Indian Journal of Pure & Applied Physics
Vol. 62, May 2024, pp. 350-362
DOI: 10.56042/ijpap.v62i5.7587

Prediction of Particulate Matter ($PM_{2.5}$) Concentrations over an Urban Region using Different Satellite

Ajay Kumar[a,b*], Sumit Singh[a,b], Amarendra Singh[c], A K Srivastava[b] & Virendra Pathak[a]

[a]Institute of Engineering and Technology, Lucknow,Uttar Pradesh 226 021, India

[b]Indian Institute of Tropical Meteorology, Ministry of Earth Sciences, New Delhi 110 060, India

[c]Indian Institute of Technology, New Delhi, 110 016 India

Received 27 December 2023; accepted 22 January 2024

The accurate estimation of ground-level particulate matter concentrations ($PM_{2.5}$) is essential for assessing air quality and its impact on human health and the environment. This study focused on estimating $PM_{2.5}$ concentrations from January 2021 to June 2023 in the city of Lucknow, India. Various models, including Bivariate Linear Regression (LR), Multiple Linear Regression (MLR), and Artificial Neural Network (ANN) predicted $PM_{2.5}$ concentrations at the station. Additionally, CALIPSO observations successfully demonstrated the vertical aerosol layer profile in the region. To improve accuracy, we incorporated Aerosol Optical Depth (AOD) data from both MODIS and VIIRS, along with meteorological parameters. The dataset was divided into two periods: 2017-2020 for estimation and January 2021-June 2023 for model training. Our findings revealed a positive correlation between model outputs, observed ground data, and meteorological parameters. For MODIS, LR, MLR, and ANN models had correlation coefficients (R) of 0.41, 0.57, and 0.66. Similarly, for VIIRS, the R-values were 0.33, 0.55, and 0.64, indicating promising agreement between model predictions and actual $PM_{2.5}$ concentrations. These findings contribute to a better understanding of air quality dynamics and can support policymakers in implementing effective measures to mitigate the adverse effects of particulate matter pollution on public health and the environment. Data sets underwent three divisions: 80% for training, and 10% each for validation and testing. ANN displayed strong correlation coefficients (R) across datasets, achieving MODIS R-values of 0.74 and 0.73 for training and overall sets, and VIIRS R- values of 0.74 and 0.72. This study highlights the significant accuracy improvement in $PM_{2.5}$ estimation by integrating meteorological, land use data, and satellite AOD. While LR and MLR methods yielded comparable outcomes, ANN emerged as a superior technique for long-term $PM_{2.5}$ estimation, holding promise for air quality monitoring and guideline adherence in diverse regions.

Keywords: Linear regression; Multiple linear regression; Artificial neural network; AOD; MODIS; ERA5; CALIPSO

1 Introduction

Air pollution is a significant environmental issue that has severe impacts on human health and the environment. Particulate matter (PM) is a type of air pollutant that consists of small particles suspended in the air, with a particle size up to 2.5 µm (*i.e.*, $PM_{2.5}$) and coarse particles, with particle size up to 10 µm (*i.e.*, PM_{10}). These particulate matter (PM) particles exhibit enhanced deleterious effects owing to their diminutive dimensions, facilitating their penetration into the deeper regions of the respiratory system. This characteristic has led to significant pollution issues in Lucknow, one of the major urban cities in India, particularly in terms of particulate pollution. Consequently, the city has been grappling with severe air pollution, which has become a critical environmental concern, particularly in urban regions[1-2]. Based on the World Air Quality Report of 2019, Lucknow was identified as the 11th most polluted city among the top 15 cities globally with poor air quality[3]. Lucknow's air pollution is mostly caused by urbanisation, industrialization, vehicle pollution emissions, and burning activities[1-2,4-5]. Numerous studies have suggested that long-term exposure to fine particles in the environment can have detrimental effects on human health[6]. These particles have been associated with the development of respiratory, cardiovascular, and other disorders[7-10]. Furthermore, the concentration levels of $PM_{2.5}$, a type of fine particle, are influenced by multiple factors[11-12]. These factors include meteorological conditions such as temperature, wind speed, total precipitation, and surface pressure, population density and construction activities. Using regular statistical models to predict $PM_{2.5}$ accurately has become difficult because there are so many different factors that influence it.

*Corresponding author: (E-mail:ajayatul1994@gmail.com)

Numerous studies have extensively examined the use of satellite data for measuring ground particulate matter (PM) concentrations, particularly by leveraging aerosol products like aerosol optical depth (AOD). AOD has been extracted from various sun-synchronous satellites such as Moderate Resolution Imaging Spectroradiometer (MODIS), and Visible Infrared Imaging Radiometer Suite (VIIRS). In this study, employed a semi-empirical model to investigate the connection between Aerosol Optical Depth (AOD) and Particulate Matter (PM), considering the meteorological parameters like influence of various factors, including boundary layer height (BLH), total precipitation (TP), relative humidity (RH), atmospheric temperature (AT), barometric pressure (BP), wind speed (WS) and wind direction (WD). Significant enhancements in the correlation between PM and AOD were achieved through the implementation of intelligent ANN methods, taking into account an additional meteorological parameter for predicting $PM_{2.5}$ - AOD relationships[13]. Moreover, utilized a multi-linear regression (MLR) approach, incorporating geostationary Satellite AOD along with various meteorological parameters, including temperature, planetary boundary layer height, wind speed, total precipitation and barometric pressure, to effectively forecast surface $PM_{2.5}$. Additionally, the researchers utilized a general linear and nonlinear regression model for forecasting purposes, and the results revealed a more robust correlation when accounting for meteorological influences[12]. This research study explores different models for $PM_{2.5}$ estimates at the Lucknow site. The models investigated include LR, MLR, and ANN models., this research integrates data from diverse sources, including polar satellites (MODIS Terra and Aqua), VIIRS (Visible Infrared Imaging Radiometer Suite), and ground-based measurements from CPCB (Central Pollution Control Board). Furthermore, the models take into account essential meteorological parameters, such as TP, PBLH, AT, WS and BP. The models are constructed and trained using ground observations gathered from 2017 to 2020. Subsequently, the validation of these models is performed using ground observations acquired between 2021 and 2023. In addition, the aerosol vertical profile over study area was also studies using CALIPSO Lidar satellite data of extinction coefficient with ground data on $PM_{2.5}$ concentration to understand the linkage between these.

