Hindawi
Scientific Programming
Volume 2019, Article ID 6780379, 12 pages
https://doi.org/10.1155/2019/6780379

Research Article

A Novel Phase Space Reconstruction- (PSR-) Based Predictive Algorithm to Forecast Atmospheric Particulate Matter Concentration

Syed Ahsin Ali Shah,[1] Wajid Aziz ,[1,2] Malik Sajjad Ahmed Nadeem,[1] Majid Almaraashi,[2] Seong-O. Shim ,[2] and Turki M. Habeebullah [3]

[1] *Department of Computer Sciences and Information Technology, University of Azad Kashmir, 13100 Azad Kashmir, Pakistan*
[2] *College of Computer Sciences and Engineering, University of Jeddah, Saudi Arabia*
[3] *Institute for Hajj and Umrah Research, Umm Al-Qura University, Makkah, Saudi Arabia*

Correspondence should be addressed to Wajid Aziz; kh_wajid@yahoo.com

Received 26 March 2019; Revised 12 June 2019; Accepted 4 July 2019; Published 25 July 2019

Academic Editor: Cristian Mateos

The prediction of atmospheric particulate matter (APM) concentration is essential to reduce adverse effects on human health and to enforce emission restrictions. The dynamics of APM are inherently nonlinear and chaotic. Phase space reconstruction (PSR) is one of the widely used methods for chaotic time series analysis. The APM mass concentrations are an outcome of complex anthropogenic contributors evolving with time, which may operate on multiple time scales. Thus, the traditional single-variable PSR-based prediction algorithm in which data points of last embedding dimension are used as a target set may fail to account for multiple time scales inherent in APM concentrations. To address this issue, we propose a novel PSR-based scientific solution that accounts for the information contained at multiple time scales. Different machine learning algorithms are used to evaluate the performance of the proposed and traditional PSR techniques for predicting mass concentrations of particulate matter up to 2.5 micron ($PM_{2.5}$), up to 10 micron ($PM_{10.0}$), and ratio of $PM_{2.5}/PM_{10.0}$. Hourly time series data of $PM_{2.5}$ and $PM_{10.0}$ mass concentrations are collected from January 2014 to September 2015 at the Masfalah air quality monitoring station (couple of kilometers from the Holy Mosque in Makkah, Saudi Arabia). The performances of various learning algorithms are evaluated using RMSE and MAE. The results demonstrated that prediction error of all the machine learning techniques is smaller for the proposed PSR approach compared to traditional approach. For $PM_{2.5}$, FFNN leads to best results (both RMSE and MAE 0.04 μgm^{-3}), followed by SVR-L (RMSE 0.01 μgm^{-3} and MAE 0.09 μgm^{-3}) and RF (RMSE 1.27 μgm^{-3} and MAE 0.86 μgm^{-3}). For $PM_{10.0}$, SVR-L leads to best results (both RMSE and MAE 0.06 μgm^{-3}), followed by FFNN (RMSE 0.13 μgm^{-3} and MAE 0.09 μgm^{-3}) and RF (RMSE 1.60 μgm^{-3} and MAE 1.16 μgm^{-3}). For $PM_{2.5}/PM_{10.0}$, FFNN is the best and accurate method for prediction (0.001 for both RMSE and MAE), followed by RF (0.02 for both RMSE and MAE) and SVR-L (RMSE 0.05 μgm^{-3} and MAE 0.04).

1. Introduction

Air pollution is one of the emerging environmental issues in the developing as well as developed countries across the globe [1]. A large amount of gaseous pollutants and other atmospheric particulate matter (APM) are being produced through immense pollution generating activities including vehicles emitting smoke and fossil fuels used for energy requirements, cooking, and different anthropogenic activities [2]. APM is reportedly one of the major causes of adverse health issues particularly which are related to human respiratory and cardiovascular systems [3].

Depending upon aerodynamic diameter, atmospheric particles can be classified into three types, namely, coarse particle fraction (CPF), fine particle fraction (FPF), and ultrafine particles (UFP). CPF comprises of diameter larger than 2.5 micrometer (μm) and up to 10 μm ($PM_{10.0}$), while FPF has diameter up to 2.5 μm ($PM_{2.5}$), and those having less than 0.1 μm ($PM_{1.0}$) diameter are UFP [4]. Crustal material, paved road dust, background sea salts, and noncatalyst equipped gasoline engines are major sources of CPF ($PM_{10.0}$), while vapor nucleation/condensation mechanisms and anthropogenic sources are responsible for FPF ($PM_{2.5}$) [5]. The lifetime of atmospheric particles, spanned from few seconds to several months, is another aspect of such particles which determines their harmfulness [4]. Beside emission sources, levels of $PM_{2.5}$ and $PM_{10.0}$ depend on the geographic characteristics and meteorological parameters including wind, relative humidity temperature, atmospheric pressure, and boundary layer height [6, 7].

Air quality can be predicted through time series analysis which in turn may be used for issuing warnings to protect the health of the public. The classical approaches which predict air pollutant concentrations are generally based on functional relationship of air quality, emissions, and metrological factors. Examples include regression and neural network techniques, which have been used to predict APM in numerous studies [8–11]. In the absence of emission data and/or metrological factors, pollutant concentration time series data are the only available information. Therefore, in such cases, linear correlation-based univariate analysis techniques including autocorrelation function and spectral analysis [8, 12] are generally used. These techniques predict time series, which have regular behavior. Contrary to linearity, the dynamics of atmospheric pollutants are complex in nature; thus, nonlinearity is inherent in the atmospheric systems. The time series data of atmospheric mass concentrations are chaotic and very sensitive to initial conditions [13, 14].

Phase space reconstruction (PSR) is the foundation of nonlinear time series analysis that allows the reconstruction of complete system dynamics using a single time series [15]. The most common approach for PSR time series is based on Takens' delay embedding theorem [16]. Using this theorem, a single vector of observations representing a chaotic system can be regenerated into multidimensional vectors series. The regenerated vectors can thus display numerous essential properties of its real time series provided that the embedding dimension is considerably large [17]. Two parameters are important for the computation of PSR, i.e., time delay (τ) and embedding dimension (m).

Numerous studies used PSR-based techniques to capture complex dynamics of particulate matter mass concentration time series [13, 14, 18–25], which were then used for prediction purpose. Li et al. [18] performed nonlinear analysis of air quality data to identify the dynamics of the ozone concentrations and to determine dimensionality of the system. Chen et al. [19] proposed a novel procedure, based on dynamical systems theory, to model and predict ozone levels by creating a multidimensional phase space map from observed ozone concentrations. The proposed model was used to make one hour to one day ahead predictions of ozone levels. Kocak et al. [20] reconstructed the attractor in the multidimensional space of the univariate ozone time series and then used local approximation to predict the ozone concentration at different stations. Chelani et al. [21] examined the predictability of chaotic time series of air pollutant (nitrogen dioxide) concentration using artificial neural networks. Chelani and Devotta [22] predicted $PM_{10.0}$ using local polynomial approximation based on the reconstructed phase space. In another study, Chelani and Devotta [23] developed a hybrid model using the combination of the autoregressive integrated moving average model, which deals with linear patterns, and nonlinear dynamical model. Using the nitrogen dioxide concentration time series, they demonstrated that the hybrid model outperforms the individual linear and nonlinear models. Kumar et al. [13] employed a correlation dimension method that uses PSR to identify nonlinearity and chaos in nitrogen dioxide and carbon mono-oxide time series. Yu et al. [24] employed PSR to air pollution index time series during past 10 years and found that $PM_{10.0}$ time series behavior is chaotic in Lanzhou, China. Saeed et al. [25] investigated chaotic behavior of $PM_{1.0}$ and $PM_{2.5}$ concentrations using PSR, largest Lyapunov exponent, and Hurst exponent and found strong chaotic behavior in the time series.

The previous studies [26–28] used last embedding dimension data points of PSR time series as the target set. Recently, the concept of multiple time scales has been introduced to study dynamics of healthy and pathological physiological systems such as regularity mechanism of cardiovascular system [29, 30], postural control [31], and gait dynamics [32]. The APM mass concentrations are an outcome of complex natural and anthropogenic contributors evolving with time, which may operate on multiple time scales. Thus, the traditional single-variable PSR algorithm [26–28] in which data points of last embedding dimension are used as a target dataset may fail to account for multiple time scales inherent in APM concentrations.

