Received: 26 April 2025 | Revised: 2 September 2025 | Accepted: 22 October 2025 | Published online: 28 July 2026

RESEARCH ARTICLE

Journal of Data Science and Intelligent Systems
2026, Vol. 00(00) 1-19
DOI: 10.47852/bonviewJDSIS62026021

BON VIEW PUBLISHING

Wavelet-Based Combinatorial Machine Learning Techniques for Forecasting Particulate Matter Concentrations

Sandip Garai[1], Debopam Rakshit[2,3,*], Arkaprava Roy[4], Ranjit Kumar Paul[3] and Ritwika Das[3]

[1] *ICAR-Indian Institute of Agricultural Biotechnology, India*

[2] *ICAR-Indian Veterinary Research Institute, India*

[3] *ICAR-Indian Agricultural Statistics Research Institute, India*

[4] *ICAR-National Institute of Biotic Stress Management, India*

Abstract: Air pollution in Delhi is so severe that, in most years, the city's average daily air quality index (AQI) consistently falls below the "Poor" category. Efficient and reliable prediction models can provide a guide to effective air pollution control. In this study, daily concentrations of $PM_{2.5}$ and PM_{10} (particle sizes ≥ 2.5 μm and ≥ 10 μm, respectively) in a residential area (Alipur), an industrial area (Okhla), and a traffic area (Pusa) of Delhi have been modeled using wavelet-based machine learning (ML) algorithms. Wavelet methodology with the help of filter bank decomposes a series into various sub-parts to gather information at multiresolution levels. Performance of various important wavelet filters namely, haar, Daubechies (d4), Coiflet (c6), best-localized Daubechies (bl14), and least asymmetric (la8) have been recorded to obtain the best combination for forecasting particulate matter (PM) levels one week and one month in advance. The proposed models have performed well in predicting PM concentration.

Keywords: air pollution, machine learning, nonlinearity, particulate matter, wavelet decomposition

1. Introduction

Exposure to polluted air is the most serious environmental health risk in the present-day world [1]. The estimated global cost attributed to the harmful impacts of fossil fuel burning is 2.9 trillion USD [2]. Among all the air pollutants, particulate matter (PM) has the most damaging consequences. It consists of microscopic solid or liquid droplets, originating from either anthropogenic or natural sources that enter the human body with the air we inhale. Particles having 2.5 to 10 μm diameter (PM_{10}) can invade deep inside one's lung, while, those smaller than 2.5 μm ($PM_{2.5}$) have the capacity to penetrate the lung barrier and enter the circulatory system. Although the estimates of mortality caused by PM span a wide range, one of the recent studies by Vohra et al. [3] shows that fossil-fuel-generated $PM_{2.5}$ causes approximately 8.7 million premature deaths worldwide. India, being the third largest emitter of greenhouse gases [4], is both a significant contributor and a victim of air pollution. In fact, none of the Indian cities has met the standard for annual $PM_{2.5}$ concentration of 5 μg m^{-3} specified in the World Health Organization's Air Quality Guideline [5].

Air pollution jeopardizes the global economy in multiple ways. Primarily, it skyrockets healthcare costs by increasing the risk of asthma, other respiratory illnesses, stroke, diabetes, potential cardiovascular diseases, impaired neurodevelopment, and cancer. Inhalation of polluted air by pregnant women can contaminate the fetal tissues with PMs [6]. This sometimes triggers the preterm birth of babies with numerous health conditions which require prolonged medical attention, sometimes throughout the lifespan of the sufferer [7]. The indirect economic repercussions of air pollution involve the loss of work hours of adults and reduced learning time of children due to sick leaves or, in extreme cases, demises. According to a global estimate, fossil-fuel-generated $PM_{2.5}$ pollution was responsible for 1.8 billion days of work absence in 2018, which, in monetary terms, was translated into 100 billion USD [2]. Air pollution-induced spikes in global temperature fuel climatic disasters like drought, destructive wildfires, heatwaves, and hurricanes which have considerable costs attached to them in terms of the perils they bring about. The total economic burden that air pollution leaves on India is equivalent to 5.4% of its gross domestic product (GDP) [2], which is 3.6 times of its healthcare budget for 2021–22. In 2019, 17.8% of deaths in this country were imputed to air pollution, 58.7% of which were caused by outdoor particulate matters [8]. It has been well documented that every year for the last few decades, the average annual $PM_{2.5}$ concentration in Delhi crosses the annual National Ambient Air Quality Standards (NAAQS) of 40 μg m^{-3}, which is set eight times higher than the WHO's standard limit [1]. The premature death and morbidity caused by air pollution compelled the capital state of India to incur an estimated implicit expense of 1207 million USD in 2019. That, on an equivalent per capita basis (62 USD), was the highest among all the states and union territories of India [8].

Consequently, a slew of measures were put into place to lower pollution levels in Indian cities. For instance, the National Clean Air Program (NCAP) was implemented by the Ministry of Environment, Forest and Climate Change, Government of India in 2019 with the aim to implement a city-, region-, and state-specific clean air action plan, to perform source apportionment studies, to strengthen air quality

*Corresponding author: Debopam Rakshit, ICAR-Indian Veterinary Research Institute and ICAR-Indian Agricultural Statistics Research Institute, India. Email: debopam.rakshit@icar.org.in

© The Author(s) 2026. Published by BON VIEW PUBLISHING PTE. LTD. This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0/).

monitoring, and to reduce PM concentration by 20–30% by 2024 in all the non-attainment cities (cities that have not been meeting the NAAQS for more than five years). Mission Lifestyle for Environment (LiFE), proposed by India at the 2021 UN Climate Change Conference, intends to bring about positive behavioral changes in each citizen to promote an environmentally friendly lifestyle. Specifically, in the National Capital Region (NCR), the Commission on Air Quality Management (CAQM) has implemented a set of emergency measures—Graded Response Action Plan (GRAP)—to prevent further deterioration of air quality beyond a certain threshold. The major interventions under GRAP include, inter alia, factory closures, the overhaul of the public transportation system, the use of greener fuel, the restricting construction and demolition activities, and prohibiting entry of medium and heavy vehicles [9]. Despite the acknowledgement of and measures taken against the problem by policymakers at multiple levels, Delhi still grapples with the issue of air pollution. Even after the adoption of a variety of interventions by the state, the concentration of $PM_{2.5}$ remained exceptionally above the NAAQS guidelines.

Prediction of future PM levels plays a crucial role in policy formulation of any country, especially those falling under low- and medium-income groups. Dimitrova et al. [4] showed that, under the current policy regime, premature deaths due to $PM_{2.5}$ exposure in India are going to increase by 5.6 million between 2010 and 2050. Nonetheless, the total premature deaths can be averted to a magnitude of 20.8 million if stringent policies are implemented which aim both at climate change mitigation and the maximum feasible reduction in air pollution.

Researchers all over the world are continuously working on several prediction measures of PM. Machine learning (ML) and deep learning (DL) algorithms are being explored by researchers and stakeholders all over the world for predicting several air pollutants including PM as well as for stochastic risk assessment [10, 11].

In various comparison studies, algorithms such as artificial neural network (ANN) [12], random forest (RF) [13], and long short-term memory (LSTM) [14] have been found to perform satisfactorily. The literature also includes applications of various extensions of these algorithms [15–17]. Ensemble ML methods have been employed for modeling and forecasting $PM_{2.5}$ concentrations [18, 19]. Studies have also explored the use of hybrid models, such as the XGBoost-GARCH-MLP model [20], as well as standalone models like XGBoost [21], for predicting $PM_{2.5}$ levels. Additionally, numerous works have integrated ML and DL algorithms with a range of input variables, including meteorological parameters [22–25], satellite data [26], other pollutants [27], and various associated factors [28], to enhance the prediction of PM concentrations.

Previously, very few studies have used decomposition-based models for PM concentration modeling. Wavelet decomposed statistical and ML algorithms such as autoregressive integrated moving average (ARIMA), ANN, support vector machine (SVM), and back propagation neural network (BPNN) have been used for forecasting daily $PM_{2.5}$ concentrations in various cities of China [29, 30] and have been found to generate better forecasting accuracy as compared to individual models. Liu et al. [31] utilized 1-hour and 24-hour $PM_{2.5}$ concentration data in wavelet packet decomposition (WPD) to create sublayers. The sub-layers in the first case (1-hour) were filtered by using a stacked auto encoder (SAE) to be utilized in bi-directional LSTM (Bi-LSTM) for forecasting, and in the second case, a direct forecast by Bi-LSTM was obtained. Forecasts were combined by a non-dominated sorting genetic algorithm for the final forecast. Wavelet-coupled with SAE and LSTM models for forecasting PM_{10} concentrations could develop better 2-step ahead forecasting accuracy than Bi-LSTM and LSTM models [32]. To the best of the authors' knowledge, the interaction of various wavelet filters with ML algorithms for the prediction of air pollutants is not yet explored. Hence, in this study, different combinations of wavelet filters and ML algorithms have been used with two rolling windows for the prediction of $PM_{2.5}$ and PM_{10}. The present work explores the trend of PM_{10} and $PM_{2.5}$ at three different sites of Delhi having different pollution signatures using different models. This study can help policymakers undertake location-specific actions based on meteorological forecasts to reduce the cost and improve the effectiveness of policy actions to reduce the effect of crop residue burning on the ambient $PM_{2.5}$ concentration in Delhi in post-monsoon seasons.

In Section 1, the introduction, brief literature review and research gap have been pointed out. In Section 2, a brief discussion about several stochastic and ML models used in this study has been discussed. The motivation and the proposed workflow have also been discussed here. Empirical performance has been evaluated in Section 3. In the end, the paper has been concluded with key findings and notes on future research work.

2. Material and Methods

Brief descriptions of various statistical and ML models for capturing the volatility and predicting daily average concentrations of $PM_{2.5}$ and PM_{10} have been given.

2.1. Autoregressive Integrated Moving Average (ARIMA)

The ARIMA model [33] can be denoted as ARIMA (p, d, q). Here, p, d and q represent the number of autoregressive (AR) terms, order of integration or differencing (I), and order of moving average (MA), respectively. If a univariate and linear time series (y_t) follows ARIMA (p, d, q), then the general representation will be as follows:

$$\varphi(L)(1-L)^d y_t = \theta(L)\varepsilon_t \tag{1}$$

where, y_t and ε_t are the actual observed value and error at time t, respectively and $\varepsilon_t \sim IID(0,\sigma^2)$. The order of differencing, d is considered an integer in the ARIMA model. $\varphi(L)$ and $\theta(L)$ are lag operator polynomials L of order p and q respectively.

2.2. Generalized Autoregressive Conditional Heteroscedasticity (GARCH)

The ARIMA model cannot capture the non-linearity present in the time series data and cannot solve the problem of conditional heteroscedasticity present in the time series data. Hence, the GARCH model [34] is used where the conditional variance is represented as a linear function of its lags as well as previous shocks. The conditional variance, h_t of a GARCH (p, q) can be represented by Equation (2).

$$h_t = a_0 + \sum_{i=1}^{q} a_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j h_{t-j} \tag{2}$$

Here, $\varepsilon_t|\psi_{t-1} \sim N(0,h_t)$ and $\varepsilon_t = \sqrt{h_t}v_t$; v_t is innovation and $v_t \sim IID(0,1)$. ψ_{t-1} is available information up to t−1 time point. $a_0 > 0$, $a_i \geq 0$ $\forall i$ and $b_j \geq 0$ $\forall j$. In the case of weekly stationary GARCH process, it satisfies

$$\sum_{i=1}^{q} a_i + \sum_{j=1}^{p} \beta_j < 1 \tag{3}$$

2.3. Stepwise Multiple Linear Regression (SMLR)

Multiple linear regression (MLR) is useful for predicting the regressed variable based on more than one independent regressors and there is a linear relationship between regressors and response variable. MLR model can be denoted as in Equation (4).

