


## Research

# Assessing particulate matter (PM<sub>2.5</sub>) concentrations and variability across Maharashtra using satellite data and machine learning techniques

Ganesh Machhindra Kunjir<sup>1,2</sup> · Suvarna Tikle<sup>3</sup> · Sandipan Das<sup>1</sup>  · Masud Karim<sup>1</sup>  · Sujit Kumar Roy<sup>4</sup>  · Uday Chatterjee<sup>5</sup> 

Received: 24 December 2024 / Accepted: 21 March 2025

Published online: 04 April 2025

© The Author(s) 2025 

## Abstract

Airborne fine particulate matter (PM<sub>2.5</sub>) is recognized globally as one of the most hazardous air pollutants due to its profound impact on human health, contributing to respiratory and cardiovascular diseases, and increasing the risk of premature mortality. The World Health Organization (WHO) attributes millions of deaths annually to PM<sub>2.5</sub> exposure, making it a critical subject of study for both environmental and public health research. In this context, the present study aims to predict PM<sub>2.5</sub> concentrations across Maharashtra, India, for the year 2023, employing machine learning models to improve spatial and temporal air quality assessments. The analysis utilizes daily station-specific datasets, incorporating PM<sub>2.5</sub> concentrations, Fine Aerosol Optical Depth (FAOD), wind components (u and v), relative humidity (RH), and air temperature (TEMP) to improve prediction accuracy. Four regression models were applied: Random Forest (RF), Multiple Linear Regression (MLR), Linear Regression (LR), and Lasso Regression, using a combination of Fine Aerosol Optical Depth (FAOD) with meteorological data from Google Earth Engine and ground-based observations from Central Pollution Control Board (CPCB) monitoring stations. The study emphasizes the importance of utilizing FAOD as a more refined metric for fine-mode aerosol concentration in PM<sub>2.5</sub> modeling, compared to conventional AOD. The RF model achieved the highest accuracy ( $R^2=0.87$ , RMSE = 12.57  $\mu\text{g}/\text{m}^3$ , MAE = 6.96  $\mu\text{g}/\text{m}^3$ ), outperforming MLR, LR, and Lasso Regression, which showed significantly lower  $R^2$  values. This highlights the RF model's effectiveness in capturing the non-linear relationships between PM<sub>2.5</sub> and its environmental factors. This study identified key PM<sub>2.5</sub> hotspots in Maharashtra, particularly in densely urbanized areas like Mumbai, Thane, and Pune, with annual PM<sub>2.5</sub> concentrations reaching 46.34  $\mu\text{g}/\text{m}^3$ , far exceeding the Indian National Ambient Air Quality Standards (NAAQS) of 40  $\mu\text{g}/\text{m}^3$ . Seasonal analysis revealed significant variability, with the highest PM<sub>2.5</sub> concentrations observed during the winter months, while levels significantly decreased during the monsoon due to higher rainfall and increased atmospheric moisture. The study identifies key PM<sub>2.5</sub> hotspots in urban areas, offering crucial insights for policymakers and urban planners to implement targeted air quality interventions. These findings support improved public health and sustainable environmental management in Maharashtra.

✉ Sandipan Das, sandipan@sig.ac.in; sandipanraj2002@gmail.com; Ganesh Machhindra Kunjir, phdgrad.ganesh.kunjir@siu.edu.in; ganeshkunjir100@gmail.com; Suvarna Tikle, sstikle@gmail.com; Masud Karim, 007karim.ai@gmail.com; Sujit Kumar Roy, sujitroy.bejoy@gmail.com; Uday Chatterjee, raj.chatterjee459@gmail.com | <sup>1</sup>Symbiosis Institute of Geoinformatics, Symbiosis International (Deemed University), Pune, Maharashtra 411016, India. <sup>2</sup>Department of Computer Science, Shri Saibaba College, Savitribai Phule Pune University, Shirdi, India. <sup>3</sup>Environmental Modeling Division, Max Planck Institute for Meteorology, Bundesstr. 53, 20146 Hamburg, Germany. <sup>4</sup>Institute of Water and Flood Management, Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh. <sup>5</sup>Department of Geography, Bhatnagar College, Dantan, Vidyasagar University, Kharagpur, West Bengal 721426, India.



**Keywords** Machine learning · Fine mode aerosol optical depth (FAOD) · Satellite remote sensing · Fine particulate matter (PM<sub>2.5</sub>)

## 1 Introduction

Airborne fine particulate matter (PM<sub>2.5</sub>) has many adverse impacts on the environment and public health, especially cardiovascular and respiratory diseases [1–4]. PM<sub>2.5</sub> pollution in recent years has been widely recognized as the most harmful air pollutants, increasing the risk of mortality and morbidity [5, 6]. The study [7], [50] on the impact of outdoor air pollution on human health provides valuable insights into the challenges posed by poor air quality, particularly in densely populated urban areas, where it presents significant public health risks. PM<sub>2.5</sub> pollution results from natural and anthropogenic sources including dust particles, emissions from vehicles [7], sea spray aerosols, industries, gas burning, and electricity generation [6, 8]. Every year, five million people worldwide die as an effect of prolonged exposure to PM<sub>2.5</sub> particles [6]. The Environmental Protection Agency in India is tasked with regularly assessing and suggesting changes to the national ambient air quality standards (NAAQS). In epidemiological study, it is important to understand the spatial and temporal properties as well as the distribution of PM<sub>2.5</sub> in order to assess the adverse impacts of air pollution on human health [3, 9, 42]. The scant and irregular distribution of air quality monitoring stations presently limits the monitoring of PM<sub>2.5</sub> contamination by ground stations. Therefore, for continuous observation and management of atmospheric PM<sub>2.5</sub> pollution, gathering information at a high temporal resolution spatial distribution of PM<sub>2.5</sub> is imperative [3, 10].

Satellite remote sensing can provide complete aerosol optical depth amplitudes on a global scale on a daily basis, and to allow predictions of terrestrial particle concentrations, satellite-derived aerosol optical depth (AOD) is routinely used to predict terrestrial PM<sub>2.5</sub> concentrations, especially aerosol optical depth (AOD) and PM<sub>2.5</sub> in areas where ground monitoring stations are not available because of its essential relationship [10–12]. Several studies have shown that AOD factors—a particulate matter including all of the limiting solar radiation in the atmospheric column—can be effectively used to predict PM<sub>2.5</sub> concentrations [3, 11, 13]. Numerous studies endorse including additional elements such as atmospheric parameters, land use features [14], and aerosol types into the AOD-PM modeling to enhance the accuracy of prediction of PM<sub>2.5</sub> based on AOD observations [15–17]. Many researchers around the world have used numerous methods to characterize the relationship between PM<sub>2.5</sub> and AOD, including multiple linear regression [18], satellite remote sensing [4, 13], chemical transport models [17], artificial neural networks [15, 19], land-use regression model (LUR) [5], geographically weighted regression (GWR) models [20] and geographically and temporarily weighted regression model [21]. In recent years, machine learning techniques have gained popularity for modeling the complex, nonlinear relationships between PM<sub>2.5</sub> and its various contributing factors, overcoming limitations associated with traditional statistical models. [17, 22–24]. Several machine learning models, which include support vector machine [25, 26], random forest [27, 28], artificial neural network (ANN) models [11], extreme gradient boosting model (XGBoost) [29], neural networks [20, 29], and Bayesian maximum entropy [30], have been utilized to estimate ground-level PM<sub>2.5</sub> concentrations. Due to several benefits, the Random Forest (RF) model has been effectively utilized in various regions worldwide [30].

Machine learning techniques are increasingly being used to estimate global PM<sub>2.5</sub> concentrations by integrating satellite data and ground-based measurements. Several studies have demonstrated the effectiveness of these approaches. For instance, studies [27, 35, 36] combined satellite-derived aerosol optical depth (AOD) with simulation data and ground-based observations to produce highly accurate global PM<sub>2.5</sub> estimates. Similarly, [37] utilized remote sensing, meteorological data, and ground-based observations to train a machine learning model for daily PM<sub>2.5</sub> estimation. Other research [37, 38] developed techniques using satellite observations and chemical transport models to generate long-term global PM<sub>2.5</sub> concentration data. Additionally, [39] employed a Random Forest model to refine grid-wise PM<sub>2.5</sub> estimations using MERRA-2 data and ground measurements, achieving high correlation across daily, monthly, and yearly scales. A comparative study [40] assessed the accuracy of six machine learning models, highlighting the superior performance of Artificial Neural Networks (ANN). These studies reinforce the potential of machine learning in integrating diverse datasets to improve global PM<sub>2.5</sub> estimation, which is crucial for environmental health research and policy-making [21]. Beyond modeling, research [41] emphasizes the need for stricter emission regulations in industrial and vehicular sectors to mitigate PM<sub>2.5</sub> pollution. Previous research has also shown that models such as Convolutional Neural Networks (CNN) and Random Forest (RF) can achieve high predictive accuracy, with R<sup>2</sup> values exceeding 0.97 and RMSE around 16% of

the standard deviation [42]. However, these models often rely on specific datasets and may struggle to fully capture the complex non-linear relationships between  $PM_{2.5}$  concentrations and environmental factors.

Using satellite data, machine learning methods have effectively estimated  $PM_{2.5}$  concentrations over many places in India. Studies show that algorithms like random forest and XGBoost significantly improve estimation accuracy when combined with meteorological data. For example, random forest achieved an  $R^2$  of 0.86 in reconstructing  $PM_{2.5}$  from MERRA-2 data, proving its utility in data-scarce regions like India. Moreover, researchers have applied diverse machine learning (ML) techniques to estimate  $PM_{2.5}$  levels [31], including support vector regression, evolutionary adaptive neuro-fuzzy systems [32], and ensemble methods. Researchers have combined satellite-derived data with chemical transport models [33, 34] and multiple predictors to improve accuracy. A comprehensive framework using ensemble averaging across four learners achieved a cross-validation  $R^2$  of 0.84 for daily  $PM_{2.5}$  estimations at high spatiotemporal resolution [35]. These approaches have revealed significant increases in  $PM_{2.5}$  levels across most Indian states since 1980 [36]. Integrating satellite data with advanced machine learning techniques offers a viable pathway for effective  $PM_{2.5}$  monitoring and management in India, particularly in regions with limited ground-based monitoring.

Many existing models focus on either short-term or long-term  $PM_{2.5}$  predictions but often struggle to maintain accuracy across different time scales. For instance, while the CNN-RF ensemble model enhances prediction accuracy by integrating feature extraction and regression techniques, it still faces challenges in adapting to varying environmental conditions. Our approach overcomes these limitations by employing a combination of machine learning techniques that enhance both flexibility and accuracy across different time intervals [43]. Although earlier research has investigated different machine learning methods for predicting  $PM_{2.5}$  levels, there is still an absence of thorough analyses that incorporate diverse data sources and account for seasonal changes in air quality. Our study addresses this shortcoming by offering a comprehensive assessment of  $PM_{2.5}$  hotspots across Maharashtra, India, while also emphasizing the seasonal trends that affect pollution rates.