In this study, we propose a novel PSR-based scientific solution that accounts for the information contained at multiple time scales to predict mass concentrations of atmospheric particulates in air. The data used in this study are collected from the Masfalah air quality monitoring station, Makkah, Saudi Arabia [6]. Previously Munir et al. [6] used these data to analyze the mass concentrations of $PM_{2.5}$ and its association with $PM_{10.0}$ and meteorology. This site is important because throughout the year, huge number of pilgrims visit Saudi Arabia to perform religious obligations using this road. Makkah is surrounded by large sandy deserts, receives little rain, and experiences high temperature throughout the year [6]. The expansion of Holy mosque, construction of railway train stations, mountain digging and construction of multistoried buildings, frequent sand and dust storms, frequent traffic jams, and congestions during the busy hours constitute the atmospheric pollution in the city [6, 7]. Millions of pilgrims visiting for Umrah and Hajj every year put additional burden on local resources and air quality. Moreover, due to the geographical characteristics and climatic conditions, $PM_{2.5}$ and $PM_{10.0}$ pollutants frequently exceed the national and international air quality standards, which is one of the major concerns in this region [6, 33]. Hence, early prediction is a managerial solution to avoid hazardous implications of atmospheric particulates on the local community as well as pilgrims.

Machine learning techniques have widely been used for classification, clustering, and association that are applied in numerous fields [34, 35]. Recently, a method of PSR of a chaotic model and support vector machine (SVM) in the field of artificial intelligence have been explored to realize the prediction of time series [36]. We used different machine learning techniques including support vector regression (SVR), random forest (RF), and feedforward neural network (FFNN) [37–39] for prediction of atmospheric particulates based on proposed and traditional settings of the target set. Root-mean-squared error (RMSE) and mean absolute error (MAE) measures are used to evaluate the performance of various learning algorithms for the prediction of atmospheric particulates by employing proposed and traditional PSR methods.

2. Materials and Methods

2.1. Datasets. The data used in this research work have been collected from the Masfalah air quality monitoring station (AQMS111) in the Holy city of Makkah, Saudi Arabia. The data were previously used by Munir et al. [6] to characterize the spatial and temporal variability of $PM_{2.5}$, $PM_{10.0}$, and their ratio $PM_{2.5}/PM_{10.0}$ in the region.

The concentrations of $PM_{2.5}$ and $PM_{10.0}$ were monitored using Aeroqual AQM 60 air quality monitoring station [6]. This device uses light scattering nephelometer and high-precision sharp cut cyclone to monitor particles and has a range of 0–2000 μgm^{-3} with an accuracy of ±2 for both $PM_{2.5}$ and $PM_{10.0}$. Hourly data collected from January 2014 to September 2015 of $PM_{2.5}$ (μgm^{-3}), $PM_{10.0}$ (μgm^{-3}), and ratio of $PM_{2.5}/PM_{10.0}$ have been used to evaluate the usefulness of the proposed modification in the PSR prediction algorithm. The quality of data is ensured by taking strict quality assurance and quality control (QA/QC) measures [6]. QA measures include careful selection of monitoring site, proper instrument installation, instrument selection, sample system design, and proper training of operators. QC is ensured by taking measures including careful selection of monitoring site, instrument calibration and its response, monitoring calibration gases, routine site visit, and data review as well as data validation and ratification. Data screening for missing values and outliers was done. Kline [40] suggested that missing data can be handled by deletion, imputation estimates or by modeling the data as a distribution for its estimation. If missing data are <5%, then any simple mechanism is acceptable for its identification and correction [41]. Both $PM_{2.5}$ and $PM_{10.0}$ data contain less than 2% missing values, and we used deletion approach for handling missing data. The outliers in the data are replaced by means of data for that specific month.

2.2. Methodology. Before describing the proposed PSR methodology, traditional PSR technique and procedures for selection of time delay τ and embedding dimension m are detailed for clarity of methodology.

2.2.1. Phase Space Reconstruction (PSR). PSR [14] theory is the base for chaotic time series. In a chaotic system, phase space can be used for the reconstruction of univariate time series. This is because in a dynamical system, whole information about the variable is present in the univariate time series. Each point of phase space represents a state of the system, while trajectory of the phase space represents the time evolution of the system according to different initial conditions.

Using Takens' time-delay embedding theorem, a phase space can be created from a one-dimensional time series [14]. This theorem is actually a way for analyzing chaotic time series. According to the theorem, if a scalar time series $T_t = \{N_1, N_2, N_3, \ldots, N_n\}$ from a chaotic system is given, then reconstruction is possible in terms of the phase space vectors $X(t)$ expressed as: $X(t) = [x(t), x(t+\tau), \ldots, x(t+(m-1)\tau)]$ where $t = 1, 2, \ldots, M$; $M = N - (m-1)\tau$. Here, τ is the time delay, m is the embedding dimension of PSR, and M is the number of phase points of reconstructed phase space. Computation of τ and m values are very essential in PSR.

The selection of τ has centered around two commonly used methods, i.e., autocorrelation function (ACF) and average mutual information (AMI) [42]. The ACF is used for estimating τ of linear time series, whereas AMI is used for estimating τ for nonlinear time series. Since mass concentration time series data of atmosphere is nonlinear in nature, we used the AMI function, which accounts for the nonlinear correlation in a specific time series to evaluate 'τ' for that time series [42]. The equation to calculate AMI is as follows:

$$I(\tau) = \sum_{t=1}^{N-\tau} P(X_t, X_{t+\tau}) \cdot \log\left(\frac{P(X_t, X_{t+\tau})}{P(X_t) \cdot P(X_{t+\tau})}\right), \quad (1)$$

where $P(X_t)$ is the probability density of X_t. $P(X_t, X_{t+\tau})$ is the joint probability density of X_t and $X_{t+\tau}$. $I(\tau)$ is a measure of the statistical dependence of the reconstruction variables. For nonmonotonous decrease of $I(\tau)$, the location of first local minimum is considered as the suitable value of τ [43]. For monotonous decrease of $I(\tau)$, either the decrease of MI to $I(t)/I(0) = 1/e$ or $I(t)/I(0) = 0.2$ can be used as the criterion for estimating time delay [43].

The false nearest neighbor (FNN) approach introduced by Kennel et al. [43] is used for computing optimal m. The FNN algorithm takes each point in the m-dimensional portrait and finds the distance $D(m)$ to its nearest neighbor and the distance $D(m+1)$ between the two points in $m+1$ dimensions. Neighbors are said to be false if the following two criteria are met [43]:

$$\begin{gathered} \delta_n > R_{\text{tol}}, \\ \frac{D(m+1)}{R_A} > R_{\text{tol}}, A_{\text{tol}}, \end{gathered} \quad (2)$$

where δ_n is the relative increase in the Euclidean distance when the dimension of PSR is increased from m to $m+1$, and it is computed as

$$\begin{aligned}\delta_n &= \sqrt{\frac{D_n^2(m+1,\tau) - D_n^2(m,\tau)}{D_n^2(m,\tau)}},\\ &= \frac{\left|x_{n+\tau m} - x^r_{n+\tau m}\right|}{D_n(m,\tau)}.\end{aligned} \tag{3}$$

The parameters R_{tol} and A_{tol} are constant thresholds, and R_A is the standard deviation of a time series. The process is repeated for dimensions and is stopped when the proportion of FNN becomes zero or necessarily small and will remain so from then onwards.

