Mathematical Biosciences and Engineering

MBE, 18(3): 1992–2009.
DOI: 10.3934/mbe.2021104
Received: 06 December 2020
Accepted: 04 February 2021
Published: 02 March 2021

http://www.aimspress.com/journal/MBE

Research article

A hybrid model for forecasting of particulate matter concentrations based on multiscale characterization and machine learning techniques

Syed Ahsin Ali Shah[1], Wajid Aziz[1,2,*], Majid Almaraashi[2], Malik Sajjad Ahmed Nadeem[1], Nazneen Habib[3] and Seong-O Shim[2]

[1] Department of Computer Science & IT, University of Azad Jammu and Kashmir, King Abdullah Campus, Muzaffarabad 13100, AJK, Pakistan
[2] College of Computer Science & Engineering, University of Jeddah, Jeddah 23890, Saudi Arabia
[3] Department of Sociology & Rural Development, University of Azad Jammu Kashmir, Muzaffarabad 13100, AJK, Pakistan

* **Correspondence:** Email: wloun@uj.edu.sa.

Abstract: Accurate prediction of particulate matter (PM) using time series data is a challenging task. The recent advancements in sensor technology, computing devices, nonlinear computational tools, and machine learning (ML) approaches provide new opportunities for robust prediction of PM concentrations. In this study, we develop a hybrid model for forecasting PM_{10} and $PM_{2.5}$ based on the multiscale characterization and ML techniques. At first, we use the empirical mode decomposition (EMD) algorithm for multiscale characterization of PM_{10} and $PM_{2.5}$ by decomposing the original time series into numerous intrinsic mode functions (IMFs). Different individual ML algorithms such as random forest (RF), support vector regressor (SVR), k-nearest neighbors (kNN), feed forward neural network (FFNN), and AdaBoost are then used to develop EMD-ML models. The air quality time series data from Masfalah air station Makkah, Saudi Arabia are utilized for validating the EMD-ML models, and results are compared with non-hybrid ML models. The PMs (PM_{10} and $PM_{2.5}$) concentrations data of Dehli, India are also utilized for validating the EMD-ML models. The performance of each model is evaluated using root mean square error (RMSE) and mean absolute error (MAE). The average bias in the predictive model is estimated using mean bias error (MBE). Obtained results reveal that EMD-FFNN model provides the lowest error rate for both PM_{10} (RMSE = 12.25 and MAE = 7.43) and $PM_{2.5}$ (RMSE = 4.81 and MAE = 3.02) using Misfalah, Makkah data whereas EMD-kNN model provides the lowest error rate for PM_{10} (RMSE = 20.56 and MAE = 12.87) and EMD-AdaBoost provides the lowest error rate for $PM_{2.5}$ (RMSE = 15.29 and MAE = 9.45) using Dehli, India data. The findings also reveal that EMD-ML models can be effectively used in forecasting PM mass concentrations and to develop rapid air quality warning systems.

Keywords: empirical mode decomposition; forecasting; hybrid forecasting model; machine learning

algorithms; particulate matter

1. Introduction

Atmospheric pollution is continuously increasing due to natural phenomena (volcanic activities, desert storms etc.) and immense anthropogenic (smoke of vehicles, industrial activities, fossil fuels for energy requirements etc.) pollution generating activities [1–3]. Air pollution has both short and long term health hazards. Irritation in the nose, eye, throat, allergic reactions, cough, and upper respiratory infections are examples of short term effects of air pollution. Cardiovascular dysfunctions, respiratory tract infections, and cancer are some of the widely putative long term effects of air pollution [4–6]. These diseases are correlated with millions of deaths globally each year [7,8]. Approximately 7 million people die due to household and environmental air pollution, 94% of which die in low and middle-income countries [9]. The maximum burden of these deaths is observed in South East Asia (2.4 million) followed by Western Pacific (2.2 million) [9].

The impact of particles within the human respiratory system and in the atmosphere is largely governed by their size and generally by their other physical properties. Their size may vary from nanometers to tens of micrometers. Based on their size, particles may be categorized as fine particles ($PM_{2.5}$ having a diameter of 2.5 micrometers (μm) or less) and coarse particles (PM_{10} having a diameter between 2.5 μm and 10 μm). Fine particles may further be categorized into ultrafine/nuclei mode (with a diameter from 0.01 μm to 0.1 μm) and accumulation mode (diameter from 0.1 μm to 1.0 μm). $PM_{2.5}$ is the most hazardous ambient air pollutant for human health [10]. High PM_{10} concentrations can cause premature death in older people with respiratory diseases and heart problems [11].

Air pollutants forecasting is an efficient way of protecting public health, as it provides an early warning against hazardous air pollutants [12]. Forecasting the levels of pollutants may be helpful to minimize the adverse health implications by reducing the exposure of these particles through timely alerts for the general public to take preventive measures. The atmospheric systems are inherently nonlinear, and pollutants are dynamically complex in nature [13], which makes the prediction of atmospheric pollutants a challenging task. The advances in digital electronics, computing, and sensor technologies led to accurate spatio-temporal monitoring and effective forecasting of atmospheric pollutants. Numerous techniques have been developed to forecast PM concentrations such as time series analysis, artificial intelligence (AI), linear or nonlinear regression, and chemical transport models [14]. However, hybrid forecasting models are more accurate and robust when compared to single forecasting models [14]. Chelani and Devotta [15] developed a hybrid model by combining the autoregressive integrated moving average model, which deals with linear patterns. The mass concentration time series data of atmospheric pollutants is an outcome of complex natural and anthropogenic activities evolving with time, which operate on multiple time scales [13]. Shah et al. [13] proposed a hybrid forecasting model based on the multiscale characterization of reconstructed phase space and machine learning (ML) techniques for the prediction of $PM_{2.5}$ and $PM_{10.0}$. Huang et al. [16] proposed empirical mode decomposition (EMD), to address the non-stationary and nonlinear behaviors present in the data which motivates practitioners and researchers to use it as an effective tool. The EMD is based on statistical modeling, which is another technique used for multiscale characterization and forecasting of nonlinear and nonstationary time series data [17–26]. In a study conducted at Xingtai in China, Zhu et al. [27] proposed two EMD based hybrid models (EMD-SVR-Hybrid and EMD-IMFs-Hybrid) to forecast air quality index (AQI) data. They compared the performance of proposed models with single forecasting models based separately on support vector regression (SVR), generalized regression neural

network (GRNN), autoregressive integrated moving average models (ARIMA), EMD-GRNN, Wavelet-SVR, and Wavelet-GRNN. They found that proposed hybrid models were superior and can be used for the forecasting of air pollution. In a study by [28], road traffic prediction was performed using EMD based convolution neural network (CNN) model. The results of the study show that prediction results of EMD based CNN model are more accurate than Lasso-BP, PCA-BP, and standard CNN models. Zhou et al. [29] developed a hybrid model (EEMD-GRENN) by utilizing ensemble EMD in combination with a regression neural network for the forecasting of $PM_{2.5}$ in Xi'an, China. They compared the proposed model (EEMD-GRNN) with ARIMA, principal component regression (PCR), multiple linear regression (MLR), and GRNN and found that the performance of the EEMD-GRNN model was much better than other models. In another study [30] proposed a novel hybrid decomposition and ensemble model by incorporating grey wolf optimizer (GWO), complementary ensemble EMD (CEEMD), and support vector regression (SVR). They compared the results of the proposed model with single AI models, hybrid decomposition ensemble model optimized by using different algorithms, and hybrid decomposition ensemble model with different decomposition methods. They achieved high prediction accuracy for $PM_{2.5}$ concentrations using the proposed model.