2 Materials and Methodology

2.1 Study Area

Lucknow (26.85˚N, 80.91˚E and 123 m above mean sea level) is a historical city and the capital of the Indian state of Uttar Pradesh (as shown in Fig. 1). With a population of approximately 3.5 million people (as of 2021), it is one of the major urban centres in northern India. Over the years, Lucknow has experienced significant urbanization and industrial growth, leading to various environmental challenges, including pollution. The major sources of air pollution in the city include vehicular emissions, industrial activities, construction dust, and waste burning. As the city continues to grow, the number of vehicles on the road has increased, leading to higher emissions of particulate matter (PM), nitrogen oxides (NOx), sulphur dioxide (SO_2), and volatile organic compounds (VOCs). Additionally, unregulated industrial emissions and construction activities contribute significantly to the worsening air quality. Lucknow experiences distinct seasons, which also influence the pattern of air pollution throughout the year. The city has a subtropical climate with hot summers and relatively mild winters. During the summer months from April to June, the average temperatures can soar up to 45°C, with humidity levels ranging from 30 to 60 percent. These conditions create favourable circumstances for the formation of ground-level harmful air pollutant. During the monsoon season from July to September,

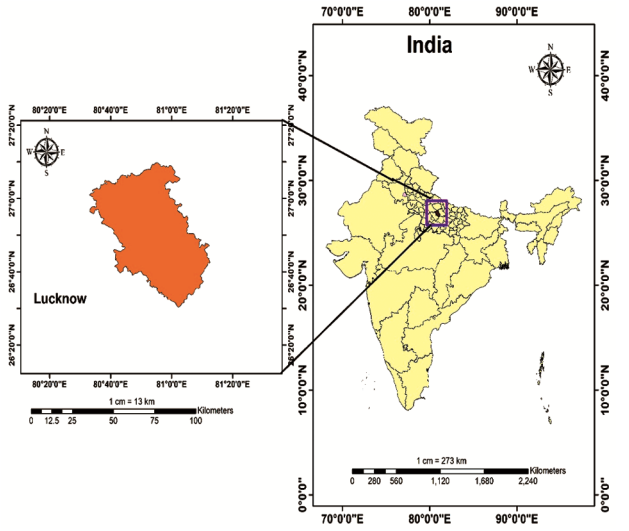

Fig. 1 — Study area map. (*The map is only intended to be used as a visual aid and do not indicate any view on the legal position of any country or territory or the delimitation frontiers or boundaries.)

Lucknow receives significant rainfall, which helps in temporarily improving air quality by washing away some pollutants from the atmosphere. However, due to high humidity levels, there may be instances of localized air pollution during this period.

2.2 Data Description

2.2.1 Measurement of surface$PM_{2.5}$

The Central Pollution Control Board (CPCB) is a statutory organization in India that was established in 1974 under the Water (Prevention and Control of Pollution) Act. It operates under the Ministry of Environment, Forest, and Climate Change. In this study the measurements of real-time data on near-surface $PM_{2.5}$ concentrations were carried out by Central Pollution Control Board (CPCB) using (Beta Attenuation Method). A total of six monitoring stations have been considered in Lucknow region to assess daily $PM_{2.5}$ concentrations from January 2017 to June 2023 from the CPCB.

2.2.2 Satellite Measurements

Satellite-derived datasets were employed for $PM_{2.5}$ forecasting, encompassing Aerosol Optical Depth (AOD) measurements obtained from MODIS (Terra and Aqua), as well as VIIRS Satellite. Additionally, meteorological parameters were incorporated in the prediction model using ERA5 measurements, discussed in the later section.

2.2.2.1 MODIS (Aqua and Terra)

MODIS (Moderate Resolution Imaging Spectroradiometer) is an advanced satellite instrument utilized for Earth observation and environmental monitoring. This instrument, known as MODIS, is installed on two NASA satellites, Aqua and Terra, both of which were launched in 1999 and 2002 respectively. These satellites have significantly advanced our comprehension of the Earth's land, ocean, and atmospheric processes.

MODIS, equipped with 36 spectral bands spanning from visible to thermal infrared, offers data at varying spatial resolutions, including 250m, 500m, and 1km. These various resolutions enable researchers to meticulously and precisely observe alterations in the Earth's surface, atmosphere, and oceans. Consequently, MODIS data is extensively applied across multiple scientific disciplines, encompassing climate science, ecology, hydrology, and atmospheric science. Furthermore, MODIS data plays a crucial role in atmospheric research. It supplies essential data on aerosols, clouds, and atmospheric gases, contributing to our understanding of atmospheric dynamics and air quality. For instance, it has been employed in investigating the dispersion of aerosols across continents, monitoring fluctuations in cloud cover, and measuring concentrations of atmospheric pollutants, such as ozone and carbon monoxide. In this particular study, Aerosol Optical Depth (AOD) was extracted from MODIS measurements using the "Deep Blue" (DB) algorithm, providing data at a spatial resolution of 10 km × 10 km at nadir. This data is widely used for estimating $PM_{2.5}$ concentrations.

2.2.2.2 VIIRS

The VIIRS instrument was launched on October 28, 2011, and has a sophisticated imaging radiometer installed aboard several polar-orbiting satellites. Developed and launched by the National Aeronautics and Space Administration (NASA) and the National Oceanic and Atmospheric Administration (NOAA) of the United States, VIIRS is a critical component of the Joint Polar Satellite System (JPSS). It represents a remarkable advancement in satellite remote sensing technology, providing a wide array of environmental data essential for various scientific, environmental, and governmental applications. It is a multi-spectral radiometer that operates in the visible and infrared spectrum. It provides valuable data for both day and night observations, enabling a comprehensive understanding of Earth's surface, atmosphere, and clouds. The instrument is capable of capturing imagery and radiometric data with high spatial resolution, allowing scientists to monitor various geophysical parameters and climate indicators.