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_p x_{pi} + \varepsilon_i \tag{4}$$

where, y_i are the responses, $x_1, x_2, \ldots, x_p$ are independent variables, $b_0, b_1, b_2, \ldots, b_p$ are regression coefficient, $\varepsilon_i \sim IID(0, \sigma^2)$, and $i = 1, 2, \ldots, n$ are the number of observations. In the case of SMLR, the model is generated by selectively adding or deleting independent variables one at a time and checking for statistical significance in an iterative way. Hence, significant features or explanatory variables can be chosen. This method avoids overfitting by removing irrelevant predictors.

2.4. Multivariate Adaptive Regression Spline (MARS)

MARS models are useful for handling non-linearity in data. It is a non-parametric model [35] that combines both regression and spline fitting to predict a continuous dependent variable. Opposite to stepwise linear regression, MARS can be considered a piecewise linear regression method. MARS algorithm is performed by moving forward and backward. Initially, a significant number of basis functions (BFs) are included in the model in the forward iterative stage leading to overfitting of the data. In the backward stage, spline functions partition the dataset into multiple piecewise linear segments and these spline functions are connected by some fixed points called knots. In this stage, the final optimum number of functions is selected by removing unnecessary ones based on the generalized cross-validation technique, thereby increasing the overall prediction accuracy of the model. The general equation of a MARS model is given in Equation (5).

$$f(x) = \delta_0 + \sum_{m=1}^{M} \delta_m h_m(X) \tag{5}$$

In Equation (5), $h_m(X)$ is spline function. δ_0, δ_m are coefficients calculated from the minimum sum of squared errors obtained from the spline function, and M denotes the number of BFs.

2.5. Principal Component Regression (PCR)

The PCR method is based on the principal component analysis (PCA) technique. Principal components (PCs) are linear combinations of original predictor variables and that they are uncorrelated. Suppose $x_1, x_2, \ldots, x_n$ are n original predictor variables, then PCs are calculated as shown in Equations (6)–(8).

$$y_1 = a_{11} x_1 + a_{12} x_2 + \cdots + a_{1n} x_n \tag{6}$$

$$y_2 = a_{21} x_1 + a_{22} x_2 + \cdots + a_{2n} x_n \tag{7}$$

$$\vdots$$

$$y_n = a_{n1} x_1 + a_{n2} x_2 + \cdots + a_{nn} x_n \tag{8}$$

PCs are represented as y_i's subjected to $cov\ (y_i, y_j) = 0 \forall$ $i \neq j = 1, 2, \ldots, n$ and the sum of square of the loadings in each PC must be unity. Among these n PCs, only a few are selected such that cumulatively they can explain around 80–90% of the variability present in the dataset. In the PCR method, regression analysis is performed using this subset of PCs as explanatory variables in place of original input variables. Two advantages of using the PCR technique are dimension reduction and overcoming the multicollinearity problem present in the dataset.

2.6. Support Vector Regression (SVR)

SVR is an important ML-based algorithm applicable for both classification and regression studies [36]. SVR is useful for predicting complex and high-dimensional continuous response variables. This method is based on structural risk minimization demonstrated by Equations (9) and (10) subject to some constraints represented in Equation (11).

$$y = k(z) = v\phi(z) + c \tag{9}$$

$$\begin{aligned} &\text{Minimize :} \\ &\tfrac{1}{2}\|v\|^2 + P\ \textstyle\sum_{b=1}^{l} \zeta_b + \zeta_b^*\ \ ;\ \ \text{where}\ \ \ P > 0 \end{aligned} \tag{10}$$

$$\begin{aligned} &\text{Subject to :} \\ &y_b - (v\phi(z_b) + c_b) \leq \varepsilon + \zeta_b \\ &(v\phi(z_b) + c_b) - y_b \leq \varepsilon + \zeta_b^* \\ &\zeta_b, \zeta_b^* \geq 0 \end{aligned} \tag{11}$$

In the above equations, $z = (z_1, z_2, \ldots, z_l)$ are input variables; y_b is predicted value of the output variable; v, c and l are constants where, $v \in R^l$ and $c \in R$; P is the penalty factor; ζ_b, ζ_b^* are slack variables, and ε is constant. In Equation (10), the first term $\frac{1}{2}\|v\|^2$ is known as the regularized term and it measures the flatness of the function. The remaining term is the empirical error estimated by Vapnik ε-insensitive loss function. In SVR, Vapnik ε-insensitive loss function is minimized to predict the output variable. If the measured values fall outside the error bound, then no solution is obtained from the model. Support vectors are the points which fall outside the error bound. Various kernel functions like the Gaussian function, Radial basis function, and Sigmoid function are used in SVR to solve complex and non-linear problems.

2.7. Random Forest (RF)

RF [37] is an important ML method which does an ensemble (bagging) of decision trees to predict the response variable from a set of inputs. From the initial training dataset, many decision trees (viz., regression trees) are generated. Each tree is generated using a bootstrap replicate of random homogeneous samples from the original dataset. From each candidate tree, predictions are obtained and finally the mean of predicted responses from all trees is calculated and denoted as an estimated dependent variable from the random forest model. Randomization reduces the correlation between trees in RF in two ways. First, a bootstrapped subset is used to train each tree. Second, each node's splitting feature is chosen at random from a subset of size m rather than from among all potential features. Because the RF algorithm creates each of the trees separately, parallelization is a breeze. It builds a complete binary tree with the maximum depth for every tree. Thus, RF performance efficiency is realized. The random forest prediction can be estimated using the following formula given in Equation (12).

$$y = \frac{1}{k} \sum_{i=1}^{k} y_i \tag{12}$$

Here, k is the total number of candidate regression trees, y_i is prediction from the ith tree and y is the final random forest prediction. The performance of the random forests is significantly influenced by the number of predictor variables used. While using fewer predictor variables might lead to higher error and lower prediction accuracy, using more predictor variables could make the algorithm more complicated. More details and application of the RF and SVR can be found in articles by Barthwal et al. [38] and Zhang et al. [39].

2.8. Artificial Neural Networks (ANN)

ANNs are very powerful ML models which are applicable when the underlying relationship between variables is unknown. It follows

a non-linear data-driven self-adaptive approach to learn, i.e., learning by example. The training process of an ANN is like the human brain. During training, predictors and corresponding outputs are provided to the model. The model can recognize inherited correlations among input variables and alter inner bounds repeatedly until it achieves the desired results. After training, the model can be used to predict the output for a new independent dataset. ANN consists of three layers, i.e., the input layer, hidden layer, and output layer. In each layer, several nodes are present. In the hidden layer, a non-linear activation function is used to handle the non-linearity present in the input dataset. The basic unit of an ANN is called a neuron. The working formulae of a network can be represented using Equation (13).

$$y = \theta\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{13}$$

Here, x_i's denotes input variables, w_i represents weightage related to ith input, b is bias, θ is activation function and y is the prediction from the model. Complex connections between different layers of an ANN need not be known. These models are robust, fault-tolerant, and have a high capability of generalization. From the erroneous, incomplete feature set, it can give true predictions. ANNs can perform data processing in a parallel manner, hence it has very high computation speed compared to other ML models. Yadav et al. [40] provided a thorough and insightful review of the application of various ANN architectures for the prediction of air pollutant concentrations.

2.9. Wavelets

The working principle of wavelet transform is to use certain high and low pass filters with a series so that it can be fragmented into several subseries containing information at different resolutions [41]. The selection of wavelet filters (haar, d4, c6, bl14, la8) was based on their complementary characteristics for time series analysis: Haar provides computational simplicity and edge detection, Daubechies (d4) offers good time-frequency localization, Coiflet (c6) provides symmetric properties, and bl14/la8 offer better frequency selectivity [42, 43]. More details about the different wavelet filters can be found in literature by Daubechies [44] and Mallat [45]. Levels (L) of decomposition are fixed according to the number of observations (N) in the series ($L \leq logN$, logarithm is of base 2). Decomposition of the series firstly creates detail and smooth or approximate coefficients. The approximate component in the next step is utilized as the original series and decomposition takes place in a similar fashion until it completes all levels of decomposition.

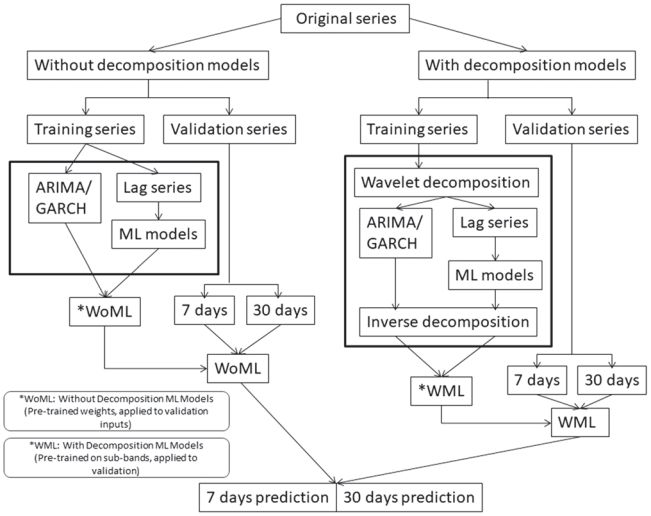
2.10. Proposed Work

To the best of our knowledge, no prior research has evaluated the performance of various wavelet filters while considering key attributes of the time series. Identifying the most effective filter for modeling and forecasting PM concentrations is a particularly compelling aspect of this study. Equally important is determining the best-performing combination of wavelet filter and statistical model for predicting average daily $PM_{2.5}$ and PM_{10} levels across different regions. This paper also introduces three hybrid models. To carry out the analysis, a range of significant wavelet filters has been applied in conjunction with multiple statistical approaches. The methodology has been clearly outlined through a series of systematic steps (Figure 1).

Step 1: data splitting. Let the original time series be $\{Y_t\}_{t=1}^{N}$, representing daily PM concentration. This series is split as:

1) **Training set:** $\{Y_t\}_{t=1}^{N-m}$
2) **Validation sets:**

Figure 1
Individual and wavelet-based combinatorial models



- $\{Y_t\}_{t=N-6}^{N}$ (last 1-week)
- $\{Y_t\}_{t=N-29}^{N}$ (last 1-month)

Step 2: base model training and evaluation. Construct lagged features for univariate time series modeling: Let $X_t = [Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}]$ be the lag vector. Train statistical models $\mathcal{M}_i$ (e.g., ARIMA, GARCH, ANN, RF and SVR) on $(X_t,\ Y_t)$ for $t = p+1, \ldots, N-m$. Obtain predictions $\widehat{Y}_t^{(i)}$ and evaluate on validation sets using performance metrics like root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE).

Step 3: wavelet decomposition. Apply discrete wavelet transform (DWT) to Y_t using different wavelet bases $w \in \{\text{haar}, d4, c6, bl14, la8\}$:

$$Y_t = A_t^{(w)} + D_t^{(w)} \tag{14}$$

where $A_t^{(w)}$ and $D_t^{(w)}$ are approximation and detail components.

Step 4: data splitting for wavelet components. Each component series $A_t^{(w)}$, $D_t^{(w)}$ is split into training and test sets analogous to Step 1.