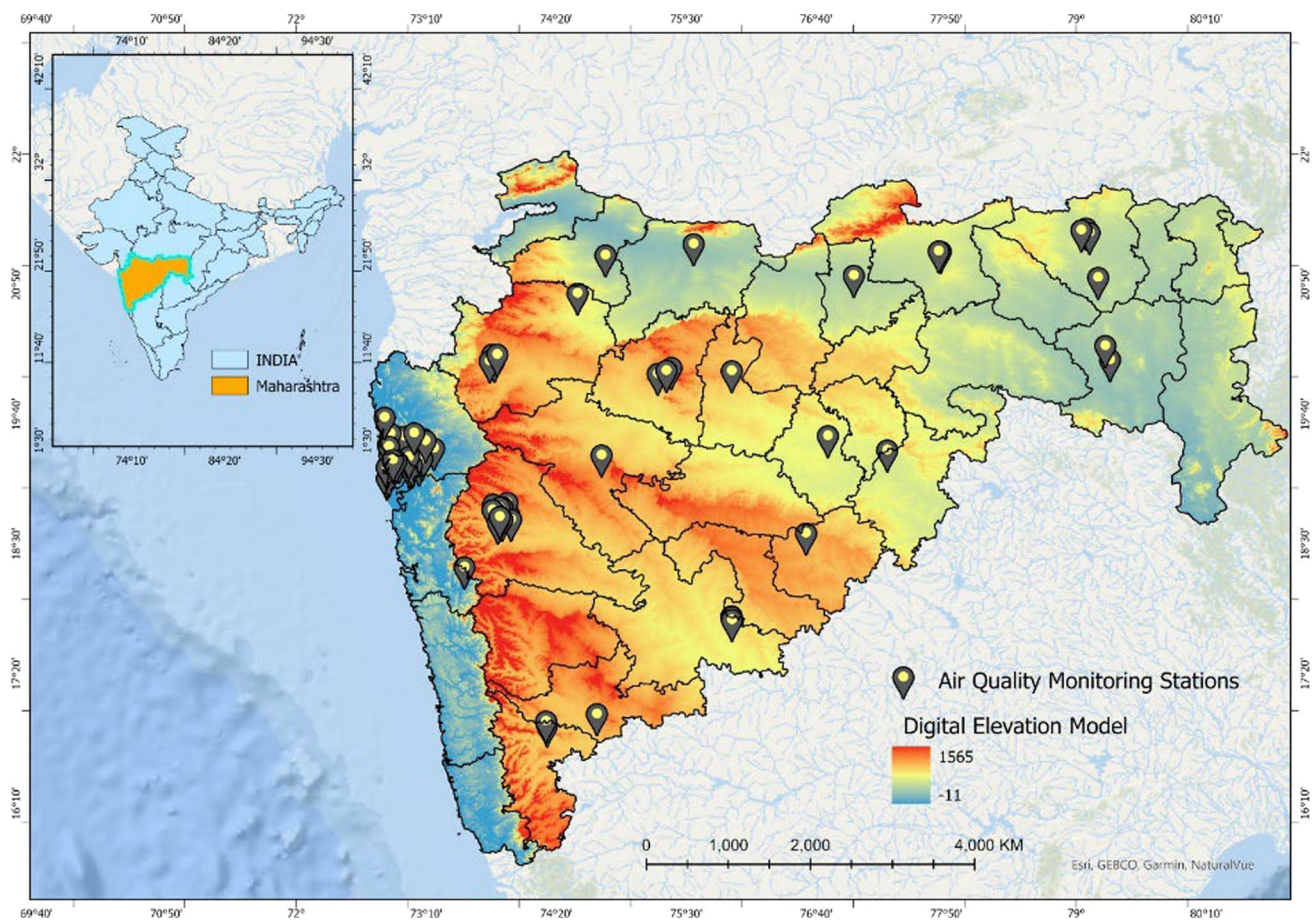
This study aims to address key research gaps in  $PM_{2.5}$  estimation by leveraging machine learning techniques, to predict fine particulate concentrations across Maharashtra at a high spatial resolution (1 km). Unlike previous studies that primarily rely on traditional AOD, this research incorporates FAOD, a more refined metric for fine-mode aerosols, alongside meteorological parameters to enhance prediction accuracy. The objectives of this study are (1) to develop an accurate machine learning-based model for  $PM_{2.5}$  estimation and (2) to analyze seasonal and spatial variations in  $PM_{2.5}$  across Maharashtra. By integrating ground-based air quality measurements with satellite-derived data [37], this study overcomes the challenge of sparse monitoring stations, providing a comprehensive spatial representation of air pollution. However, some limitations exist, such as potential uncertainties in FAOD-derived estimates, reduced model accuracy during monsoon due to high humidity affecting aerosol measurements, and the lack of real-time emission sources like traffic and industrial activities in the model. Despite these challenges, the findings offer valuable insights for policymakers, urban planners, and environmental managers, enabling targeted interventions to mitigate air pollution and improve public health in Maharashtra.

## 2 Study area

The state of Maharashtra is in the western part of Peninsular India, with latitudinal extends from 15°33'46" N to 22° 02'13" N and 72°38'45" E to 80°53'17" E longitude (Fig. 1). This Indian state is bordered by several others state boundaries: to the north, Gujarat and Madhya Pradesh; to the northwest, the Arabian Sea; to the south, Karnataka and Goa; to the southeast, Telangana; and to the east, Chhattisgarh. Spanning a vast area of 3.07 lakh km<sup>2</sup>, Maharashtra is renowned as India's financial hub and a prominent hub for industry and commerce. According to the 2011 census (<http://www.censusindia.gov.in/>), this state is the third most populous in the country, with a population of 1,123,743,333 people.

The population density is 365 persons per square kilometer. Maharashtra is host to numerous notable cities, including as Mumbai, Pune, Nagpur, and Nashik. The state has been categorized into coastal Konkan, Western Maharashtra, Marathwada, North Maharashtra, and Vidarbha administrative divisions. The five states, including Madhya Pradesh, Chhattisgarh, Telangana, Karnataka and Goa, Maharashtra is in physical contact with the study area, along a coastline of 720 kms stretching across the Arabian Sea. The geography of Maharashtra is largely characterized by the Deccan Plateau, Sahyadri (the Western Ghats) and Konkan, or the coast of the Sahyadri's to the west. These physical features contribute to the state's distinct geography and weather patterns. The Western Ghats have an average elevation between 1000–1200 m, with the highest





**Fig. 1** Study area with air quality monitoring stations

peak being Kalsubai, which reaches 1646 m. The Western Ghats also encompass multiple hill stations, such as Matheran and Mahabaleshwar, and are characterized by a series of steep escarpments called 'ghats', which slope steeply towards the coastal zone. The Western Ghats act as a natural barrier, dividing the coastal region of Konkan from the Deccan Plateau. The Konkan coastal region represents a narrow strip of land sandwiched between the Western Ghats and the Arabian Sea. The coastal belt is approximately 50 km wide, which gradually declines from north to south. Multiple river creeks intersect the shoreline and the various Sahyadri branches reach the shoreline. The rivers of the Konkan region such as, the Ulhas, the Savitri, the Vashishti, and the Shastri flow quickly into the Arabian Sea.

The southwest monsoon season in Maharashtra, occurring from June to September, constitutes 88.4% of the state's total annual precipitation. The area exhibits a customary tropical monsoonal climate, distinguished by humid summers and arid winters. The western slopes of the Ghats have significant rainfall, ranging from 2000 to 4000 mm annually. However, when one moves across the Ghats to the east, there is a reduction in rainfall, and the foothills on the eastern side see virtually no rainfall due to rain shadow effect. The mean yearly temperature in Maharashtra typically varies between 25 and 30 degrees Celsius. The CPCB (Central Pollution Control Board) runs a system of stations for monitoring air quality throughout the state of Maharashtra in order to evaluate the magnitude of air pollution. The CPCB oversees the national air quality monitoring program (NAMP) through various monitoring agencies. The CPCB monitors various data, such as particulate matter (PM) concentrations, real-time data from monitoring stations on air quality, a live air quality index, and other pertinent information. The Maharashtra Central Pollution Control Board (CPCB) network monitoring stations are integral components of a comprehensive nationwide network of monitoring stations. This network serves as a valuable source of data for many projects, including graded response action plans and initiatives aimed at air quality control.

### 3 Materials

#### 3.1 Ground PM<sub>2.5</sub> measurements

The daily average concentrations of PM<sub>2.5</sub> were obtained from the portal of the Central Pollution Control Board (CPCB) (<https://cpcb.nic.in>). 90 air quality monitoring stations data in Maharashtra were collected from the official CPCB website for the period of January 1st, 2023, to December 31st, 2023 as per availability of the required data. The region's air quality monitoring stations are dispersed unevenly. Due to data limitations, many monitoring stations outside of urban centers lacked sufficient corresponding datasets. Consequently, the available ground-truth data are primarily focused in urban and industrial regions, hence creating a spatial bias in the dataset. This bias towards urban areas indicates an emphasis on infrastructure in places with higher pollution risk, thus neglecting rural and less-industrialized areas in estimations. As depicted in Fig. 1, half of the monitoring stations are in Mumbai and Pune, while the remaining stations are spread out among other important cities. Ground PM<sub>2.5</sub> levels were monitored using beta gauge attenuation monitors (BAM-1020; Met One Instrument) that provide hourly average concentrations. The monitoring process followed calibration and rigorous quality checks by India's National Ambient Air Quality Standards (NAAQS). In addition, a few stations have missing hourly data. We obtained data before and after the observed period to apply as substitute values. The mean value of the replaced data was then computed to fill in the missing values. As per a quality-controlled approach, the missing data and outliers brought about by incorrect detection and natural variables were eliminated [21].

#### 3.2 MODIS AOD product

The study employs Fine Aerosol Optical Depth (FAOD) as a critical variable to understand aerosol concentration, specifically for its relevance in assessing PM<sub>2.5</sub> levels. FAOD is a more refined measure compared to the conventional Aerosol Optical Depth (AOD) because it specifically focuses on the fine-mode aerosols, which are directly related to particulate matter like PM<sub>2.5</sub>. Fine-mode aerosols, which have a radius of less than 1 micron, are primarily produced by combustion processes such as biomass burning and industrial or auto pollution. These particles, often considered anthropogenic in origin, are of particular concern due to their potential for deeper penetration into the human respiratory system and their larger impact on human health, earth's radiation budget, cloud processes, and climate (NASA, 2023). The formula for FAOD can be expressed as:

$$FAOD = AOD \times FineModeFraction \quad (1)$$

where AOD is the total Aerosol Optical Depth and Fine Mode Fraction refers to the proportion of the aerosol load that corresponds to fine-mode particles. The FAOD values used in this study were extracted from the MODIS MCD19A2 dataset, which provides both AOD and Fine Mode Fraction at a global scale. For the study period from January 1, 2023, to December 31, 2023, daily FAOD values were extracted for CPCB stations across the region, enabling a detailed temporal and spatial analysis. This approach, which emphasizes FAOD over AOD, addresses the previous limitation in the study and provides a more accurate reflection of fine particulate matter in the atmosphere, aligning the analysis with current research standards and guidelines for PM<sub>2.5</sub> modeling.

FAOD is considered a more relevant metric for understanding fine particulate matter because it specifically captures the impact of fine-mode aerosols (particles with a radius of less than 1 micron), which have a higher health risk due to their ability to penetrate the lungs. This is consistent with the international standard for air quality, where PM<sub>2.5</sub> mass, a direct result of fine-mode aerosols, is the primary indicator for evaluating air quality and its potential health effects. By using FAOD in this study, the research more accurately aligns with current aerosol research methodologies and offers better precision in estimating PM<sub>2.5</sub> concentrations [43].