In this study, the EMD algorithm is combined with ML algorithms (random forest (RF), support vector regressor (SVR) with linear and radial kernels, k-nearest neighbors (kNN), feed forward neural network (FFNN), and AdaBoost) to develop EMD-ML models (EMD-RF, EMD-SVR-L, EMD-SVR-R, EMD-kNN, EMD-FFNN, and EMD-AdaBoost) to forecast two types of PMs (PM_{10} and $PM_{2.5}$) concentrations. To evaluate and compare the algorithms, monthly PM concentrations (PM_{10} and $PM_{2.5}$) have been predicted. In EMD-ML models, EMD is employed to decompose original PMs time series data into several intrinsic mode functions (IMFs). Then the spearman coefficient correlation is used to select the IMFs having a strong correlation with the original time series and finally, ML algorithms are used to forecast monthly PMs concentrations using selected IMFs. Hourly averaged data from Masfalah air quality monitoring station of duration from January 2014 to September 2015 and hourly averaged data from Dehli city, India of duration from January 2018 to December 2019 have been used. Single forecasting models using simple RF, SVR-L, SVR-R, kNN, FFNN, and AdaBoost algorithms alone are also developed to forecast monthly PM (PM_{10} and $PM_{2.5}$) concentrations of Masfalah air quality monitoring station using input data of pollutants (CO, NO_2, and CO_2) and meteorological parameters (temperature (Temp), wind speed (WS), and relative humidity (RH)). The results indicate that the EMD-ML models outperform the single models. EMD-FFNN model provides the lowest error rate for both PM_{10} (RMSE = 12.25 and MAE = 7.43) and $PM_{2.5}$ (RMSE = 4.81 and MAE = 3.02) using Misfalah, Makkah data whereas EMD-kNN model provides the lowest error rate for PM_{10} (RMSE = 20.56 and MAE = 12.87) and EMD-AdaBoost provides the lowest error rate for $PM_{2.5}$ (RMSE = 15.29 and MAE = 9.45) using Dehli, India data. Therefore, EMD-ML models can be used in forecasting complex time series and to develop rapid air quality warning systems.

The rest of the paper is organized as follows: First, we describe in detail the datasets used in this study along with the EMD-ML models' flowchart and algorithm and other ML algorithms. Then the results of the study are presented and discussed followed by the conclusion section.

2. Materials and method

2.1. Data set

The datasets used in this study were collected from the Masfalah air quality monitoring station (AQMS111) and were previously used by researchers [31]. The monitoring station is situated in the Holy city of Makkah, Saudi Arabia. The reason for selecting the Masfalah site is that it is very near to

the Holy Mosque (Al-Haram), a very busy area surrounded by shops and residential houses. The road near the monitoring station is very busy which emits almost all sorts of air pollutants. High levels of air pollutants pose a potential risk to the local residents, workers, and visitors. Therefore, it is important to monitor air quality in this area and carry out air quality health risk assessment. Hourly data from January 2014 to September 2015 monitored using Aeroqual AQM60 environmental station are used in this study. The data includes air pollutants (nitrogen dioxide (NO_2) (μg/m3), carbon monoxide (CO) (mg/m3), and carbon dioxide (CO_2) (PPM)), particulate matters (PM_{10} (μg/m3) and $PM_{2.5}$ (μg/m3)) and meteorological parameters (temperature (Temp) (°C), wind speed (WS) (m/s) and relative humidity (RH) (%)).

Strict quality assurance and quality control (QA/QC) measures are taken to ensure data quality [31]. The QA measures comprise a selection of monitoring site, correct instrument deployment, instrument selection, design of sample system, and appropriate training of operators. QC is maintained by steps such as calibration of the instrument and its response, routine site visits, monitoring calibration gases, data review, data testing, and authorization.

Missing values and extreme pollutant cases (outliers) have been screened. According to [32] missing data can be handled by modeling the data as a distribution for its estimation, by deletion, and by imputation estimates. If data contains missing values < 5%, then any method can be used for the identification and correction of data [33]. Datasets used in this study contain missing values < 2%, and the deletion method has been used for handling missing data. Outliers present in the data are replaced with the mean value of specific month data. The outliers were identified by computing the z-score. The data values having a z-score greater than 2 standard deviation from the mean position were considered outliers. We use mean for imputing new value to handle extreme pollutant cases.

The second datasets used in this study were obtained from an online source [34] and is collected from Dehli city, India. The datasets contain PMs (PM_{10} and $PM_{2.5}$) concentrations data of duration from January 2018 to December 2019 and are utilized for validating the EMD-ML models.

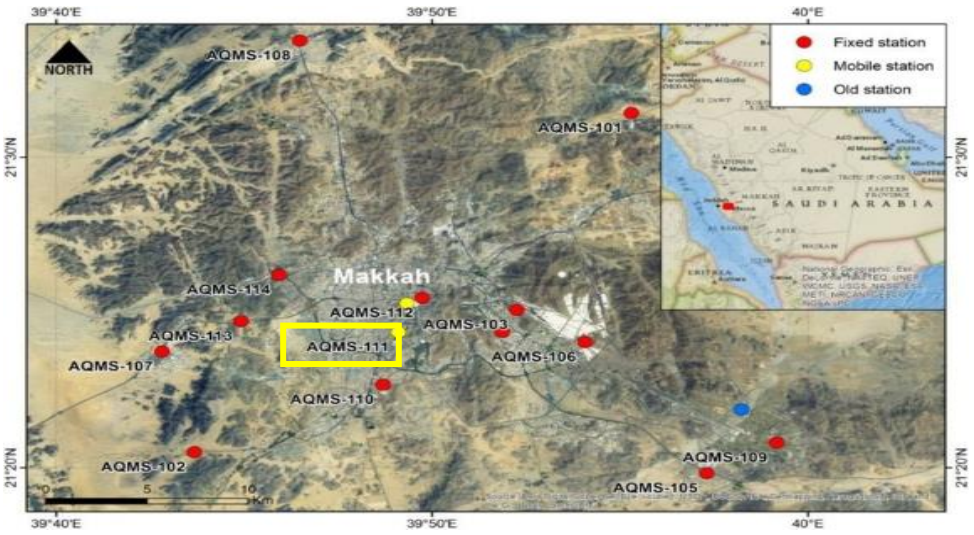

Figure 1. Air quality and meteorological monitoring sites map in Makkah, Saudi Arabia, AQMS 111 represents the site where the data used in this study were collected [35].

2.2. Empirical mode decomposition

Huang et al. [16] proposed the EMD method to decompose non-linear and non-stationary signals

into various IMFs and a residual. Each IMF component of the original signal must satisfy two conditions. (a) The total number of zero-crossing and extrema must be equal or vary at most by one. (b) At all points, the envelope mean value defined by both local minima and local maxima must be zero. The steps involved in the EMD algorithm are as follows.

Step 1: Identify all local minima and maxima of input time series data $X(T)$. By using cubic spline interpolation, generate lower envelop $E_{min(T)}$ using local minima and upper envelope $E_{max(T)}$ using local maxima.

Step 2: Compute the mean of lower and upper envelopes $M(T) = (E_{min(T)} + E_{max(T)})/2$.

Step 3: Compute the candidate IMF $G(T)$ by subtracting envelopes mean $M(T)$ from original input time series data $X(T)$. If $G(T)$ satisfied the above mentioned conditions of IMF, $G(T)$ is considered as ith IMF and residual $R(T)$ is substituted for the original time series data $X(T)$ as $R(T) = X(T) - G(T)$.

Step 4: If candidate IMF $G(T)$ does not meet the above mentioned conditions of IMF, replace the original input time series data $X(T)$ with $G(T)$.

Step 5: Repeat step (1–4) until the residual $R(T)$ becomes a constant value or monotonic function, or there is no more IMF to extract from residual $R(T)$.

2.2.1. Hybrid EMD-ML models

Hybrid EMD-ML models are developed by incorporating traditional EMD, correlated IMFs, and ML algorithms (RF, SVR-L, SVR-R, kNN, FFNN, and AdaBoost) for improved forecasting. For this purpose IMF components (generated through EMD) selected using the spearman correlation coefficient are used to predict each original time series. The whole process in the development of each EMD-ML model (EMD-RF, EMD-SVR-L, EMD-SVR-R, EMD-kNN, EMD-FFNN, and EMD-AdaBoost) is illustrated in Figure 2.

2.3. Learning algorithms

In this section, five learning algorithms used in this study are explained.

2.3.1. Feed-forward neural network (FFNN)

The artificial neural network (ANN) concept is based on a biological neural network of the human brain. The ANN is a computer model used to recognize relations or patterns among data [36]. Two main components of the ANN are a set of nodes and node links.