Step 5: feature engineering for wavelet components. For each wavelet-transformed component, construct lag series:

$$X_t^{(w)} = \left[A_{t-1}^{(w)}, D_{t-1}^{(w)}, \ldots, A_{t-p}^{(w)}, D_{t-p}^{(w)}\right] \tag{15}$$

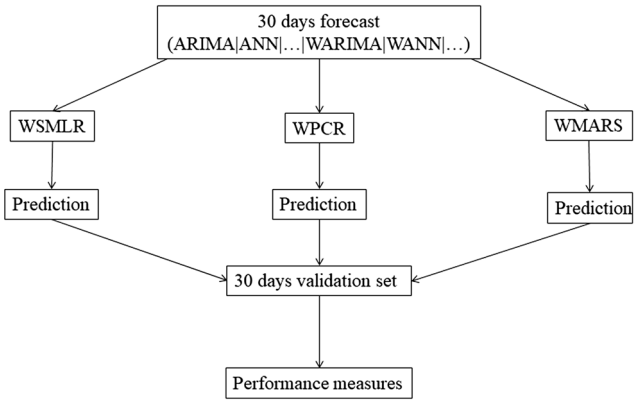
Train models $\mathcal{M}_i^{(w)}$ using these features.

Step 6: wavelet-based model prediction and evaluation. Predict $\hat{Y}_t^{(w,i)}$ using trained models on validation data. Evaluate and compare performance.

Step 7: selection of best wavelet-model combination. Determine optimal combination (w^*, $\mathcal{M}^*$) such that:

$$(w^*, \mathcal{M}^*) = \arg\min_{w, \mathcal{M}_i} \text{RMSE}\left(\hat{Y}_t^{(w,i)}, Y_t\right) \tag{16}$$

Figure 2
Wavelet-based hybrid models



Step 8: development of hybrid models. Define hybrid model inputs as:

$$Z_t = \left[\hat{Y}_t^{(1)}, \hat{Y}_t^{(2)}, \ldots, \hat{Y}_t^{(w,i)}, \ldots\right] \tag{17}$$

Train new models $\mathcal{H}_j$ (e.g., wavelet-stepwise multiple linear regression (WSMLR), wavelet-principal component regression (WPCR), and wavelet-multivariate adaptive regression spline (WMARS)) on (Z_t, Y_t) to obtain final predictions:

$$\hat{Y}_t^{\text{hybrid}} = \mathcal{H}_j(Z_t) \tag{18}$$

Overfitting was addressed through multiple strategies: (1) time series cross-validation using rolling windows, (2) separate validation sets for 7-day and 30-day predictions, (3) regularization in the underlying ML algorithms (SVR, RF), and (4) the ensemble nature of hybrid models which inherently reduces overfitting risk.

Performance of hybrid models (Figure 2) is then compared with singular and wavelet-based models. Here, the 7-day forecast values have been extracted from the 30-day forecast values.

3. Results and Discussion

Data for $PM_{2.5}$ and PM_{10} concentration (μg m^{-3}) was obtained at 24-hour interval from the Central Pollution Control Board (CPCB) website[1] for three Delhi Pollution Control Committee (DPCC)-regulated air quality monitoring stations situated at Alipur, Okhla Phase-II, and Pusa Road. These three sampling locations represent the northern (Alipur), central (Pusa), and southern (Okhla Phase-II) Delhi. There are multiple sources of pollution in Alipur, namely, household and vehicular emissions, and biomass burning. The major source of pollution in Okhla Phase-II is industrial emission, and that in Pusa is vehicular emission. Locations of these three weather monitoring stations have been shown in Figure 3. The collected data spanned over the period from 01 April 2019 to 30 November 2022 (1340 data points).

The descriptive statistics of the $PM_{2.5}$ and PM_{10} concentration data for the selected locations are given in Table 1. As observed for $PM_{2.5}$, the mean and median concentration followed the sequence Alipur > Okhla > Pusa. For PM_{10}, the mean concentration followed

[1] https://app.cpcbccr.com/ccr/#/caaqmdashboard-all/caaqm-landing/caaqmcomparison-data.

Figure 3
Map showing the locations of the sampling stations (Alipur, Pusa, and Okhla Phase-II) in Delhi

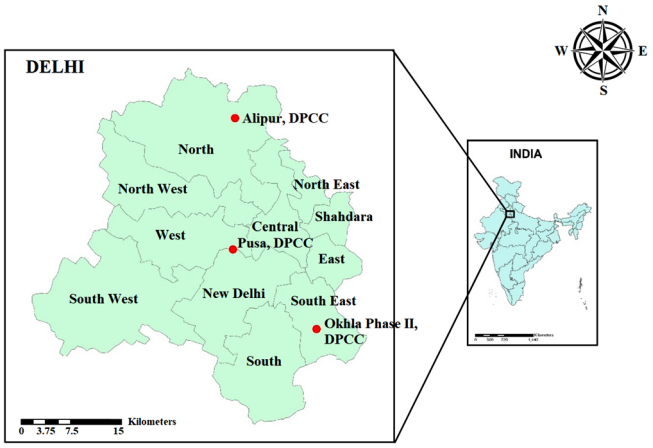


Table 1
Descriptive statistics of $PM_{2.5}$ and PM_{10} concentration data

	Alipur		Okhla		Pusa	
Statistics	**$PM_{2.5}$**	**PM_{10}**	**$PM_{2.5}$**	**PM_{10}**	**$PM_{2.5}$**	**PM_{10}**
Mean	99.44	197.86	95.56	206.65	89.79	195.48
Median	72.00	176.80	63.21	187.85	62.09	180.57
Minimum	4.81	10.04	6.18	9.73	3.44	9.97
Maximum	712.11	758.40	551.19	740.19	570.61	726.86
S.D.	82.41	126.45	86.52	124.76	78.61	118.76
C.V. (%)	82.88	63.91	90.54	60.37	87.55	60.75
Skewness	1.82	0.80	1.83	0.90	1.81	0.77
Kurtosis	4.93	0.43	3.79	0.66	4.09	0.52

the sequence Okhla > Alipur > Pusa. The corresponding date on which the minimum and maximum concentrations were realized is given in Table 2. The C.V. percentage for PM_{10} concentration was highest for Alipur, and for Okhla and Pusa, it was almost the same. All the $PM_{2.5}$ concentration series exhibited more skewness and kurtosis than the corresponding locations' PM_{10} concentration series.

The time plots of $PM_{2.5}$ and PM_{10} concentrations for all the selected locations are visualized in Figure 4. From the time plot, it can be seen that very high concentrations of both $PM_{2.5}$ and PM_{10} can be found from late October to early January in all three locations. This period coincides with the time of lowest precipitation and densest air which attributes to slow air circulation. Relatively low concentrations can be found from early July to mid-September. During this period, the concentration of $PM_{2.5}$ and PM_{10} generally lies below the daily safe limit of 60 and 100 μg m^{-3}, respectively, as recommended by Indian NAAQS. The month of August is the safest with respect to PM concentration as the maximum precipitation occurred in this month which forces down and coagulates the particulate matters to the ground. The presence of high moisture during monsoon also decreases the air density which makes the air circulation faster.

The normality of the selected series is tested using the Shapiro-Wilk test [46]. For this test, the null hypothesis is, H_0: the series follows the normal distribution against the alternative hypothesis, H_1: the series does not follow the normal distribution. It is seen (Table 3) that none

Table 2
The date of realization of minimum and maximum concentration

	Alipur		Okhla		Pusa	
	$PM_{2.5}$	PM_{10}	$PM_{2.5}$	PM_{10}	$PM_{2.5}$	PM_{10}
Minimum	9th October 2022	17th August 2019	9th October 2022	9th October 2022	17th August 2019	17th August 2019
Maximum	3rd November 2019	3rd November 2019	3rd November 2019	9th November 2020	3rd November 2019	9th November 2020

Figure 4
Time plot of $PM_{2.5}$ concentrations at (a) Alipur, (b) Okhla, and (c) Pusa; and PM_{10} concentrations at (d) Alipur, (e) Okhla, and (f) Pusa

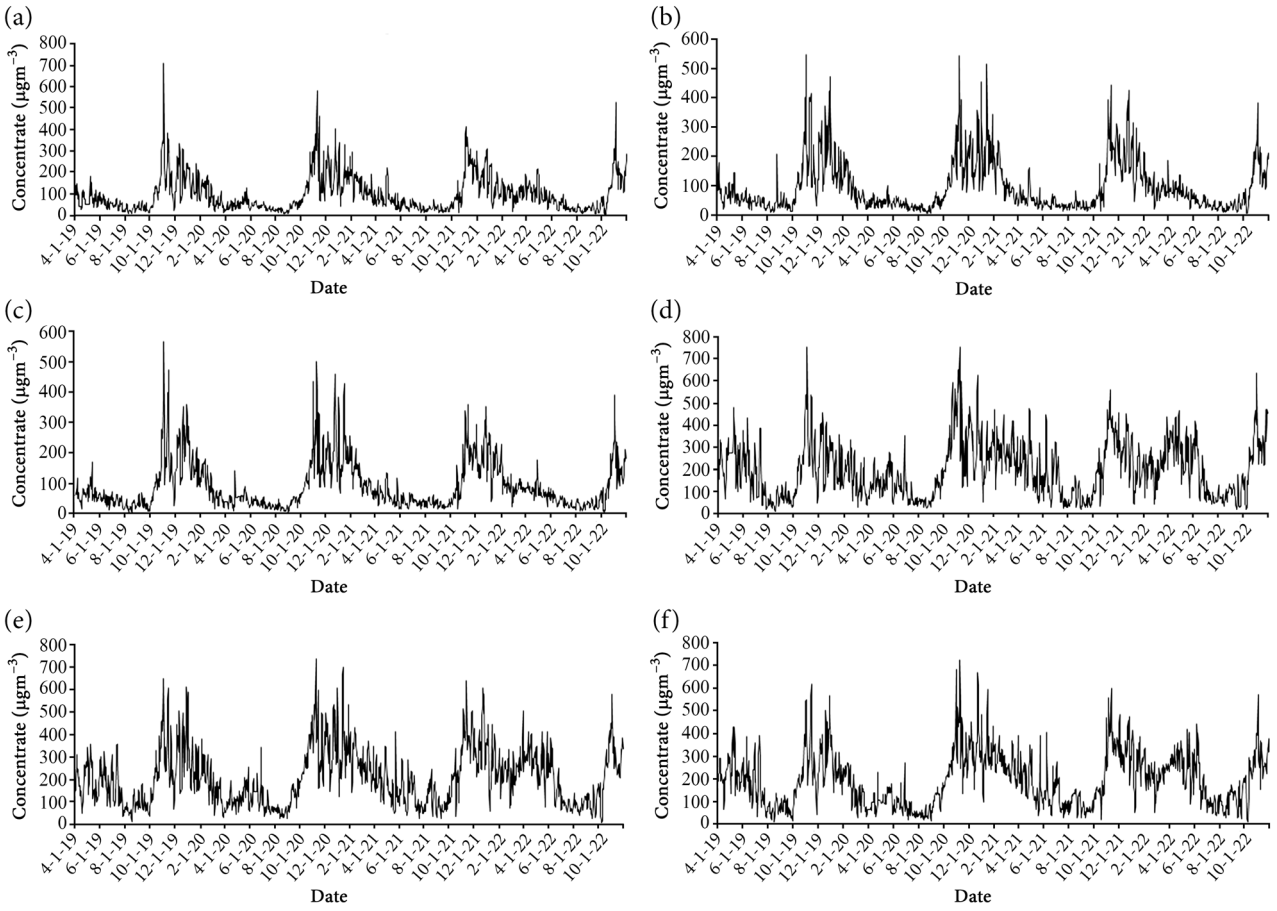


Table 3
Test for normality (Shapiro-Wilk test)

	Alipur		Okhla		Pusa	
	$PM_{2.5}$	PM_{10}	$PM_{2.5}$	PM_{10}	$PM_{2.5}$	PM_{10}
Test statistic	0.831	0.943	0.802	0.939	0.820	0.952

Note: The p-value for all the test statistics is <0.001.

