time series

model for predicting air pollutants in India and China respectively. The performance of this model used in Samal et al. (2019) was better than this study. The RMSE value ( $= 3.54 \mu g m^{-3}$ ) of this model was much satisfactory for SPM simulation. Unlike this study, Ye (2019) found the RMSE values for  $PM_{2.5}$  and  $PM_{10}$  prediction were  $10.34 \mu g m^{-3}$  and  $15.5 \mu g m^{-3}$  respectively.

## L-SVM and M-SVM

From Table 2, it is clear that both SVM models are comparable in terms of prediction metrics and they showed good performances in the prediction of daily mean  $PM_{2.5}$  and  $PM_{10}$  concentrations. In particular, M-SVM performed better than L-SVM. During the training period, M-SVM showed higher  $R^2$  and lower RMSE values than the L-SVM. For  $PM_{2.5}$  prediction of MS-1, MS-2, MS-3, and MS-4, the RMSE values of the M-SVM were  $8.89 \mu g m^{-3}$ ,  $10.6 \mu g m^{-3}$ ,  $9.89 \mu g m^{-3}$ , and  $10.2 \mu g m^{-3}$ , respectively, whereas for L-SVM, they were  $8.57 \mu g m^{-3}$ ,  $12.3 \mu g m^{-3}$ ,  $10.7 \mu g m^{-3}$ , and  $10 \mu g m^{-3}$ , respectively. Like  $PM_{2.5}$ , M-SVM showed better performance for  $PM_{10}$  prediction. In terms of  $R^2$  value, for  $PM_{10}$  prediction, among the MS-1, MS-2, MS-3, and MS-4, the highest value was experimented in MS-3 (L-SVM = 0.93 and M-SVM = 0.94) for both models. The lowest MAE value ( $= 4.87 \mu g m^{-3}$ ) was found in MS-1 during the M-SVM execution. Over-fitting was controlled in this study during the model execution. Generally, over-fitting occurs when the results of testing are greater than the validation (Mehdipour et al. 2018). Singh et al. (2013) used an SVM model for predicting urban air quality in India where the RMSE value was  $9.14 \mu g m^{-3}$  and  $9.22 \mu g m^{-3}$  during testing and training period, respectively. However, in Tehran, Mehdipour et al. (2018) experimented much lower RMSE value ( $0.0501 \mu g m^{-3}$  and  $0.519 \mu g m^{-3}$  during testing and training respectively) using SVM models to predict  $PM_{2.5}$  concentration.

## ANN

For the MS-2 and MS-4, the best prediction result was obtained using ANN. To predict the  $PM_{2.5}$ , the lowest RMSE value ( $= 9.42 \mu g m^{-3}$ ) and MAE ( $= 4.93 \mu g m^{-3}$ ) were found in MS-2 during the test period. On the other hand, for  $PM_{10}$  prediction of MS-1, MS-2, MS-3, and MS-4, the RMSE values of the ANN were  $13.8 \mu g m^{-3}$ ,  $14.2 \mu g m^{-3}$ ,  $14.7 \mu g m^{-3}$ , and  $14.9 \mu g m^{-3}$ , respectively. Özdemir and Taner (2014) used multiple linear regression and ANN to predict the  $PM_{10}$  concentration in Turkey. The accuracy of the back-propagation feed-forward ANN (BPNN) with two hidden layers for the urban and industrial zone was 87% and 49% respectively. Feng et al. (2015) studied ANN with a trajectory model and wavelet model to improve the forecasting performance of

**Table 2** Performance metrics ( $R^2$ , RMSE, and MAE) of the models during testing and training period. Here, MS-1, MS-2, MS-3, and MS-4 are representing Dhaka, Chattogram, Rajshahi, and Sylhet respectively. The values in italics indicate the best results among the models