The feed forward neural network (FFNN) is the simplest form of ANN. In FFNN, data/input flow in one direction only. The FFNN has multiple processing elements (neurons). The neurons are linked to each other through weights. The FFNN comprises of input, hidden, and output layer(s). At the input layer, various input parameters are passed, also the aggregated weighted values are applied to hidden layer neurons. The hidden layer(s) is the intermediated layer between the input and output layers. It performs intermediate calculations. The aggregated weighted values computed at the hidden layer are applied to the output layers. The output layer produces the final output. The output Y obtained (at the output layer) is given as:

$$Y = \omega\{\beta_0 + \sum_{i=1}^{j} \beta_i \, \Phi(\alpha_{i0} + \sum_{k=1}^{l} \alpha_{ik} \, A_k)\} \quad (1)$$

where $(\beta_0, \beta_1, \ldots\ldots\ldots, \beta_j, \alpha_{10}, \ldots\ldots\ldots, \alpha_{jl})$ are weight and bias parameters, respectively. Φ and ω represent the activation functions that are applied at the hidden layer as well as the output layer. A_k are

the input values for each input neuron k. We used 100 neurons in the hidden layer with the logistic activation function to develop the FFNN model.

2.3.2. Adaptive boosting (AdaBoost)

Adaptive boosting (AdaBoost) is the first effective boosting algorithm proposed by [37]. AdaBoost produces weak learners by adjusting each weak learner's weights adaptively. AdaBoost raises the weight of misclassified samples after training a weak learner such that these samples contribute more in the next weak learner training set. The AdaBoost predictions are made by majority voting of the weak learners' outcomes. Therefore, AdaBoost mainly works by generating expanding diversity that can enhance prediction performance.

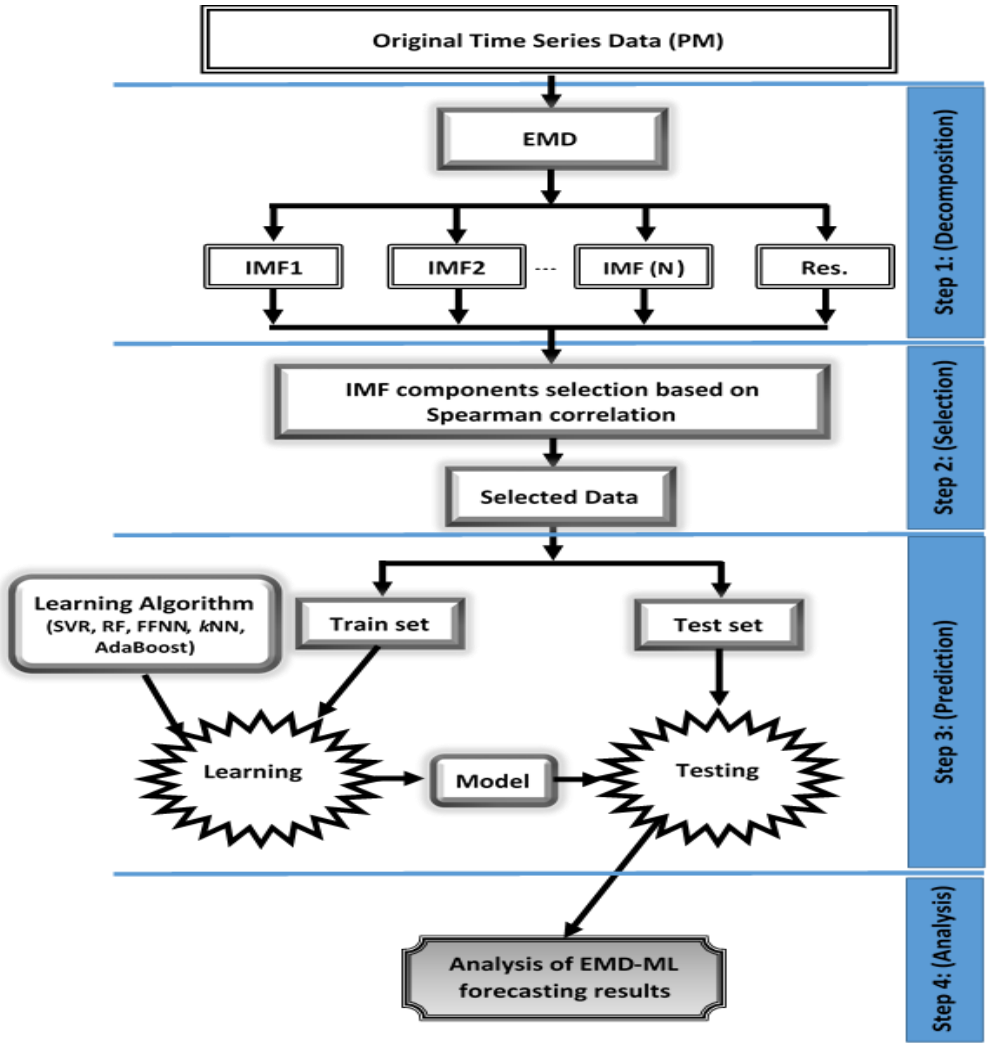

Figure 2. Block diagram of the EMD-ML forecasting model.

2.3.3. Random forest (RF)

Random forest (RF) is a type of ensemble learning algorithm, proposed by [38]. The RF algorithm depends on the classification and regression trees (CART) model. The aim of CART is to learn the relation between a dependent (X) and a series of predictor (Y) variables. The RF algorithm is built on a multitude of decision trees, which are then aggregated into a forest. First, each tree is constructed

according to the bagging method on a random sample of the observations. Secondly, a random collection of features is chosen to separate nodes for each forest tree (feature sampling). Eventually, the trees are aggregated in order to use the model for prediction. This is achieved by averaging the results. In this study 10 number of trees are used to construct RF predictive model.

2.3.4. k-nearest neighbor (kNN)

k-Nearest Neighbor (kNN) [39] algorithm is based on distance function (e.g. Euclidean distance) and is used to classify data with respect to their k nearest neighbor. Let $[(x_1, y_1), \ldots, (x_l, y_l)]$, be a training set, the kNN regression model prediction is defined as $f_{\mathrm{kNN}}(x') = \frac{1}{k}\sum_{i \in Nk(x')} y_i$. $Nk(x')$ contains k-nearest neighbors indices of x'. Bailey and AJ [40] introduces a distance-weighted variant method to smooth down the prediction function by weighting the prediction with the similarity $\Delta(x', x_i)$ of the nearest patterns $x_i$with $i \in Nk(x')$ to the target x' as

$$f_{\mathrm{wk}NN}(x') = \sum_{i \in Nk(x')} \frac{\Delta(x', x_i)}{\sum_{j \in Nk(x')} \Delta(x', x_i)} y_i \tag{2}$$

where the model $f_{\mathrm{wk}NN}$ introduces a continuous output. The contribution of patterns closer to the target in the prediction should be more than other patterns. The similarity in term of the distance between patterns can be defined as:

$$\Delta(x', x_i) = \frac{1}{||x' - x_i||^2} \tag{3}$$

In this study, k = 3 is used to construct the kNN model.

2.3.5. Support vector regressor (SVR)

Let $[(x_1, y_1), \ldots, (x_l, y_l)]$ be a set of training data, where each $x_i \subset R^n$ denotes the input samples along with conforming target value $y_i \subset R$ for $i = 1, \ldots, l$ (l is the size of training data) [41]. The generic form of SVR estimating function is:

$$f\,(x) \;=\; (w \cdot \Phi(x)) \;+\; b \tag{4}$$

In the above equation, $w \subset R^n$, $b \subset R$ and Φ represents the non-linear transformation from R^n to high dimensional space. The objective is to identify the w and b in order to determine the values of x by minimizing the regression risk.

$$R_{reg}\,(f) = C \sum_{i=0}^{l} \Gamma\,(f(x_i) - y_i) + \frac{1}{2}||w||^2 \tag{5}$$

C is a constant, Γ represents a cost function. In terms of data points, vector w can be written as:

$$w = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)\Phi(x_i) \tag{6}$$

The generic equation using Eqs 4 and 6 can be rewritten as:

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)\big(\Phi(x_i).\Phi(x)\big) + b \tag{7}$$

$$= \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) k(x_i, x) + b \tag{8}$$

$k(x_i, x)$ indicates the kernel function.

The dot product in Eq (7) can be replaced with kernel function $k(x_i, x)$. The mathematical representations of kernel functions used in this study are as follows.

$$\text{Linear:} \qquad k(x_1, x_2) = {x_1}^T x_2 \tag{9}$$

$$\text{Radial:} \qquad k(x_1, x_2) = e^{-\gamma||x_1 - x_2||^2} \tag{10}$$

SVR with a linear kernel is termed as SVR-L and SVR with the radial kernel is termed as SVR-R.