Table 4
Test for stationarity

	Alipur		Okhla		Pusa	
	$PM_{2.5}$	PM_{10}	$PM_{2.5}$	PM_{10}	$PM_{2.5}$	PM_{10}
ADF	-4.07	-4.42	-3.65	-4.04	-3.58	-3.85
	(0.01)	(0.01)	(0.03)	(0.01)	(0.03)	(0.02)
PP	-145.90	-165.95	-137.58	-163.93	-126.34	-150.60
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)

Note: The p-values are in parentheses.

of the selected series follows normal distribution at a 1% level of significance.

The stationarity of the selected series is tested using the Augmented Dickey-Fuller (ADF) test [47], and Phillips-Perron (PP) test [48]. The null hypothesis for the ADF and PP tests is, H_0: unit root is present in the time series data against the alternative hypothesis, H_1: unit root is not present in the time series data. From Table 4, it is seen that all the selected series are stationary.

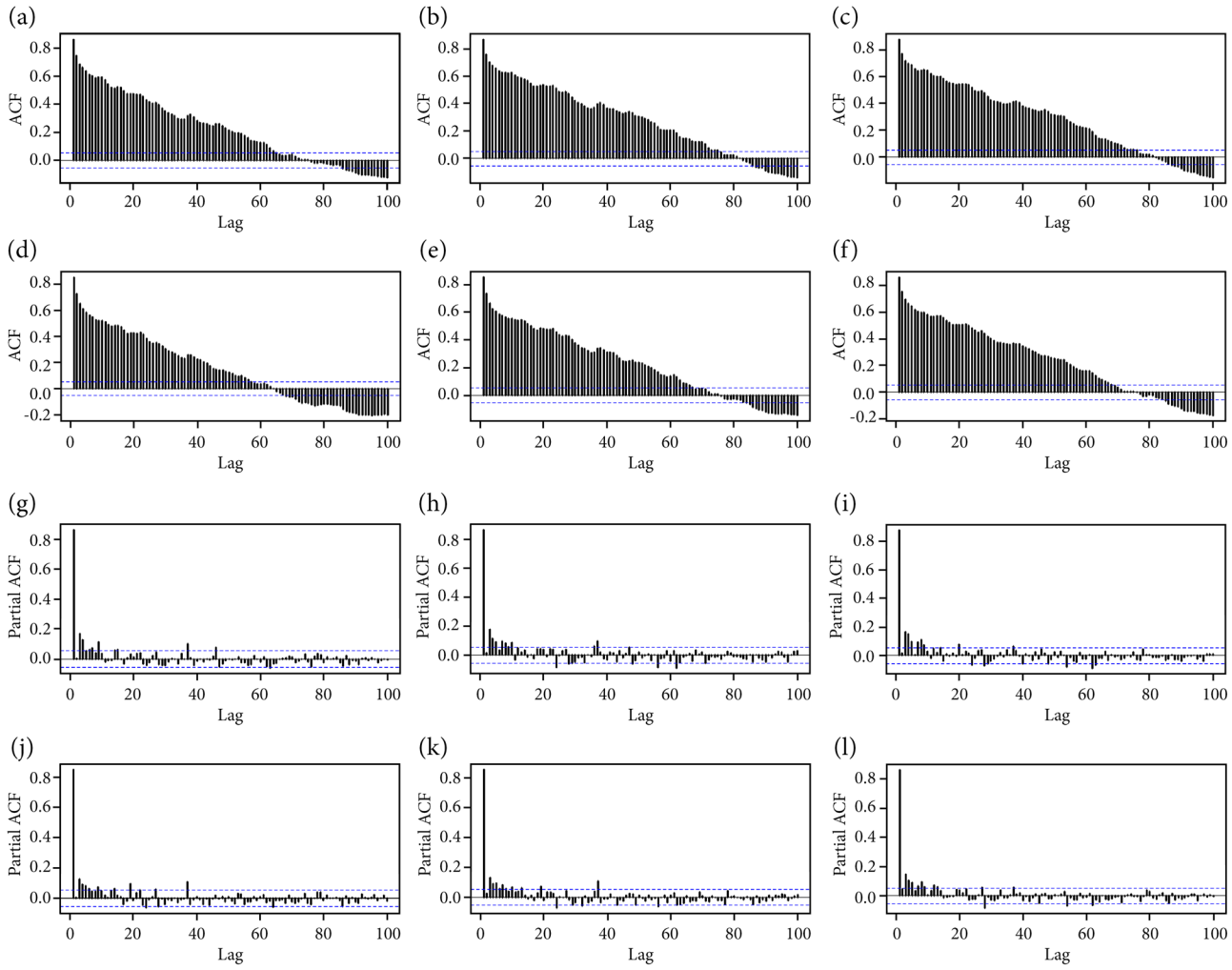
The presence of nonlinearity of the selected time series is tested using the Broock-Dechert-Scheinkman (BDS) test [49]. The null hypothesis of this test is, H_0: the time series is independently and identically distributed (IID) against the alternative hypothesis, H_1: the time series is not independently and identically distributed (IID). It is seen that (Table 5) the null hypothesis is rejected for

Table 5
Test for linearity: BDS test

	$PM_{2.5}$		PM_{10}	
	Embedding dimension			
	2	3	2	3
		Alipur		
eps [1]	71.74***	97.44***	173.26***	249.43***
eps [2]	55.94***	62.09***	99.75***	113.73***
eps [3]	47.25***	48.24***	64.90***	66.82***
eps [4]	41.48***	40.74***	51.16***	49.93***
		Okhla		
eps [1]	64.93***	87.03***	160.58***	231.77***
eps [2]	53.21***	59.39***	88.84***	101.61***
eps [3]	44.40***	45.33***	58.88***	60.67***
eps [4]	37.06***	36.21***	47.18***	46.36***
		Pusa		
eps [1]	72.54***	99.50***	214.68***	319.92***
eps [2]	58.55***	65.72***	109.17***	128.32***
eps [3]	47.39***	48.42***	68.92***	71.45***
eps [4]	38.37***	37.44***	51.50***	50.22***

Note: *** indicates p-value <0.01.

Figure 5
ACF plots of $PM_{2.5}$ concentration (a) Alipur, (b) Okhla, (c) Pusa; ACF plots of PM_{10} concentration (d) Alipur, (e) Okhla, (f) Pusa; PACF plots of $PM_{2.5}$ concentration (g) Alipur, (h) Okhla, (i) Pusa; PACF plots of PM_{10} concentration (j) Alipur, (k) Okhla, (l) Pusa



all the selected time series. Hence, all the series exhibit nonlinear patterns.

In order to establish statistical dependencies among the realizations of the data sets, the autocorrelation function (ACF) and partial autocorrelation function (PACF) are obtained. Figure 5 depicts the ACF and PACF plots of the selected time series. From the ACF plots of each instance, it can be seen that autocorrelations are significant for a large lag value. This is known as hyperbolic decay and it indicates the presence of long-term persistence. This long-term persistence phenomenon is known as the long-memory property. Rakshit et al. [50] noticed the same phenomenon.

The performance of the used model is evaluated based on three popularly used error functions namely RMSE, MAE, and MAPE. Tables 6–8 represent the prediction performance of daily

Table 6
Forecasting performance of the selected models in model validation set for $PM_{2.5}$ concentration at Alipur

	1-week prediction			1-month prediction		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
ARIMA	141.041	81.986	37.705	81.821	50.683	26.516
GARCH	157.739	102.913	53.944	92.705	62.226	38.147
ANN	110.591	58.809	23.230	67.122	39.754	20.691
SVR	117.228	70.175	27.532	70.932	43.930	23.238
RF	114.100	80.821	31.263	70.864	49.297	26.102
SMLR	118.906	75.013	26.817	71.015	44.875	23.080
MARS	107.102	69.743	26.302	67.711	43.344	22.357
PCR	118.906	75.013	26.817	71.015	44.875	23.080
WARMAhaar	129.585	87.672	35.964	78.507	51.849	26.288
WARMAd4	148.573	101.126	48.441	94.845	62.163	42.044
WARMAc6	232.331	202.335	202.925	142.722	114.673	116.840
WARMAbl14	244.139	214.297	248.966	147.859	118.561	129.722
WARMAla8	244.139	214.297	248.966	147.859	118.561	129.722
WGARCHhaar	138.031	92.129	40.579	87.222	61.579	27.144
WGARCHd4	141.752	93.404	42.393	86.108	58.071	26.460
WGARCHc6	284.400	259.502	627.866	191.001	170.544	411.229
WGARCHbl14	248.254	217.766	268.485	158.171	132.091	171.047
WGARCHla8	248.254	217.766	268.485	158.171	132.091	171.047
WANNhaar	134.587	75.929	33.417	78.701	47.431	24.070
WANNd4	151.046	105.576	51.292	84.913	56.188	29.884
WANNc6	277.580	254.239	551.392	182.728	161.421	324.988
WANNbl14	269.370	241.681	481.684	183.012	162.260	394.622
WANNla8	269.370	241.681	481.684	183.012	162.260	394.622
WSVRhaar	90.137	64.588	19.782	43.972	18.201	**6.412**
WSVRd4	173.039	138.464	64.559	121.990	100.731	54.163
WSVRc6	265.122	229.509	1273.660	168.620	136.619	349.827
WSVRbl14	267.710	234.926	451.254	167.415	137.746	473.554
WSVRla8	267.710	234.926	451.254	167.415	137.746	473.554
WRFhaar	60.496	42.471	13.263	30.883	16.332	**6.726**
WRFd4	175.789	144.181	65.524	125.685	103.398	52.114
WRFc6	263.180	221.296	727.180	165.393	130.308	221.326
WRFbl14	265.884	233.599	423.237	167.292	136.375	299.005
WRFla8	265.884	233.599	423.237	167.292	136.375	299.005
WSMLR	**9.741**	**7.675**	**2.849**	**9.121**	**7.526**	**4.059**
WMARS	**7.392**	**6.100**	**2.614**	**6.986**	**5.478**	**2.936**
WPCR	38.142	26.969	**8.344**	23.555	17.164	**9.087**

Note: Values less than 10 have been emboldened.

Table 7
Forecasting performance of the selected models in model validation set for $PM_{2.5}$ concentration at Okhla

	1-week prediction			1-month prediction		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
ARIMA	103.760	73.466	41.266	61.992	44.453	27.287
GARCH	102.240	70.406	38.888	65.414	47.994	26.823
ANN	64.070	42.314	17.542	46.775	33.010	18.631
SVR	79.862	57.198	24.444	51.396	36.219	20.836
RF	89.143	67.595	27.327	56.451	39.595	21.864
SMLR	74.233	49.332	20.876	48.356	33.810	20.468
MARS	70.135	45.976	18.656	47.234	32.758	18.357
PCR	74.233	49.332	20.876	48.356	33.810	20.468
WARMAhaar	83.911	63.074	29.804	58.555	47.389	26.083
WARMAd4	178.196	164.653	188.049	106.068	82.793	91.573
WARMAc6	178.296	157.468	179.655	105.817	81.358	89.583
WARMAbl14	182.229	165.543	196.471	107.871	83.721	94.576
WARMAla8	182.229	165.543	196.471	107.871	83.721	94.576
WGARCHhaar	86.014	64.811	30.940	75.042	62.275	29.736
WGARCHd4	83.485	64.324	29.912	77.259	64.692	30.181
WGARCHc6	228.402	215.950	596.815	154.688	139.702	403.739
WGARCHbl14	201.322	186.411	286.942	129.575	111.538	177.975
WGARCHla8	201.322	186.411	286.942	129.575	111.538	177.975
WANNhaar	84.613	64.519	29.949	73.073	59.895	28.635
WANNd4	82.558	71.451	32.855	73.915	62.701	29.644
WANNc6	225.591	209.630	586.002	148.802	131.969	333.200
WANNbl14	211.646	195.766	384.543	146.899	132.217	344.792
WANNla8	211.646	195.766	384.543	146.899	132.217	344.792
WSVRhaar	10.194	**8.072**	**3.499**	**5.626**	**3.827**	**2.323**
WSVRd4	115.348	102.733	45.989	90.787	78.784	47.289
WSVRc6	220.191	200.199	768.704	139.563	113.251	217.946
WSVRbl14	208.468	188.197	446.805	133.269	107.341	292.856
WSVRla8	208.468	188.197	446.805	133.269	107.341	292.856
WRFhaar	22.120	18.592	**7.341**	13.661	10.176	**5.709**
WRFd4	107.802	95.335	42.727	86.025	74.015	43.367
WRFc6	214.496	194.226	556.591	133.869	107.037	165.974
WRFbl14	206.217	186.572	395.551	131.642	104.866	282.107
WRFla8	206.217	186.572	395.551	131.642	104.866	282.107
WSMLR	**3.187**	**2.901**	**1.267**	**2.987**	**2.407**	**1.530**
WMARS	**3.552**	**3.171**	**1.392**	**3.589**	**2.960**	**2.074**
WPCR	**4.666**	**3.809**	**1.625**	**3.306**	**2.645**	**1.642**

Note: Values less than 10 have been emboldened.