			PM <sub>2.5</sub>				PM <sub>10</sub>			
			MS-1	MS-2	MS-3	MS-4	MS-1	MS-2	MS-3	MS-4
Training	L-SVM	$R^2$	0.91	0.87	0.93	0.92	0.92	0.87	0.93	0.92
		RMSE	8.57	12.3	10.7	10.0	12.8	13.7	13.6	14.5
		MAE	4.98	7.29	5.72	5.92	7.78	7.63	8.37	9.24
	M-SVM	$R^2$	0.89	0.90	0.94	0.93	0.92	0.89	0.94	0.92
		RMSE	8.89	10.6	9.89	10.2	13.8	12.8	15.2	13.4
		MAE	4.78	5.59	5.01	5.97	8.62	7.54	9.21	8.27
	GPR	$R^2$	0.91	0.92	0.94	0.93	0.93	0.87	0.95	0.92
		RMSE	7.68	11.3	8.82	10.8	12.5	13.2	11.7	13.2
		MAE	3.59	6.87	4.27	6.34	7.61	7.98	6.76	7.65
	RFR	$R^2$	0.92	0.89	0.94	0.93	0.91	0.86	0.94	0.93
		RMSE	9.22	11.6	9.97	9.83	14.0	13.5	12.3	13.6
		MAE	4.35	7.59	5.89	4.74	8.58	8.01	7.24	7.02
	ANN	$R^2$	0.90	0.92	0.93	0.94	0.91	0.87	0.94	0.91
		RMSE	11.9	9.40	12.6	8.82	12.9	13.7	14.2	14.3
		MAE	6.22	4.92	7.24	3.87	7.65	8.01	8.39	7.67
	PROPHET	$R^2$	0.72	0.74	0.78	0.75	0.68	0.59	0.76	0.77
		RMSE	27.3	22.4	23.4	21.8	51.6	44.2	34.1	34.1
		MAE	22.6	18.9	17.9	17.5	42.7	41.8	27.3	27.1
Test	L-SVM	$R^2$	0.86	0.82	0.89	0.87	0.85	0.81	0.87	0.88
		RMSE	9.52	13.1	11.2	10.2	13.1	13.8	13.9	14.7
		MAE	4.94	7.57	5.88	6.31	7.95	7.87	8.39	9.21
	M-SVM	$R^2$	0.88	0.89	0.90	0.87	0.88	0.88	0.89	0.88
		RMSE	9.23	10.7	9.92	10.3	14.1	13.1	15.7	13.9
		MAE	4.87	5.61	5.24	6.07	8.79	7.91	9.71	8.68
	GPR	$R^2$	0.91	0.88	0.92	0.89	0.90	0.83	0.91	0.88
		RMSE	7.68	11.2	8.72	10.2	12.8	13.4	12.1	13.5
		MAE	3.59	6.89	4.17	6.47	7.62	8.04	6.89	7.74
	RFR	$R^2$	0.91	0.82	0.92	0.91	0.90	0.86	0.89	0.91
		RMSE	9.49	11.9	9.89	9.94	14.2	13.6	13.4	13.7
		MAE	4.57	7.82	5.21	4.82	8.62	8.22	7.87	7.57
	ANN	$R^2$	0.87	0.88	0.89	0.90	0.90	0.86	0.89	0.88
		RMSE	11.8	9.42	12.7	8.73	13.8	14.2	14.7	14.9
		MAE	6.52	4.93	7.64	3.98	7.68	8.17	8.62	8.07
	PROPHET	$R^2$	0.54	0.62	0.69	0.68	0.52	0.51	0.54	0.61
		RMSE	30.5	29.3	27.4	27.8	61.6	54.4	47.1	45.1
		MAE	24.3	23.7	21.4	21.3	51.0	42.6	39.5	38.7

PM<sub>2.5</sub> in China. Using the ANN model particularly, they found RMSE values ranged from 28.63–36.78  $\mu\text{g m}^{-3}$ .