2.2.2. Proposed Methodology. The whole procedure of PSR-based prediction is illustrated in Figure 1.

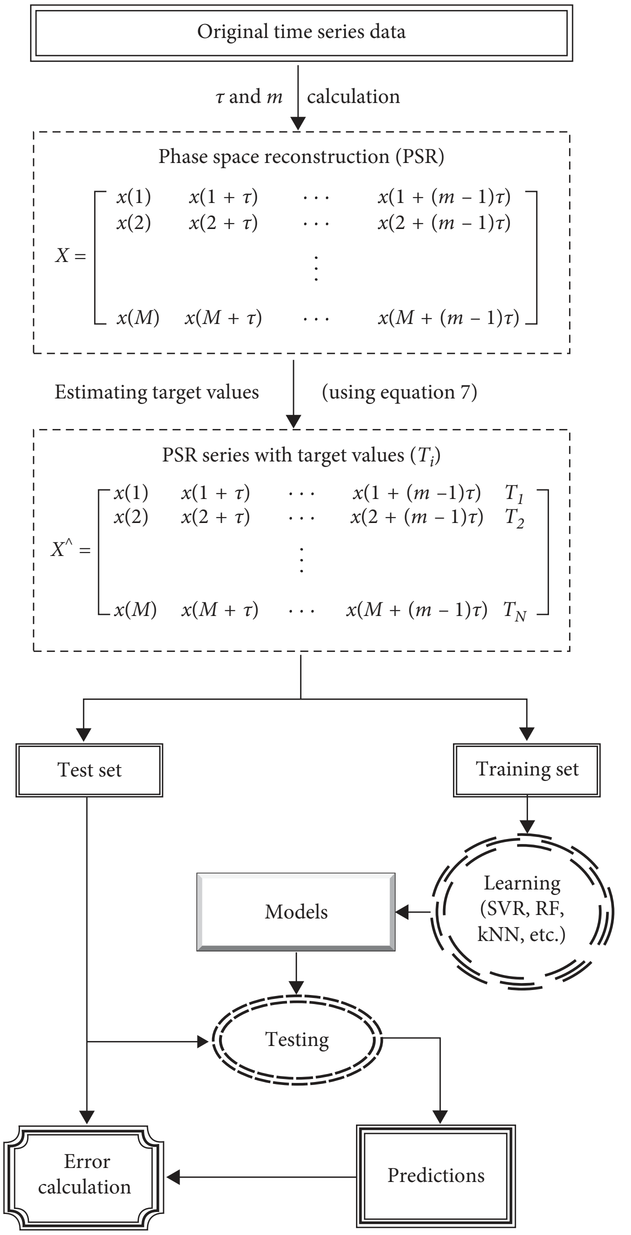

Original time series data
τ and m calculation
Phase space reconstruction (PSR)
X =
x(1) x(1 + τ) · · · x(1 + (m − 1)τ)
x(2) x(2 + τ) · · · x(2 + (m − 1)τ)
x(M) x(M + τ) · · · x(M + (m − 1)τ)
Estimating target values (using equation 7)
PSR series with target values (T_i)
X^ =
x(1) x(1 + τ) · · · x(1 + (m −1)τ) T_1
x(2) x(2 + τ) · · · x(2 + (m − 1)τ) T_2
x(M) x(M + τ) · · · x(M + (m − 1)τ) T_N
Test set
Training set
Learning (SVR, RF, kNN, etc.)
Models
Testing
Error calculation
Predictions

Figure 1: Proposed methodology for PSR-based prediction.

Step 1 (PSR). One-dimensional time series have been projected to higher dimensions using the PSR method to generate high-dimensional series:

$$X(t) = [x(t), x(t+\tau), \ldots, x(t+(m-1)\tau)], \tag{4}$$

where $t = 1, 2, \ldots, M$ and $M = N - (m-1)\tau$.

The parameters τ and m have been determined by using AMI and FNN methods, respectively. The input and output (target) samples can be represented by the matrixes following X and Y, respectively, in the following forms:

$$X = \begin{bmatrix} x(1) & x(1+\tau) & \ldots & x(1+(m-1)\tau) \\ x(2) & x(2+\tau) & \ldots & x(2+(m-1)\tau) \\ & & \cdots\cdots\cdots & \\ & & \cdots\cdots\cdots & \\ & & \cdots\cdots\cdots & \\ x(M) & x(M+\tau) & \ldots & x(M+(m-1)\tau) \end{bmatrix},$$

$$Y = \begin{bmatrix} x(2+(m-1)\tau) \\ x(3+(m-1)\tau) \\ \cdots\cdots\cdots \\ \cdots\cdots\cdots \\ \cdots\cdots\cdots \\ x(M+1+(m-1)\tau) \end{bmatrix}. \tag{5}$$

The last embedding dimension data points $[Y(t) = (t+1+(m-1)\tau)]$ as the target set have been used in numerous studies [24–26]. The concept of multiple time scales has been used in various studies [26–29]; therefore, in this study, it is proposed to use the concept of multiple time scales for the computation of the target set in order to get a better prediction of PSR series. Thus, the target values can be represented as

$$Y = \begin{bmatrix} \frac{1}{m}[\text{Sum}(x(1), x(1+\tau), \ldots, x(1+(m-1)\tau))] \\ \frac{1}{m}[\text{Sum}(x(2), x(2+\tau), \ldots, x(2+(m-1)\tau))] \\ \cdots\cdots\cdots \\ \cdots\cdots\cdots \\ \cdots\cdots\cdots \\ \frac{1}{m}[\text{Sum}(x(M), x(M+\tau), \ldots, x(M+(m-1)\tau))] \end{bmatrix}. \tag{6}$$

Both of input (X) and target (Y) of reconstructed series are divided into two sets, namely, the training set and the test set. The training set consists of the reconstructed series from January 2014 to August 2015, while the test set comprises September 2015 reconstructed series.

Step 2 (prediction). The regression model was built for the settings mentioned in step 1 using different learning algorithms (linear and radial SVRs, RF, and FFNN).

Step 3 (results). For the evaluation of prediction models, RMSE and MAE were computed.

In traditional PSR prediction, last embedding dimension of reconstructed time series is used as the target set, whereas in the proposed approach, target set data values T_i are computed using the following equation:

$$T_i = \frac{1}{m} \sum_{i=1}^{M+(m-1)\tau} x_i, \qquad (7)$$

where $i = 1, 2, \ldots, M$ and $M = N - (m-1)\tau$, m is the embedding dimension and τ is the time delay. Equation (7) is used in numerous studies [26–29] for constructing coarse grained series at multiple time series. In these studies, original time series has been divided in to nonoverlapping windows and then each window is averaged for constructing multiscale time series. Therefore, in this study, we used the same approach (equation (7)) after transforming the original time series into higher dimension instead of original time series (i.e., each row of PSR series is averaged for computation of the target set at various m).

2.2.3. Support Vector Regression (SVR). Consider a set of training data $\{(x_1, y_1), \ldots, (x_l, y_l)\}$, where each $x_i \varepsilon R^n$ represents the input samples with corresponding target value $y_i \varepsilon R$ for $i = 1, \ldots, l$ (l represents training data size) [34]. The generic SVR estimating function takes the following form:

$$f(x) = (w \cdot \Phi(x)) + b, \qquad (8)$$

where $w \varepsilon R^n$, $b \varepsilon R$, and Φ represent a nonlinear transformation from R^n to high-dimensional space. The objective is to find the values of w and b such that values of x can be determined by minimizing the regression risk:

$$R_{\text{reg}}(f) = C \sum_{i=0}^{l} \Gamma(f(x_i) - y_i) + \frac{1}{2}\|w\|^2, \qquad (9)$$

where C is a constant, Γ represents a cost function, and vector w can be written (in terms of data points) as

$$w = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) \Phi(x_i). \qquad (10)$$

Using equations (10) and (8), the generic equation can be rewritten as

$$\begin{aligned} f(x) &= \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)(\Phi(x_i) \cdot \Phi(x)) + b, \\ &= \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) k(x_i, x) + b, \end{aligned} \qquad (11)$$

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

2.2.4. Random Forest (RF). RF [38] is an ensemble approach that relies on classification and regression trees (CART) models. The purpose of CART is to learn the relationship between a dependent (Y) and a set of predictor variables (P). The learning algorithm employs recursive partitioning which splits P variables to create homogenous grouping of Y. The recursive partitioning continues until the subset of Y (at each node) has the same value. RF differs from the CART procedure by (a) employing bootstrap resampling [44], and (b) random variable selection. Consider a regression tree which is made up of splits and nodes. In RF, a random subset of P is used to determine the split for each node. For continuous variables, the ensemble estimate is the mean of the predicted values across trees mean $(\acute{Y})$ and the variance across trees is var $(\acute{Y})$.

2.2.5. Feedforward Neural Network (FFNN). Neural networks are computing models used for recognition of pattern or relation among data [38]. Neural networks comprise of two main components: set of nodes and links between nodes.

The FFNN possesses a massive number of processing elements called neurons. These neurons are interlinked through weights. Neurons have input, output, and hidden layer(s). The summation of weighted values at the input layer is applied to each of hidden layer neurons. Similarly sum weighted values at the hidden layer is applied to the output layers. The output Y obtained (at the output layer) is given as

$$Y = \omega \left\{ \beta_0 + \sum_{i=1}^{j} \beta_i \Phi \left(\alpha_{i0} + \sum_{k=1}^{l} \alpha_{ik} A_k \right) \right\}, \qquad (12)$$