#### 3.3 Meteorological data

The European Centre for medium-range weather forecasting' ERA5 product offers hourly values of atmospheric parameters; it is the fifth-generation atmospheric reanalysis of the global climate dataset. The ERA5 data have seen extensive use because to their superior spatial and temporal resolution compared to the National Center for Environmental Prediction. To develop machine learning model for PM<sub>2.5</sub> concentration, we sourced meteorological data from the ERA5 dataset

(Copernicus, <https://cds.climate.copernicus.eu/>), focusing on key variables such as Relative Humidity, Temperature, and wind components (U and V). Atmosphere:  $0.25^\circ \times 0.25^\circ$  for reanalysis;  $0.5^\circ \times 0.5^\circ$  for ensemble products. The dataset was processed and extracted using Google Earth Engine platform.

### 3.3.1 Relative humidity

To calculate relative humidity (RH), we used air temperature (T) and dewpoint temperature (Td) data from the ERA5-Land Hourly dataset provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). Relative humidity, a critical parameter for understanding atmospheric moisture content, is calculated using the widely recognized Magnus formula. The calculation involves determining the saturation vapor pressure ( $e_s$ ) and the actual vapor pressure ( $e$ ). Saturation vapor pressure was derived using the equation:

$$e_s = 6.112 \times \exp \left[ \frac{(17.67 \times T)}{T + 243.5} \right] \quad (2)$$

while actual vapor pressure was computed using:

$$e = 6.112 \times \exp \left[ \frac{17.67 \times T_d}{T_d + 243.5} \right] \quad (3)$$

where T and Td are in Celsius. Relative humidity is then obtained using the formula:

$$RH = 100 \times \left( \frac{e}{e_s} \right) \quad (4)$$

The air and dewpoint temperatures from the ERA5-Land dataset were converted from Kelvin to Celsius before application of these formulas. The dataset was processed in Google Earth Engine, and daily averages of relative humidity for specific monitoring locations were extracted and linked with PM<sub>2.5</sub> concentration measurements for further analysis. The Magnus formula used for this computation is an empirically validated approximation derived from the Clausius–Clapeyron equation, as discussed by Sonntag (1990) and further refined by Alduchov and Eskridge (1996). These sources confirm the accuracy of the method for meteorological applications, ensuring robust calculations for studies involving PM<sub>2.5</sub> and related air quality indices [41].

### 3.3.2 Temperature

The air temperature (T) data at 2 m above the surface was extracted from the ERA5-Land Hourly dataset provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). This dataset offers hourly global coverage with a spatial resolution of approximately 11 km. For this analysis, daily mean temperature values were computed by averaging hourly data for each day, covering the period from 2023-01-01 to 2023-12-31. Using Google Earth Engine, the data was spatially reduced for specific monitoring station locations using the mean temperature for the corresponding grid cells. The temperature values, initially in Kelvin, were converted to Celsius to facilitate compatibility with standard climatological analyses and subsequent integration with PM<sub>2.5</sub> concentration measurements.

### 3.3.3 U and V component of wind

The u-component and v-component of wind at 10 m above the surface were extracted from the ERA5-Land Hourly dataset provided by ECMWF. These variables represent the east–west (u-component) and north–south (v-component) directional components of wind velocity, expressed in meters per second (m/s). The analysis involved calculating daily mean values by averaging the hourly data for each day, ensuring temporal consistency with the study period. Data was processed for the period 2023-01-01 to 2023-12-31. The data was spatially reduced using Google Earth Engine to compute mean values for specific station locations. These components provide essential insights into wind patterns and directions,

forming a basis for calculating resultant wind speed and direction, which are critical parameters in atmospheric studies, particularly for understanding pollutant dispersion and PM<sub>2.5</sub> concentration dynamics.

## 4 Methodology

### 4.1 Data processing and integration

The study utilized daily ground-based PM<sub>2.5</sub> concentration data, satellite-derived Fine Aerosol Optical Depth (FAOD) and various meteorological parameters, to model PM<sub>2.5</sub> concentrations. The data processing involved extracting daily CPCB network monitoring station-specific datasets for PM<sub>2.5</sub>, FAOD, which was calculated from MODIS Terra AOD gridded Level 2 product produced daily at 1 km resolution and fine mode fraction, wind components (u and v), relative humidity (RH), and air temperature (TEMP) extracted from ERA5 dataset by using google earth engine tools for the year 2023 (from January 1st to December 31st). The datasets were merged using Python, aligning the data from multiple sources for each station location across the entire study period. This integrated dataset, which represents the spatial and temporal variability of air quality and meteorological factors, was subsequently processed for machine learning model development (Fig. 2). The preprocessing steps included handling missing data and outliers, where records were flagged for inconsistencies, particularly when discrepancies arose between ground-based PM<sub>2.5</sub> measurements and the corresponding satellite or meteorological data. These anomalies were addressed to ensure data quality and consistency. After merging the data, the dataset underwent scaling, ensuring that every feature has an equal impact on the model's performance. Scaling aids normalizing meteorological data because it might vary greatly in range (for example, temperature in degrees vs. humidity as a percentage), feature engineering, and preparation for machine learning models like random forest regressor, multilinear, linear regression and lasso regression. The entire workflow was implemented using Python 3.10 for data manipulation and modeling, while spatial analysis and visualization were carried out using ArcGIS Pro. This integrated and cleaned dataset was then used to train models to predict PM<sub>2.5</sub> concentrations with high temporal and spatial accuracy. Models was trained with 80% dataset from the total dataset containing 12563 number of records and remaining 20% dataset were used for testing the model, as shown in the methodology section of the study.

### 4.2 Model development

#### 4.2.1 Random forest (RF) model

RF model is an ensemble-based decision tree approach that combines multiple decision trees trained on randomly selected subgroup of training samples [6]. Random forest objectively evaluates feature relevance during classification and can handle issues associated with a significant volume of missing data. Furthermore, the RF model outperforms the standard models in handling large amounts of data without requiring any dimensionality reduction. Classification trees were employed to select the most effective trees for predictive purposes. Random forest involves many classification trees, where all variables are involved in each tree as independent features for classification. To conduct this study, hyperparameter tuning was conducted using Grid Search Cross-Validation, optimizing *n\_estimators* (100), *max\_depth* (20), *min\_samples\_split* (5), *min\_samples\_leaf* (2), and *max\_features* ('sqrt') for better model performance. The final predictions produced by the RF model have been intended to be calculated using the mean of the outcomes from all separate trees.

$$\widehat{PM_{2.5}} = \frac{1}{B} \sum_{b=1}^B f_b(x) \quad (5)$$

where  $f_b(x)$  is nothing but the output from every tree  $b$ .

The standard deviation of predictions from each individual tree can be calculated by Eq. (6) to evaluate prediction uncertainty, especially for regression tasks.

$$\sigma = \sqrt{\frac{\sum_{b=1}^B (f_b(x') - \hat{f})^2}{B - 1}} \quad (6)$$

where  $x'$  is the input for which predictions are made,  $\hat{f}$  is the average prediction across all trees.



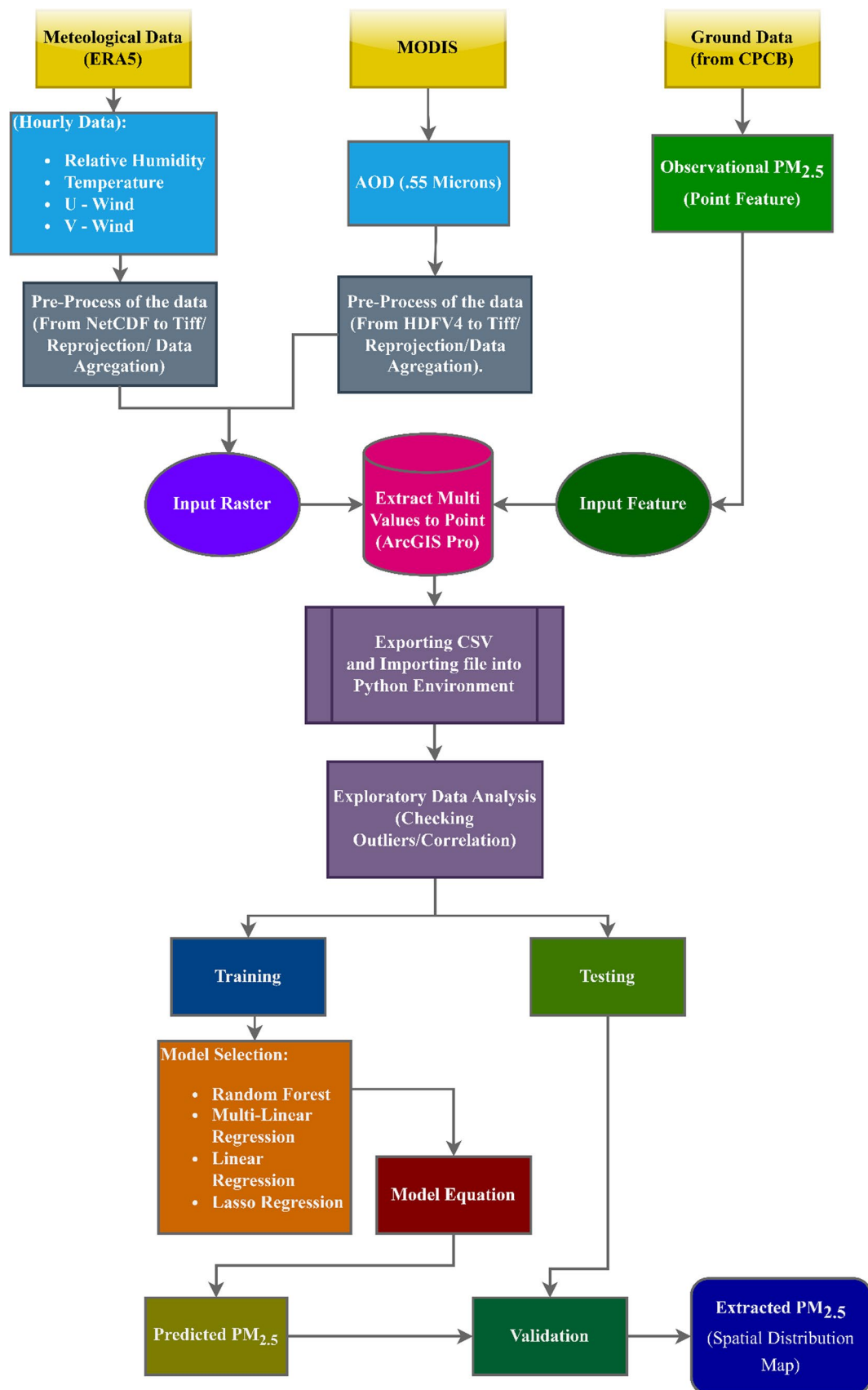


Fig. 2 Methodology flowchart for estimation of PM<sub>2.5</sub>



#### 4.2.2 Multiple linear regression (MLR)

The multiple linear regression model is a statistical method used to examine the correlation between a dependent variable and two or more independent variables. The multiple linear regression model is employed to construct a model that explains a dependent variable  $y$  in terms of numerous independent variables. Determine the estimated value of the dependent variable  $y$  when the values of the independent variables are given. Multilinear regression does not have traditional hyperparameters like decision trees or ensemble methods, but model performance can be improved through feature selection, regularization, and interaction terms. The model is commonly estimated using python software, and the results are evaluated by analyzing the regression coefficients, their standard errors, and the  $p$ -values associated with each coefficient. The model can also be utilized to detect and manage confounding variables in the analysis. The model is represented by the following formula:

$$PM_{2.5} = \beta_0 + \beta_1(FAOD) + \beta_2(Temp) + \beta_3(U - Wind) + \beta_4(V - Wind) + \beta_5(RH) + \epsilon \quad (7)$$

The dependent variable in this case is  $PM_{2.5}$ , which represents the concentration of particulate matter having a diameter of  $2.5 \mu g/m^3$  or smaller.  $\beta_0$  is the  $y$ -intercept, which denotes the  $PM_{2.5}$  value when all independent variables are set to 0. The regression coefficients  $\beta_1, \beta_2, \beta_3, \beta_4$ , and  $\beta_5$  indicate the amount of change in  $PM_{2.5}$  that occurs when there is a one-unit change in the corresponding independent variable while keeping all other variables constant. FAOD refers to the Fine mode Aerosol Optical Depth, which quantifies the concentration of aerosol particles present in the vertical column of the atmosphere. Temperature refers to the degree of hotness or coldness of an object or environment. RH represents the relative humidity.