## RFR

RFR performed the best results for MS-4 to predict PM<sub>10</sub>. During the testing period, RFR showed high  $R^2$ , and lower RMSE, and MAE values than the L-SVM, M-SVM, and

PROPHET. For PM<sub>2.5</sub> prediction of MS-1, MS-2, MS-3, and MS-4, the RMSE values of the RFR were ranged from 9.2 to 11.6  $\mu\text{g m}^{-3}$  and 9.59 to 11.9  $\mu\text{g m}^{-3}$  during training and testing period respectively. Like PM<sub>2.5</sub>, the performance of RFR was better than the SVMs. In terms of  $R^2$ , RMSE, and MAE value, for PM<sub>10</sub> prediction, among the MS-1, MS-2, MS-3, and MS-4, the highest value was experimented in MS-3 during model execution ( $R^2 = 0.91$ , RMSE = 13.7  $\mu\text{g m}^{-3}$ , MAE = 7.57  $\mu\text{g m}^{-3}$ ).

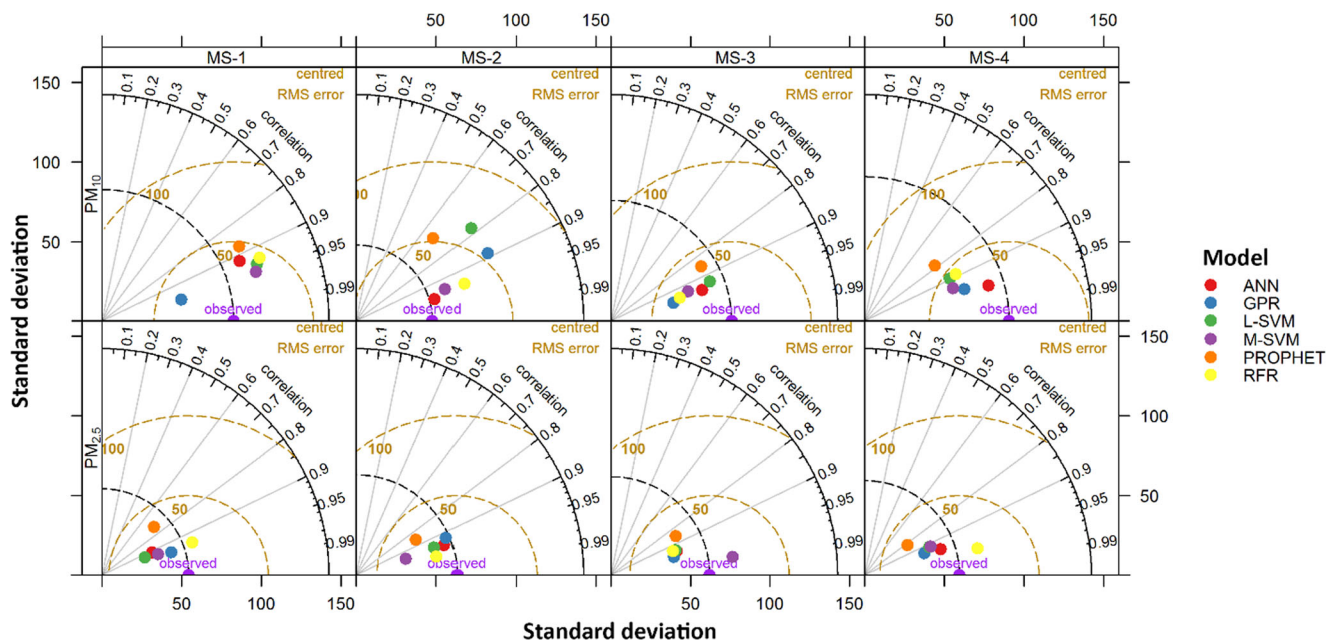


Fig. 1 Taylor diagram of validation results of machine learning models

In China, Hu et al. (2017) experimented with RF and the  $R^2$  value was 0.80 on average. However, Joharestani et al. (2019) used 23 features including meteorological variables, geographic data, and ground measured concentration data to predict  $PM_{2.5}$  in Tehran. In that study, the  $R^2$  and RMSE values varied from 0.66 to 0.78 and 14.47 to 15.30  $\mu g m^{-3}$ , respectively.