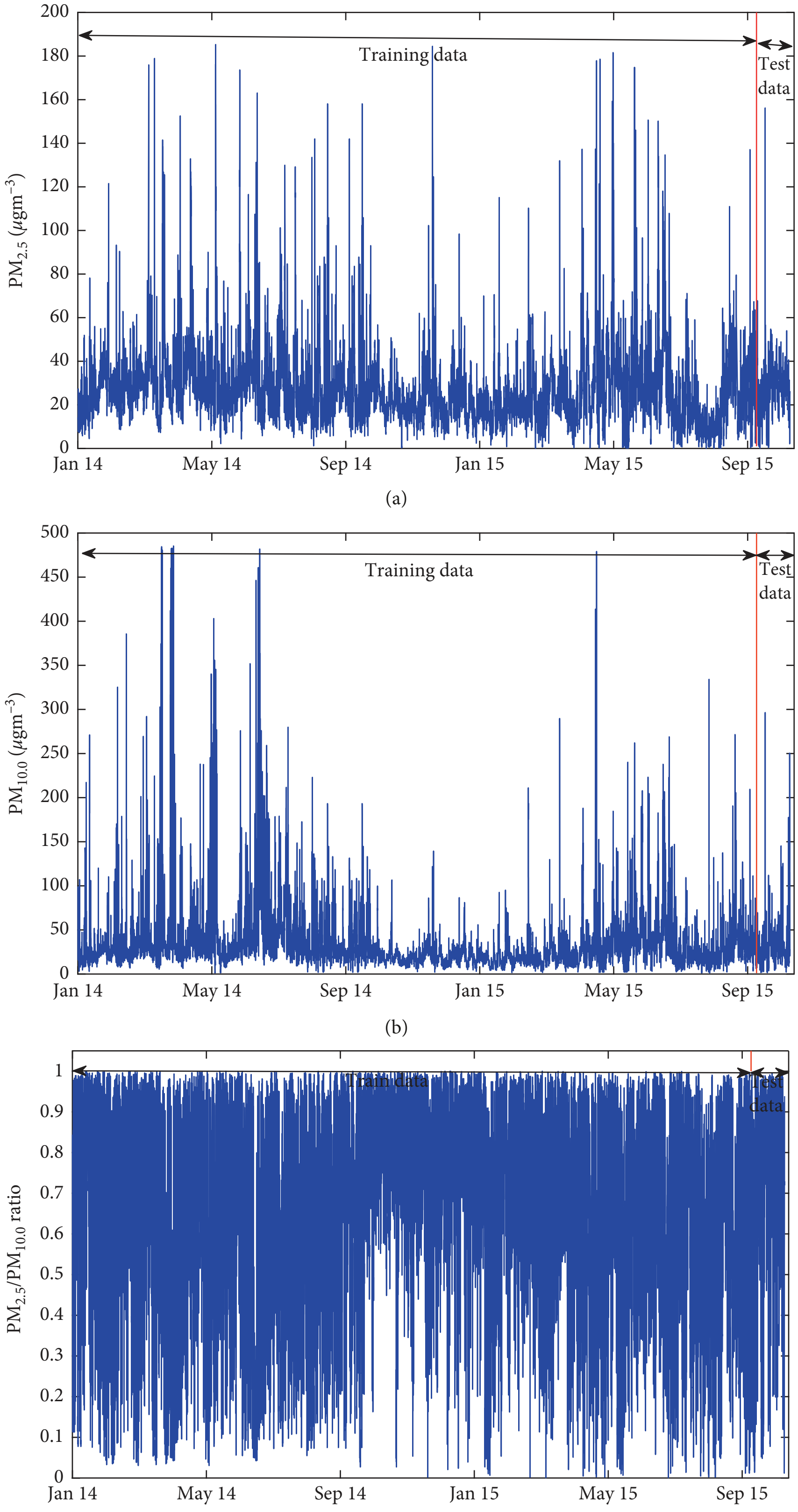
where $(\beta_0, \beta_1, \ldots, \beta_j, \alpha_{10}, \ldots, \alpha_{jl})$ are the bias and weight parameters. Φ and ω are the activation functions applied at hidden and output layer, respectively. A_k indicates the input value at the input neuron k.

2.2.6. Performance Evaluation Measures. The root-mean-squared error (RMSE) and mean absolute error (MAE) have widely been used to measure the performance of predicted models. The range of both measures is from 0 to ∞, and their lowest values show that the performance of the predicted model is better. RMSE can be calculated by taking square root of mean squared error (MSE). It can provide the complete scenario of the error distribution. MAE can be calculated by taking average of absolute differences between the actual and predicted values. Mathematically, RMSE and MAE can be calculated using equations (13) and (14), respectively:

$$\text{RMSE} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (X_t - P_t)^2}, \qquad (13)$$

$$\text{MAE} = \frac{1}{T} \sum_{t=1}^{T} (|X_t - P_t|), \qquad (14)$$

where X_t represents the target (expected) values and P_t is the model's predicted values.

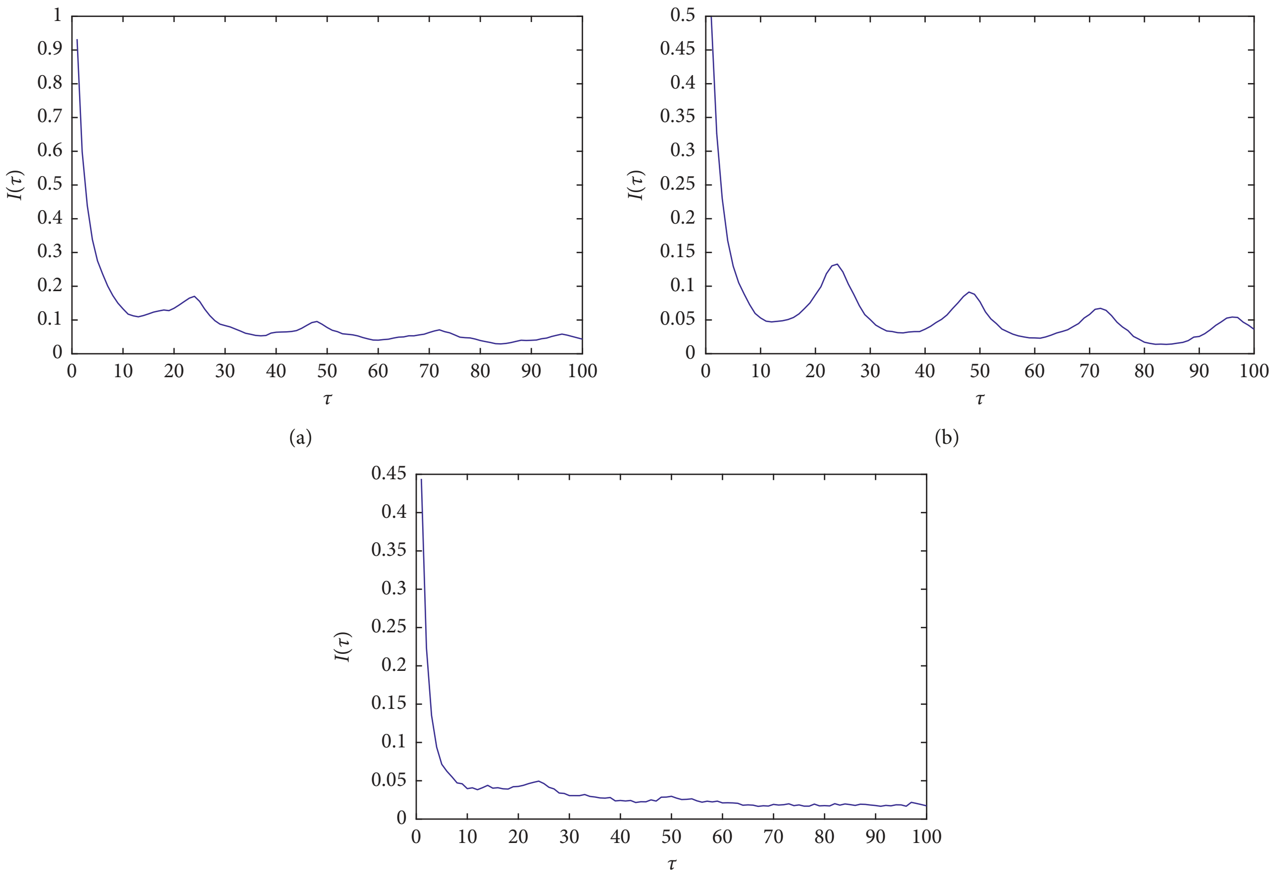


(a)

(b)

(c)

Figure 2: Hourly data of $PM_{2.5}$ (μgm^{-3}), $PM_{10.0}$ (μgm^{-3}), and $PM_{2.5}/PM_{10.0}$ ratio from January 2014 to September 2015 in Makkah.