2.2.2.3 CALIPSO

The Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP), a dual-wavelength polarization-sensitive lidar system, is a critical instrument aboard the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations (CALIPSO) satellite. CALIOP's operational capability relies on laser pulses generated by a Neodymium-doped Yttrium Aluminium Garnet (Nd:YAG) laser source, emitting at wavelengths of 532 nm and 1064 nm. This advanced instrument is specifically engineered to capture and analyze the range-resolved backscattered radiation originating from the Earth's atmosphere. This capability enables the acquisition of highly detailed vertical profiles of aerosols and cloud formations on a global scale. CALIOP operates within

a recurring measurement cycle of 16 days, facilitating systematic observations of the Earth's atmosphere. The present study employs data from Level 3 version 4 of CALIOP's aerosol profile product, spanning the period from January 2017 to June 2023. This dataset provides a spatial resolution of 2° × 5°, a vertical resolution of 60 meters, and extends its detection height range from sea level to 12.1 kilometres, comprising a total of 208 discrete layers for analysis. In this research, our attention is directed toward two key parameters: Aerosol Optical Depth (AOD) and Extinction Coefficient at 532 nm. These parameters serve as focal points for obtaining valuable insights into the characteristics and behavior of aerosols and clouds during the specified time frame.

2.2.3 Meteorological data

Meteorological data, including Atmospheric temperature, Wind speed, Barometric pressure, planetary boundary layer height and Total precipitation were obtained from ERA5. Atmospheric temperature is the measurement of the average kinetic energy of air molecules, governing various atmospheric processes. Wind speed, influenced by pressure and temperature gradients, represents the velocity of air movement. Barometric pressure, on the other hand, signifies the weight of the air column above a location and impacts the development and movement of weather systems. The planetary boundary layer height indicates the vertical extent of the lower atmosphere, where pollutants mix and disperse. Lastly, total precipitation, including rain, snow, and hail, provides critical insights into the hydrological cycle's impact on ecosystems and water resources. Together, these parameters offer valuable insights into the complex interplay of atmospheric processes, proving vital for understanding climate phenomena and weather-related events. Table 1 displays the correlation matrix, elucidating the interrelationships among key meteorological parameters, AOD, and ground-measured $PM_{2.5}$ concentrations. The table showcases correlation coefficients, ranging from -1 to 1, providing insights into the strength and direction of associations between study variables. Notably, AOD exhibits a substantial positive correlation (0.39) with $PM_{2.5}$, indicating a noteworthy connection. Additionally, wind speed (WS) and air temperature (AT) reveal weak negative correlations, while surface pressure (SP) displays weak positive correlations with both AOD and $PM_{2.5}$.

2.2.3.3 ERA5

ERA5, the latest iteration of ECMWF atmospheric reanalysis, represents the fifth generation of this global climate analysis. Spanning from January 1940 up to the current date, ERA5 is meticulously crafted by the Copernicus Climate Change Service (C3S) in collaboration with ECMWF. The ERA5-Interim reanalysis is made readily available on a daily temporal scale, presenting data with a spatial resolution of 0.1°x0.1° per hour. This reanalysis process involves the assimilation of radiosonde observations and remote sensing data in a consistent manner[14]. A multitude of contemporary studies have substantiated that ERA-Interim offers a dependable evaluation of the dominant macroscopic atmospheric conditions and adeptly captures meteorological fluctuations.

2.3 Model Estimationof PM2.5

This section presents the sequential approach (as shown in Fig. 2) proposed in this study for evaluating $PM_{2.5}$ concentrations, utilizing data from MODIS, VIIRS (AOD), CALIPSO (Extinction Coefficient), and ERA5 (meteorological parameter). The first step involves data collection, where we acquire datasets from various web-based platforms, such as NASA's 'Giovanni' and the Indian Space Research Organization's (ISRO) 'Meteorological and Oceanographic Satellite Data Archival Centre

Table 1 — Correlation matrix with meteorological parameters

	MODIS	VIIRS	WS	AT	SP	BLH	TP	$PM_{2.5}$
MODIS	1.00							
VIIRS	0.84	1.00						
WS	-0.21	-0.22	1.00					
AT	-0.26	-0.13	0.20	1.00				
SP	0.20	0.07	-0.25	-0.87	1.00			
BLH	-0.25	-0.17	0.57	0.69	-0.54	1.00		
TP	0.05	0.15	0.06	0.12	-0.32	-0.11	1.00	
$PM_{2.5}$	0.39	0.29	-0.26	-0.51	0.57	-0.27	-0.32	1.00

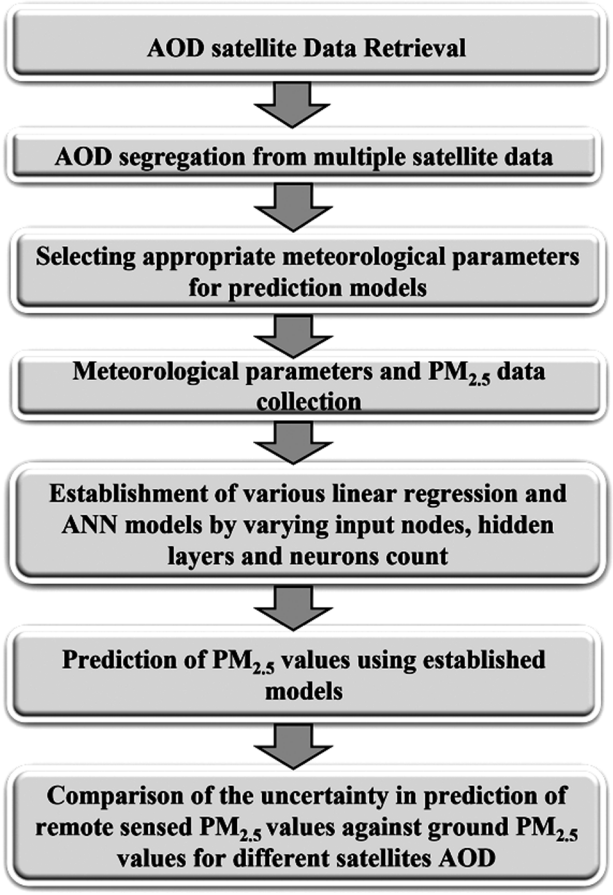

Fig. 2 — Methodology of current study.