## GPR

Among the models in this study, GPR showed the best performance particularly for both  $PM_{2.5}$  and  $PM_{10}$  concentration for

MS-1 ( $R^2 = 0.91$ , RMSE = 7.68  $\mu g m^{-3}$ , MAE = 3.59  $\mu g m^{-3}$  for  $PM_{2.5}$ ;  $R^2 = 0.90$ , RMSE = 12.8  $\mu g m^{-3}$ , MAE = 7.62  $\mu g m^{-3}$  for  $PM_{10}$ ) and MS-3 ( $R^2 = 0.92$ , RMSE = 8.72  $\mu g m^{-3}$ , MAE = 4.17  $\mu g m^{-3}$  for  $PM_{2.5}$ ;  $R^2 = 0.91$ , RMSE = 12.1  $\mu g m^{-3}$ , MAE = 6.89  $\mu g m^{-3}$  for  $PM_{10}$ ) both in training and testing period. During training, the  $R^2$ , RMSE, and MAE values were ranged from 0.91 to 0.94, 7.68 to 11.3  $\mu g m^{-3}$ , and 3.59 to 6.87  $\mu g m^{-3}$  for  $PM_{2.5}$ , respectively, and 0.87 to 0.95, 12.5 to 11.2  $\mu g m^{-3}$ , and 6.76 to 7.61  $\mu g m^{-3}$  for  $PM_{10}$ , respectively. The worst performance of GPR was found for  $PM_{10}$  prediction in MS-2 (Table 2). A study in

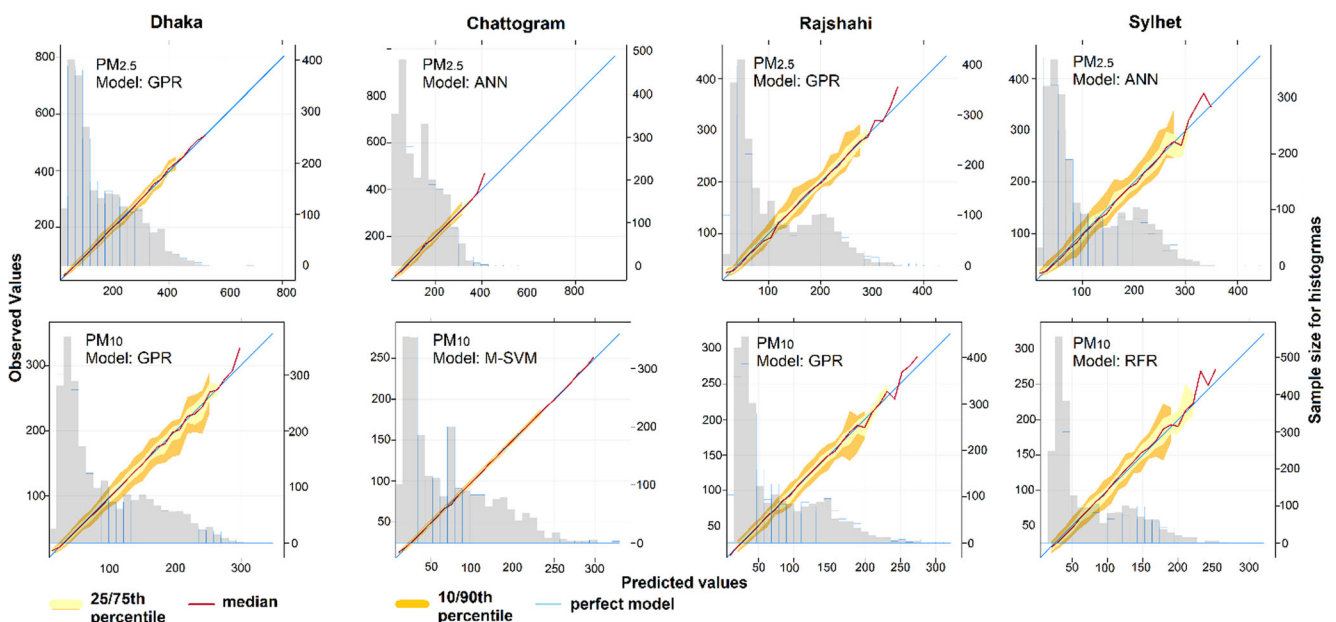


Fig. 2 Proposed models for predicting and monitoring the concentration of  $PM_{2.5}$  and  $PM_{10}$  across the metropolitan areas in Bangladesh