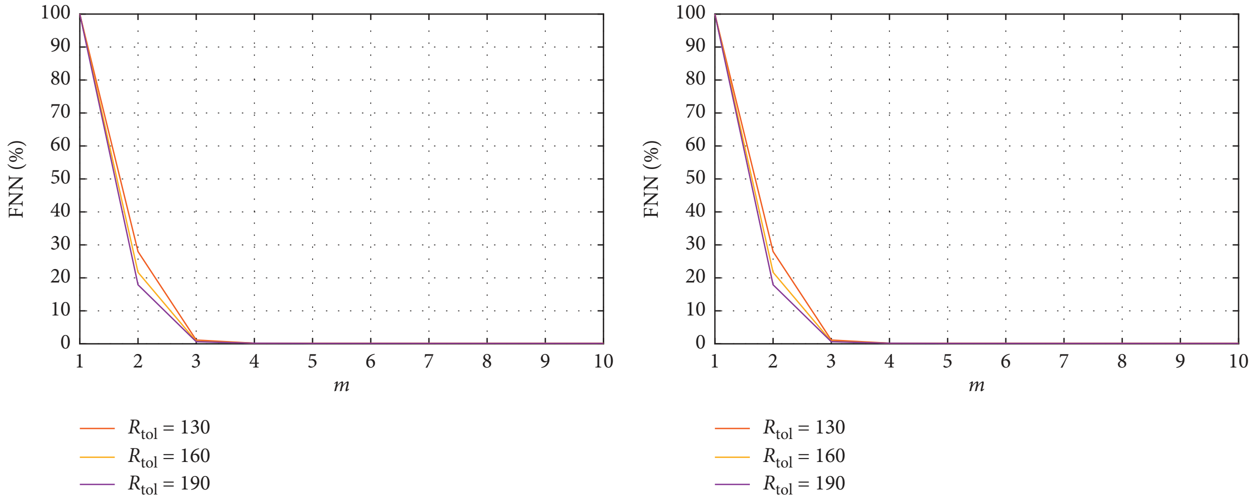


(a)

(b)

(c)

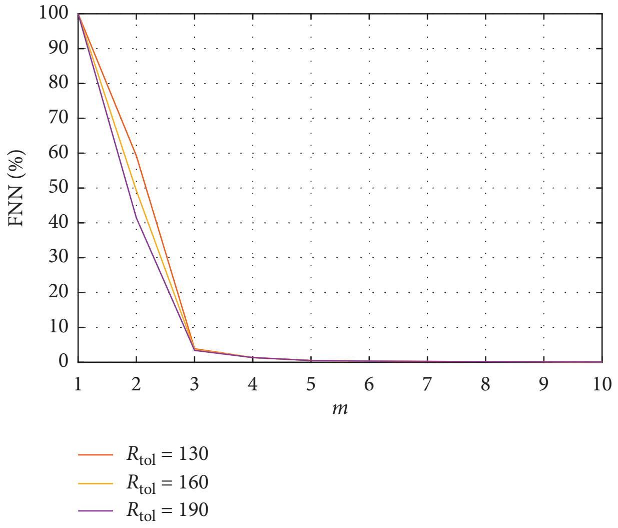
FIGURE 3: Optimal τ, for particulates (a) $PM_{2.5}$, (b) $PM_{10.0}$, and (c) $PM_{2.5}/PM_{10.0}$ ratio using the AMI method.



(a)

(b)

FIGURE 4: Continued.



(c)

FIGURE 4: FNN curves with various thresholds for determining the m of particulates (a) $PM_{2.5}$, (b) $PM_{10.0}$, and (c) $PM_{2.5}/PM_{10.0}$ ratio.

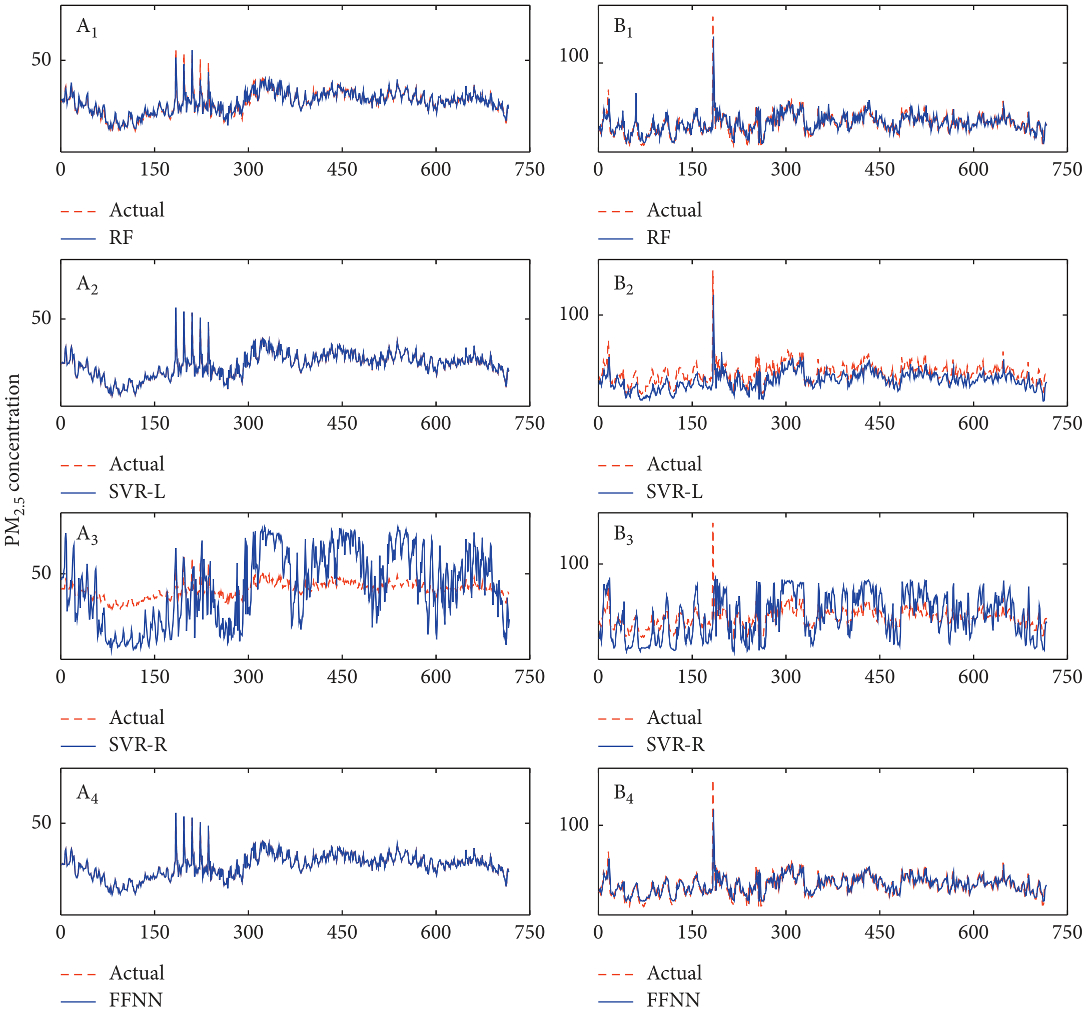


FIGURE 5: Actual and predicted values using proposed setting of the target set (A1 to A4) and traditional setting of the target set (B1 to B4) for $PM_{2.5}$ (μgm^{-3}) PSR series.