(MOSDAC)'. NASA's Giovanni platform allows users to access and visualize Earth science data from different satellite missions and numerical models. Likewise, MOSDAC serves as a central repository for ISRO's satellite mission data products, including satellite imagery, weather data, and oceanography data. Additionally, we extract in-situ $PM_{2.5}$ data from the continuous ambient air quality monitoring stations (CAAQMS) network of CPCB. Moreover, meteorological parameters are obtained from the ERA-5 reanalysis platform, and Extinction Coefficient parameters are obtained using CALIPSO satellite. In the second phase (data pre-processing), the data sets undergo a thorough quality check to identify and eliminate any missing or erroneous observations. Moving on to the third phase, we conduct temporal and spatial analyses to examine the variation in AOD and $PM_{2.5}$ concentrations across the designated study area. Proceeding to the fourth phase, we employ three distinct techniques, namely linear regression (LR), multiple regression (MR), and artificial neural network (ANN), to construct the model, incorporating both independent and dependent variables. Through this model, we can effectively assess $PM_{2.5}$ concentrations, providing valuable insights for air quality management and the formulation of public health policies. Finally, the model's validity is assessed using various statistical measures such as bias, root mean square error, and coefficient of determination, among others.

2.3.1 Linear regressions (LR)

Linear regression models are used to establish the connection between input parameters and output parameters. Their primary purpose is to analyse how input parameters impact the response of output parameters. Numerous studies have employed linear and multiple linear regression models to estimate $PM_{2.5}$ levels. Nevertheless, basic linear regression models fail to adequately account for measurement uncertainties and can result in significant inaccuracies[15]. The study identifies errors in the estimation of ground-level $PM_{2.5}$ concentrations, specifically when relying solely on AOD as the predictor (as shown in Equation 1 and Fig. 3(a)). As a result, simple linear regression is not utilized in this research to estimate $PM_{2.5}$ levels.

$$PM_{2.5}\,(\text{estimated}) = \beta_0 AOD_0 + \beta_1 \quad \ldots (1)$$

Where AOD represents aerosol optical depth and β represents the regression coefficients obtained through the least square's method during the regression of ground-based PM concentration and AOD measurements.

Using mathematical models were formulated to estimate the concentrations of $PM_{2.5}$. In which Eq.1(a) for MODIS and 1(b) for VIIRS.

$$Y = 87 + 56.54AOD \quad (\text{MODIS}) \quad \ldots 1(a)$$

$$Y = 79.54 + 75.55AOD \ (\text{VIIRS}) \quad \ldots 1(b)$$

2.3.2 Multiple linear regressions (MLR)

The utilization of multiple linear regression (MLR) models can effectively address these concerns and significantly enhance the reliability of the model. In this particular investigation, the input parameters consist of satellite AOD and meteorological variables, while the objective is to estimate the output variable, $PM_{2.5}$. However, these calculations are impacted by the vertical distribution of aerosols in conjunction with various meteorological factors. In Equation 2 demonstrates the integration of MODIS and VIIRS AOD with meteorological variables into multi-regression models(as shown in Fig. 3(b)),

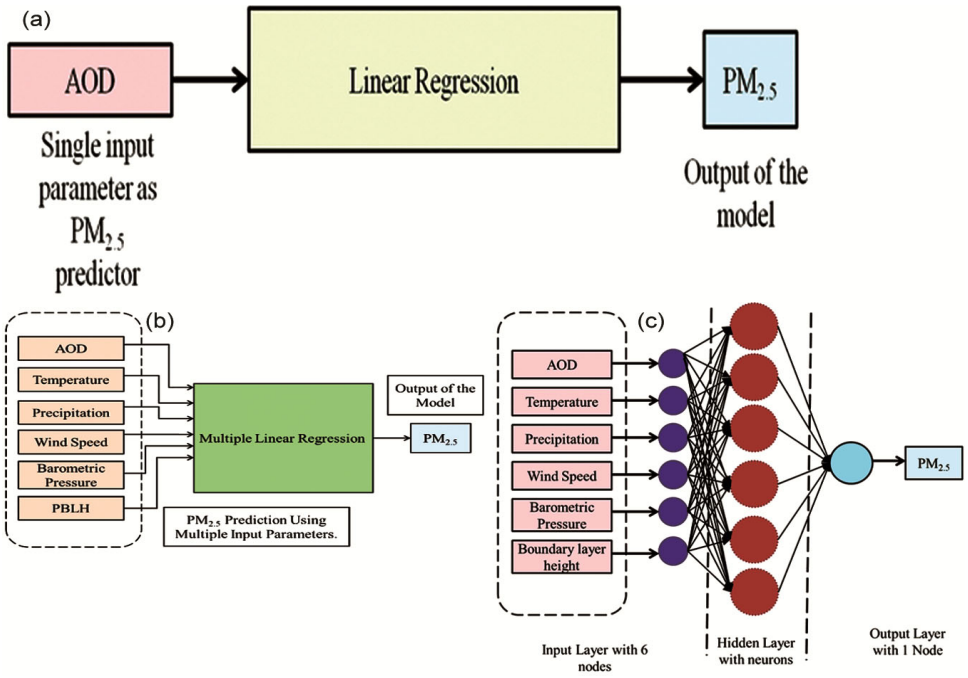

Fig. 3 — Architecture of different model used to estimate $PM_{2.5}$.

aiming to enhance the accuracy of estimating $PM_{2.5}$ concentrations at ground level[16]

$$PM_{2.5}(estimated) = \beta_0 + \beta_1 AOD_d + \beta_2(\text{AT}) + \beta_3(\text{WS}) + \beta_4(\text{BP}) + \beta_5(\text{PBL}) + \beta_6(BP) \quad \ldots (2)$$

Where, β_0 *and* β_{1-6} are the regression co-efficient associated with the daily measurements of AOD (MODIS, VIIRS). Furthermore, AT represents Atmospheric Temperature, WS stands for Wind Speed, BP indicates Barometric Pressure, PBL refers to Planetary Boundary Layer Height, and TP represents Total Precipitation.