2.4. Evaluation measures

The root mean square error (RMSE) and mean absolute error (MAE) is the most commonly used measures for evaluating the performance of predictive models. The range of both measures is from 0 to ∞, lowest values show that the predicted model's performance is better. The RMSE can be determined by taking the square root of mean square error (MSE) and can provide a complete error distribution scenario. MAE is the average of absolute differences between the actual and predicted values. Mean bias error (MBE) is also used to estimate the average bias in the model or average forecasting error. MBE represents the systematic error of the forecasting model to over or under forecast. The positive value of MBE represents the over-forecast of the model whereas the negative value represents the under-forecast of the model. The mathematical equations used for computing RMSE, MAE, and MBE are given below.

$$RMSE = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(X_T - P_T)^2} \tag{11}$$

$$MAE = \frac{1}{T}\sum_{i=1}^{T}(|X_T - P_T|) \tag{12}$$

$$MBE = \frac{1}{T}\sum_{i=1}^{T}(P_T - X_T) \tag{13}$$

where X_T represents the target (expected) values and P_T is the model's predicted values.

3. Results

In the first phase of each of the EMD-ML models, EMD is used to extract the data characteristics of PM_{10} and $PM_{2.5}$ time series by decomposing the historical data as presented in Figure 2 and discussed before. EMD algorithm is applied on both PMs (PM_{10} and $PM_{2.5}$) time-series data of Misfalah, Makkah, and 14 IMFs along-with a single residual has been generated for each of the time-series data. Similarly, the EMD algorithm is applied on both PMs (PM_{10} and $PM_{2.5}$) time-series data of Dehli, India, and 11 IMFs along-with a single residual for PM_{10} and 12 IMFs along-with a single residual for $PM_{2.5}$ have been generated. Figure 3(a) plots the decomposed IMFs and residuals of original PMs time series data of Misfalah, Makkah, and Figure 3(b) plots the decomposed IMFs and residuals of original PMs time series data of Dehli, India. IMF1 (with the highest frequency), represents the high time variant of the original data, and the residual (with the lowest frequency) represents the trend of the original data.

IMF components of each time series data might have a strong or weak correlation with the original data. To find out the correlation between IMF components and original data, Spearman correlation coefficient is computed and coefficient values of both PMs (PM_{10} and $PM_{2.5}$) IMFs and residual are summarized in Table 1. The bold values in the table indicate a weak correlation (values less than 0.15).

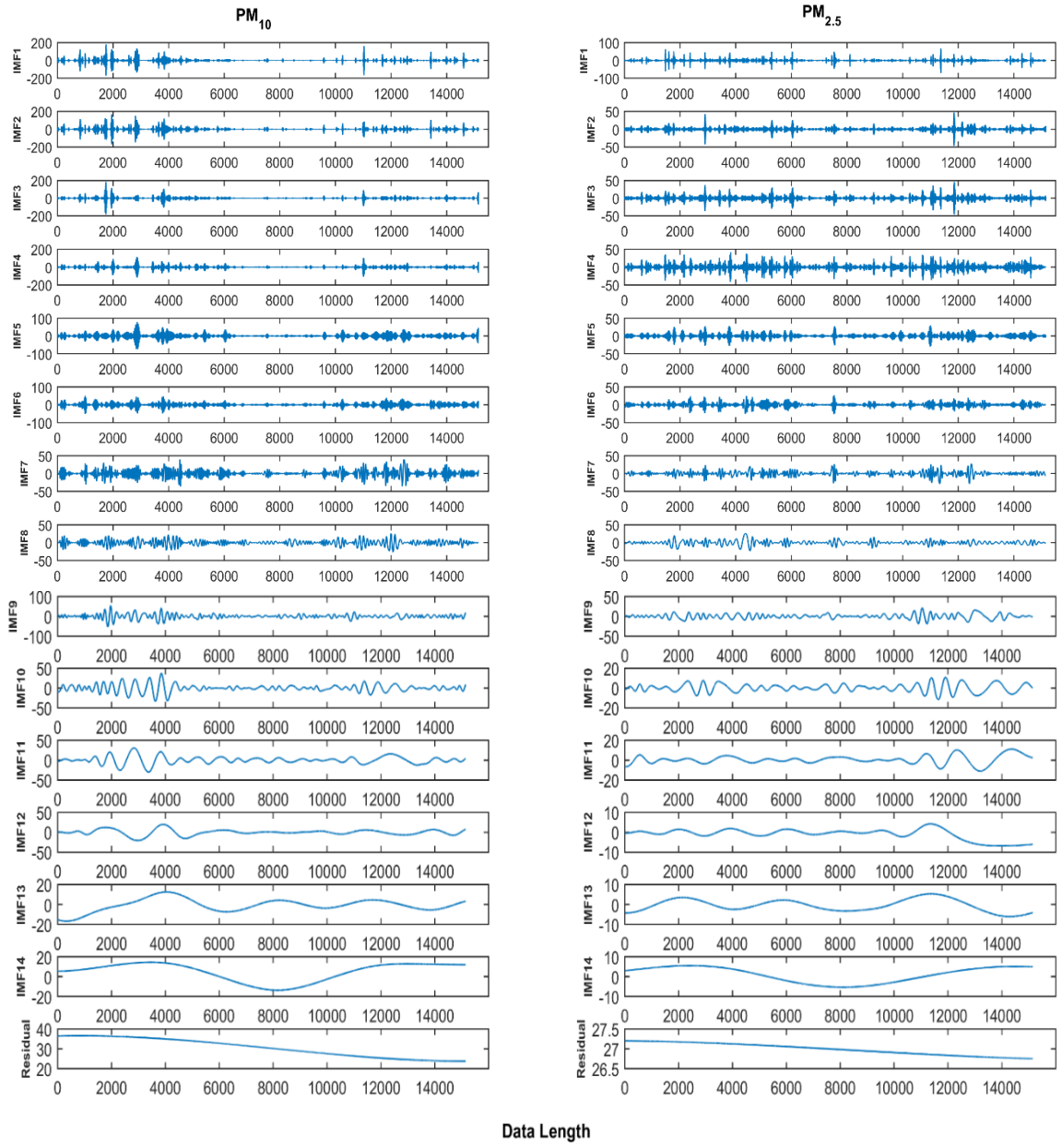

Figure 3(a). Decomposed IMFs and residuals of original PM_{10} and $PM_{2.5}$ time series data of Misfalah, Makkah.

Table 1 shows that for PM_{10} time series data of Misfalah, Makkah, IMF3-IMF11 and IMF13-IMF14 have a strong correlation with original data and for $PM_{2.5}$ data of Misfalah, Makkah, IMF2-IMF14 and residual have a strong correlation with original data. For PM_{10} data of Dehli, India, IMF2-IMF8, and IMF10-IMF11 have a strong correlation with original data and for $PM_{2.5}$ data of Dehli, India, IMF3-IMF4, and IMF6-IMF12 have a strong correlation with original data.

Therefore, in the second phase of EMD-ML models, only these IMFs are given to ML algorithms for the prediction of each time-series data.