$PM_{2.5}$ concentrations at two different prediction windows by various models for Alipur, Okhla, and Pusa, respectively. To explain a wavelet-based combinatorial model, shortcuts have been used. For example, WARMAhaar represents a wavelet-based autoregressive moving average (ARMA) model using the haar filter. WSMLR, WPCR, and WMARS represent wavelet based-hybrid models using SMLR, PCR, and MARS models, explained in Section 2.

Tables 9–11 represent the prediction performance of daily PM_{10} concentrations at two different prediction windows by various models for Alipur, Okhla, and Pusa, respectively.

Table 8
Forecasting performance of the selected models in model validation set for $PM_{2.5}$ concentration at Pusa

	1-week prediction			1-month prediction		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
ARIMA	104.624	70.991	41.849	62.663	43.422	27.804
GARCH	120.759	89.198	61.024	76.507	55.589	45.171
ANN	65.177	48.794	20.585	53.043	39.079	23.456
SVR	63.463	41.511	18.645	51.697	37.130	22.543
RF	69.690	46.220	19.037	56.544	39.429	21.834
SMLR	75.201	57.253	24.316	54.353	39.490	24.645
MARS	64.402	47.243	19.727	51.396	37.576	21.986
PCR	75.201	57.253	24.316	54.353	39.490	24.645
WARMAhaar	90.136	69.455	35.145	59.369	46.633	26.215
WARMAd4	172.367	149.022	179.040	106.751	83.985	98.415
WARMAc6	164.091	142.473	149.325	102.894	81.698	89.538
WARMAbl14	177.564	156.512	195.659	108.671	85.719	102.197
WARMAla8	177.564	156.512	195.659	108.671	85.719	102.197
WGARCHhaar	93.761	71.897	37.247	68.597	55.740	28.947
WGARCHd4	93.813	72.189	37.381	68.422	55.417	28.810
WGARCHc6	216.566	200.243	538.538	149.426	134.412	367.301
WGARCHbl14	187.306	167.152	242.028	122.935	104.384	158.423
WGARCHla8	187.306	167.152	242.028	122.935	104.384	158.423
WANNhaar	87.110	68.335	32.523	76.208	63.897	30.383
WANNd4	91.738	82.102	42.807	75.453	64.474	31.879
WANNc6	213.632	194.526	513.352	145.243	129.267	322.304
WANNbl14	200.443	180.664	344.896	143.332	128.638	326.037
WANNla8	200.443	180.664	344.896	143.332	128.638	326.037
WSVRhaar	18.709	14.139	**5.937**	15.121	**9.306**	**5.610**
WSVRd4	118.272	108.352	54.824	86.605	74.522	47.581
WSVRc6	214.828	193.359	618.996	135.606	108.344	191.911
WSVRbl14	205.958	178.287	584.055	134.665	110.555	335.847
WSVRla8	205.958	178.287	584.055	134.665	110.555	335.847
WRFhaar	26.866	18.129	**7.506**	15.879	**9.475**	**5.493**
WRFd4	114.807	105.132	51.798	85.571	72.134	44.119
WRFc6	209.756	187.186	541.067	132.404	105.003	170.151
WRFbl14	201.893	175.543	418.130	132.561	107.862	331.092
WRFla8	201.893	175.543	418.130	132.561	107.862	331.092
WSMLR	13.740	10.783	**4.468**	**8.951**	**6.956**	**4.414**
WMARS	10.994	**9.649**	**4.211**	**7.683**	**5.953**	**3.470**
WPCR	13.398	10.793	**4.246**	10.159	**8.312**	**5.446**

Note: Values less than 10 have been emboldened.

In the next step, the ranking of the used models based on their forecasting performance in terms of RMSE, MAE, and MAPE has been done using the technique for order preference by similarity to ideal solution (TOPSIS) approach [51]. It is a multi-criteria decision-making method that ranks the competitive models based on their shortest distance from the positive ideal solution and the longest distance from the negative ideal solution. It computes the relative closeness of a model to the ideal solution based on both the distances. This relative closeness ranges between 0 and 1. In cases where a tie is observed, the tied models are assigned the same rank, while

Table 9
Forecasting performance of the selected models in model validation set for PM_{10} concentration at Alipur

	1-week prediction			1-month prediction		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
ARIMA	145.501	91.606	26.943	105.197	71.816	24.384
GARCH	136.065	88.376	24.703	95.423	69.584	20.165
ANN	109.870	70.574	17.716	79.517	55.614	17.610
SVR	114.347	80.462	19.487	80.284	57.645	17.512
RF	124.464	81.359	20.107	84.598	60.477	18.345
SMLR	114.645	68.690	16.823	82.221	55.493	17.865
MARS	125.602	90.845	22.835	84.844	59.780	18.570
PCR	114.645	68.690	16.823	82.221	55.493	17.865
WARMAhaar	143.234	92.783	26.974	110.690	75.628	26.634
WARMAd4	165.505	120.218	38.020	128.121	89.207	35.205
WARMAc6	256.461	225.905	111.334	178.537	148.823	75.160
WARMAbl14	283.034	255.637	148.480	189.106	158.444	85.957
WARMAla8	283.034	255.637	148.480	189.106	158.444	85.957
WGARCHhaar	139.667	105.371	29.198	100.947	80.918	22.449
WGARCHd4	146.102	107.614	30.874	100.158	76.575	22.142
WGARCHc6	340.464	318.099	287.644	254.305	233.644	208.630
WGARCHbl14	269.262	234.900	125.879	189.701	150.288	84.791
WGARCHla8	269.262	234.900	125.879	189.701	150.288	84.791
WANNhaar	142.640	84.864	24.956	108.010	72.231	25.353
WANNd4	168.404	127.420	40.360	124.008	89.290	33.672
WANNc6	319.204	299.374	232.277	206.009	176.791	113.079
WANNbl14	328.922	301.843	306.110	269.166	250.842	306.833
WANNla8	328.922	301.843	306.110	269.166	250.842	306.833
WSVRhaar	29.131	20.131	**5.064**	15.592	**8.966**	**2.459**
WSVRd4	195.325	176.745	48.205	159.137	130.022	40.954
WSVRc6	278.756	219.870	160.465	187.112	145.677	67.185
WSVRbl14	313.420	274.905	237.434	221.654	182.545	184.631
WSVRla8	313.420	274.905	237.434	221.654	182.545	184.631
WRFhaar	49.454	40.689	**9.294**	29.347	19.108	**4.953**
WRFd4	189.577	169.430	47.810	153.783	125.920	39.570
WRFc6	284.839	228.606	171.896	184.690	144.215	68.387
WRFbl14	311.216	272.834	227.173	218.750	182.567	209.335
WRFla8	311.216	272.834	227.173	218.750	182.567	209.335
WSMLR	11.353	**9.990**	**2.604**	**8.197**	**6.466**	**1.946**
WMARS	**7.159**	**5.786**	**1.398**	**6.804**	**5.573**	**1.776**
WPCR	13.400	10.947	**2.631**	10.478	**8.227**	**2.568**

Note: Values less than 10 have been emboldened.

the subsequent rank reflects a unit gap, consistent with the standard competition ranking rule. Tables 12 and 13 represent the TOPSIS rank of the competitive models for forecasting $PM_{2.5}$ and PM_{10}, respectively.

It can be observed in Tables 6–11 that the WSVRhaar model is the best performer among individual as well as wavelet-based statistical and ML models in every circumstance except daily average $PM_{2.5}$ concentration prediction for Alipur. WRFhaar is the best model to predict the average daily $PM_{2.5}$ concentration for Alipur.

The hybrid models, viz., WSMLR, WPCR, and WMARS, are used in this study to predict daily average PM concentrations. The results (Tables 6 and 9) also indicate that the WMARS is the best performer

Table 10
Forecasting performance of the selected models in model validation set for PM_{10} concentration at Okhla

	1-week prediction			1-month prediction		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
ARIMA	137.763	101.668	30.905	83.289	63.248	20.956
GARCH	125.241	89.455	25.689	87.496	69.064	20.558
ANN	91.670	67.285	16.278	57.811	40.328	12.260
SVR	99.996	74.351	17.575	60.838	41.596	12.302
RF	92.852	68.148	15.945	57.419	38.189	11.194
SMLR	95.797	65.343	15.352	60.308	40.926	12.708
MARS	90.129	65.790	15.757	57.283	40.280	12.243
PCR	95.797	65.343	15.352	60.308	40.926	12.708
WARMAhaar	128.944	93.874	27.270	84.326	67.863	21.988
WARMAd4	248.771	226.686	113.848	146.599	113.754	56.180
WARMAc6	241.557	224.137	108.774	144.133	113.250	55.148
WARMAbl14	261.823	243.840	131.568	153.306	120.037	61.996
WARMAla8	261.823	243.840	131.568	153.306	120.037	61.996
WGARCHhaar	109.447	93.368	24.425	111.185	97.279	25.483
WGARCHd4	114.846	94.355	25.510	103.417	88.782	24.119
WGARCHc6	343.659	329.473	331.141	237.460	219.141	222.086
WGARCHbl14	265.509	243.504	134.423	167.566	136.896	79.035
WGARCHla8	265.509	243.504	134.423	167.566	136.896	79.035
WANNhaar	115.818	81.969	22.762	81.448	66.956	20.531
WANNd4	129.667	102.609	29.573	84.996	70.306	21.730
WANNc6	335.640	321.417	306.990	213.801	190.110	157.295
WANNbl14	310.299	290.084	229.889	223.655	207.224	195.744
WANNla8	310.299	290.084	229.889	223.655	207.224	195.744
WSVRhaar	**7.912**	**5.335**	**1.342**	**5.202**	**3.573**	**1.113**
WSVRd4	172.602	157.875	42.441	131.512	109.226	34.132
WSVRc6	282.481	233.968	257.516	183.581	144.438	84.165
WSVRbl14	293.097	265.052	194.353	194.452	153.183	483.807
WSVRla8	293.097	265.052	194.353	194.452	153.183	483.807
WRFhaar	32.711	27.350	**6.458**	17.697	11.607	**3.304**
WRFd4	168.438	153.759	42.664	127.880	106.194	33.275
WRFc6	292.635	248.686	279.520	183.920	144.149	88.423
WRFbl14	291.050	263.931	186.239	193.029	153.987	196.139
WRFla8	291.050	263.931	186.239	193.029	153.987	196.139
WSMLR	**5.572**	**4.939**	**1.318**	**3.709**	**2.956**	**0.962**
WMARS	**5.932**	**4.087**	**1.089**	**4.271**	**3.090**	**1.012**
WPCR	**8.281**	**6.660**	**1.659**	**5.156**	**4.006**	**1.270**