The objective of this multiple linear regression model is to determine the regression coefficients ( $\beta_0, \beta_1, \beta_2, \dots, \beta_5$ ) that provide the most accurate fit to the observed data. This will enable the prediction of  $PM_{2.5}$  concentrations based on the independent variable values. The model assumes a linear relationship between the dependent variable ( $PM_{2.5}$ ) and the independent variables and that the errors ( $\epsilon$ ) follow a normal distribution with a mean of 0 and constant variance.

**4.2.2.1 Linear regression (LR)** Linear Regression (LR) is a fundamental machine learning algorithm utilized for predictive analysis. Linear regression is a statistical technique that uses a linear equation to represent the connection between a dependent variable (also known as the target variable) and one or more independent variables (also known as input features). The objective of linear regression is to determine the optimal line of best fit that minimizes the discrepancy between the projected values and the actual values. The linear regression model can be represented as:

$$y = \beta_0 + \beta_1(FAOD) + \epsilon \quad (8)$$

where,  $y$  is the target variable ( $PM_{2.5}$ ),  $\beta_0$  is the intercept or constant term,  $\beta_1$  is the coefficient of the input features FAOD,  $\epsilon$  is the error term, which represents the random variation in the data.

**4.2.2.2 Lasso regression** Lasso regression (LR) is a variant of linear regression that includes regularization techniques to mitigate the problem of overfitting. In this study, hyperparameter tuning focused on optimizing the regularization parameter (alpha) to control feature selection and prevent overfitting. The model was trained using  $\alpha=0.1$  with  $\max\_iter=1500$ , ensuring convergence. Lasso regression can be employed to forecast  $PM_{2.5}$  concentrations by identifying the most significant input variables and reducing the coefficients of less important variables. This strategy can enhance the model's performance by reducing the influence of noise and extraneous data. Within the realm of air quality modeling, Lasso regression proves to be particularly advantageous in pinpointing the pivotal components that impact  $PM_{2.5}$  concentrations.

$$\text{Minimize} \left( \sum_{i=1}^n y_i - \hat{y}_i^2 + \lambda \sum_{j=1}^p |\beta_j| \right) \quad (9)$$

where,  $y_i$  is the observed value for the  $i$ -th observation.  $\hat{y}_i$  is the predicted value for the  $i^{\text{th}}$  observation, calculated as:

$$\hat{y}_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} \quad (10)$$

where,  $n$  is the number of observations,  $p$  is the number of predictors  $\beta_0$  is the intercept of the model,  $\beta_j$  are the coefficients for each predictor  $x_j$ ,  $\lambda$  (lambda) is a tuning parameter that controls the strength of the penalty. A larger value of  $\lambda$  results in more coefficients being pushed towards zero, effectively performing variable selection.

### 4.3 Model evaluation

The model's performance was assessed using some statistical metrics, including the coefficient of determination ( $R^2$ ), means absolute errors (MAE), root mean squared error (RMSE), and mean square error (MSE).

#### (a) Mean square error (MSE)

A model's predictive accuracy can be evaluated by calculating its mean square error (MSE), where lesser MSE values resemble to higher predictive accuracy.

$$MSE = \frac{1}{n} \sum_{i=1}^n ((y_o)_i - (y_p)_i)^2 \quad (11)$$

#### (b) Mean absolute error (MAE)

An array of actual and predicted  $PM_{2.5}$  concentration values are two continuous variables, and the average magnitude of errors between them is represented by the Mean Absolute Error (MAE), or simply MAE. Equation (12) is used in its computation.

$$MAE = \frac{1}{n} \sum_{i=1}^n |(y_p)_i - (y_o)_i| \quad (12)$$

#### (c) Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) is the second measure used for comparison. The square root of the average of the squared differences between the expected and actual values in a dataset is used to calculate this metric, which is represented by the sign RMSE. Equation (13) is used in its computation.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n ((y_p)_i - (y_o)_i)^2} \quad (13)$$

#### (d) Coefficient of determination ( $R^2$ -score)

The  $R^2$ -score, often known as the coefficient of determination or just  $R^2$ , is the last evaluation criterion that is used. Along a random forest regression model, this metric evaluates the proximity of predicted values to matching actual values. Equation (14) describes the relationship that is used in its calculation.

$$R^2 = \frac{(\text{cov}(y_o, y_p))^2}{\sigma_{y_o}^2 \sigma_{y_p}^2} \quad (14)$$

where  $y_o$  is observed values of  $PM_{2.5}$  and  $y_p$  is predicted value of  $PM_{2.5}$  by random forest regression model.  $\sigma_{y_o}^2, \sigma_{y_p}^2$  are variances of observed values of  $PM_{2.5}$  and predicted value of  $PM_{2.5}$  respectively.

## 5 Results and discussion

### 5.1 Data overview

The descriptive statistics for FAOD,  $PM_{2.5}$  and relevant meteorological parameters utilized in model development are shown in Tables 1. The average  $PM_{2.5}$  concentration was  $46.22 \mu\text{g}/\text{m}^3$  and the FAOD was 335.50 annually. The seasonal averages for  $PM_{2.5}$  and FAOD were as follows: Winter— $75.02 \mu\text{g}/\text{m}^3$ , 517.09; Summer— $43.65 \mu\text{g}/\text{m}^3$ , 384.89; Monsoon— $22.38 \mu\text{g}/\text{m}^3$ , 29.26; post-monsoon— $62.86 \mu\text{g}/\text{m}^3$ , 662.25. Maximum  $PM_{2.5}$  levels were reported in winter, whereas the lowermost were detected in the monsoon period. The seasonal changes in (FAOD) and  $PM_{2.5}$ , particularly the low

concentrations of  $PM_{2.5}$  with lower values of  $fAOD$  during the monsoon period, might be associated to the availability of water vapor in the atmosphere during the time of the monsoon.

## 5.2 Weather parameters and fine particulate matter ( $PM_{2.5}$ )

The Pearson's correlation graph of fine particulate matter ( $PM_{2.5}$ ) and environmental parameters in Maharashtra, India demonstrates the existence of notable statistical relationships. There is a significant negative correlation between  $PM_{2.5}$  levels and Relative Humidity(RH) ( $r = -0.55$ ,  $p < 0.005$ ), Rainfall(RF) ( $r = -0.31$ ,  $p < 0.005$ ), and Wind Speed(WS) ( $r = -0.23$ ,  $p < 0.005$ ) ([38]) based on meteorological parameters in Maharashtra (Fig. 3). Moreover, it has a substantial positive correlation with Wind Direction (WD) ( $r = 0.42$ ,  $p < 0.001$ ), as well as temperature ( $r = 0.19$ ,  $p < 0.005$ ). The correlation between relative humidity and temperature is inversely related to  $PM_{2.5}$  levels, indicating that higher humidity and temperatures are related to lower concentrations of  $PM_{2.5}$  ([38]). The investigation uncovered the inverse correlations between  $PM_{2.5}$  concentrations and both RF and WS. This finding indicates a possible relationship between heightened RF and Wind Velocity (WV) in Maharashtra, leading to a successive decrease in  $PM_{2.5}$  levels. The examination authenticates the substantial effect of atmospheric variables on  $PM_{2.5}$  levels, in line with previous scholarly investigations. [18, 39]. The results indicated above support and add to the information previously available from other studies, emphasizing the significance of taking geographic variation into account when examining.

## 5.3 Models fitting and evaluation

Figure 4 displays an annual scatter plot comparing the measured and estimated  $PM_{2.5}$ , illustrating the Random Forest (RF) model's best fit for Maharashtra, India in 2023. The RF model achieved  $R^2$ , MSE, MAE, and RMSE values of 0.87, 157.91, 6.96  $\mu g/m^3$ , and 12.57  $\mu g/m^3$ , respectively, demonstrating its accurate approximation of the training set values.

The results shown in Table 2 indicate that the feature selection approaches and algorithmic implementations used in this study produced positive effects. The Random Forest (RF) model had superior performance, with an  $R^2$  value of 0.87, indicating that it explained 87% of the variability in the training dataset. The  $R^2$  values for the Multilinear regression models surpassed 0.41, whereas the Linear Regression (LR) with single variable and lasso regression model exhibited a comparatively lower  $R^2$  of 0.18 and 0.41, respectively. In general, incorporating the estimates of  $PM_{2.5}$  pollution values improved the overall accuracy of the models. The coefficient of determination ( $R^2$ ) was found to be the most effective indicator of how well the regression equation fit the data. Therefore, the RF model is considered the most suitable for  $PM_{2.5}$  retrieval modeling.

Figure 4 presents the comparative examination of the monitoring data and the findings obtained from best fitting the model random forest model. Typically, when the levels of  $PM_{2.5}$  are relatively low, the scatter plot shows a stronger correlation with the 1:1 line. Nevertheless, when the concentrations are above 150  $\mu g/m^3$ , the anticipated outcomes from the multilinear regression, lasso regression (LR), and Linear Regression (LR) models have a tendency to underestimate the observed values. This discovery indicates that these three models do not possess the capability to reliably forecast high levels of  $PM_{2.5}$  concentration. Despite the Random Forest (RF) regression model showing improved fitting performance in the high-concentration region, Observable changes in  $R^2$  and Root Mean Square Error (RMSE) are evident when comparing the four ML models. This phenomenon can be explained by the fact that there is a higher relationship between  $PM_{2.5}$  values in monitoring stations that are close to each other. This relationship tends to weaken beyond a distance of 100 km.