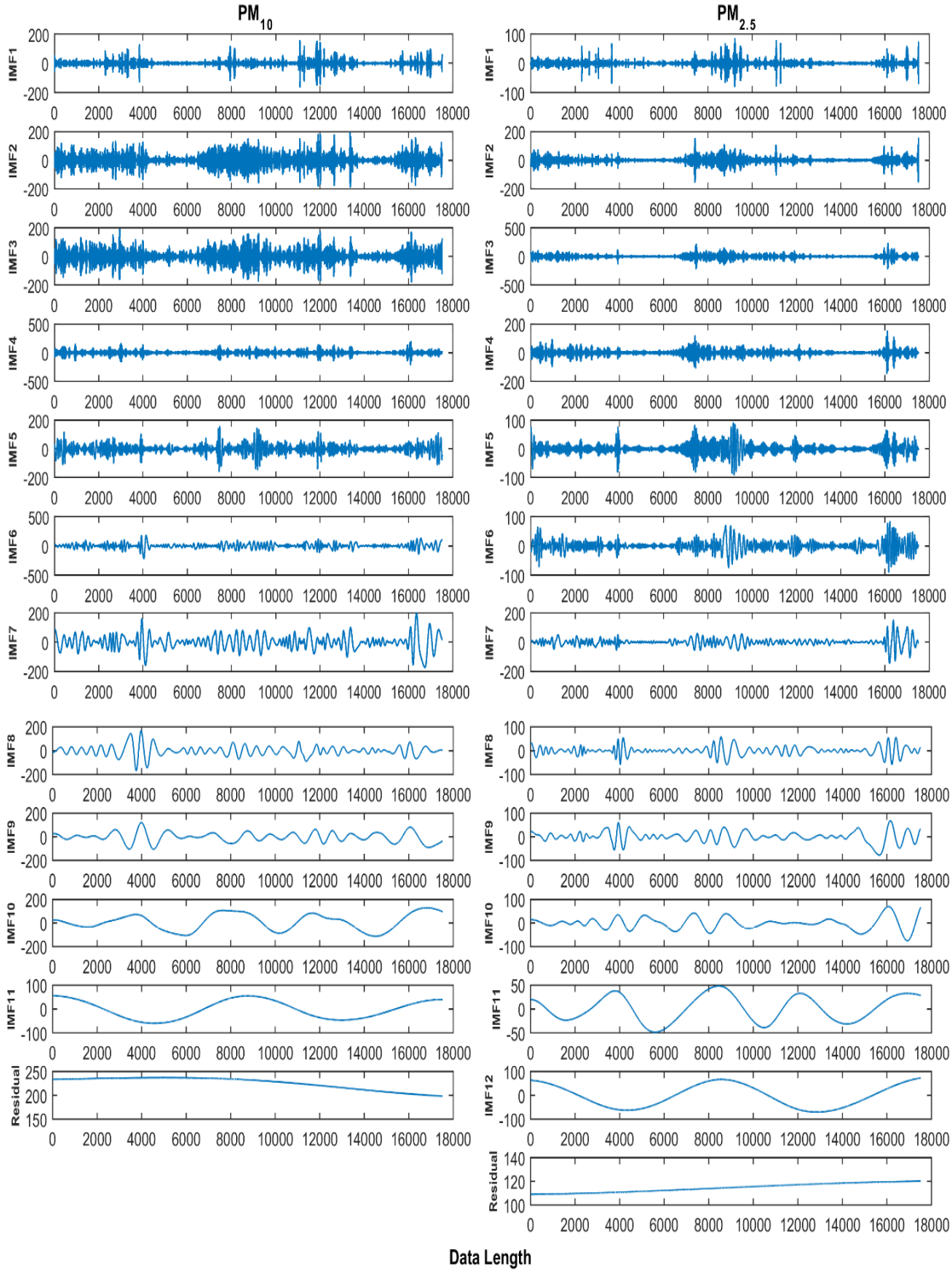

Figure 3(b). Decomposed IMFs and residuals of original PM10 and PM2.5 time series data of Dehli, India.

To forecast both PM_{10} and $PM_{2.5}$ time series of Misfalah, Makkah, selected IMFs data (length of each IMF is the same as the original data) and original data are organized according to the following settings.

Setting 1: The selected IMFs and original data are divided into two sets namely the train-set and the test-set. The train-set consists of the selected IMFs and original data from January 2014 to August 2015, while the test-set comprises of selected IMFs and original data from September 2015.

Table 1. Spearman correlation coefficient values of IMFs and residuals for both PMs.

IMFs	Misfalah, Makkah Data		Dehli, India Data	
	PM_{10}	$PM_{2.5}$	PM_{10}	$PM_{2.5}$
IMF1	0.11	0.14	0.06	0.04
IMF2	0.07	0.17	0.19	0.13
IMF3	0.16	0.19	0.26	0.27
IMF4	0.25	0.32	0.19	0.18
IMF5	0.29	0.26	0.19	0.14
IMF6	0.21	0.20	0.26	0.18
IMF7	0.19	0.23	0.31	0.22
IMF8	0.17	0.26	0.20	0.17
IMF9	0.22	0.24	0.14	0.17
IMF10	0.22	0.24	0.58	0.22
IMF11	0.21	0.27	0.47	0.53
IMF12	0.09	0.15	-------	0.63
IMF13	0.18	0.18	-------	-------
IMF14	0.35	0.22	-------	-------
Res.	0.09	0.16	-0.04	-0.07

Setting 2: The selected IMFs and original data are organized in 10-fold cross-validation (CV) fashion. In 10-fold CV the samples are randomly partitioned into 10 equal portions; nine of which are used as train-set and one as test-set. The procedure is repeated 10 times so each portion is used once for validation.

The design of the ML algorithms (RF, SVR-L, SVR-R, kNN, FFNN, and AdaBoost) follows the configurations detailed in the section learning algorithms and hybrid EMD-ML models. RMSE, MAE, and MBE measures are computed to evaluate the performances of learning algorithms used for forecasting PM_{10} and $PM_{2.5}$ time series. The exemplary plots of actual and predicted values of PM_{10} and $PM_{2.5}$ using EMD-ML models and single forecasting models (RF, SVR-L, SVR-R, kNN, FFNN, and AdaBoost) according to both settings of data mentioned above are illustrated in Figure 4. The prediction curve produced by each of the EMD-ML models using setting 1 fits better at many points and followed the trend of actual values in a quite better way for both PM_{10} and $PM_{2.5}$ concentration as compared to single forecasting models. As the trend of predicted values using each of the EMD-ML models is quite closer to actual values which clearly showed that the hybrid EMD-ML models can better forecast PMs concentrations. Among all EMD-ML models, EMD-FFNN model using setting 1 provides better prediction of both PMs.

Similarly for forecasting both PM_{10} and $PM_{2.5}$ time series of Dehli, India, selected IMFs data (length of each IMF is the same as the original data) and original data are organized according to setting 1 and setting 2, but in setting 1 the train-set consists of the selected IMFs and original data from January 2018 to November 2019, while the test-set comprises of selected IMFs and original data of December 2019.

The design of the ML algorithms (RF, SVR-L, SVR-R, kNN, FFNN, and AdaBoost) follows the configurations detailed in the section learning algorithms and hybrid EMD-ML models. RMSE, MAE, and MBE measures are computed to evaluate the performances of learning algorithms used for forecasting PM_{10} and $PM_{2.5}$ time series. The exemplary plots of actual and predicted values of PM_{10} and $PM_{2.5}$ using EMD-ML models according to both setting 1 and setting 2 are illustrated in Figure 5. The prediction curve produced by each of the EMD-ML models using setting 2 fits better

at many points and followed the trend of actual values in a quite better way for both PM_{10} and $PM_{2.5}$ concentrations. As the trend of predicted values using each of the EMD-ML models is quite.

The scatter plot of observed and predicted values (prediction has been done using EMD-ML closer to actual values which clearly showed that the hybrid EMD-ML models can better forecast PMs concentrations. Among all EMD-ML models, the EMD-kNN model using setting 2 for PM_{10} and EMD-AdaBoost model using setting 2 for $PM_{2.5}$ provides better prediction.models) of both PM_{10} and $PM_{2.5}$ concentrations are presented in Figure 6. The figure shows a good agreement between observed and predicted values.

In Table 2, prediction results of both PM_{10} and $PM_{2.5}$ using setting 1 and setting 2 in terms of RMSE, MAE, and MBE based on EMD-ML models and single forecasting models are presented for Misfalah, Makkah. It is clear from the table that EMD-ML models using setting 1 produced lower errors against forecasted values of both PM_{10} and $PM_{2.5}$ as compared to setting 2 and traditional single forecasting models. The lowest error rate in terms of RMSE and MAE for both PM_{10} (RMSE = 12.26 and MAE = 7.43) and $PM_{2.5}$ (RMSE = 4.81 and MAE = 3.02) have been achieved using the EMD-FFNN model. In the case of single models the lowest error rate in terms of RMSE and MAE for PM_{10} (RMSE = 22.18 and MAE = 11.98) has been achieved using the RF model and setting 1 and for $PM_{2.5}$ (RMSE = 11.88 and MAE = 8.28) has been achieved using FFNN model and setting 1. The results clearly show that hybrid models with setting 1 of data are the robust choice for the prediction of PM concentrations of Misfalah, Makkah.

MBE represents the systematic error of the forecasting model to over or under forecast. The positive value of MBE represents that the predictive model is overestimated and vice versa. The MBE values present in Table 2 are considerably better showing no bias for models RF, kNN, FFNN, and AdaBoost. The SVR-L and SVR-R models exhibited the highest MBE values showing model bias which needs to be filtered out.