Note: Values less than 10 have been emboldened.

for Alipur for both $PM_{2.5}$ and PM_{10} concentration prediction irrespective of prediction windows. For Okhla (Tables 7 and 10), WSMLR is the best performer for $PM_{2.5}$ concentration prediction for both prediction windows. For this location, WMARS is the best performer for 1-week ahead prediction and WSMLR for 1-month ahead prediction of PM_{10} concentration. For Pusa, WMARS is the best performer for $PM_{2.5}$ concentration prediction for both prediction windows (Table 8). On the other hand, for PM_{10} concentration prediction, WSMLR is the best performer for 1-week ahead prediction and WMARS for 1-month ahead prediction (Table 11). The same has been reflected in the TOPSIS rank

Table 11
Forecasting performance of the selected models in model validation set for PM_{10} concentration at Pusa

	1-week prediction			1-month prediction		
Models	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**
ARIMA	121.126	88.817	28.163	79.371	59.916	20.619
GARCH	119.099	90.891	28.320	80.884	64.271	20.658
ANN	70.971	52.750	13.797	65.203	46.270	15.800
SVR	74.712	59.726	15.915	66.425	47.573	16.301
RF	87.720	66.008	17.214	69.759	49.859	16.712
SMLR	90.261	72.774	18.388	69.336	50.524	17.023
MARS	84.229	67.096	17.669	66.689	48.497	16.130
PCR	90.261	72.774	18.388	69.336	50.524	17.023
WARMAhaar	119.181	100.324	30.162	82.846	66.585	22.077
WARMAd4	105.970	96.539	27.865	82.863	67.562	22.941
WARMAc6	220.940	196.549	108.688	139.015	113.402	60.032
WARMAbl14	210.354	181.278	93.161	139.496	113.188	60.426
WARMAla8	210.354	181.278	93.161	139.496	113.188	60.426
WGARCHhaar	113.563	99.316	28.588	94.789	79.439	22.982
WGARCHd4	112.913	99.333	28.323	96.116	80.528	23.105
WGARCHc6	255.249	231.637	154.486	175.488	152.469	102.761
WGARCHbl14	180.588	137.855	58.580	115.296	84.239	37.390
WGARCHla8	180.588	137.855	58.580	115.296	84.239	37.390
WANNhaar	109.097	91.343	26.146	82.478	67.637	20.667
WANNd4	106.384	97.879	28.196	79.090	65.983	20.852
WANNc6	248.723	219.529	145.699	137.785	98.059	54.946
WANNbl14	224.569	193.288	115.346	178.580	160.078	121.940
WANNla8	224.569	193.288	115.346	178.580	160.078	121.940
WSVRhaar	**5.477**	**4.808**	**1.329**	11.042	**5.972**	**2.093**
WSVRd4	155.834	143.516	38.676	116.831	99.442	33.389
WSVRc6	230.377	187.195	108.783	157.607	127.975	52.459
WSVRbl14	211.590	171.538	97.333	156.924	118.411	127.281
WSVRla8	211.590	171.538	97.333	156.924	118.411	127.281
WRFhaar	27.604	21.511	**5.940**	20.360	11.957	**4.012**
WRFd4	144.323	133.293	37.283	112.701	94.814	31.651
WRFc6	238.056	193.507	119.400	155.305	123.399	52.452
WRFbl14	211.161	171.917	95.623	155.861	117.791	181.649
WRFla8	211.161	171.917	95.623	155.861	117.791	181.649
WSMLR	**3.125**	**2.384**	**0.687**	**7.168**	**5.236**	**1.960**
WMARS	**5.602**	**4.826**	**1.449**	**5.695**	**4.536**	**1.599**
WPCR	**7.448**	**6.915**	**1.864**	**8.994**	**6.746**	**2.437**

Note: Values less than 10 have been emboldened.

also (Tables 12 and 13). In the next step, the hierarchical clustering using the Ward's method has been done based on the scaled values of the RMSE, MAE, and MAPE of the competitive models. The heatmaps of these clusters are given in Figures 6 and 7 for forecasting $PM_{2.5}$ and PM_{10}, respectively. The results of the cluster analysis are also homogenous with the TOPSIS ranks.

The key highlight of this study is that the 1-week and 1-month prediction windows of the model's validation phase align with the

Table 12
TOPSIS rank of the models for forecasting $PM_{2.5}$

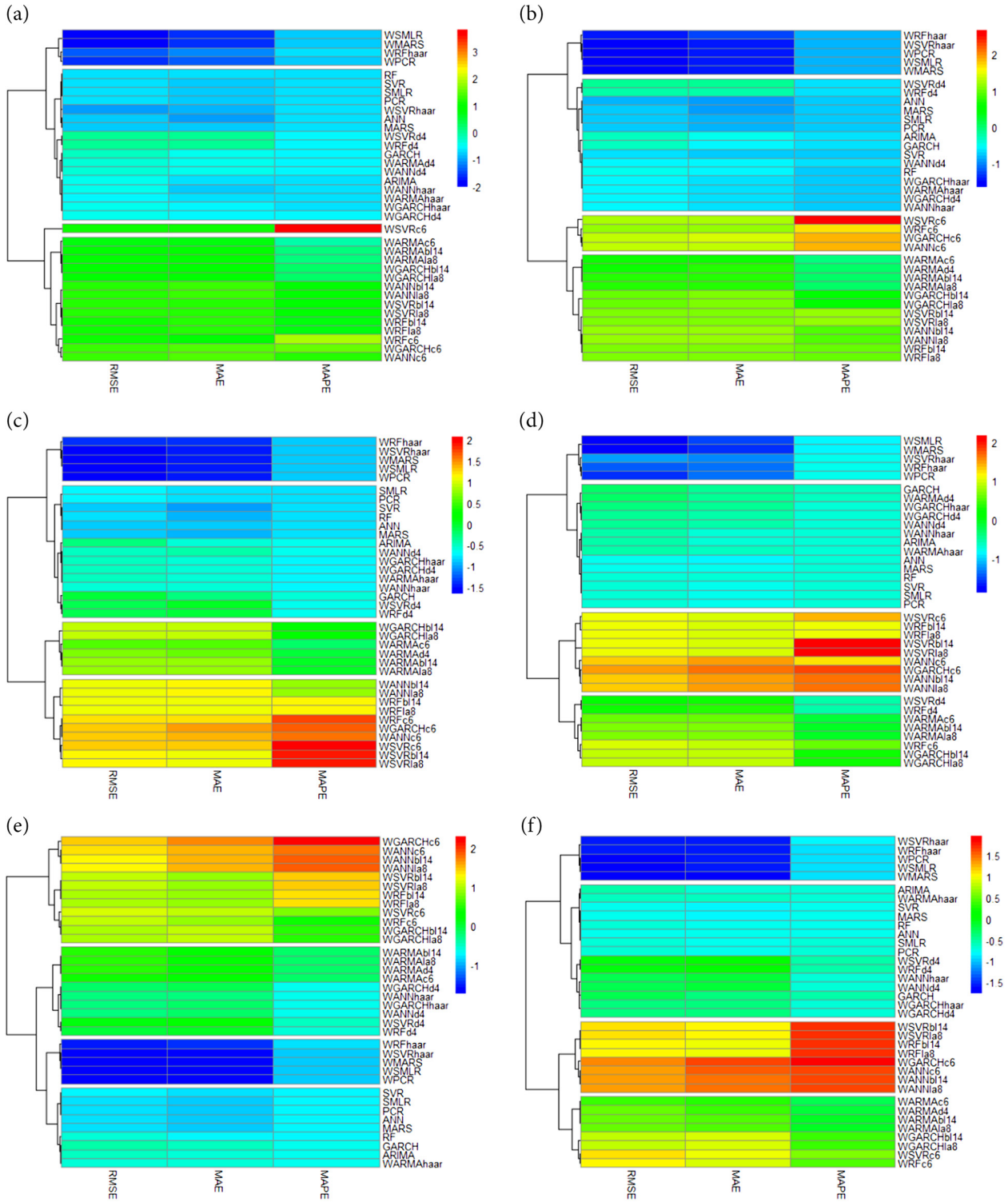
	Alipur		Okhla		Pusa	
Models	7-day	30-day	7-day	30-day	7-day	30-day
ARIMA	14	14	18	13	17	13
GARCH	19	18	17	14	18	16
ANN	6	6	6	6	8	8
SVR	8	8	10	10	6	6
RF	11	11	16	11	9	11
SMLR	9	9	8	8	10	9
MARS	7	7	7	7	7	7
PCR	9	9	8	8	10	9
WARMAhaar	13	13	11	12	13	12
WARMAd4	17	19	22	22	22	22
WARMAc6	22	22	21	21	21	21
WARMAbl14	23	23	23	23	23	23
WARMAla8	23	23	23	23	23	23
WGARCHhaar	15	17	14	17	14	15
WGARCHd4	16	16	12	18	15	14
WGARCHc6	34	36	35	36	33	36
WGARCHbl14	25	25	25	26	25	25
WGARCHla8	25	25	25	26	25	25
WANNhaar	12	12	13	15	12	17
WANNd4	18	15	15	16	16	18
WANNc6	33	31	34	33	31	33
WANNbl14	31	32	27	34	27	34
WANNla8	31	32	27	34	27	34
WSVRhaar	5	5	4	4	4	4
WSVRd4	20	20	20	20	20	20
WSVRc6	36	30	36	28	36	28
WSVRbl14	29	34	31	31	34	31
WSVRla8	29	34	31	31	34	31
WRFhaar	4	4	5	5	5	5
WRFd4	21	21	19	19	19	19
WRFc6	35	27	33	25	32	27
WRFbl14	27	28	29	29	29	29
WRFla8	27	28	29	29	29	29
WSMLR	2	2	1	1	3	2
WMARS	1	1	2	3	1	1
WPCR	3	3	3	2	2	3

Table 13
TOPSIS rank of the models for forecasting PM_{10}

	Alipur		Okhla		Pusa	
Models	7-day	30-day	7-day	30-day	7-day	30-day
ARIMA	15	14	18	12	15	12
GARCH	12	12	15	16	16	13
ANN	6	6	7	8	6	6
SVR	9	9	11	11	7	7
RF	10	11	8	6	9	9
SMLR	7	7	9	9	10	10
MARS	11	10	6	7	8	8
PCR	7	7	9	9	10	10
WARMAhaar	14	17	16	14	19	16
WARMAd4	18	19	22	22	13	17
WARMAc6	22	24	21	21	31	25
WARMAbl14	26	27	23	23	26	26
WARMAla8	26	27	23	23	26	26
WGARCHhaar	16	16	13	18	18	18
WGARCHd4	17	13	14	17	17	19
WGARCHc6	34	34	36	34	36	30
WGARCHbl14	23	25	25	25	22	20
WGARCHla8	23	25	25	25	22	20
WANNhaar	13	15	12	13	12	15
WANNd4	19	18	17	15	14	14
WANNc6	33	29	35	29	35	24
WANNbl14	35	35	32	32	32	33
WANNla8	35	35	32	32	32	33
WSVRhaar	4	4	3	3	2	4
WSVRd4	21	21	20	20	21	23
WSVRc6	25	23	31	27	30	29
WSVRbl14	31	30	29	35	28	31
WSVRla8	31	30	29	35	28	31
WRFhaar	5	5	5	5	5	5
WRFd4	20	20	19	19	20	22
WRFc6	28	22	34	28	34	28
WRFbl14	29	32	27	30	24	35
WRFla8	29	32	27	30	24	35
WSMLR	2	2	2	1	1	2
WMARS	1	1	1	2	3	1
WPCR	3	3	4	4	4	3

first week and the entire month of November 2022, respectively. This timeframe is marked by exceptionally high and highly variable concentrations of both $PM_{2.5}$ and PM_{10}. Earlier research on modeling PM concentrations has shown limited success during such severe pollution episodes, with prediction accuracy decreasing as the forecast period extends [52]. The performance measures of the best-performing wavelet decomposed hybrid model indicate that the RMSE, MAE, and MAPE values are considerably small for both the prediction windows. Here, wavelet decomposition plays a crucial role in capturing the hidden signal in the time series data. A highly accurate prediction can be helpful for policymakers to take pre-emptive actions promptly.