Using mathematical models were formulated to estimate the concentrations of $PM_{2.5}$. In which Eq.2(a) for MODIS and 2(b) for VIIRS.

$PM_{2.5}$ = -786.82 + 41.40AOD -17.66WS-4.20AT +1.34SP +0.05BLH -116.80TP MODIS ... 2(a)

$PM_{2.5}$ = -768.87+ 62.38AOD -17.93WS-3.83AT +1.311SP +0.03BLH -118.62TP VIIRS ... 2(b)

2.3.3 Artificial Neural Network

Artificial neural networks (ANN) combined with regression models are designed to evaluate the comparative analysis of estimated $PM_{2.5}$ levels. The utilization of ANN is widely acknowledged for its advantages over conventional regression methods, owing to its efficient computations, superior generalization capabilities, and reduced reliance on prior knowledge[17]. The provided diagram illustrates an Artificial Neural Network (ANN) consisting of three layers: an input layer, a hidden layer, and an output layer, with two neurons in each layer (2-2-1 configuration). In this network, the output value of each neuron is determined by calculating the weighted sum of inputs from the neurons in the previous layer. The weights (w) used for these calculations are initially assigned random values drawn from either a uniform or normal distribution. The diagram specifically depicts this process for neuron N3, as represented by Eq.3.

$$N_3 = \text{f}(x_1 w_{13} + x_2 w_{23}) \quad \ldots (3)$$

Where f(.) is the activation function. In the given context, where N represents neurons and x represents input vectors, the output is calculated at N5. Subsequently, the calculated output is compared to the expected value, and the resulting error is computed. Since the initial weights were randomly assigned, the current output might not align with the expected value. Consequently, this error is then propagated in reverse to determine the adjustments needed for each weight. Eq. 3(a) allows us to calculate the error (σ). The First term $Out_{N5}(1\text{-}Out_{N5})$ represents the derivative of the sigmoid activation function in the output layer. The second term $(Target_{N5} - Out_{N5})$ calculates the disparity between the current and desired performance.

$$\sigma_{N5} = Out_{N5}(1 - Out_{N5})\,(Target_{N5} - Out_{N5}) \quad \ldots (3a)$$

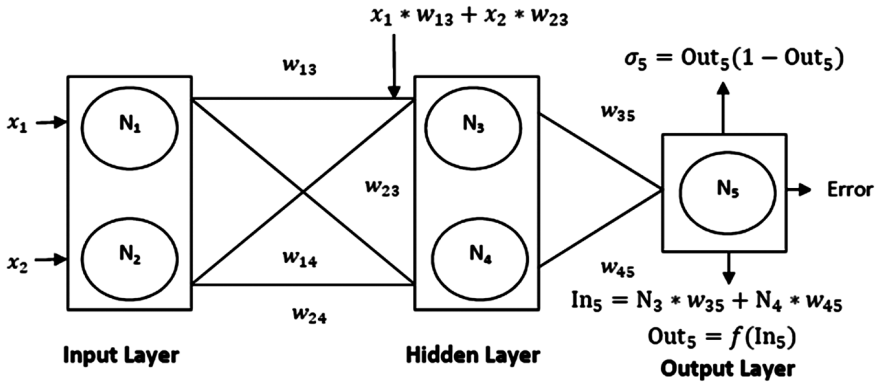

Fig. 4 — Schematics of ANN with error back propagation algorithm. "In" represents input and "Out" represents the output.

As a result, the calculation for weight change is determined using the Eq. 3(b).

$$\Delta w_{35} = \sigma_{N5} * Out_{N3} * \text{Learning rate } (\pi) + w_{35} \quad \ldots (3b)$$

By employing this approach, the neural network's weights and biases are adjusted to enhance its performance, while simultaneously minimizing the cost function. The mean squared error (MSE) serves as the cost function and Sigmoid function serves as activation function. The ANN network, as illustrated in (Fig. 4 and Fig. 3(c)), is designed based on the error back propagation algorithm.

2.4 Evaluation of estimation performance

The performances of Linear Regression(LR), Multiple Linear Regression(MLR) and Artificial Neural Network (ANN) were evaluated using three distinct statistical measures throughout the training and validation phases. These measures include the mean bias (MB), normalized mean bias (NMB), coefficient of correlation (R), mean square error (MSE), mean absolute error (MAE), Root Mean Square Error (RMSE), and index of agreement (IOA). The mathematical expressions for these measures are presented in Table 2.

3 Results and Discussion

3.1 Time series analysis of AOD

The time series plots of AOD values for each instrument reveal distinct patterns and trends, offering an in-depth understanding of the aerosol dynamics observed by each sensor as show in (Fig. 5). Specifically, both Calipso and MODIS demonstrate relatively higher AOD values compared to VIIRS, reflecting the varying capabilities and resolutions of these instruments in capturing aerosol concentrations. Notably, specific months exhibit prominent peaks and troughs in AOD levels for each instrument. For Calipso, the highest AOD value (1.16) was recorded in November 2017 while MODIS exhibited its peak (1.03) in December 2019. On the other hand, VIIRS demonstrated its highest AOD value (0.75) in May 2017. Conversely, the lowest AOD values for Calipso, MODIS, and VIIRS were observed in August 2017 (0.17), September 2019 (0.58), and July 2017 (0.30), respectively.