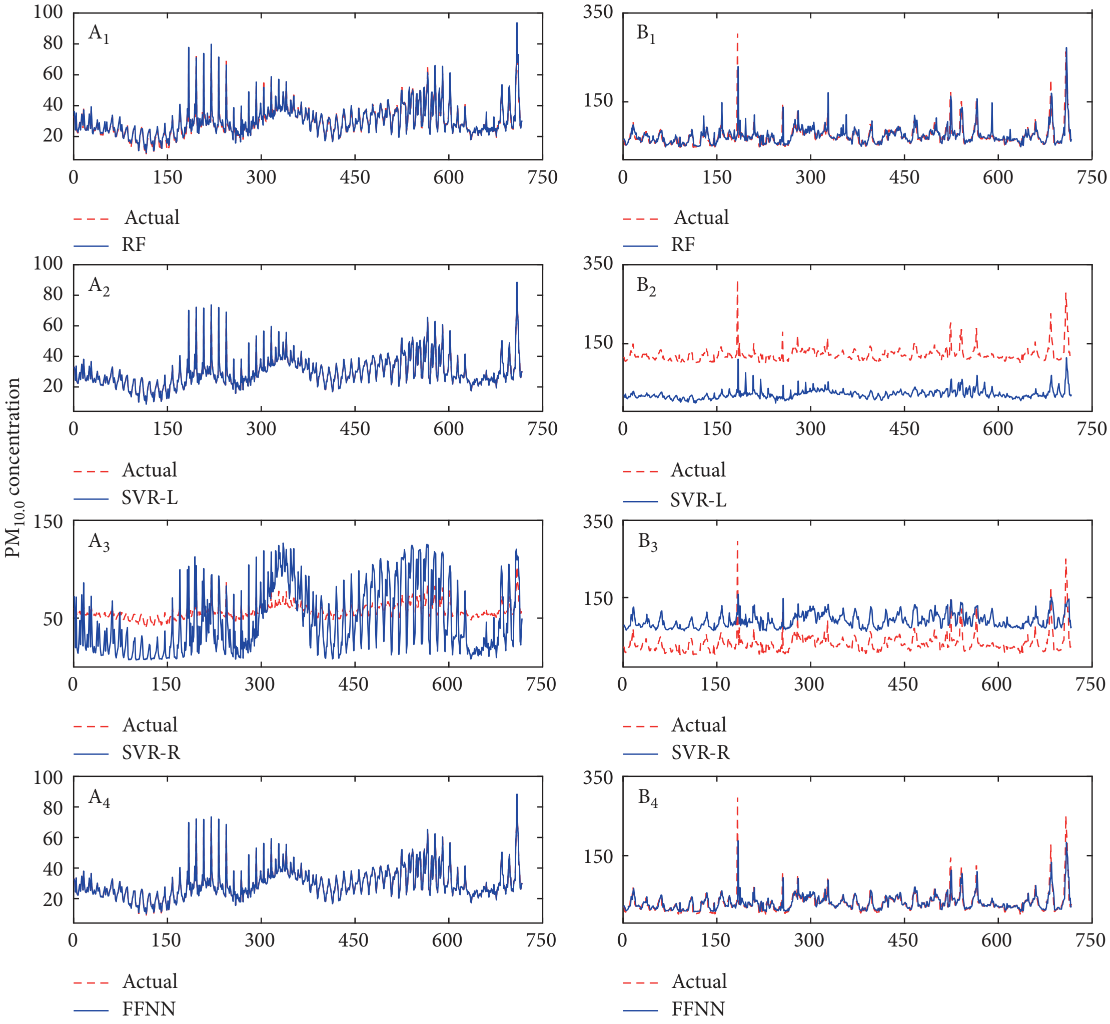


Figure 6: Actual and predicted values using proposed setting of the target set (A1 to A4) and traditional setting of the target set (B1 to B4) for $PM_{10.0}$ (μgm^{-3}) PSR series.

3. Results and Discussion

The original time series of $PM_{2.5}$, $PM_{10.0}$, and ratio of $PM_{2.5}/PM_{10.0}$ are shown in Figure 2.

First of all, phase space is reconstructed using equation (4). The selection of two parameters τ and m are important for PSR. AMI has been used for the computation of τ. In Figure 3, the AMI is plotted against varying τ, for getting the optimal value of τ for $PM_{2.5}$, $PM_{10.0}$, and $PM_{2.5}/PM_{10.0}$ ratio. The presence of chaos in atmospheric particulates reveals that time series of mass concentrations $PM_{2.5}$ and $PM_{10.0}$ can be described and predicted even if the source information is univariate time series.

Figure 3 depicts that the value of $I(\tau)$ decreases nonmonotonically with increasing τ for $PM_{2.5}$, $PM_{10.0}$, and $PM_{2.5}/PM_{10.0}$ ratio. Hence, the value of τ at which first minimum of $I(\tau)$ occurred is taken as the optimal τ. The optimal τ for $PM_{2.5}$ is 13, for $PM_{10.0}$ is 12, and for $PM_{2.5}/PM_{10.0}$ ratio is 10.

FNN approach is used to find the optimal minimum embedding (m). For any given m, the proportion of the identified FNN for all the neighbors was computed for the given τ. The percentages of the FNN are plotted as a function of the m. A zero FNN percentage indicates the minimum m.

The results of FNN approach for determining the optimum m of $PM_{2.5}$, $PM_{10.0}$, and $PM_{2.5}/PM_{10.0}$ ratio using various values for the threshold parameters R_{tol} and A_{tol} are shown in Figure 4. The value of the parameter R_{tol} is varied from 130 to 190 with a step size of 30, and $A_{\text{tol}} = 0.2 \times R_{\text{tol}}$ is used. m obtained for $PM_{2.5}$ is 5 and for $PM_{10.0}$ and $PM_{2.5}/PM_{10.0}$ ratio is 6. The higher values of m show that the mass concentration time series of $PM_{2.5}$ and $PM_{10.0}$ have dominant degrees of freedom, which indicates that atmospheric particulate dynamics are complex in nature.

Based on the values of parameters τ and m, phase space is reconstructed for $PM_{2.5}$, $PM_{10.0}$, and $PM_{2.5}/PM_{10.0}$ ratio and prediction models (using different machine learning algorithms including RF, linear, and radial SVRs and FFNN) are built using traditional and proposed settings. The predicted values for the next 1 month (i.e., September 2015) are obtained by using different learning models (RF,

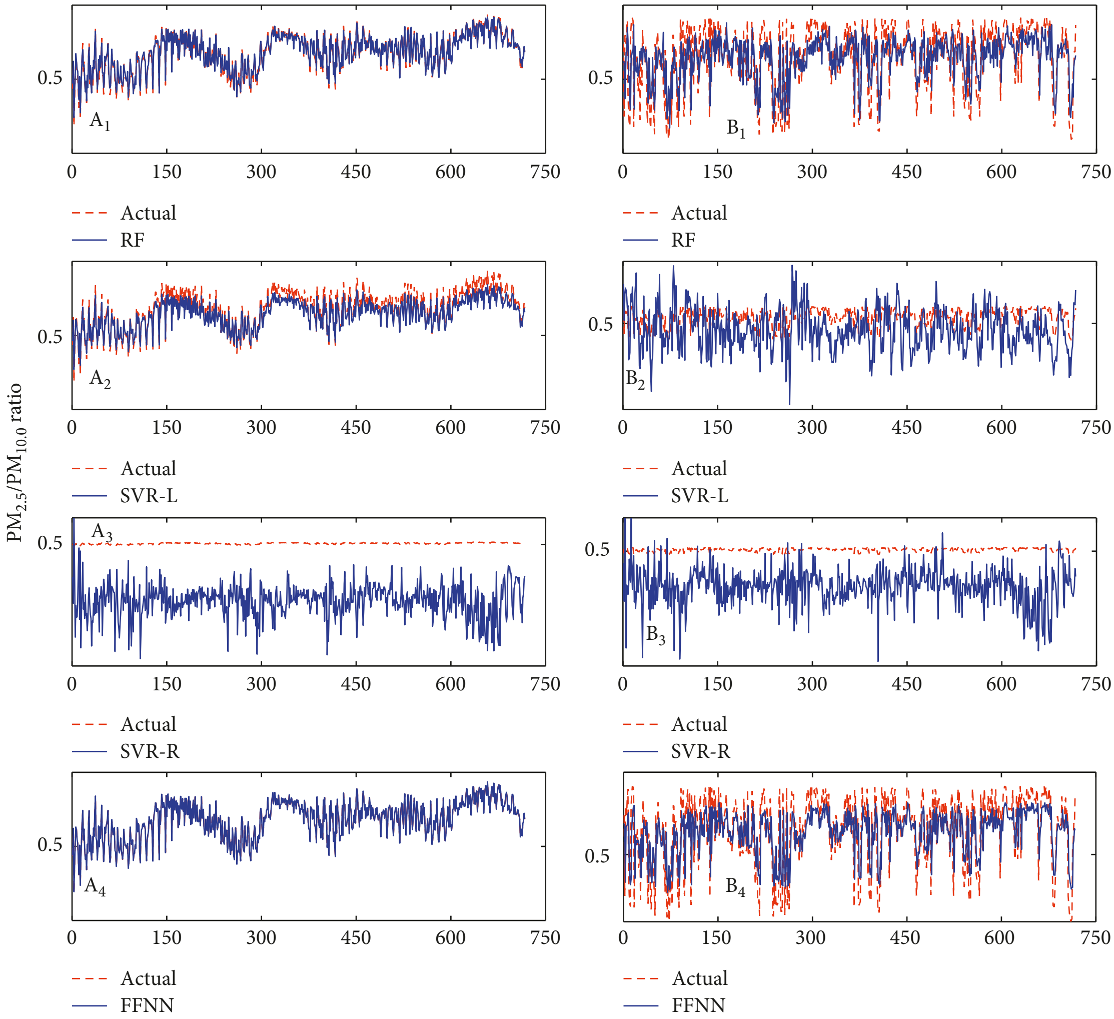


Figure 7: Actual and predicted values using proposed setting of the target set (A1 to A4) and traditional setting of the target set (B1 to B4) for $PM_{2.5}/PM_{10.0}$ ratio PSR series.

linear, and radial SVRs and FFNN). For both training and testing data, traditional and proposed settings of the target set from PSR series have been used. Target and predicted values of $PM_{2.5}$ are shown in Figure 5. It is clear from the figure that results of the proposed PSR technique are robust compared to the traditional PSR method for all learning algorithms. Learning algorithms FFNN, SVR-L, and RF show perfect overlap of the predicted and actual values for proposed settings. Figure 6 shows the prediction results of $PM_{10.0}$ for different learning algorithms using proposed and traditional settings. The results revealed that like $PM_{2.5}$, learning algorithms FFNN, SVR-L, and RF for $PM_{10.0}$ also showed perfect overlap between predicted and actual values. In case of both $PM_{10.0}$ and $PM_{2.5}$, prediction results of SVR-R are modest for both proposed and traditional settings. It can be observed from Figure 6 that SVR-L shows model underestimation and SVR-R shows model overestimation for traditional PSR settings. This may be due to that fact that in the case of traditional PSR, the target set is the last embedding PSR, whereas in the case of the proposed PSR method, the target set is the row average of the reconstructed phase space data point. The averaging process yielded better prediction and avoided model over and underestimation. Figure 7 shows the prediction results for the joined dataset (i.e., the ratio $PM_{2.5}/PM_{10.0}$). The proposed PSR showed a better prediction result, with FFNN and RF showing almost perfect overlap. The SVR-R showed model overestimation for both traditional and proposed settings.