Figure 6
The heatmap of the hierarchical clusters of the models for forecasting the $PM_{2.5}$ concentrations: 7-day forecasts for (a) Alipur, (b) Okhla, (c) Pusa; and 30-day forecasts for (d) Alipur, (e) Okhla, and (f) Pusa

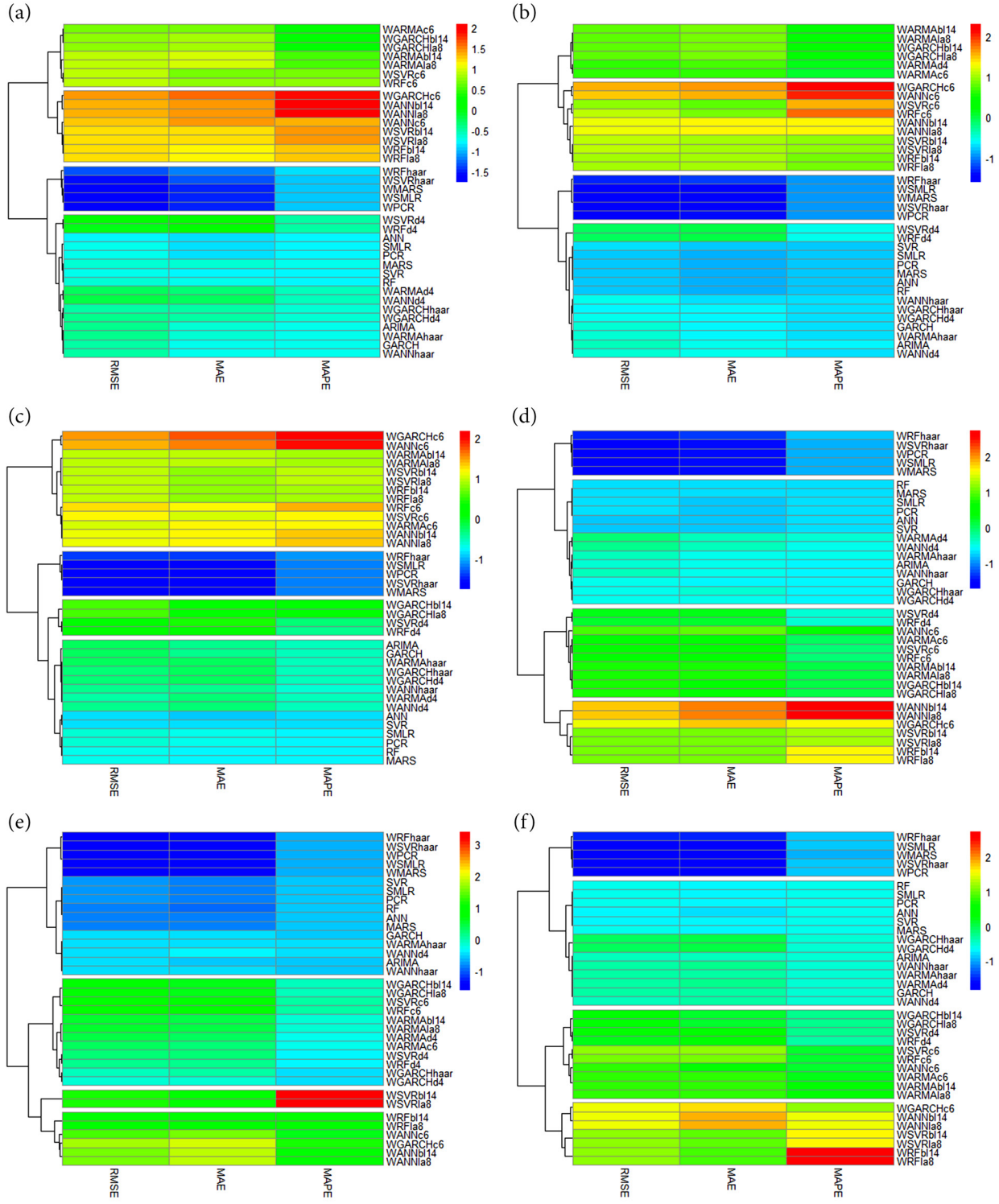


4. Conclusion

This research article attempts to model the concentration of $PM_{2.5}$ and PM_{10} in three different locations with different primary sources of pollution. The aim of this study is to evaluate the forecasting performance of various statistical and ML models, as well as wavelet decomposed hybrid models, for 1-week and 1-month horizons in the model validation set. This study focuses on the post-monsoon season in Delhi, India, which is notorious for its high levels of air pollution caused by crop residue burning and higher air density.

This study finds that the wavelet decomposed hybrid models are the most efficient algorithms for predicting $PM_{2.5}$ and PM_{10}

Figure 7
The heatmap of the hierarchical clusters of the models for forecasting the PM_{10} concentrations: 7-day forecasts for (a) Alipur, (b) Okhla, (c) Pusa; and 30-day forecasts for (d) Alipur, (e) Okhla, and (f) Pusa



concentrations in Delhi during the post-monsoon season. The predictive accuracy measures of the error functions of these hybrid models are at a minimal level, which suggests that they can provide accurate predictions of PM concentrations in Delhi. This can directly guide policymakers in implementing location-specific measures informed by meteorological forecasts, helping to lower costs and enhance the effectiveness of interventions aimed at mitigating the impact of crop residue burning on ambient $PM_{2.5}$ levels in Delhi during the post-monsoon season.

The issue of crop residue burning is a major contributor to air pollution in India. Every year, farmers burn crop residues for ensuring hassle-free field preparation for the next crop, which releases a significant amount of smoke and particulate matter into the atmosphere

[53–55]. This practice is particularly prevalent in North India, where the post-monsoon season exacerbates the pollution problem. As a result, the study's findings can be of great significance for policymakers who are struggling to reduce air pollution levels in Delhi. Since the best-performing models worked well across three areas with different pollution signatures, it is expected that they will also perform effectively in other regions that share similar socio-economic, cultural, environmental, and climatic conditions with our study sites.

The future scope of this study involves extending research to more comprehensively capture the impact of crop residue burning on air pollution. While the current work is limited to the post-monsoon season, examining its effects across all seasons would provide a more complete understanding. Again, this study is based on the daily PM concentration data. More accurate prediction model can be developed using a higher frequency time series data like hourly data. Additionally, since the present study is only based on the time series data of PM, future research could explore the effects of the causal variates of pollution like traffic intensity, construction, and fuel consumption, as well as meteorological factors such as wind direction, temperature, and humidity to more accurately forecast the PM concentration.

Furthermore, the use of other decomposition techniques, such as empirical mode decomposition (EMD), for forecasting PM concentrations could be explored. EMD has been shown to perform well in time-series forecasting, and incorporating it into the wavelet decomposed hybrid models could improve the accuracy of the predictions. Again, evaluation against CNN-LSTM networks that can capture both spatial patterns and temporal dependencies in PM concentration sequences could be done. Transformer-based models with attention mechanisms could potentially identify the most effective temporal patterns. Investigation of wavelet-enhanced DL models could be done, where wavelet decomposition serves as a preprocessing step for CNN-LSTM or Transformer architectures. This could combine the multi-resolution analysis capabilities of wavelets with the pattern recognition strengths of DL. The inclusion of feature selection methods and optimization techniques could also improve the models' predictive accuracy.

In conclusion, this study provides valuable insights into forecasting $PM_{2.5}$ and PM_{10} concentrations in Delhi during the post-monsoon season. The findings can be used to inform policymakers on location-specific actions to reduce air pollution levels in Delhi caused by crop residue burning. Future research can expand on the current study by analyzing the impact of crop residue burning on air pollution throughout the year, incorporating additional variables, and exploring other decomposition techniques and ML methods to improve the accuracy of the models.

Acknowledgment

Research was supported by the Indian Council of Agricultural Research, Department of Agricultural Research and Education, Government of India. The authors are thankful to Arshad Reza for generating Figure 3 using GIS.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data are available from the corresponding author upon reasonable request.

Author Contribution Statement

Sandip Garai: Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Debopam Rakshit:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Arkaprava Roy:** Conceptualization, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Ranjit Kumar Paul:** Conceptualization, Methodology, Validation, Investigation, Resources, Writing – original draft, Writing – review & editing, Supervision, Project administration. **Ritwika Das:** Conceptualization, Validation, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization.

References

[1] Das, M., Das, A., Sarkar, R., Mandal, P., Saha, S., & Ghosh, S. (2021). Exploring short term spatio-temporal pattern of $PM_{2.5}$ and PM_{10} and their relationship with meteorological parameters during COVID-19 in Delhi. *Urban Climate*, *39*, 100944. https://doi.org/10.1016/j.uclim.2021.100944

[2] Myllyvirta, L., & Thieriot, H. (2020). *11,000 air pollution-related deaths avoided in Europe as coal, oil consumption plummet*. Centre for Research on Energy and Clean Air. https://energyandcleanair.org/wp/wp-content/uploads/2020/04/CREA-Europe-COVID-impacts.pdf

[3] Vohra, K., Vodonos, A., Schwartz, J., Marais, E. A., Sulprizio, M. P., & Mickley, L. J. (2021). Global mortality from outdoor fine particle pollution generated by fossil fuel combustion: Results from GEOS-Chem. *Environmental Research*, *195*, 110754. https://doi.org/10.1016/j.envres.2021.110754

[4] Dimitrova, A., Marois, G., Kiesewetter, G., KC, S., Rafaj, P., & Tonne, C. (2021). Health impacts of fine particles under climate change mitigation, air quality control, and demographic change in India. *Environmental Research Letters*, *16*(5), 054025. https://doi.org/10.1088/1748-9326%2Fabe5d5

[5] World Health Organization. (2024). *Ambient (outdoor) air pollution*. https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[6] Bongaerts, E., Lecante, L. L., Bové, H., Roeffaers, M. B. J., Ameloot, M., Fowler, P. A., & Nawrot, T. S. (2022). Maternal exposure to ambient black carbon particles and their presence in maternal and fetal circulation and organs: An analysis of two independent population-based observational studies. *The Lancet Planetary Health*, *6*(10), e804–e811. https://doi.org/10.1016/S2542-5196(22)00200-5

[7] Perera, F. (2018). Pollution from fossil-fuel combustion is the leading environmental threat to global pediatric health and equity: Solutions exist. *International Journal of Environmental Research and Public Health*, *15*(1), 16. https://doi.org/10.3390/ijerph15010016

[8] India State-Level Disease Burden Initiative Air Pollution Collaborators. (2021). Health and economic impact of air pollution in the states of India: The Global Burden of Disease Study 2019. *The Lancet Planetary Health*, *5*(1), e25–e38. https://doi.org/10.1016/S2542-5196(20)30298-9

[9] Chowdhury, S., Dey, S., Guttikunda, S., Pillarisetti, A., Smith, K. R., & di Girolamo, L. (2019). Indian annual ambient air quality standard is achievable by completely mitigating emissions from household sources. *Proceedings of the National Academy of Sciences*, *116*(22), 10711–10716. https://doi.org/10.1073/pnas.1900888116