3.2 Monthly variation of observed and estimation $PM_{2.5}$

Monthly variation of observed and estimated $PM_{2.5}$ concentrations (as shown in Fig. 6). $PM_{2.5}$ concentration patterns exhibited significant variability around the mean monthly estimated value with a standard deviation of 83.28 ± 32.12 μg/m^3. Moreover, the monthly mean ground-based $PM_{2.5}$ concentration was consistently measured at 83.28 μg/m^3. Conducting a detailed regression analysis, we observed the highest averaged estimated $PM_{2.5}$ concentrations of 109.99, 168.92, and 111.11 μg/m^3 in April 2020, January 2021, and January 2023, respectively. These findings underscore the temporal fluctuations in $PM_{2.5}$ levels over the study duration. To enhance our comprehension, we conducted a more in-depth exploration of the modelling of $PM_{2.5}$ concentrations, employing Modis AOD and meteorological parameters across three distinct models: BV, MLR, and ANN. This modelling approach yielded intriguing results, with the maximum estimated $PM_{2.5}$ values recorded as 153.70 μg/m^3 in July 2021, 184.41 μg/m^3 in January 2021, and 187.80 μg/m3 in December 2022. Moreover, utilizing VIIRS data, we identified the highest $PM_{2.5}$ concentrations as 155.07 μg/m^3 in November 2022, 185.83 μg/m^3 in January 2021, and 194.84 μg/m^3 in November 2021. By integrating ground-based and model-derived $PM_{2.5}$ concentrations, we gained comprehensive insights into

Table 2 — Evaluation Metrics for estimating Performance

Fraction of predictions within a factor of 2	$FAC2 = 0.5 \leq \frac{P_I}{O_i} \leq 2.0$
Mean Bias	$MB = \frac{1}{n}\sum_{i=1}^{n} P_i - O_i$
Mean Gross Error	$MGE = \frac{1}{n}\sum_{i=1}^{n} \lvert P_i - O_i \rvert$
Normalized Mean Bias	$NMB = \frac{\sum_{i=1}^{n} P_i - O_i}{\sum_{i=1}^{n} O_i}$
Normalized Mean Gross Error	$NMGE = \frac{\sum_{i=1}^{n} \lvert P_i - O_i \rvert}{\sum_{i=1}^{n} O_i}$
Root Mean Squared Error	$RMSE = \sqrt{\left(\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}\right)}$
Correlation Coefficient	$r = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(\frac{P_i - \bar{P}}{\sigma_P}\right)\left(\frac{O_i - \bar{O}}{\sigma_O}\right)$
Coefficient of Efficiency	$COE = 1.0 - \left(\frac{\sum_{i=1}^{n} \lvert P_i - O_i \rvert}{\sum_{i=1}^{n} \lvert O_i - \bar{O} \rvert}\right)$
Index of agreement	$IOA = \begin{cases} 1.0 - \frac{\sum_{i=1}^{n} \lvert P_i - O_i \rvert}{2\sum_{i=1}^{n} \lvert O_i - \bar{O} \rvert}, & when \sum_{i=1}^{n} \lvert P_i - O_i \rvert \leq 2\sum_{i=1}^{n} \lvert O_i - \bar{O} \rvert \\ \frac{2\sum_{i=1}^{n} \lvert O_i - \bar{O} \rvert}{\sum_{i=1}^{n} \lvert P_i - O_i \rvert} - 1.0, & when \sum_{i=1}^{n} \lvert P_i - O_i \rvert > 2\sum_{i=1}^{n} \lvert O_i - \bar{O} \rvert \end{cases}$

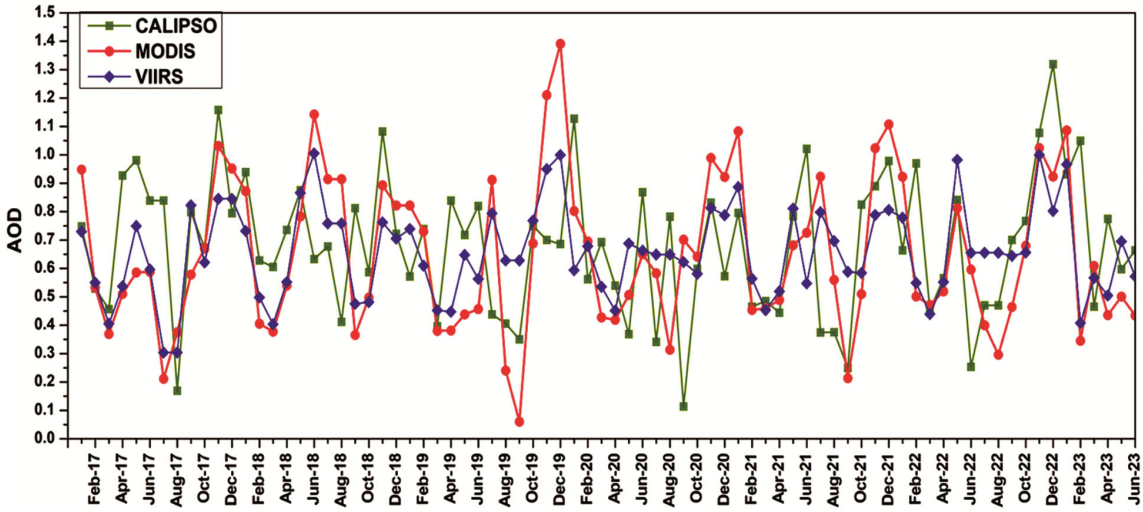

Fig. 5 — Time series variation of mean monthly AOD during January 2017 to June 2023 (The brown,red and blueline represents monthly mean of CALIPSO, MODIS AND VIIRS)

the spatiotemporal dynamics of $PM_{2.5}$ level in the study region. These findings provide invaluable information to understand air quality patterns and formulate effective pollution control strategies.

3.3 Model estimation and Accuracy analysis

3.3.1 Bivariate linear regression model

In this study, mathematical models were formulated to estimate the concentrations of $PM_{2.5}$, a type of particulate matter, based on relevant variables. Through rigorous evaluation, the model exhibiting the coefficient of corelation (R = 0.16) for MODIS and For VIIRS (R=0.11) was selected as the optimal solution (refer to Eq.1 (a) & (b)).