Prediction errors between actual and predicted values in terms of RMSE and MAE are presented in Table 1. The table compares the performances of different machine learning algorithms using the proposed and traditional settings. The results depict that prediction error of all the machine learning techniques is smaller for the proposed PSR approach compared to traditional approach.

For $PM_{2.5}$, FFNN leads to best results (both RMSE and MAE 0.04 μgm^{-3}), followed by SVR-L (RMSE 0.01 μgm^{-3} and MAE 0.09 μgm^{-3}) and RF (RMSE 1.27 μgm^{-3}). For $PM_{10.0}$, SVR-L leads to best results (both RMSE and MAE 0.06 μgm^{-3}), followed by FFNN (RMSE 0.13 μgm^{-3} and MAE 0.09 μgm^{-3}). For $PM_{2.5}/PM_{10.0}$, FFNN is the best and

Table 1: Performance of the predicted models in terms of RMSE and MAE.

	$PM_{2.5}$ (μgm^{-3})				$PM_{10.0}$ (μgm^{-3})				$PM_{2.5}/PM_{10.0}$			
Models	Using proposed setting of the target set		Using traditional setting of the target set		Using proposed setting of the target set		Using traditional setting of the target set		Using proposed setting of the target set		Using traditional setting of the target set	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
RF	1.27	0.86	8.82	5.07	1.60	1.16	19.86	9.80	0.02	0.02	0.21	0.16
SVR-L	0.10	0.09	12.65	9.93	0.06	0.06	141.5	140.0	0.05	0.04	0.71	0.59
SVR-R	24.04	21.08	21.85	18.43	37.80	34.17	62.17	60.39	10.14	9.74	5.58	5.02
FFNN	0.04	0.04	8.03	4.46	0.13	0.09	18.05	8.57	0.001	0.001	0.20	0.16

accurate predictor (0.001 for both RMSE and MAE), followed by RF (0.02 for both RMSE and MAE) and SVR_L (RMSE 0.05 μgm^{-3} and MAE 0.04).

Due to the geographical characteristics and climatic conditions, $PM_{2.5}$ and $PM_{10.0}$ pollutants frequently exceed the national and international air quality standards in Makkah region [5, 32]. These particles are very tiny, and their exposure is associated with adverse health effects. According to World Health Organization (WHO), reduction in annual $PM_{10.0}$ concentration from 70 μgm^{-3} to 20 μgm^{-3} is associated with 15% reduction in deaths [3]. Exposure of these pollutants not only affects health of local community but also affects millions of pilgrims visiting Makkah annually. The current study can have implications to predict these pollutants to provide managerial solutions for the prevention and/or mitigating adverse health implications.

4. Conclusion

The traditional PSR prediction method generally uses the data points of last embedding dimensions of PSR series (single scale) as the target set. APM mass concentrations are an outcome of complex natural and anthropogenic contributors evolving with time that may operate on multiple time scales. This study has proposed a novel PSR-based scientific solution that accounts for the information contained at multiple time scales. The optimal embedding dimension of $PM_{2.5}$ is 5; for $PM_{10.0}$ and $PM_{2.5}/PM_{10.0}$ ratio, it is 6. The higher values of embedding dimensions reveal the chaotic behavior of both atmospheric particulates. Different machine learning algorithms are used to realize the prediction of APM mass concentrations using proposed and traditional PSR techniques. Performance of various learning algorithms is evaluated using RMSE and MAE. The results demonstrated that the proposed modification in PSR approach provided better prediction of APMs compared to traditional approach. The robust prediction is obtained using the FFNN learning model using the proposed modification in the PSR algorithm. The good prediction results indicate the usefulness of the proposed PSR approach and the suitability of the various machine learning approaches in combination for predicting atmospheric particulates mass concentrations. The proposed technique can be used for analyzing and prediction of interbeat interval time series, EEG time series, human gait dynamics, and financial time series data.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was partly funded by the King Abdulaziz City for Science and Technology (KACST) (grant no. 13-ENES2373-10)

References

[1] J. Colls, *Air Pollution*, Taylor and Francis, London, UK, 2nd edition, 2002.

[2] A. M. F. Mohammed, O. A. Attala, and T. M. Habeebullah, "PM_{10} and their bio-contamination in Makkah Saudi Arabia-case study," *International Journal of Biosensors & Bioelectronics*, vol. 2, no. 1, pp. 1–6, 2017.

[3] WHO, *Ambient (Outdoor) Air Quality and Health*, 2018, https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health.

[4] T. M. Habeebullah, "Health impacts of PM_{10} using AirQ2. 2.3 model in Makkah," *Journal of Basic and Applied Sciences*, vol. 8, pp. 259–268, 2013.

[5] S. K. M. Hassan, "Atmospheric polycyclic aromatic hydrocarbons and some heavy metals in suspended particulate matter in urban, industrial and residential areas in Greater Cairo", Ph.D. dissertation, Cairo University, Egypt, 2006.

[6] S. Munir, T. M. Habeebullah, A. M. F. Mohammed, E. A. Morsy, M. Rehan, and K. Ali, "Analysing $PM_{2.5}$ and its association with PM_{10} and meteorology in the arid climate of Makkah, Saudi Arabia," *Aerosol and Air Quality Research*, vol. 17, no. 2, pp. 453–464, Feb 2017.

[7] S. Munir, T. M. Habeebullah, A. R. Seroji, S. S. Gabr, A. M. F. Mohammed, and E. A. Morsy, "Quantifying temporal trends of atmospheric pollutants in Makkah (1997–2012)," *Atmospheric Environment*, vol. 77, pp. 647–655, 2013.

[8] E. Harnandez, F. Martin, and F. Valero, "Statistical forecast models for daily air particulate iron and lead concentrations for Madrid, Spain," *Atmos. Environ*, vol. 26, no. 1, pp. 107–116, 1992.

[9] P. Pérez, A. Trier, and J. Reyes, "Prediction of $PM_{2.5}$ concentrations several hours in advance using neural networks in

Santiago, Chile," *Atmospheric Environment*, vol. 34, no. 8, pp. 1189–1196, 2000.

[10] A. B. Chelani, D. G. Gajghate, S. M. Tamhane, and M. Z. Hasan, "Statistical modeling of ambient air pollutants in Delhi," *Water, Air, and Soil Pollution*, vol. 132, no. 3-4, pp. 315–331, 2001.

[11] A. B. Chelani, D. G. Gajghate, and M. Z. Hasan, "Prediction of ambient PM_{10} and toxic metals using artificial neural networks," *Journal of the Air & Waste Management Association*, vol. 52, no. 7, pp. 805–810, 2002.

[12] U. Schlink, O. Herbarth, and G. Tetzlaff, "A component time-series model for SO_2 data: forecasting, interpretation and modification," *Atmospheric Environment*, vol. 31, no. 9, pp. 1285–1295, 1997.

[13] U. Kumar, A. Prakash, and V. K. Jain, "Characterization of chaos in air pollutants: a Volterra-Wiener-Korenberg series and numerical titration approach," *Atmospheric Environment*, vol. 42, no. 7, pp. 1537–1551, 2008.

[14] G. A. Salini and P. Pérez, "A study of the dynamic behaviour of fine particulate matter in Santiago, Chile," *Aerosol and Air Quality Research*, vol. 15, no. 1, pp. 154–165, 2015.

[15] E. Bradley and H. Kantz, "Nonlinear time-series analysis revisited," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 25, no. 9, p. 097610, 2015.

[16] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence*, vol. 898, pp. 366–381, Lecture Notes in Mathematics, Springer, New York, NY, USA, 1981.

[17] C. Frazier and K. M. Kockelman, "Chaos theory and transportation systems: instructive example," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1897, no. 1, pp. 9–17, 2004.

[18] I.-F. Li, P. Biswas, and S. Islam, "Estimation of the dominant degrees of freedom for air pollutant concentration data: applications to ozone measurements," *Atmospheric Environment*, vol. 28, no. 9, pp. 1707–1714, 1994.

[19] J.-L. Chen, S. Islam, and P. Biswas, "Nonlinear dynamics of hourly ozone concentrations," *Atmospheric Environment*, vol. 32, no. 11, pp. 1839–1848, 1998.