In Table 3, EMD-ML models based prediction results of both PM_{10} and $PM_{2.5}$ using setting 1 and setting 2 in terms of RMSE, MAE, and MBE are presented for Dehli, India. It is clear from the table that EMD-ML models using setting 2 produced lower errors against forecasted values of both PM_{10} and $PM_{2.5}$ as compared to setting 1. The lowest error rate in terms of RMSE and MAE for PM_{10} (RMSE = 20.56 and MAE = 12.87) and $PM_{2.5}$ (RMSE = 15.29 and MAE = 9.45) have been achieved using EMD-kNN and EMD-AdaBoost models respectively. The results clearly show that hybrid models with setting 2 of data are the robust choice for the prediction of PM concentrations of Dehli, India data.

MBE represents the systematic error of the forecasting model to over or under forecast. The positive value of MBE represents that the predictive model is overestimated and vice versa. The MBE values present in Table 3 are considerably better showing no bias for models each model against setting 2. Various EMD-ML models for PM_{10} and $PM_{2.5}$ using setting 1, exhibited the highest MBE values showing model bias which needs to be filtered out.

The feasibility of EMD-ML models (EMD-RF, EMD-SVR-L, EMD-SVR-R, EMD-kNN, EMD-FFNN, EMD-AdaBoost) lies in the following two points. First, the PMs (PM_{10} and $PM_{2.5}$) concentrations, which are non-stationary and non-linear, can be decomposed into various IMFs using the EMD algorithm. Thus, IMFs having a strong correlation with original data can be used

Table 2. Performance of the predictive models using Misfalah, Makkah data in terms of RMSE, MAE, and MBE.

Model	setting 1			setting 2		
	RMSE	MAE	MBE	RMSE	MAE	MBE
			PM_{10} using EMD-ML models			
RF	13.89	9.64	2.34	19.30	8.20	-1.91
SVR-R	58.99	57.73	56.21	52.75	49.09	71.89
kNN	14.29	9.78	-0.72	20.39	7.70	-3.13
FFNN	12.26	7.43	2.36	17.91	8.38	-2.65
AdaBoost	13.49	8.75	0.96	21.54	7.74	-3.51
			$PM_{2.5}$ using EMD-ML models			
RF	8.88	6.48	-0.51	6.77	4.08	-0.44
SVR-L	19.63	18.49	-40.79	18.18	14.15	18.75
SVR-R	10.44	8.54	31.04	10.81	7.24	103.65
kNN	9.66	6.95	-16.20	5.89	3.46	-0.57
FFNN	4.81	3.02	-0.96	5.28	3.10	0.22
AdaBoost	7.84	5.99	4.49	6.35	3.59	-0.61
			PM_{10} using single forecasting models			
RF	22.18	11.98	0.12	27.74	11.93	-0.02
SVR-L	66.19	64.49	-49.29	63.03	58.69	-3.30
SVR-R	75.66	73.83	28.21	75.56	72.43	3.43
kNN	22.65	12.47	-0.24	29.82	12.24	-0.18
FFNN	22.74	11.57	0.30	28.05	12.97	0.04
AdaBoost	23.11	11.79	-0.69	27.21	9.94	-0.29
			$PM_{2.5}$ using single forecasting models			
RF	14.44	9.50	-1.32	11.88	7.13	4.34
SVR-L	48.69	46.63	-18.48	56.30	50.50	-46.26
SVR-R	42.01	37.38	7.85	38.11	33.55	36.26
kNN	13.75	9.52	1.72	13.23	7.57	2.63
FFNN	11.88	8.28	0.13	13.45	8.48	3.30
AdaBoost	12.87	8.84	-1.89	11.91	6.28	3.57

Table 3. Performance of the predictive models using Dehli, India data in terms of RMSE, MAE, and MBE.

Model	setting 1			setting 2		
	RMSE	MAE	MBE	RMSE	MAE	MBE
			PM_{10} using EMD-ML models			
RF	74.97	60.89	35.53	28.93	19.40	0.20
SVM-L	63.10	54.40	52.47	58.34	46.57	0.57
SVM-R	121.82	103.66	66.64	163.47	142.88	3.55
kNN	82.45	69.46	57.73	20.56	12.87	0.13
FNN	95.03	88.84	88.79	37.89	28.37	0.31
AdaBoost	83.38	67.13	50.61	23.27	15.47	0.16
			$PM_{2.5}$ using EMD-ML models			
RF	68.02	53.04	0.57	17.48	10.89	0.05
SVM-L	59.62	48.51	53.65	51.06	43.22	-0.39
SVM-R	70.97	60.02	67.47	126.00	116.08	44.50
kNN	65.61	50.52	-0.86	16.00	9.02	-0.25
FNN	45.64	32.61	0.12	27.69	17.66	0.21
AdaBoost	67.59	53.57	-2.97	15.29	9.45	-0.76

as input for EMD-ML models. Second, the EMD-ML models are well suited for time-series data prediction and have achieved significant results in various fields like wind speed forecasting [19], Chinese currency exchange rates forecasting [20], energy time series forecasting [21], rotating machinery structural faults detection [24], sudden cardiac death (SCD) prediction [26], and air quality index forecasting [27]. In comparison with the study of [27] which suggests EMD-SVR-Hybrid as an optimal predictive model for the forecasting of daily AQI with RMSE = 24.46 and MAE = 18.10, the performance of EMD-ML models used in this study is quite better (optimal predictive model is EMD-FFNN with RMSE = 12.26 and MAE = 7.43 for forecasting PM_{10} and RMSE = 4.81 and MAE = 3.02 for forecasting $PM_{2.5}$). Similarly, in comparison with [29] study which utilizes ensemble EMD in combination with regression neural network (EEMD-GRNN) for the forecasting of $PM_{2.5}$, with RMSE = 29.41 and MAE = 19.80, the performance of EMD-ML models used in this study is quite better for forecasting both PM_{10} and $PM_{2.5}$.

In general, EMD-ML models can be better than single forecasting models for the prediction of PMs (PM_{10} and $PM_{2.5}$) concentrations. The results of the current study verify the validity and feasibility of EMD-ML models.

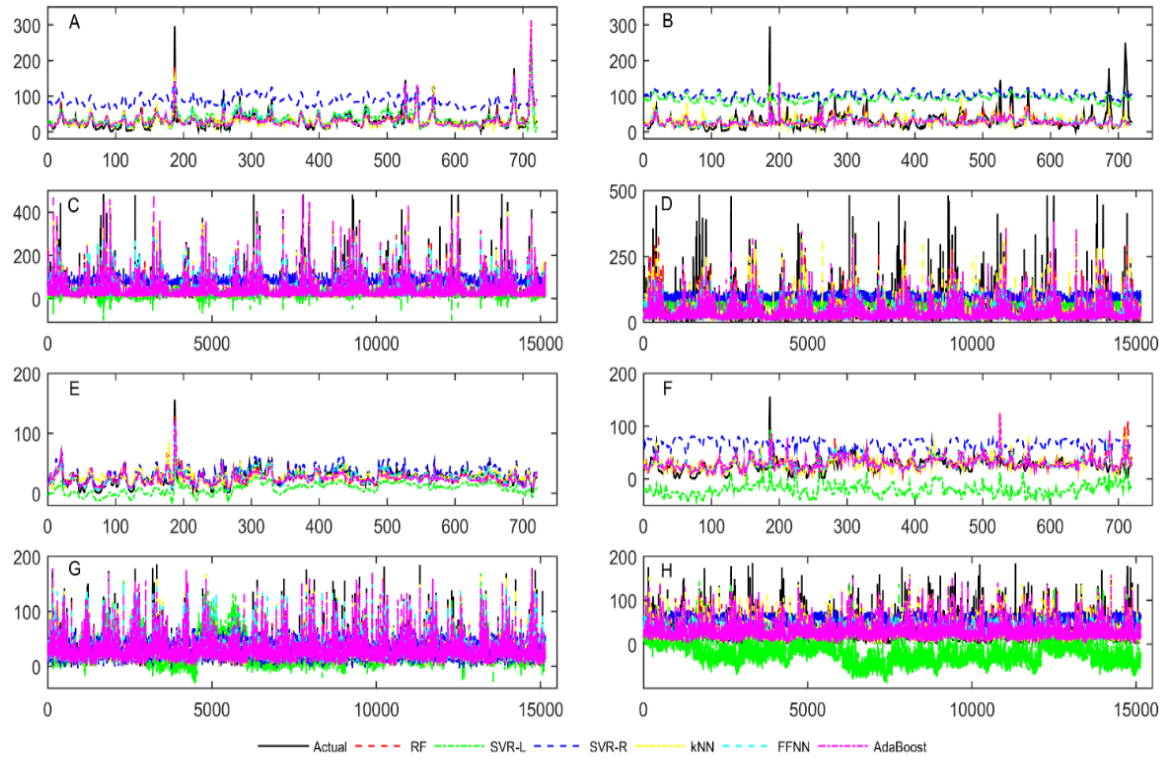

Figure 4. Comparison of actual and predicted curves for predicting A) PM_{10} time series data using each of EMD-ML model and setting 1, B) PM_{10} time series data using each single forecasting models and setting 1, C) PM_{10} time series data using each of EMD-ML model and setting 2, D) PM_{10} time series data using each single forecasting models and setting 2, E) $PM_{2.5}$ time series data using each of EMD-ML model and setting 1, F) $PM_{2.5}$ time series data using each single forecasting models and setting 1, G) $PM_{2.5}$ time series data using each of EMD-ML model and setting 2, H) $PM_{2.5}$ time series data using each single forecasting models and setting 2.