3.3.2 Multi -variate linear regression model-

Multiple linear regression (MLR) was employed to predict the $PM_{2.5}$ concentrations in the investigated study. Subsequently, the optimal model exhibiting the coefficient of corelation

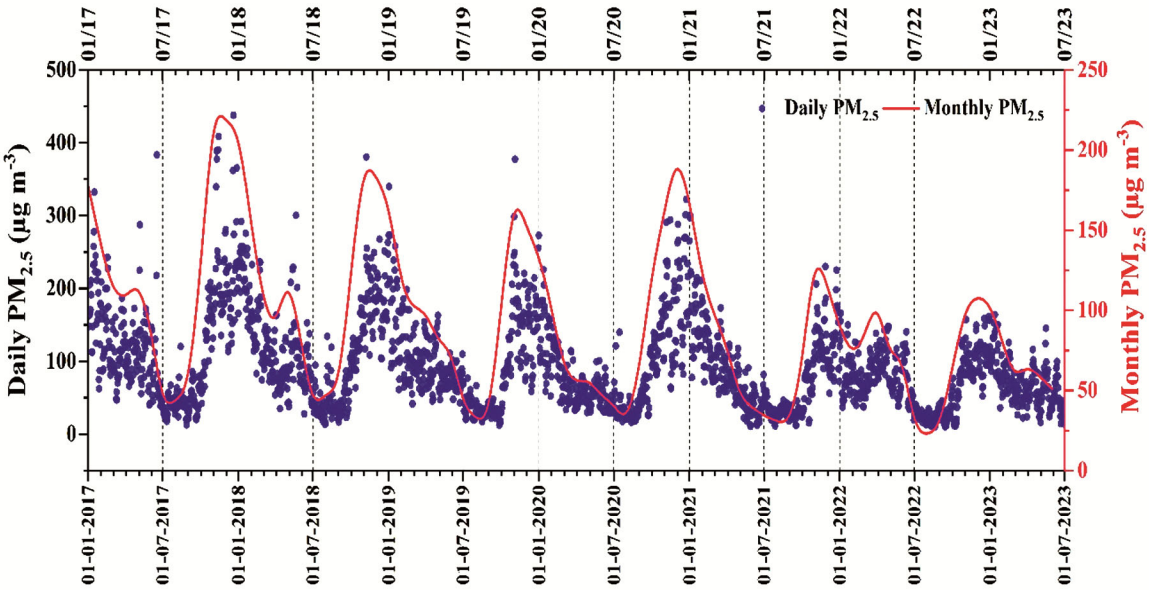

Fig. 6 — Daily and Monthly temporal variation of observed $PM_{2.5}$. during Jan 2017 to June 2023.

(R = 0.41) for MODIS and for VIIRS (R=0.40) (refer to Eq.2 (a) &2 (b).)

In this model, MODIS AOD and VIIRS AOD, along with temperature, wind speed, precipitation, pressure, and PBL, were utilized as independent parameters**.**

3.3.3 ANN Model

In this Model we provided with a training data set spanning from 2017 to 2020 and tasked with making predictions for the period between January 2021 and June 2023 as shown in (Fig. 7). In order to achieve optimal performance using ANN, we conducted experiments with different pairs of transfer functions for both the hidden and output layers. The neuron numbers in the hidden layers were systematically altered during these tests after evaluating various combinations, we selected the most appropriate pairings that demonstrated correlation coefficients (R) for regression analysis of training, validation, test, and all data as 0.75, 0.74, 0.67, and 0.73, respectively, for the MODIS model. Similarly, for the VIIRS model, the corresponding R- values were 0.74, 0.67, 0.66, and 0.72, with the highest R being 0.74. The R-value demonstrated a strong and positive linear correlation between input and output vectors. Hence, within this investigation, the proposed ANN model demonstrates its capability to effectively manage such unpredictable variations, rendering it highly regarded and deemed satisfactory. Furthermore, in Figure, the pattern of network training is depicted, revealing a consistent decline in the mean squared error (MSE) between the target and estimated output as the training progresses. The plotted data indicates a balanced state, free from both under-fitting and over-fitting. To ensure optimal results and prevent these issues, the convergence criterion is implemented, dictating when the network training should cease. In our proposed approach, the training set is divided into three subsets: training, validation, and test sets. The convergence is determined by evaluating the validation accuracy after each epoch. If the validation accuracy does not improve for three consecutive iterations, we adjust the learning rate and continue training. If, after six consecutive checks, the accuracy does not show improvement, and the learning rate becomes exceedingly small, we halt the training, concluding that convergence has been achieved. further, as depicted in the figure, convergence is attained at the 9th epoch for the MODIS model and at the 10th epoch for the VIIRS model. Afterward, the ensuing epochs exhibit only marginal performance changes after the 9th and 10th epochs, respectively.

3.4 Daily estimation of $PM_{2.5}$

The results obtained from the statistical analysis of the $PM_{2.5}$ prediction models using MODIS and VIIRS AOD data, along with meteorological parameters, reveal valuable insights into the accuracy and effectiveness of each model. The metrics used to

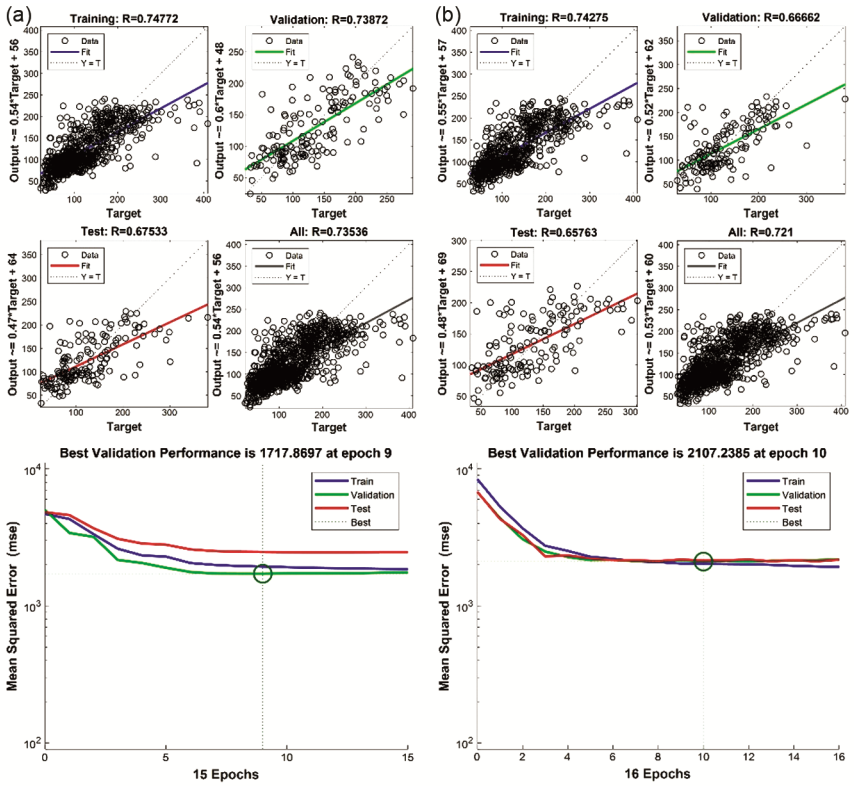

Fig. 7 — The correlation (R) values for training, validation, test, and all data regression analysis. a) MODIS b) VIIRS.