Tehran, i.e., Mehdipour et al. (2018), used Bayesian network to predict  $PM_{2.5}$  where the final RMSE value was 0.1077. Figure 1 represents the overall validation results of the models.

### Model comparison and proposed model

Figure 2 represents the selection of the best model for the monitoring stations. The training results showed that the over-fitting was controlled perfectly in this study. From Fig. 1 and Table 2, it was clear that for  $PM_{2.5}$  and  $PM_{10}$ , the PROPHET time series model performed worse than the machine learning models. However, GPR showed the best performance among all the models, particularly in MS-1 and MS-3. Therefore, GPR was selected as the best model for the prediction of both  $PM_{2.5}$  and  $PM_{10}$  in MS-1 and MS-3. In contrast, ANN well performed only for  $PM_{2.5}$  prediction in MS-2 and MS-4. Unlike  $PM_{2.5}$ , RFR and M-SVM were selected for  $PM_{10}$  for MS-4 and MS-2 respectively. The results of the models are further compared in terms of exploratory variables. Initially, the study used only meteorological variables to predict particulate matters. The initial results of the model validation using only meteorological parameters are presented in the supplementary section (Table S6). Finally, when the chemical species such as  $NO_x$ ,  $SO_2$ ,  $CO$ , and  $O_3$  were fed into the machine learning models, the study found more meaningful results than before. The use of source pollutants in the models decreases the RMSE values for the models. Therefore, in terms of developing the machine learning models, the study recommends the use of more source pollutants and meteorological variables to reveal more fruitful results.

### Conclusion

Machine learning provides reliable forecasting of atmospheric pollution. This research, therefore, explores the application of ML models in the management of PM and air quality in Bangladesh. Five models, i.e., ANN, L-SVM, M-SVM, RFR, and GPR were used in fulfilling the purpose of the study with a comparison of time series model namely PROPHET. Among them, for Dhaka and Rajshahi, GPR showed the best results in terms of  $R^2$ , RMSE, and MAE evaluation metrics. Therefore, the study recommended using GPR to predict the concentration of both  $PM_{2.5}$  and  $PM_{10}$  in those two stations. However, to predict the  $PM_{2.5}$  and  $PM_{10}$  concentration in Chattogram and Sylhet, the study referred to individual models. The proposed model for  $PM_{2.5}$  for Chattogram and Sylhet was ANN, whereas for  $PM_{10}$ , the models were M-SVM and L-SVM respectively. However, the study recommends using data for a longer period to examine the performance of the models. Moreover, the hybrid models could be

an option to compare it with these models. After all, the obtained results from this study revealed that the machine learning can offer convenient information that the government officials and policy makers of different countries can utilize it to issue early alerts of atmospheric pollution incidents and accordingly protect the citizens from exposure.

**Acknowledgments** We would like to thank the Department of Environment, Government of Bangladesh for providing the data and gratefully acknowledge Shajedul Islam from the Department of Environmental Science and Disaster Management, NSTU, for his support in data preparation.