We evaluated the accuracy and reliability of our model by validating the predicted  $PM_{2.5}$  concentrations against observed values from monitoring stations. This validation involved comparing the average predicted  $PM_{2.5}$  levels with

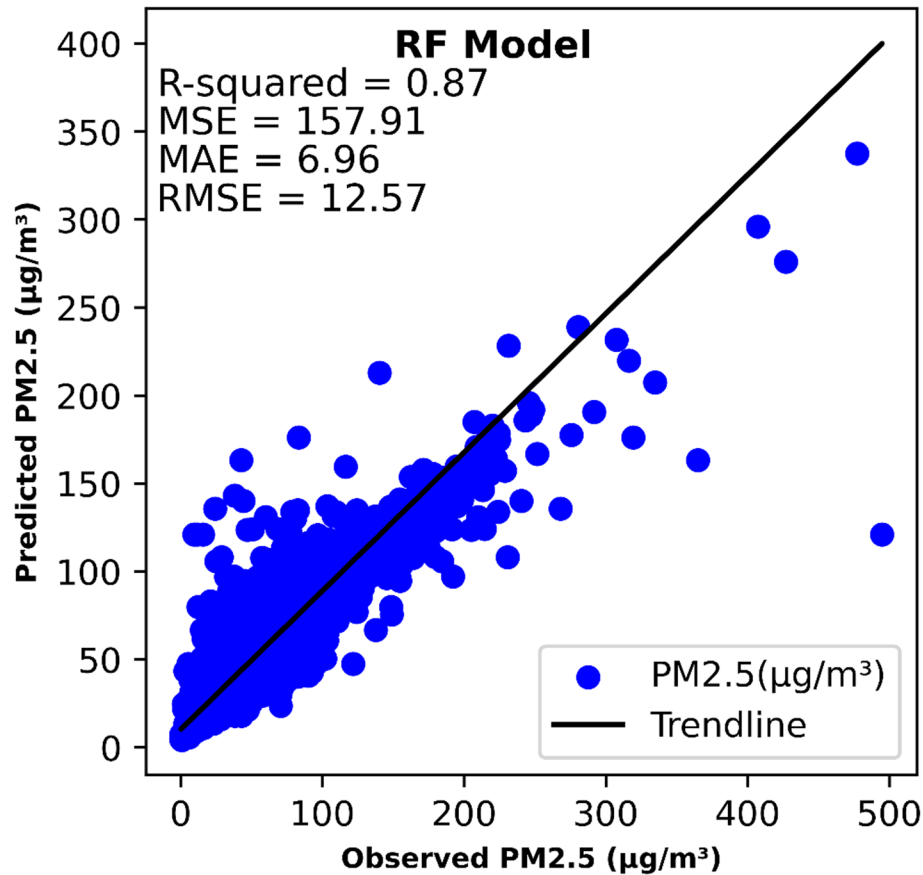
**Table 1** Statistical summary of predictor variables of CPCB AQI station considered for the RF model in year 2023

Parameter	Min	Max	Standard Deviation	Mean	Number of Rows
FAOD	0	2794	468.55	335.50	12563
PM2.5	0.53	427	34.98	46.22	12563
RH	18.14	97.91	16.97	67.80	12563
Temp	17.25	36.52	2.66	26.12	12563
U-Wind	-4.17	9.29	2.05	1.15	12563
V-Wind	-5.93	10.44	1.48	-0.15	12563

**Fig. 3** Correlation analysis between the PM<sub>2.5</sub> concentration and meteorological variables used in the study



**Fig. 4** Annual scatter plot between the observed and predicted pm<sub>2.5</sub>, by using a random forest regressor model





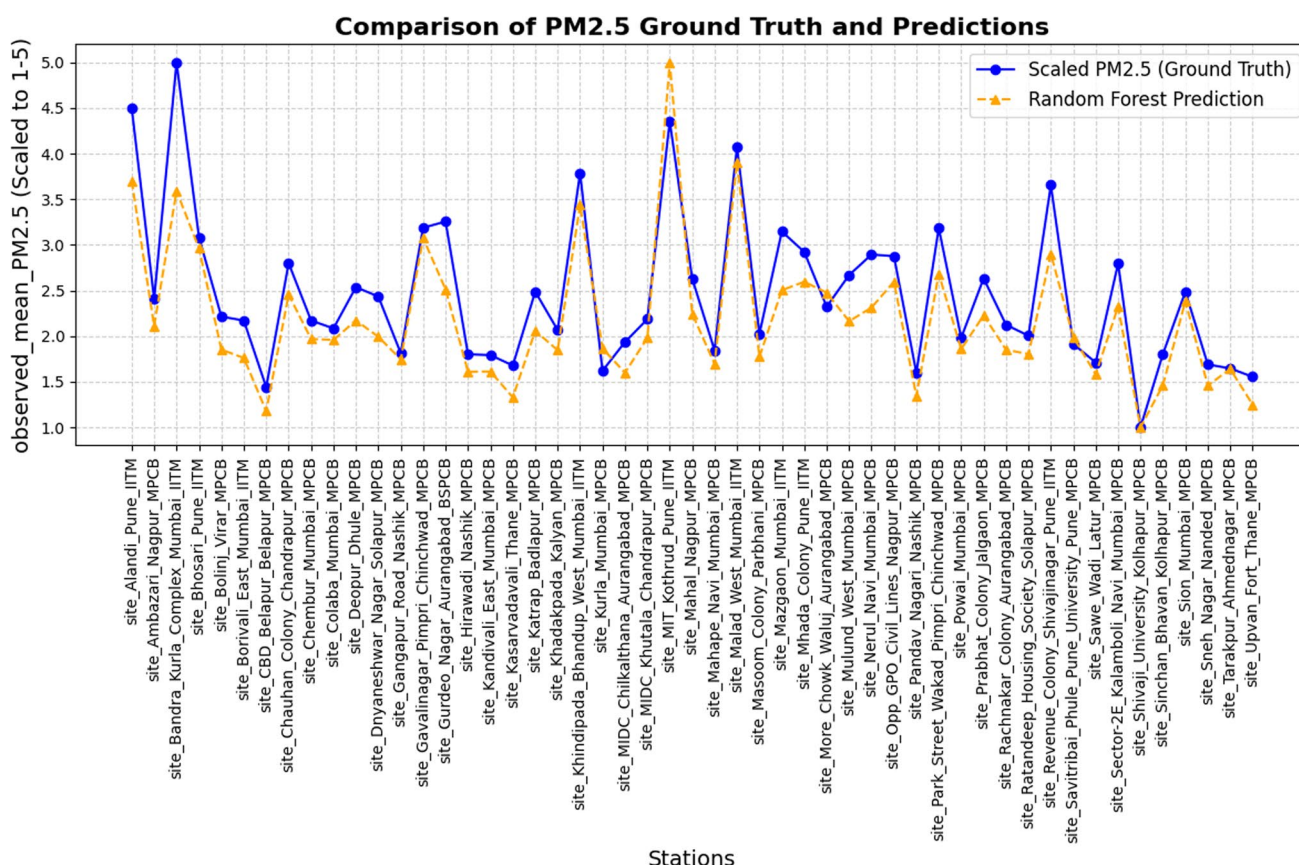
**Table 2** Validation index score of different models used in the study

Model	R2	MSE	MAE	RMSE
RF	0.87	157.91	6.96	12.57
Multilinear regression	0.41	742.25	17.47	27.24
LR	0.18	1008.36	22.41	31.75
Lasso (LR)	0.40	742.25	17.48	27.24

actual measurements across different locations. As shown in Fig. 5, the results indicate a strong correlation between predicted and observed values, with data points from each station illustrating how well the model aligns with real-world measurements. By using scaled values, we enhanced the clarity of comparisons, allowing for a better assessment of the model's performance across various sites. The findings suggest that the Random Forest model provided highly accurate predictions, with many stations showing a close match between estimated and actual values. This validation is crucial not only to confirm the effectiveness of our predictive approach but also to identify areas where the model performs exceptionally well and where further refinement may be needed. Our results demonstrate the effectiveness of machine learning techniques, particularly Random Forest, in accurately predicting  $PM_{2.5}$  concentrations by integrating Fine Aerosol Optical Depth (FAOD) data with meteorological parameters.

#### 5.4 Spatiotemporal variability of predicted $PM_{2.5}$

We assessed the effect of climatic conditions on the satellite data sets of different pollutants that were processed and assessed for four seasons, i.e., winter, summer, and monsoon, and post-monsoon, to analyze the effect of climatic conditions as well on the concentration level over the study area. Google Earth Engine was used to process satellite imagery,

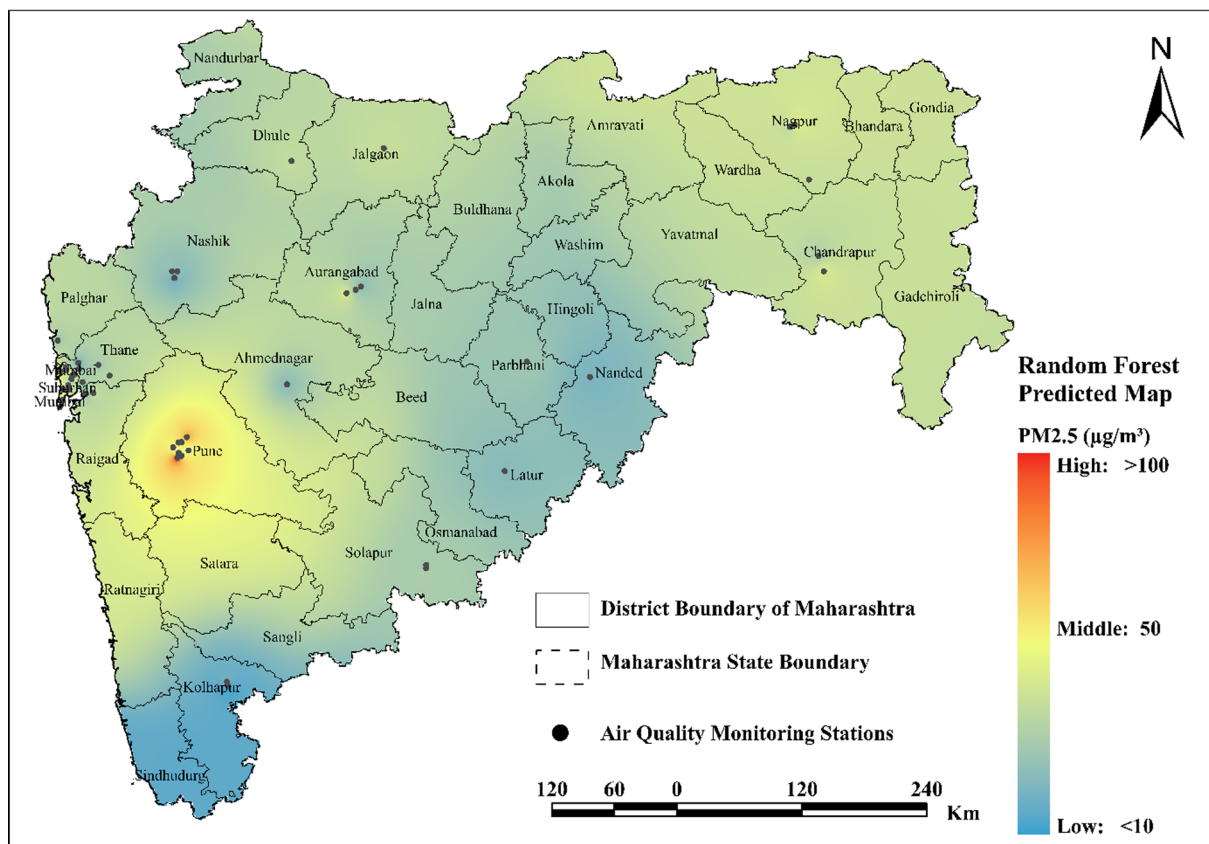


**Fig. 5** Validation of annual mean of estimated  $PM_{2.5}$  concentrations with an annual mean of ground  $PM_{2.5}$  collected from CPCB network monitoring stations across Maharashtra, India in 2023