[20] K. Koçak, L. Şaylan, and O. Şen, "Nonlinear time series prediction of O_3 concentration in Istanbul," *Atmospheric Environment*, vol. 34, no. 8, pp. 1267–1271, 2000.

[21] A. B. Chelani, R. N. Singh, and S. Devotta, "Nonlinear dynamical characterization and prediction of ambient nitrogen dioxide concentration," *Water, Air, and Soil Pollution*, vol. 166, no. 1–4, pp. 121–138, 2005.

[22] A. B. Chelani and S. Devotta, "Nonlinear analysis and prediction of coarse particulate matter concentration in ambient air," *Journal of the Air & Waste Management Association*, vol. 56, no. 1, pp. 78–84, 2006.

[23] A. Chelani and S. Devotta, "Air quality forecasting using a hybrid autoregressive and nonlinear model," *Atmospheric Environment*, vol. 40, no. 10, pp. 1774–1780, 2006.

[24] B. Yu, C. Huang, Z. Liu, H. Wang, and L. Wang, "A chaotic analysis on air pollution index change over past 10 years in Lanzhou, northwest China," *Stochastic Environmental Research and Risk Assessment*, vol. 25, no. 5, pp. 643–653, 2011.

[25] S. Saeed, W. Aziz, M. Rafique, I. Ahmad, K. J. Kearfott, and S. Batoolb, "Quantification of non-linear dynamics and chaos of ambient particulate matter concentrations in Muzaffarabad city," *Aerosol and Air Quality Research*, vol. 17, no. 3, pp. 849–856, Mar 2017.

[26] Q. Yan, S. Wang, and B. Li, "Forecasting uranium resource price prediction by extreme learning machine with empirical mode decomposition and phase space reconstruction," *Discrete Dynamics in Nature and Society*, vol. 2014, Article ID 390579, 10 pages, 2014.

[27] L. Shen, Y. Wen, and X. Li, "Improving prediction accuracy of cooling load using EMD, PSR and RBFNN," *Journal of Physics: Conference Series*, vol. 887, no. 1, article 012016, 2017.

[28] H. Z. Liu, S. L. Wang, and J. Y. Liu, "LS-SVM prediction model based on phase space reconstruction for dam deformation," *Advanced Materials Research*, vol. 663, pp. 55–59, 2013.

[29] M. Costa, A. L. Goldberger, and C. K. Peng, "Multiscale entropy analysis of biological signals," *Physical Review E*, vol. 71, no. 2, article 021906, 2005.

[30] I. Awan, W. Aziz, I. H. Shah et al., "Studying the dynamics of interbeat interval time series of healthy and congestive heart failure subjects using scale based symbolic entropy analysis," *PLoS One*, vol. 13, no. 5, Article ID e0196823, 2018.

[31] M. Costa, C. K. Peng, A. L. Goldberger, and J. M. Hausdorff, "Multiscale entropy analysis of human gait dynamics," *Physica A: Statistical Mechanics and its Applications*, vol. 330, no. 1-2, pp. 53–60, 2003.

[32] M. Costa, A. A. Priplata, L. A. Lipsitz et al., "Noise and poise: enhancement of postural complexity in the elderly with a stochastic-resonance-based therapy," *Europhysics Letters (EPL)*, vol. 77, no. 6, article 68008, 2007.

[33] N. Othman, M. Z. Mat-Jafri, and L. H. San, "Estimating particulate matter concentration over arid region using satellite remote sensing: a case study in Makkah, Saudi Arabia," *Modern Applied Science*, vol. 4, no. 11, 2010.

[34] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*, Elsevier, Amsterdam, Netherlands, 2011.

[35] B. Çığşar and D. Ünal, "Comparison of data mining classification algorithms determining the default risk," *Scientific Programming*, vol. 2019, Article ID 8706505, 8 pages, 2019.

[36] H. Zhang, L. Zhou, and J. Zhu, "Study on financial time series prediction based on phase space reconstruction and support vector machine (SVM)," *American Journal of Applied Mathematics*, vol. 3, no. 3, pp. 112–117, 2015.

[37] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," *Advances in Neural Information Processing Systems*, vol. 9, pp. 155–161, 1997.

[38] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[39] G. P. Zhang, *Neural Networks for Time-Series Forecasting*, Springer, Berlin, Heidelberg, 2012.

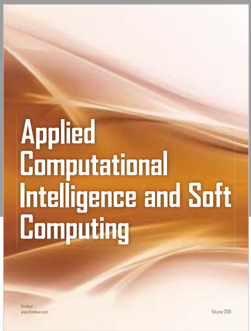
[40] P. Kline, *The New Psychometrics: Science, Psychology and Measurement*, Routledge, Abingdon, UK, 1998.

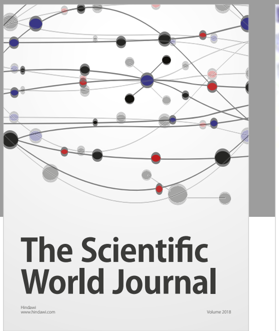
[41] A. Olinsky, S. Chen, and L. Harlow, "The comparative efficacy of imputation methods for missing data in structural equation modeling," *European Journal of Operational Research*, vol. 151, no. 1, pp. 53–79, 2003.

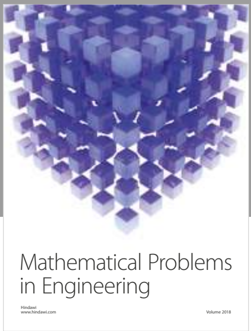
[42] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," *Physical Review A*, vol. 33, no. 2, pp. 1134–1140, 1986.

[43] M. B. Kennel, R. Brown, and H. D. I. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," *Physical Review A*, vol. 45, no. 6, pp. 3403–3411, 1992.

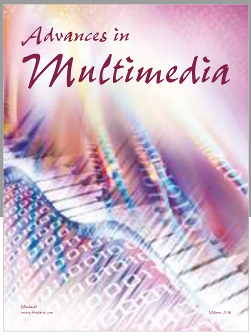
[44] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, CRC Press, Boca Raton, FL, USA, 1994.

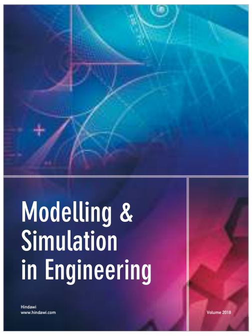

Applied Computational Intelligence and Soft Computing

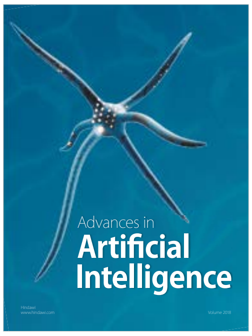

The Scientific World Journal


Mathematical Problems in Engineering

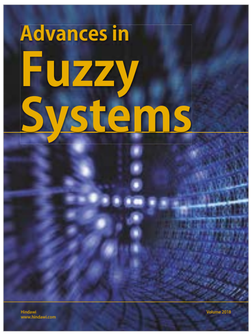
Journal of Engineering

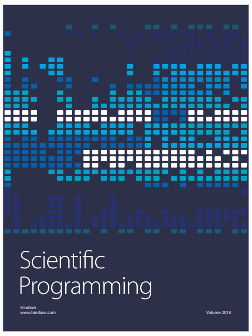

Advances in Multimedia


Modelling & Simulation in Engineering

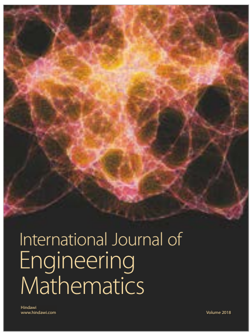

Advances in Artificial Intelligence

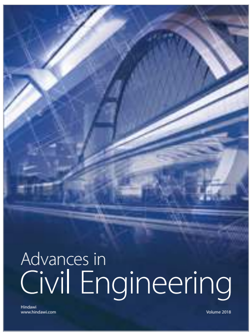
International Journal of Reconfigurable Computing

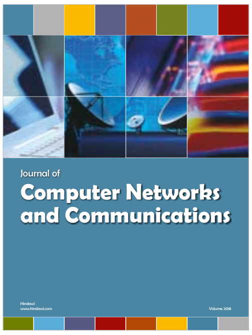

Advances in Fuzzy Systems


Scientific Programming

Advances in Human-Computer Interaction


International Journal of Engineering Mathematics


Advances in Civil Engineering

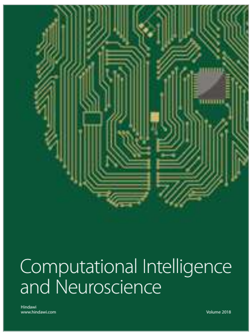

Journal of Computer Networks and Communications

Journal of Robotics

International Journal of Computer Games Technology

International Journal of Biomedical Imaging

Journal of Electrical and Computer Engineering


Computational Intelligence and Neuroscience