### References

- Bai Y, Li Y, Wang X, Xie J, Li C (2016) Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions. *Atmos Pollut Res* 7(3):557–566. <https://doi.org/10.1016/j.apr.2016.01.004>
- Barzeghar V, Sarbakhsh P, Hassanvand MS et al (2020) Long-term trend of ambient air  $PM_{10}$ ,  $PM_{2.5}$ , and  $O_3$  and their health effects in Tabriz city, Iran, during 2006–2017. *Sustain Cities Soc* 54:101988. <https://doi.org/10.1016/j.scs.2019.101988>
- Begum BA, Hopke PK (2018) Ambient air quality in Dhaka Bangladesh over two decades: impacts of policy on air quality. *Aerosol Air Qual Res* 18:1910–1920. <https://doi.org/10.4209/aaqr.2017.11.0465>
- Begum BA, Biswas SK, Hopke PK (2011) Key issues in controlling air pollutants in Dhaka, Bangladesh. *Atmos Environ* 45(40):7705–7713. <https://doi.org/10.1016/j.atmosenv.2010.10.022>
- Brokamp C, Jandarov R, Rao MB, LeMasters G, Ryan P (2017) Exposure assessment models for elemental components of particulate matter in an urban environment: a comparison of regression and random forest approaches. *Atmos Environ* 151:1–11. <https://doi.org/10.1016/j.atmosenv.2016.11.066>
- Chen J, Chen H, Wu Z, Hu D, Pan JZ (2017) Forecasting smog-related health hazard based on social media and physical sensor. *Infor Syst* 64:281–291. <https://doi.org/10.1016/j.is.2016.03.011>
- Daly A, Zannetti P (2007) Air pollution modeling—an overview. In: Zannetti P (ed) *Ambient air pollution*. The EnviroCopl Institute, California, pp 15–28 <http://home.iitk.ac.in/~anubha/Modeling.pdf>
- Feng X, Li Q, Zhu Y, Hou J, Jin L, Wang J (2015) Artificial neural networks forecasting of  $PM_{2.5}$  pollution using air mass trajectory based geographic model and wavelet transformation. *Atmos Environ* 107:118–128. <https://doi.org/10.1016/j.atmosenv.2015.02.030>
- Gomez-Sanchis J, Martín-Guerrero JD, Soria-Olivas E et al (2006) Neural networks for analysing the relevance of input variables in the prediction of tropospheric ozone concentration. *Atmos Environ* 40(32):6173–6180. <https://doi.org/10.1016/j.atmosenv.2006.04.067>
- Hu X, Belle JH, Meng X, Wildani A, Waller LA, Strickland MJ, Liu Y (2017) Estimating  $PM_{2.5}$  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>
- Islam MM, Sharmin M, Ahmed F (2020) Predicting air quality of Dhaka and Sylhet divisions in Bangladesh: a time series modeling