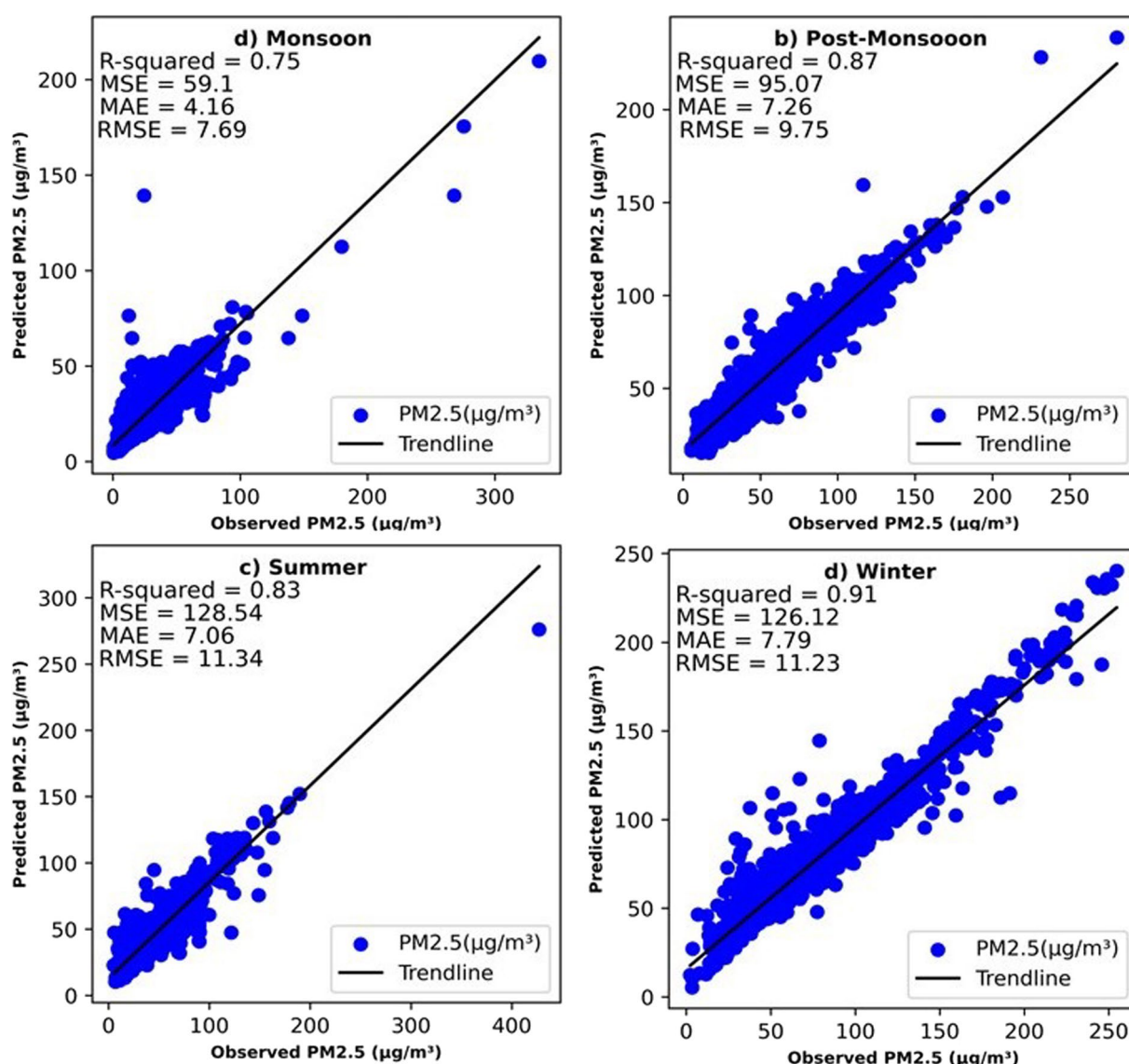
while ArcGIS 10.3 was used to prepare the maps. The RF model was used to construct a yearly average satellite-derived map of  $PM_{2.5}$  for Maharashtra, India shown in Fig. 6. The map has a grid resolution of 1 km. The mean annual  $PM_{2.5}$  concentration was computed as  $46.22 \mu\text{g}/\text{m}^3$ , surpassing the Indian National Ambient Air Quality Standards (NAAQS) limit of  $40 \mu\text{g}/\text{m}^3$ . Spatially, high concentrations of  $PM_{2.5}$  were predominately observed in the western, and moderate in the northwestern of study area. In particular, the maximum yearly mean  $PM_{2.5}$  was mainly concentrated in Mumbai, Navi Mumbai, and Thane. The dense population and the associated industrial, transportation, and residential emissions in the area probably cause high  $PM_{2.5}$  at regional levels.

### 5.5 Seasonal model performance and fluctuations of $PM_{2.5}$

Figure 7 shows scatter plots for each season, showing the comparison between the measured and estimated  $PM_{2.5}$  values utilizing the random forest model. The seasons are classified into four discrete categories: winter (December–February), summer (March–May), monsoon (June–September), and post-monsoon (October–November). In the winter, the model attained the highest levels of accuracy, as indicated by the  $R^2$ , MAE, MSE and RMSE values of 0.91,  $7.76 \mu\text{g}/\text{m}^3$ , 126.12 and  $11.23 \mu\text{g}/\text{m}^3$ , respectively (Table 3). In contrast, the lowest accuracy was found during the monsoon, with  $R^2$ , MSE, RMSE, and MAE values of 0.75, 59.1,  $7.69 \mu\text{g}/\text{m}^3$ , and  $4.16 \mu\text{g}/\text{m}^3$ , respectively. The random forest technique was utilized in 2023 to create maps at a spatial resolution of 1 km that estimate the concentrations of  $PM_{2.5}$  in Maharashtra for every season (Fig. 8). The  $PM_{2.5}$  levels reached their highest point during winter, while they were at their lowest during the monsoon season. The elevated levels of  $PM_{2.5}$  pollution witnessed in the winter can be attributed to atmospheric constancy and lowest temperatures, which generate conditions that are less favorable for the scattering of  $PM_{2.5}$ . The model shows strong concordance between the anticipated output and observed values at CPCB sites, both in the summer and winter (Table 4). Nevertheless, the model forecasts areas in western Maharashtra during the summer and eastern Maharashtra during the post-monsoon season where greater values are expected.



**Fig. 6** Annual mean satellite-based  $pm_{2.5}$  estimated map for Maharashtra, India at a 1 km grid resolution by using random forest regression model

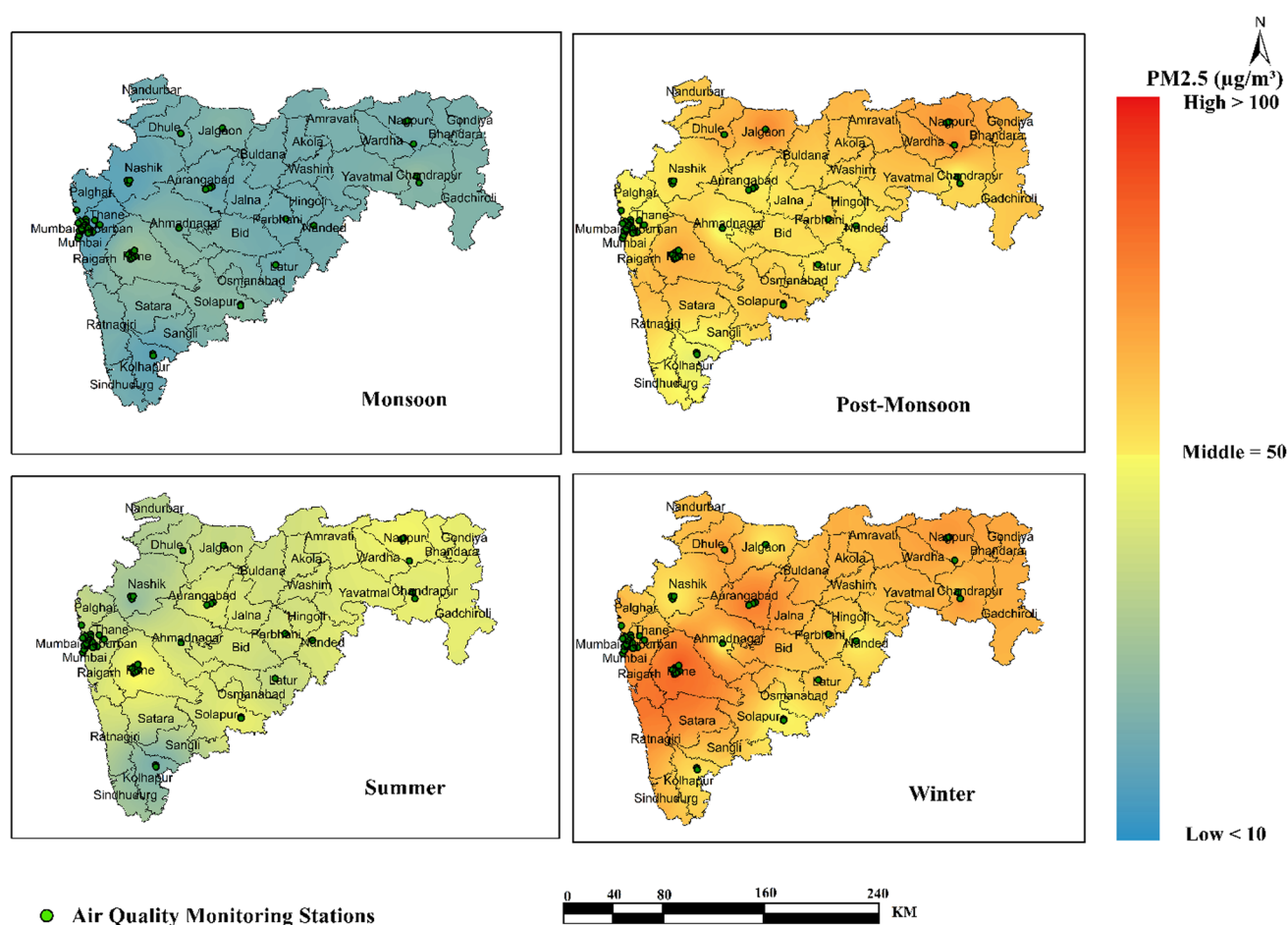


**Fig. 7** Scatter plots for each season comparing the measured and estimated  $pm_{2.5}$  values using the random forest model. **a** monsoon (June–September), **b** post-monsoon (October–November), **c** summer (March–May), and **d** winter (December–February)

During the monsoon season,  $PM_{2.5}$  concentrations are generally lower due to frequent and intense rainfall, which effectively removes pollutants from the atmosphere. High humidity and strong winds further aid in pollutant dispersion, creating a natural cleansing mechanism that significantly reduces particulate matter buildup across Maharashtra.  $PM_{2.5}$  levels rise notably during the post-monsoon season, driven by multiple factors. Festive activities, particularly Diwali, contribute to increased pollution from firecracker emissions and heightened vehicular traffic. Additionally, weaker winds and reduced precipitation during this period lead to atmospheric stagnation, trapping pollutants near the surface and exacerbating air quality issues. The summer season is also characterized by elevated  $PM_{2.5}$  levels,

**Table 3** RF Model performance in each season in year 2023

Metrics used	Monsoon	Post-Monsoon	Summer	Winter
$R^2$	0.75	0.87	0.83	0.91
MSE	59.1	95.07	128.54	126.12
MAE	4.16	7.26	7.06	7.79
RMSE	7.69	9.75	11.34	11.23



**Fig. 8** Mean seasonal spatial distributions of estimated  $PM_{2.5}$  concentrations across Maharashtra, India in 2023

influenced by dust storms, vehicular emissions, and biomass burning, including agricultural waste disposal. High temperatures and dry conditions facilitate the suspension and long-range transport of particulate matter, further amplifying pollution levels. Interestingly, contrary to typical seasonal patterns, winter  $PM_{2.5}$  concentrations in this study are observed to be lower than in the post-monsoon and summer seasons. While winter is generally associated with higher pollution due to atmospheric inversion trapping pollutants, regional factors such as stronger sea breezes in coastal areas, occasional rainfall events, or stricter pollution control measures may have contributed to improved pollutant dispersion and lower  $PM_{2.5}$  levels in 2023.