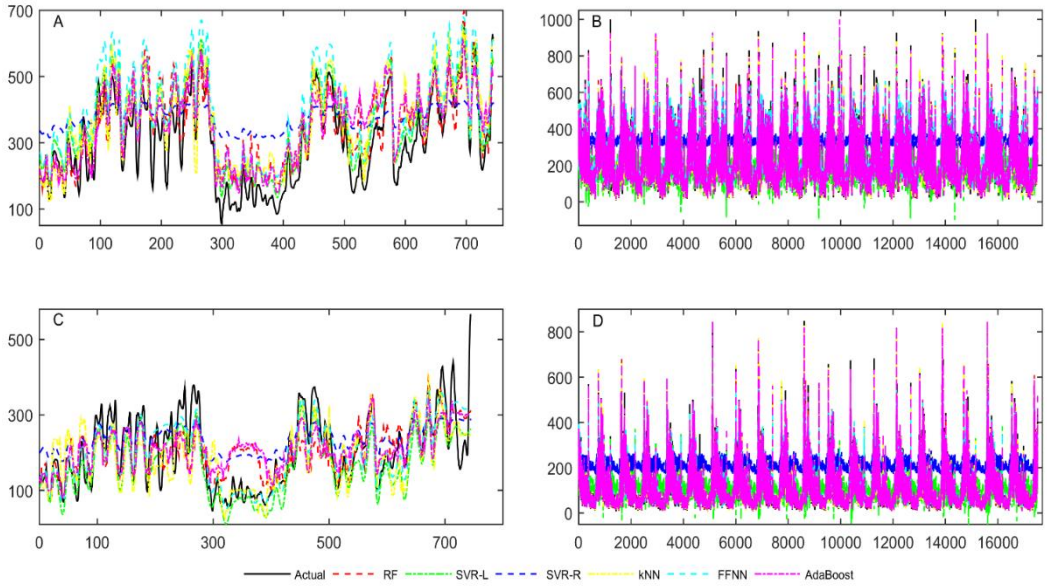

Figure 5. Comparison of actual and predicted curves for predicting A) PM_{10} time series data using each of EMD-ML models and setting 1, B) PM_{10} time series data using each of EMD-ML models and setting 2, C) $PM_{2.5}$ time series data using each of EMD-ML models and setting 1, D) $PM_{2.5}$ time series data using each of EMD-ML models and setting 2.

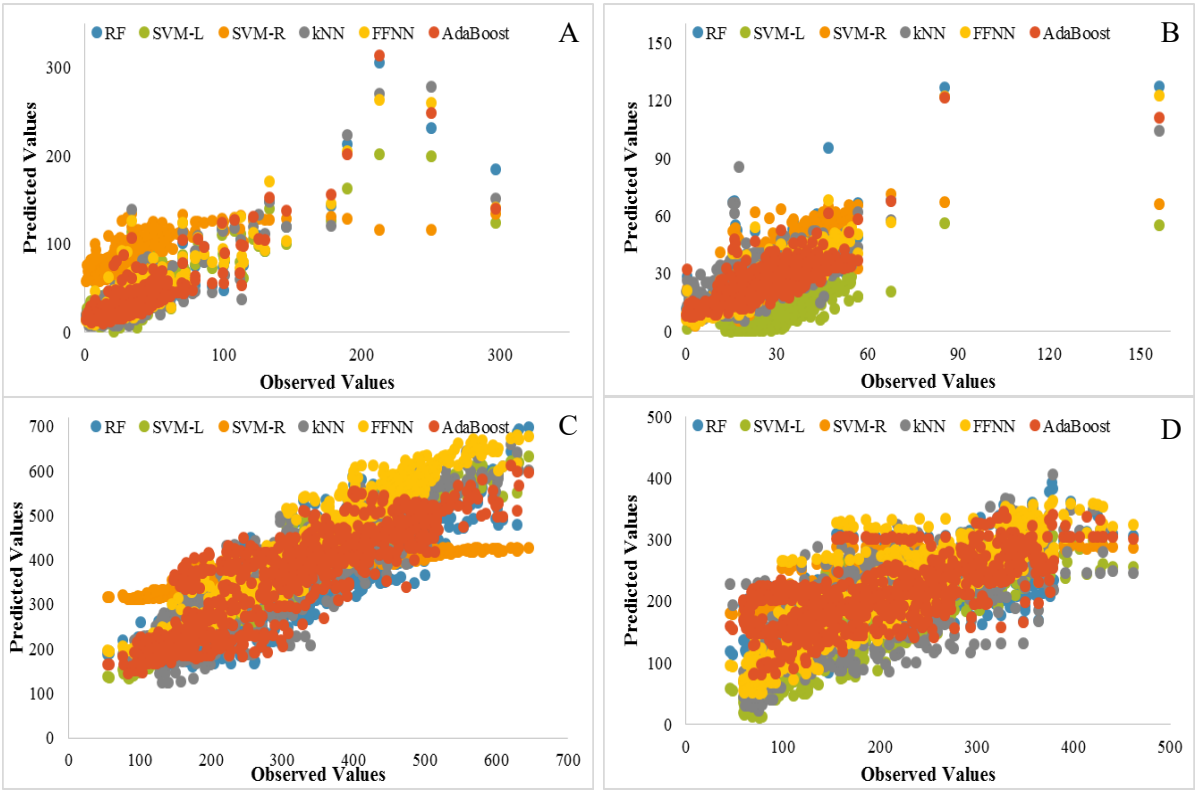

Figure 6. A) Scatter plot of observed and predicted values of Misfalah, Makkah PM_{10} time series, B) Scatter plot of observed and predicted values of Misfalah, Makkah $PM_{2.5}$ time series, C) Scatter plot of observed and predicted values of Dehli, India PM_{10} time series, D) Scatter plot of observed and predicted values of Dehli, India $PM_{2.5}$ time series.

4. Conclusions

In this study, the EMD algorithm was applied to address the trends and random behavior of time series data to enhance the accuracy of PM forecasting. This study attempted to improve ML model prediction by coupling them with the EMD procedure. The EMD algorithm is used for multiscale characterization of PM_{10} and $PM_{2.5}$ by decomposing the original time series into numerous IMFs. We used Spearman's correlation coefficient to select strong correlated IMFs of PM_{10} and $PM_{2.5}$ to build a predictive model. The air quality time series data from Masfalah air station Makkah, Saudi Arabia, and Dehli city, India are utilized for the validation of the developed hybrid model. Firstly, the EMD based predictive models are applied to predict monthly PMs (PM_{10} and $PM_{2.5}$). For the hybridized models, the original time series data are decomposed into fourteen IMFs and one residual for the PMs modeling process. The non-hybridized RF, SVR-L, SVR-R, kNN, FFNN, and AdaBoost models are also applied to forecast monthly PM (PM_{10} and $PM_{2.5}$) using input data of pollutants (CO, NO_2, and CO_2), PMs (PM_{10} and $PM_{2.5}$), and meteorological factors (Temp, WS, and RH). The results demonstrated that correlated IMFs incorporated in EMD-ML models provide more prediction abilities of PMs and should be recommended to forecast PMs concentrations.

The EMD-ML models have accomplished good predictive performance and can be applied for the prediction of other pollutants present in the air as well as for other time-series data such as biological signals, financial time series, and energy time series. Other versions of EMD such as ensemble EMD (EEMD), complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN), and multivariate EMD (MEMD) can also be used instead of EMD in EMD-ML models.