[10] Tang, D., Zhan, Y., & Yang, F. (2024). A review of machine learning for modeling air quality: Overlooked but important issues. *Atmospheric Research*, *300*, 107261. https://doi.org/10.1016/j.atmosres.2024.107261

[11] Zhou, S., Wang, W., Zhu, L., Qiao, Q., & Kang, Y. (2024). Deep-learning architecture for $PM_{2.5}$ concentration prediction: A review. *Environmental Science and Ecotechnology*, *21*, 100400. https://doi.org/10.1016/j.ese.2024.100400

[12] Masood, A., & Ahmad, K. (2020). A model for particulate matter ($PM_{2.5}$) prediction for Delhi based on machine learning approaches. *Procedia Computer Science*, *167*, 2101–2110. https://doi.org/10.1016/j.procs.2020.03.258

[13] Palanichamy, N., Haw, S.-C., S, S., Murugan, R., & Govindasamy, K. (2022). Machine learning methods to predict particulate matter $PM_{2.5}$. *F1000Research*, *11*, 406. https://doi.org/10.12688/f1000research.73166.1

[14] Kristiani, E., Lin, H., Lin, J.-R., Chuang, Y.-H., Huang, C.-Y., & Yang, C.-T. (2022). Short-term prediction of $PM_{2.5}$ using LSTM deep learning methods. *Sustainability*, *14*(4), 2068. https://doi.org/10.3390/su14042068

[15] Bai, Y., Zeng, B., Li, C., & Zhang, J. (2019). An ensemble long short-term memory neural network for hourly $PM_{2.5}$ concentration forecasting. *Chemosphere*, *222*, 286–294. https://doi.org/10.1016/j.chemosphere.2019.01.121

[16] Xiao, F., Yang, M., Fan, H., Fan, G., & Al-qaness, M. A. A. (2020). An improved deep learning model for predicting daily PM2.5 concentration. *Scientific Reports*, *10*(1), 20988. https://doi.org/10.1038/s41598-020-77757-w

[17] Chae, S., Shin, J., Kwon, S., Lee, S., Kang, S., & Lee, D. (2021). PM10 and PM2.5 real-time prediction models using an interpolated convolutional neural network. *Scientific Reports*, *11*(1), 11952. https://doi.org/10.1038/s41598-021-91253-9

[18] Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., ..., & Schwartz, J. (2019). An ensemble-based model of $PM_{2.5}$ concentration across the contiguous United States with high spatiotemporal resolution. *Environment International*, *130*, 104909. https://doi.org/10.1016/j.envint.2019.104909

[19] Ejohwomu, O. A., Shamsideen Oshodi, O., Oladokun, M., Bukoye, O. T., Emekwuru, N., Sotunbo, A., & Adenuga, O. (2022). Modelling and forecasting temporal $PM_{2.5}$ concentration using ensemble machine learning methods. *Buildings*, *12*(1), 46. https://doi.org/10.3390/buildings12010046

[20] Dai, H., Huang, G., Zeng, H., & Zhou, F. (2022). $PM_{2.5}$ volatility prediction by XGBoost-MLP based on GARCH models. *Journal of Cleaner Production*, *356*, 131898. https://doi.org/10.1016/j.jclepro.2022.131898

[21] Peng, J., Han, H., Yi, Y., Huang, H., & Xie, L. (2022). Machine learning and deep learning modeling and simulation for predicting PM2.5 concentrations. *Chemosphere*, *308*, 136353. https://doi.org/10.1016/j.chemosphere.2022.136353

[22] Kleine Deters, J., Zalakeviciute, R., Gonzalez, M., & Rybarczyk, Y. (2017). Modeling $PM_{2.5}$ urban pollution using machine learning and selected meteorological parameters. *Journal of Electrical and Computer Engineering*, *2017*(1), 5106045. https://doi.org/10.1155/2017/5106045

[23] Pak, U., Ma, J., Ryu, U., Ryom, K., Juhyok, U., Pak, K., & Pak, C. (2020). Deep learning-based PM2.5 prediction considering the spatiotemporal correlations: A case study of Beijing, China. *Science of the Total Environment*, *699*, 133561. https://doi.org/10.1016/j.scitotenv.2019.07.367

[24] Zhang, Z., Wu, L., & Chen, Y. (2020). Forecasting $PM_{2.5}$ and PM_{10} concentrations using GMCN (1, *N*) model with the similar meteorological condition: Case of Shijiazhuang in China. *Ecological Indicators*, *119*, 106871. https://doi.org/10.1016/j.ecolind.2020.106871

[25] Zender-Świercz, E., Galiszewska, B., Telejko, M., & Starzomska, M. (2024). The effect of temperature and humidity of air on the concentration of particulate matter-PM2.5 and PM10. *Atmospheric Research*, *312*, 107733. https://doi.org/10.1016/j.atmosres.2024.107733

[26] Gregório, J., Gouveia-Caridade, C., & Caridade, P. J. S. B. (2022). Modeling PM2.5 and PM10 using a robust simplified linear regression machine learning algorithm. *Atmosphere*, *13*(8), 1334. https://doi.org/10.3390/atmos13081334

[27] Zaman, N. A. F. K., Kanniah, K. D., Kaskaoutis, D. G., & Latif, M. T. (2021). Evaluation of machine learning models for estimating $PM_{2.5}$ concentrations across Malaysia. *Applied Sciences*, *11*(16), 7326. https://doi.org/10.3390/app11167326

[28] Pan, S., Du, S., Wang, X., Zhang, X., Xia, L., Liu, J., ..., & Wei, Y. (2019). Analysis and interpretation of the particulate matter (PM10 and PM2.5) concentrations at the subway stations in Beijing, China. *Sustainable Cities and Society*, *45*, 366–377. https://doi.org/10.1016/j.scs.2018.11.020

[29] Cheng, Y., Zhang, H., Liu, Z., Chen, L., & Wang, P. (2019). Hybrid algorithm for short-term forecasting of $PM_{2.5}$ in China. *Atmospheric Environment*, *200*, 264–279. https://doi.org/10.1016/j.atmosenv.2018.12.025

[30] Wang, P., Zhang, G., Chen, F., & He, Y. (2019). A hybrid-wavelet model applied for forecasting $PM_{2.5}$ concentrations in Taiyuan city, China. *Atmospheric Pollution Research*, *10*(6), 1884–1894. https://doi.org/10.1016/j.apr.2019.08.002

[31] Liu, H., Duan, Z., & Chen, C. (2020). A hybrid multi-resolution multi-objective ensemble model and its application for forecasting of daily PM2.5 concentrations. *Information Sciences*, *516*, 266–292. https://doi.org/10.1016/j.ins.2019.12.054

[32] Qiao, W., Wang, Y., Zhang, J., Tian, W., Tian, Y., & Yang, Q. (2021). An innovative coupled model in view of wavelet transform for predicting short-term PM10 concentration. *Journal of Environmental Management*, *289*, 112438. https://doi.org/10.1016/j.jenvman.2021.112438

[33] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control*. USA: John Wiley & Sons.

[34] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

[35] Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, *19*(1), 1–67. https://doi.org/10.1214/aos/1176347963

[36] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

[37] Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[38] Barthwal, A., Acharya, D., & Lohani, D. (2023). Prediction and analysis of particulate matter ($PM_{2.5}$ and PM_{10}) concentrations using machine learning techniques. *Journal of Ambient Intelligence and Humanized Computing*, *14*(3), 1323–1338. https://doi.org/10.1007/S12652-021-03051-W

[39] Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational data mining techniques for student performance prediction: Method review and comparison analysis. *Frontiers in Psychology*, *12*, 698490. https://doi.org/10.3389/fpsyg.2021.698490

[40] Yadav, V., Yadav, A. K., Singh, V., & Singh, T. (2024). Artificial neural network an innovative approach in air pollutant prediction for environmental applications: A review. *Results in Engineering*, *22*, 102305. https://doi.org/10.1016/j.rineng.2024.102305

[41] Percival, D. B., & Walden, A. T. (2000). *Wavelet methods for time series analysis*. UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511841040

[42] Amolins, K., Zhang, Y., & Dare, P. (2007). Wavelet based image fusion techniques—An introduction, review and comparison. *ISPRS Journal of Photogrammetry and Remote Sensing*, *62*(4), 249–263. https://doi.org/10.1016/j.isprsjprs.2007.05.009

[43] Taswell, C. (2000). Constraint-selected and search-optimized families of Daubechies wavelet filters computable by spectral factorization. *Journal of Computational and Applied Mathematics*, *121*(1–2), 179–195. https://doi.org/10.1016/S0377-0427(00)00340-X

[44] Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, *41*(7), 909–996. https://doi.org/10.1002/cpa.3160410705

[45] Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*(7), 674–693. https://doi.org/10.1109/34.192463

[46] Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

[47] Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*(366a), 427–431. https://doi.org/10.1080/01621459.1979.10482531

[48] Phillips, P. C. B., & Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, *75*(2), 335–346. https://doi.org/10.1093/biomet/75.2.335

[49] Broock, W. A., Scheinkman, J. A., Dechert, W. D., & LeBaron, B. (1996). A test for independence based on the correlation dimension. *Econometric Reviews*, *15*(3), 197–235. https://doi.org/10.1080/07474939608800353

[50] Rakshit, D., Roy, A., Atta, K., Adhikary, S., & Vishwanath. (2022). Modeling temporal variation of particulate matter concentration at three different locations of Delhi. *International Journal of Environment and Climate Change*, *12*(11), 1831–1839. https://doi.org/10.9734/ijecc/2022/v12i1131191

[51] Chakraborty, S. (2022). TOPSIS and modified TOPSIS: A comparative analysis. *Decision Analytics Journal*, *2*, 100021. https://doi.org/10.1016/j.dajour.2021.100021

[52] Khan, A., Suryanarayanan, U., Ganguly, T., & Ganesan, K. (2022). *Improving air quality management through forecasts: A case study of Delhi's air pollution of winter 2021*. Council on Energy, Environment and Water. https://www.ceew.in/sites/default/files/ceew-research-on-air-quality-management-delhi-winter-pollution-case-study-2021.pdf

[53] Bag, K., Roy, A., Tigga, P., Didawat, R. K., Kumar, P., & Dey, A. (2022). Real-time monitoring of crop residue burning through satellite remote sensing. In M. Nemichandrappa & R. R. Halidoddi (Eds.), *Advanced innovative technologies in agricultural engineering for sustainable agriculture* (Vol. 4, pp. 55–79). AkiNik Publications.

[54] Dutta, A., Patra, A., Hazra, K. K., Nath, C. P., Kumar, N., & Rakshit, A. (2022). A state of the art review in crop residue burning in India: Previous knowledge, present circumstances and future strategies. *Environmental Challenges*, *8*, 100581. https://doi.org/10.1016/j.envc.2022.100581

[55] Sen, M., Roy, A., Rani, K., Nalia, A., Das, T., Tigga, P., ..., & Das, A. (2024). Crop residue: Status, distribution, management, and agricultural sustainability. In V. S. Meena, A. Rakshit, M. D. Meena, M. Baslam, I. R. Fattah, S. S. Lam, & J. S. Kaba (Eds.), *Waste management for sustainable and restored agricultural soil* (pp. 167–201). Academic Press. https://doi.org/10.1016/B978-0-443-18486-4.00017-8

How to Cite: Garai, S., Rakshit, D., Roy, A., Paul, R. K., & Das, R. (2026). Wavelet-Based Combinatorial Machine Learning Techniques for Forecasting Particulate Matter Concentration. *Journal of Data Science and Intelligent Systems*. https://doi.org/10.47852/bonviewJDSIS62026021