The seasonal variation in  $PM_{2.5}$  levels is influenced by a complex interaction of meteorological conditions, human activities, and regional factors. Notably, the observed anomalies, such as lower concentrations in winter, underscore the need for continuous and localized air quality monitoring to better understand pollution dynamics. These findings

**Table 4** Statistical summary of estimated  $PM_{2.5}$  by RF model and CPCB observed  $PM_{2.5}$  in each season in the year 2023, in bracket, CPCB observed value statistics is given, i.e., **model predicted  $PM_{2.5}$  statistics (CPCB Observed  $PM_{2.5}$  statistics)**

Descriptor	Min	Max	Standard Deviation	Mean	Number of Rows
$PM_{2.5}$	<b>4.50</b> (0.53)	<b>292.84</b> (427)	<b>29.42</b> (34.98)	<b>46.34</b> (46.22)	<b>10050</b> (12563)
Winter	<b>20.40</b> (2.27)	<b>292.07</b> (377)	<b>21.93</b> (40.18)	<b>44.31</b> (75.02)	<b>10050</b> (2628)
Summer	<b>10.79</b> (5)	<b>271.32</b> (327)	<b>28.18</b> (27.83)	<b>74.97</b> (43.64)	<b>10050</b> (1982)
Monsoon	<b>4.50</b> (0.44)	<b>209.71</b> (290)	<b>11.60</b> (17.22)	<b>22.60</b> (22.38)	<b>10050</b> (5014)
Post-Monsoon	<b>13.57</b> (5.32)	<b>237.84</b> (280)	<b>21.38</b> (27.13)	<b>62.91</b> (62.86)	<b>10050</b> (2939)



highlight the importance of integrating climatic, regulatory, and anthropogenic factors into air quality management strategies to support informed decision-making and improve environmental outcomes.

As shown in Table 3, the performance of the Random Forest (RF) model varied across seasons. It exhibited the highest accuracy in the post-monsoon season ( $R^2 = 0.87$ ;  $RMSE = 9.75 \mu\text{g}/\text{m}^3$ ) and the lowest accuracy during the monsoon season ( $R^2 = 0.75$ ;  $RMSE = 7.69 \mu\text{g}/\text{m}^3$ ). This decline in performance during the monsoon is likely due to the increased water vapor content, which affects the FAOD– $\text{PM}_{2.5}$  relationship. The seasonal variation in model performance reflects the challenges of predicting  $\text{PM}_{2.5}$  concentrations when meteorological parameters such as relative humidity and precipitation play a dominant role. During the monsoon, FAOD values tend to be higher due to enhanced light scattering by water vapor, while  $\text{PM}_{2.5}$  concentrations decrease significantly, making their relationship less predictable. Despite this, the model effectively captures key seasonal trends, even though its accuracy is comparatively lower during the monsoon season.

## 5.6 Discussions

This study examines the seasonal and spatial variability of  $\text{PM}_{2.5}$  concentrations in Maharashtra, India, emphasizing the crucial role of meteorological factors in shaping air quality. The analysis reveals that  $\text{PM}_{2.5}$  levels peak in winter due to atmospheric stability and lower temperatures, whereas they decrease during the monsoon season as rainfall and high humidity facilitate the removal of particulates from the air. The strong correlations between  $\text{PM}_{2.5}$  and meteorological parameters reinforce the importance of incorporating these factors into air quality management strategies.

The Random Forest (RF) model effectively predicted  $\text{PM}_{2.5}$  levels, demonstrating strong performance, particularly in summer. However, its accuracy declined during the monsoon season due to complex weather interactions that affect the FAOD– $\text{PM}_{2.5}$  relationship. The RF model's ability to capture non-linear relationships among variables makes it well-suited for air quality predictions. However, while random forest models excel at handling complex datasets, they can be less interpretable compared to simpler models like multiple linear regression and linear regression. Additionally, we address challenges such as overfitting in more complex models and the assumptions associated with linear regression methods. High  $\text{PM}_{2.5}$  concentrations were observed in western Maharashtra, particularly in densely populated and industrialized areas like Mumbai and Pune, where emissions from industries and vehicles worsen pollution levels. Our findings show  $\text{PM}_{2.5}$  hotspot zones over Mumbai and Pune with higher  $\text{PM}_{2.5}$  concentrations. These findings suggest that targeted mitigation strategies, particularly during winter, could significantly improve air quality. Furthermore, the study underscores the need to account for regional and seasonal variations when developing air quality models. By analyzing seasonal  $\text{PM}_{2.5}$  trends across Maharashtra, this study provides a comprehensive understanding of air pollution dynamics in the region, offering valuable insights for policymakers and environmental management initiatives.

This study demonstrates the effectiveness of machine learning, particularly the Random Forest (RF) model, in accurately predicting  $\text{PM}_{2.5}$  concentrations by integrating FAOD, meteorological parameters, and ground-based measurements. The high spatial resolution (1 km) and seasonal analysis provide valuable insights into pollution patterns, supporting targeted interventions. However, limitations include lower model accuracy during the monsoon ( $R^2 = 0.75$ ) due to high humidity affecting FAOD, the absence of real-time emission data from traffic and industries, and potential spatial biases from unevenly distributed monitoring stations. Additionally, the model's applicability beyond Maharashtra remains untested, and deep learning techniques such as LSTMs and CNNs were not explored. Future improvements should integrate real-time emissions, refine monsoon-season predictions, and explore deep learning approaches for enhanced accuracy and broader applicability.

The study findings of our research fits in well with other research, especially in terms of the dispersion of  $\text{PM}_{2.5}$  over India, saw those larger metropolitan areas, especially those with high industrial and transport emissions, consistently showed the highest  $\text{PM}_{2.5}$  concentrations. Another study in Delhi, India, used machine learning and deep learning to develop a model to predict the  $\text{PM}_{2.5}$  concentrations in Delhi. This research accentuates the importance of urbanization and industrial activity on  $\text{PM}_{2.5}$  atmospheric concentrations ([35]). On the other hand, a study in China found that the COVID-19 shutdown resulted in a big drop in  $\text{PM}_{2.5}$  levels as all industries and transportation came to a halt. The drop in  $\text{PM}_{2.5}$  during the lockdown period shows the impact of human activity on air quality. A study in India looked at the air quality of several cities and found big fluctuations in  $\text{PM}_{2.5}$  levels throughout the year. It was found that  $\text{PM}_{2.5}$  was higher in winter and lower in monsoon season. [4, 36, 40]. Overall, this study provides valuable insights into the interplay between meteorology and air pollution in Maharashtra, offering a basis for more effective air quality management and future research focused on refining predictive models.

## 6 Conclusion

This study presents a comprehensive analysis of PM<sub>2.5</sub> concentration patterns in Maharashtra, India, for the year 2023, using a high-resolution (1 km) approach. By integrating data from air quality monitoring stations, MODIS Fine Aerosol Optical Depth (FAOD), and meteorological parameters within a Random Forest (RF) machine learning model, we provide a more precise and spatially detailed estimation of PM<sub>2.5</sub> levels. A key advancement of this research lies in the application of FAOD as a refined metric for fine-mode aerosol concentration, improving upon traditional methods that primarily rely on AOD. The results indicate persistently high PM<sub>2.5</sub> concentrations in western, northwestern, and central Maharashtra, with an annual average of 46.22 µg/m<sup>3</sup>. Seasonal variations were evident, with a significant decline in PM<sub>2.5</sub> levels during the monsoon and elevated concentrations in winter, largely driven by atmospheric stability and lower temperatures. These findings emphasize the importance of considering regional and seasonal variability when developing air quality management strategies. Looking ahead, a deeper understanding of the spatiotemporal dynamics of PM<sub>2.5</sub> pollution will be essential for refining predictive models and implementing effective mitigation strategies. Future research should focus on integrating additional environmental and anthropogenic factors to enhance model accuracy and better assess the long-term impacts of air pollution. By leveraging advanced modeling techniques, policymakers and environmental agencies can develop targeted interventions to reduce the adverse effects of PM<sub>2.5</sub> on public health and the environment.

## 7 Limitations and future aspects

This study faces limitations such as FAOD-based PM<sub>2.5</sub> estimation uncertainties, especially during monsoons, the absence of real-time emission data, and lower model accuracy in high-humidity conditions. Spatial biases exist due to uneven monitoring station distribution, and the model's generalizability remains untested beyond Maharashtra. Additionally, deep learning methods like LSTMs and CNNs were not explored. Future work should integrate real-time emissions, enhance monsoon predictions, expand to other regions, and incorporate advanced deep learning techniques. Developing a real-time PM<sub>2.5</sub> forecasting system and analyzing long-term trends could further support air quality management and policy planning. In the future, understanding the spatiotemporal patterns of PM<sub>2.5</sub> pollution will help improve model performance, enabling more effective mitigation of its impact on public health and the environment. This can be achieved through targeted interventions and the incorporation of additional influencing factors.

**Acknowledgements** The authors express gratitude to the Land Processes Distributed Active Archive Center (LPDAAC) and Level-1 and Atmosphere Archive & Distribution System (LAADS) for supplying MODIS data. The authors express sincere gratitude to Central Pollution Control Board (CPCB), Government of India for supplying the ground PM<sub>2.5</sub> measurements and meteorological data. We also sincerely thank the anonymous reviewers and editors for their time and valuable suggestions.

**Author contributions** G.M.K contributed to conceptualization, data curation, methodology, visualization, and writing the original draft. S. D supervised the visualization, Investigation, writing, review and editing; S.T developed the conceptualization, data curation, and methodology. M. K contributed to data curation, formal analysis; S.K.R contributed in investigation, writing, review and editing of the manuscript. U. C supervised the data curation, and formal analysis.

**Funding** Open access funding provided by Symbiosis International (Deemed University). This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Data availability** The datasets used in this study are publicly available and can be accessed through the following sources: 1. Fine Aerosol Optical Depth (FAOD): MODIS/Terra (Collection 061) data were obtained from the NASA Land Processes Distributed Active Archive Center (LP DAAC) portal <https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/products/MCD19A2>. 2. Air quality data, including PM<sub>2.5</sub> concentrations, were obtained from the Central Pollution Control Board (CPCB) online portal: <https://airquality.cpcb.gov.in/ccr/#/caaqm-dashboard-all/caaqm-landing/data>. 3. Hourly data for temperature, U-Wind, V-Wind, and relative humidity were obtained from the ERA5-Land reanalysis dataset provided by the European Centre for Medium-Range Weather Forecasts through the Google Earth Engine data catalog: [https://developers.google.com/earth-engine/datasets/catalog/ECMWF\\_ERA5\\_LAND\\_HOURLY](https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_LAND_HOURLY).