Conflict of interest

The authors declare no conflict of interest in this paper.

References

1. B. Chen, H. Kan, Air pollution and population health: A global challenge, *Environ. Health Prev. Med.*, **13** (2008), 94–101.
2. R. Habre, B. Coull, E. Moshier, J. Godbold, A. Grunin, A. Nath, et al., Sources of indoor air pollution in New York city residences of asthmatic children, *J. Expo. Sci. Environ. Epidemiol.*, **24** (2014), 269–278.
3. D. L. Robinson, Air pollution in Australia: Review of costs sources and potential solutions, *Health Promot. J. Austr.*, **16** (2005), 213–220.
4. H. S. Rumana, R. C. Sharma, V. Beniwal, A. K. Sharma, A retrospective approach to assess human health risks associated with growing air pollution in urbanized area of Thar Desert, western Rajasthan, India, *J. Environ. Health Sci. Eng.*, **12** (2014), 23.
5. S. Yamamoto, R. Phalkey, A. Malik, A systematic review of air pollution as a risk factor for cardiovascular disease in South Asia: Limited evidence from India and Pakistan, *Int. J. Hyg. Environ. Health*, **217** (2014), 133–144.
6. W. Zhang, C. N. Qian, Y. X. Zeng, Air pollution: A smoking gun for cancer, *Chin. J. Cancer*, **33** (2014), 173.
7. H. Kan, B. Chen, N. Zhao, S. J. London, G. Song, G. Chen, et al., Part 1: A time-series study of ambient air pollution and daily mortality in Shanghai, China, *Res. Rep. Health. Eff. Inst.*, **154** (2010), 17–78.

8. K. Vermaelen, G. Brusselle, Exposing a deadly alliance: Novel insights into the biological links between COPD and lung cancer, *Pulm. Pharmacol. Ther.*, **26** (2013), 544–554.
9. WHO., Burden of disease from the joint effects of household and ambient air pollution for 2016, *Soc. Environ. Determ. Health Dep.: Geneva, Switzerland*, **7** (2018).
10. C A. Pope III, D. W. Dockery, Health effects of fine particulate air pollution: Lines that connect, *J. Air Waste Manag. Assoc.*, **56** (2006), 709–742.
11. R. Shad, M. S. Mesgari, A. Shad, Predicting air pollution using fuzzy genetic linear membership kriging in GIS, *Comput. Environ. Urban Syst.*, **33** (2009), 472–481.
12. J. G. Titus, Greenhouse Effect, Sea Level Rise, and Barrier Islands: Case Study of Long Beach Island, New Jersey, 1990.
13. S. A. A. Shah, W. Aziz, M. S. A. Nadeem, M. Almaraashi, S. O. Shim, T. M. Habeebullah, A novel phase space reconstruction (PSR) based predictive algorithm to forecast atmospheric particulate matter concentration, *Sci. Program.*, 2019.
14. J. Zhu, P. Wu, H. Chen, L. Zhou, Z. Tao, A hybrid forecasting approach to air quality time series based on endpoint condition and combined forecasting model, *Int. J. Eviron. Res. Pub. Health*, **15** (2018), 1941.
15. A. B. Chelani, S. Devotta, Air quality forecasting using a hybrid autoregressive and nonlinear model, *Atmos. Environ.*, **40** (2006), 1774–1780.
16. N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, et al., The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, *Proc. Math. Phys. Eng. Sci.*, **454** (1998), 903–995.
17. Q. Chen, D. Wen, X. Li, D. Chen, H. Lv, J. Zhang, et al., Empirical mode decomposition based long short-term memory neural network forecasting model for the short-term metro passenger flow, *PloS one*, **14** (2019), 222365.
18. O. K. Cura, S. K. Atli, H. S. Türe, A. Akan, Epileptic seizure classifications using empirical mode decomposition and its derivative, *Bio. Med. Eng. OnLine*, **19** (2020), 1–22.
19. J. Song, J. Wang, H. Lu, A novel combined model based on advanced optimization algorithm for short-term wind speed forecasting, *Appl. Energy*, **215** (2018), 643–658.
20. J. N. Wang, J. Du, C. Jiang, K. K. Lai, Chinese currency exchange rates forecasting with EMD-based neural network, *Complexity*, 2019.
21. W. Xu, H. Hu, W. Yang, Energy time series forecasting based on empirical mode decomposition and FRBF-AR model, *IEEE Access*, **7** (2019), 36540–36548.
22. L. Yu, Z. Wang, L. Tang, A decomposition-ensemble model with data-characteristic-driven reconstruction for crude oil price forecasting, *Appl. Energy*, **156** (2015), 251–267.
23. X. Zhang, J. Wang, A novel decomposition-ensemble model for forecasting short-term load-time series with multiple seasonal patterns, *Appl. Soft Comput.*, **65** (2018), 478–494.
24. Z. Guan, Z. Liao, K. Li, P. Chen, A precise diagnosis method of structural faults of rotating machinery based on combination of empirical mode decomposition, sample entropy and deep belief network, *Sensors*, **19** (2019), 591.
25. X. B. Jin, N. X. Yang, X. Y. Wang, Y. T. Bai, T. L. Su, J. L. Kong, Hybrid deep learning predictor for smart agriculture sensing based on empirical mode decomposition and gated recurrent unit group model, *Sensors*, **20** (2020), 1334.
26. O. Vargas-Lopez, J. P. Amezquita-Sanchez, J. J. De-Santiago-Perez, J. R. Rivera-Guillen, M. Valtierra-Rodriguez, M. Toledano-Ayala, et al., A new methodology based on EMD and nonlinear measurements for sudden cardiac death detection, *Sensors*, **20** (2020), 9.

27. S. Zhu, X. Lian, H. Liu, J. Hu, Y. Wang, J. Che, Daily air quality index forecasting with hybrid models: A case in China, *Environ. Pollut.*, **231** (2017), 1232–1244.
28. K. Pholsena, L. Pan, Z. Zheng, Mode decomposition based deep learning model for multi-section traffic prediction, *World Wide Web*, **23** (2020), 2513–2527.
29. Q. Zhou, H. Jiang, J. Wang, J. Zhou, A hybrid model for PM2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network, *Sci. Total Environ.*, **496** (2014), 264–274.
30. M. Niu, Y. Wang, S. Sun, Y. Li, A novel hybrid decomposition-and-ensemble model based on CEEMD and GWO for short-term PM2.5 concentration forecasting, *Atmos. Environ.*, **134** (2016), 168–180.
31. S. Munir, T. M. Habeebullah, A. M. Mohammed, E. A. Morsy, M. Rehan, K. Ali, Analysing PM2.5 and its association with PM10 and meteorology in the arid climate of Makkah, Saudi Arabia, *Aerosol Air Qual. Res.*, **17** (2016), 453–464.
32. T. M. Habeebullah, S. Munir, E. A. Morsy, A. M. Mohammed, Spatial and temporal analysis of air pollution in Makkah, the Kingdom of Saudi Arabia, *2010 5th Int. Conf. Environ. Sci. Tech.*, *IPCBEE*, 2010, 65–70.
33. P. Kline, The new psychometrics: Science, psychology and measurement, *Psychol. Press*, 1998.
34. A. Olinsky, S. Chen, L. Harlow, The comparative efficacy of imputation methods for missing data in structural equation modeling, *Eur. J. Oper. Res.*, **151** (2003), 53–79.
35. Vopani, Air Quality Data in India (2015-2020), Version 12, Available from https://www.kaggle.com/rohanrao/air-quality-data-in-india/version/12.
36. G. P. Zhang, Neural networks for time-series forecasting, *Springer Berlin Heidelberg*, 2012.
37. Y. Freund, R. E. Schapire, A short introduction to boosting, *J. Jpn. Soc. Artif. Intell.*, **14** (1999), 771–780.
38. L. Breiman, Random forests, *Mach. Learn.*, **45** (2001), 5–32.
39. E. Fix, J. Hodges, Discriminatory analysis: Nonparametric discrimination consistency properties, *USAF School Avi. Med. Project*, (1952), 21–49.
40. T. Bailey, A note on distance-weighted k-nearest neighbor rules, *IEEE Trans. Syst. Man, Cybernet.*, **8** (1978), 311–313.
41. H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, V. Vapnik, Support vector regression machines, *Proc. Adv. Neural Inf. Process. Syst.*, (1997), 155–161.

AIMS Press