- approach. *Air Qual Atmos Health* 13:607–615. <https://doi.org/10.1007/s11869-020-00823-9>
- Jiménez PA, Dudhia J (2013) On the ability of the WRF model to reproduce the surface wind direction over complex terrain. *J Appl Meteorol Climatol* 52:1610–1617. <https://doi.org/10.1175/JAMC-D-12-0266.1>
- Joharestani MZ, Cao C, Ni X, Bashir B, Talebiesfandarani S (2019) PM<sub>2.5</sub> prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. *Atmosphere* 10(7):373. <https://doi.org/10.3390/atmos10070373>
- Kang GK, Gao JZ, Chiao S et al (2018) Air quality prediction: big data and machine learning approaches. *Int J Environ Sci Develop* 9(1):8–16. <https://doi.org/10.18178/ijesd.2018.9.1.1066>
- Kayes I, Shahriar SA, Hasan K et al (2019) The relationships between meteorological parameters and air pollutants in an urban environment. *Global J Environ Sci Manag* 5(3):265–278. <https://doi.org/10.22034/gjesm.2019.03.01>
- Kim KH, Kabir E, Kabir S (2015) A review on the human health impact of airborne particulate matter. *Environ Int* 74:136–143. <https://doi.org/10.1016/j.envint.2014.10.005>
- Li C, Wang Z, Li B, Peng ZR, Fu Q (2019) Investigating the relationship between air pollution variation and urban form. *Build Environ* 147:559–568. <https://doi.org/10.1016/j.buildenv.2018.06.038>
- Lu HY, Mwangi JK, Wang LC, Wu YL, Tseng CY, Chang KH (2016) Atmospheric PM<sub>2.5</sub> characteristics and long-term trends in Tainan city, southern Taiwan. *Aerosol Air Qual Res* 16(10):2488–2511. <https://doi.org/10.4209/aaqr.2016.07.0332>
- Manju A, Kalaiselvi K, Dhananjayan V, Palanivel M, Banupriya GS, Vidhya MH, Panjakumar K, Ravichandran B (2018) Spatio-seasonal variation in ambient air pollutants and influence of meteorological factors in Coimbatore, southern India. *Air Qual Atmos Health* 11(10):1179–1189. <https://doi.org/10.1007/s11869-018-0617-x>
- Marshall JD, Nethery E, Brauer M (2008) Within-urban variability in ambient air pollution: comparison of estimation methods. *Atmos Environ* 42:1359–1369. <https://doi.org/10.1016/j.atmosenv.2007.08.012>
- Mehdipour V, Stevenson DS, Memarianfard M, Sihag P (2018) Comparing different methods for statistical modeling of particulate matter in Tehran, Iran. *Air Qual Atmos Health* 11(10):1155–1165. <https://doi.org/10.1007/s11869-018-0615-z>
- Orioli R, Cremona G, Ciancarella L, Solimini AG (2018) Association between PM<sub>10</sub>, PM<sub>2.5</sub>, NO<sub>2</sub>, O<sub>3</sub> and self-reported diabetes in Italy: a cross-sectional, ecological study. *PLoS One* 13(1):e0191112. <https://doi.org/10.1371/journal.pone.0191112>
- Özdemir U, Taner S (2014) Impacts of meteorological factors on PM<sub>10</sub>: artificial neural networks (ANN) and multiple linear regression (MLR) approaches. *Environ Foren* 15(4):329–336. <https://doi.org/10.1080/15275922.2014.950774>
- Papacharalampous GA, Tyralis H (2018) Evaluation of random forests and Prophet for daily streamflow forecasting. *Adv Geosci* 45:201–208. <https://doi.org/10.5194/adgeo-45-201-2018>
- Rybarczyk Y, Zalakeviciute R (2018) Machine learning approaches for outdoor air quality modelling: a systematic review. *Appl Sci* 8(12):2570. <https://doi.org/10.3390/app8122570>
- Salnikov VG, Karatayev MA (2011) The impact of air pollution on human health: focusing on the Rudnyi Altay industrial area. *Am J Environ Sci* 7(3):286–294. <https://doi.org/10.3844/ajessp.2011.286.294>
- Samal KKR, Babu KS, Das SK, Acharaya A (2019) Time series based air pollution forecasting using SARIMA and Prophet model. In proceedings of the 2019 international conference on information technology and computer communications, pp 80–85. <https://doi.org/10.1145/3355402.3355417>
- Shimadera H, Kojima T, Kondo A (2016) Evaluation of air quality model performance for simulating long-range transport and local pollution of PM<sub>2.5</sub> in Japan. *Adv Meteorol* 2016:5694251. <https://doi.org/10.1155/2016/5694251>
- Singh KP, Gupta S, Rai P (2013) Identifying pollution sources and predicting urban air quality using ensemble learning methods. *Atmos Environ* 80:426–437. <https://doi.org/10.1016/j.atmosenv.2013.08.023>
- Sportisse B (2007) A review of current issues in air pollution modeling and simulation. *Comput Geosci* 11:159–181. <https://doi.org/10.1007/s10596-006-9036-4>
- WHO (2018) WHO global ambient air quality database (update 2018). World Health Organization. <https://www.who.int/airpollution/data/cities/en/>
- Wu X, Nethery RC, Sabath BM, Braun D, Dominici F (2020) Exposure to air pollution and COVID-19 mortality in the United States. *medRxiv*. <https://doi.org/10.1101/2020.04.05.20054502>
- Ye Z (2019) Air pollutants prediction in Shenzhen based on ARIMA and Prophet method. In E3S web of conferences, EDP sciences, 136: p05001. <https://doi.org/10.1051/e3sconf/201913605001>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.