## Declarations

**Ethics approval and consent to participate** Not applicable.

**Consent for publication** Not applicable.

**Competing interests** The authors declare no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

## References

1. Chu HJ, Bilal M. PM 2.5 mapping using integrated geographically temporally weighted regression (GTWR) and random sample consensus (RANSAC) models. *Environ Sci Pollut Res*. 2019;26(2):1902–10. <https://doi.org/10.1007/s11356-018-3763-7>.
2. Cohen AJ, et al. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *Lancet*. 2017;389(10082):1907–18. [https://doi.org/10.1016/S0140-6736\(17\)30505-6](https://doi.org/10.1016/S0140-6736(17)30505-6).
3. Xu X, Zhang C, Liang Y. Review of satellite-driven statistical models PM2.5 concentration estimation with comprehensive information. *Atmos= Environ*. 2021. <https://doi.org/10.1016/j.atmosenv.2021.118302>.
4. Zhang G, Shi Y, Xu M. Evaluation of LJ1–01 nighttime light imagery for estimating monthly PM2.5 Concentration: a comparison with NPP-VIIRS nighttime light data. *IEEE J Sel Top Appl Earth Obs Remote Sens*. 2020. <https://doi.org/10.1109/JSTARS.2020.3002671>.
5. Yao F, Si M, Li W, Wu J. A multidimensional comparison between MODIS and VIIRS AOD in estimating ground-level PM2.5 concentrations over a heavily polluted region in China. *Sci Total Environ*. 2018;618:819–28. <https://doi.org/10.1016/j.scitotenv.2017.08.209>.
6. Bai Y, Wu L, Qin K, Zhang Y, Shen Y, Zhou Y. A geographically and temporally weighted regression model for ground-level PM2.5 estimation from satellite-derived 500 m resolution AOD. *Remote Sens*. 2016. <https://doi.org/10.3390/rs8030262>.
7. Lakra AR, Gautam S, Samuel C, Blaga R. College bus commuter exposures to air pollutants in Indian city: the urban-rural transportation exposure study. *Geosystems Geoenviron*. 2025;4(1): 100346. <https://doi.org/10.1016/j.geogeo.2024.100346>.
8. Tan H, Chen Y, Wilson JP, Zhang J, Cao J, Chu T. An eigenvector spatial filtering based spatially varying coefficient model for PM2.5 concentration estimation: a case study in Yangtze River Delta region of China. *Atmos Environ*. 2020;223: 117205. <https://doi.org/10.1016/J.ATMOSENV.2019.117205>.
9. Ma J, et al. Evaluation on the surface PM2.5 concentration over China mainland from NASA's MERRA-2. *Atmos Environ*. 2021. <https://doi.org/10.1016/j.atmosenv.2020.117666>.
10. Zheng T, Bergin MH, Hu S, Miller J, Carlson DE. Estimating ground-level PM2.5 using micro-satellite images by a convolutional neural network and random forest approach. *Atmos Environ*. 2020. <https://doi.org/10.1016/j.atmosenv.2020.117451>.
11. Hu X, et al. Estimating ground-level PM2.5 concentrations in the southeastern U.S. using geographically weighted regression. *Environ Res*. 2013;121(December):1–10. <https://doi.org/10.1016/j.envres.2012.11.003>.
12. Chen G, et al. A machine learning method to estimate PM2.5 concentrations across China with remote sensing, meteorological and land use information. *Sci Total Environ*. 2018;636:52–60. <https://doi.org/10.1016/j.scitotenv.2018.04.251>.
13. Muthukumar P, et al. Predicting PM2.5 atmospheric air pollution using deep learning with meteorological data and ground-based observations and remote-sensing satellite big data. *Air Qual Atmos Heal*. 2022;15(7):1221–34. <https://doi.org/10.1007/s11869-021-01126-3>.
14. Jodhani KH, et al. Synergizing google earth engine and earth observations for potential impact of land use/ land cover on air quality. *Results Eng*. 2024. <https://doi.org/10.1016/j.rineng.2024.102039>.
15. Haque M, Sartelli M, McKimm J, Bakar MA. Health care-associated infections—an overview. *Infect Drug Resist*. 2018;11:2321–33. <https://doi.org/10.2147/IDR.S177247>.
16. Xu Y, et al. Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level PM2.5. *Environ Pollut*. 2018;242:1417–26. <https://doi.org/10.1016/j.envpol.2018.08.029>.
17. Liu Z, et al. Characteristics of PM2.5 mass concentrations and chemical species in urban and background areas of China: Emerging results from the CARE-China network. *Atmos Chem Phys*. 2018;18(12):8849–71. <https://doi.org/10.5194/acp-18-8849-2018>.
18. Kayes I, Shahriar SA, Hasan K, Akhter M, Kabir MM, Salam MA. The relationships between meteorological parameters and air pollutants in an urban environment. *Glob J Environ Sci Manag*. 2019;5(3):265–78. <https://doi.org/10.22034/gjesm.2019.03.01>.
19. Cakir S, Sita M. Evaluating the performance of ANN in predicting the concentrations of ambient air pollutants in Nicosia. *Atmos Pollut Res*. 2020;11(12):2327–34. <https://doi.org/10.1016/j.apr.2020.06.011>.
20. Feng X, Li Q, Zhu Y, Hou J, Jin L, Wang J. Arti ficial neural networks forecasting of PM 2.5 pollution using air mass trajectory based geographic model and wavelet transformation. *Atmos Environ*. 2015. <https://doi.org/10.1016/j.atmosenv.2015.02.030>.
21. Li L. A robust deep learning approach for spatiotemporal estimation of Satellite AOD and PM2.5. *Remote Sens*. 2020;12(2):1–27. <https://doi.org/10.3390/rs12020264>.
22. Reichstein M, et al. Deep learning and process understanding for data-driven Earth system science. *Nature*. 2019;566(7743):195–204. <https://doi.org/10.1038/s41586-019-0912-1>.
23. Nabipour N, et al. Estimating biofuel density via a soft computing approach based on intermolecular interactions. *Renew Energy*. 2020. <https://doi.org/10.1016/j.renene.2020.01.140>.

24. Hanbin Z, Xiong Q, Yin L, Zhang J, Zhu Y, Liang S. CFD-based reduced-order modeling of fluidized-bed biomass fast pyrolysis using artificial neural network. *Renew Energy*. 2020. <https://doi.org/10.1016/j.renene.2020.01.057>.
25. Weizhen H, et al. Using support vector regression to predict PM<sub>10</sub> and PM<sub>2.5</sub>. *IOP Conf Ser Earth Environ Sci*. 2014. <https://doi.org/10.1088/1755-1315/17/1/012268>.
26. De Hoogh K, et al. Modelling daily PM<sub>2.5</sub> concentrations at high spatio-temporal resolution across Switzerland. *Environ Pollut*. 2018. <https://doi.org/10.1016/j.envpol.2017.10.025>.
27. Xin Fang BZ, Shenxin Li L, Xiong. Remote sensing. *Remote Sens*. 2022;1:16.
28. Wang Z, Zhong S, He HD, Peng Z-R. Fine-scale variations in PM<sub>2.5</sub> and black carbon concentrations and corresponding influential factors at an urban road intersection. *Build Environ*. 2018. <https://doi.org/10.1016/j.buildenv.2018.04.042>.
29. Chen S, Yuval, Broday DM. Re-framing the Gaussian dispersion model as a nonlinear regression scheme for retrospective air quality assessment at a high spatial and temporal resolution. *Environ Model Softw*. 2020. <https://doi.org/10.1016/j.envsoft.2019.104620>.
30. Jiang Q, Christakos G. Space-time mapping of ground-level PM<sub>2.5</sub> and NO<sub>2</sub> concentrations in heavily polluted northern China during winter using the Bayesian maximum entropy technique with satellite data. *Air Qual Atmos Heal*. 2018;11(1):23–33. <https://doi.org/10.1007/s11869-017-0514-8>.
31. Masood A, Ahmad K. A model for particulate matter (PM<sub>2.5</sub>) prediction for Delhi based on machine learning approaches. *Procedia Comput Sci*. 2020;167(2019):2101–10. <https://doi.org/10.1016/j.procs.2020.03.258>.
32. Nabipour N, et al. Estimating biofuel density via a soft computing approach based on intermolecular interactions. *Renew Energy*. 2020;152:1086–98. <https://doi.org/10.1016/j.renene.2020.01.140>.
33. Cao JJ, et al. Winter and summer PM<sub>2.5</sub> chemical compositions in fourteen Chinese cities. *J Air Waste Manag Assoc*. 2012;62(10):1214–26. <https://doi.org/10.1080/10962247.2012.701193>.
34. Zhan Y, et al. Spatiotemporal prediction of continuous daily PM<sub>2.5</sub> concentrations across China using a spatially explicit machine learning algorithm. *Atmos Environ*. 2017;155:129–39. <https://doi.org/10.1016/j.atmosenv.2017.02.023>.
35. Ruidas D, Pal SC. Potential hotspot modeling and monitoring of PM<sub>2.5</sub> concentration for sustainable environmental health in Maharashtra, India. *Sustain Water Resour Manag*. 2022;8(4):1–22. <https://doi.org/10.1007/s40899-022-00682-5>.
36. Kumar K, Pande BP. Air pollution prediction with machine learning: a case study of Indian cities. *Int J Environ Sci Technol*. 2023;20(5):5333–48. <https://doi.org/10.1007/s13762-022-04241-5>.
37. Jodhani KH, et al. Unveiling seasonal fluctuations in air quality using google earth engine: a case study for Gujarat, India. *Top Catal*. 2024;67(15–16):961–82. <https://doi.org/10.1007/s11244-024-01957-1>.
38. Bose A, Chowdhury IR. Towards cleaner air in Siliguri: a comprehensive study of PM<sub>2.5</sub> and PM<sub>1.0</sub> through advance computational forecasting models for effective environmental interventions. *Atmos Pollut Res*. 2024. <https://doi.org/10.1016/J.APR.2023.101976>.
39. Islam N, Toha TR, Islam MM, Ahmed T. Spatio-temporal Variation of meteorological influence on PM<sub>2.5</sub> and PM<sub>10</sub> over major urban cities of Bangladesh. *Aerosol Air Qual Res*. 2023;23(1):1–20. <https://doi.org/10.4209/aaqr.220082>.
40. Unik M, Sitanggang IS, Syaafina L, Jaya INS. PM<sub>2.5</sub> estimation using machine learning models and satellite data: a literature review. *Int J Adv Comput Sci Appl*. 2023;14(5):359–70. <https://doi.org/10.14569/IJACSA.2023.0140538>.
41. Sonntag D. Important new values of the physical constants of 1986, vapour pressure formulations based on the ITS-90, and psychrometer formulae. *Z Meteorol*. 1990;40(5):340–4.
42. Nasar-u-Minallah M, Zainab M, Jabbar M. Exploring mitigation strategies for smog crisis in Lahore: a review for environmental health, and policy implications. *Environ Monit Assess*. 2024;196:1269. <https://doi.org/10.1007/s10661-024-13336-0>.
43. Nasar-u-Minallah M, Jabbar M, Zia S, et al. Assessing and anticipating environmental challenges in Lahore, Pakistan: future implications of air pollution on sustainable development and environmental governance. *Environ Monit Assess*. 2024;196:865. <https://doi.org/10.1007/s10661-024-12925-3>.

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