Table 3 — Statistical analysis of different model performance

Model	N	FAC2	MB	MGE	NMB	NMGE	RMSE	R	COE	IOA
MODIS BV	587	0.782	33.7	41.1	0.37	0.451	50.1	0.414	-0.344	0.328
MODIS MV	587	0.782	36.4	44	0.4	0.483	52.1	0.566	-0.437	0.281
MODIS ANN	587	0.843	36.1	41.9	0.397	0.461	52	0.657	-0.371	0.314
VIIRS BV	587	0.748	38.7	45.4	0.425	0.499	54.7	0.326	-0.485	0.257
VIIRS MV	587	0.767	40	47	0.44	0.516	55.4	0.551	-0.536	0.232
VIIRS ANN	587	0.802	40.1	45.9	0.441	0.504	56.8	0.637	-0.499	0.251

validate the predicted $PM_{2.5}$ values against observed data provide a comprehensive assessment of model performance are shown in Table 3. Among the different approaches tested, the ANN models consistently demonstrated superior predictive capabilities, as evidenced by their higher coefficient of correlation (R) and index of agreement (IOA) values. As shown in Fig. 8. The MODIS ANN model achieved an R- value of 0.657 and an IOA of 0.314, while the VIIRS ANN model achieved an R- value of 0.637 and an IOA of 0.251. These results highlight the ANN models' ability to capture complex non-linear relationships between the input features and $PM_{2.5}$ levels, making them a promising choice for air quality prediction in the Lucknow region. Comparing the bi-variate and multi-variate models, it becomes evident that incorporating multiple variables significantly improves the accuracy of $PM_{2.5}$ predictions. The MODIS multi-variate model achieved an R- value of 0.566 and an IOA of 0.281, while the VIIRS multi-variate model achieved an R- value of 0.551 and an IOA of 0.232. This improvement is reflected in the lower root mean squared error (RMSE) for the multi-variate models, with the MODIS multi-variate model at 52.1 and the VIIRS multi-variate model at 55.4, indicating a better fit to the observed data compared to the bi-variate models. Both MODIS and VIIRS AOD data exhibit reasonably good correlations with observed $PM_{2.5}$ levels, with MODIS showing slightly better performance in general. The MODIS bi-variate Model achieved an R- value of 0.414 and an IOA of

0.328, while the VIIRS bi-variate model achieved an R- value of 0.326 and an IOA of 0.257. However, the ANN models using VIIRS AOD data still produce commendable results, indicating the potential usefulness of both satellite-based datasets for air quality estimation. Consistent with recent studies[18-20], our findings align with the superior performance of ANN models over traditional regression methods in $PM_{2.5}$ concentration prediction using remote sensing data. Similar research[21] utilizing MODIS AOD data, similarly observed that the ANN model outperformed traditional regression approaches. This study affirms the effectiveness of the ANN model in $PM_{2.5}$ prediction, especially with remote sensing data, as indicated by higher R- values compared to LR and MLR models.

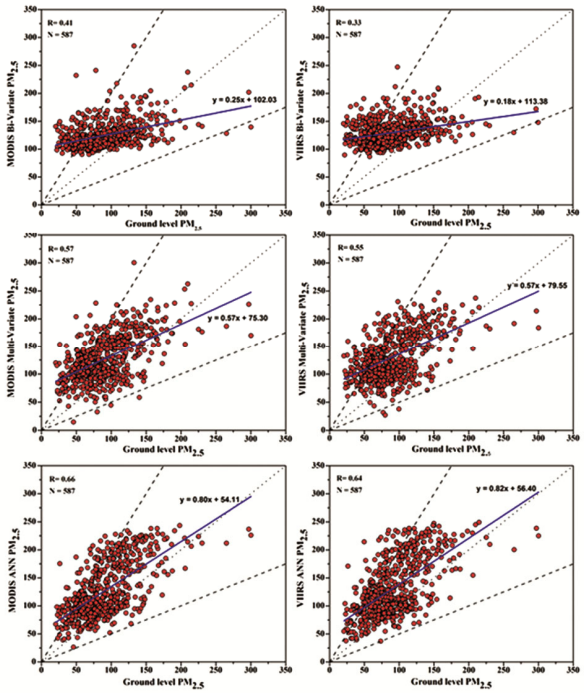

Fig. 8 — Scatter plots of estimated $PM_{2.5}$ concentration against CPCB measured values for Lucknow during Jan 2021 to June 2023 (The blue, dashed and dotted line represent the linear regression).

3.5 Temporal variability of Observed $PM_{2.5}$

Daily and monthly variation of observed $PM_{2.5}$ concentrations in the city of Lucknow, located in Uttar Pradesh, exhibited cyclic and abrupt fluctuations. The daily $PM_{2.5}$ concentration exhibited a trend characterized by elevated levels during the end of summer season and onset of winter, while it showed lower levels during the monsoon season (Fig. 8). The daily maximum value occurred on 20th December 2017 at 437.93 µg/m^3. While the minimum was on 20th August 2022 at 9.45µg/m^3(As shown in Fig. 9). Monthly average $PM_{2.5}$ concentrations in Lucknow city exhibited a U-shaped pattern. This pattern was characterized by a noticeable downward trend during the initial months of the year, from January to April. During this period, the city experienced relatively lower levels of $PM_{2.5}$ pollution.

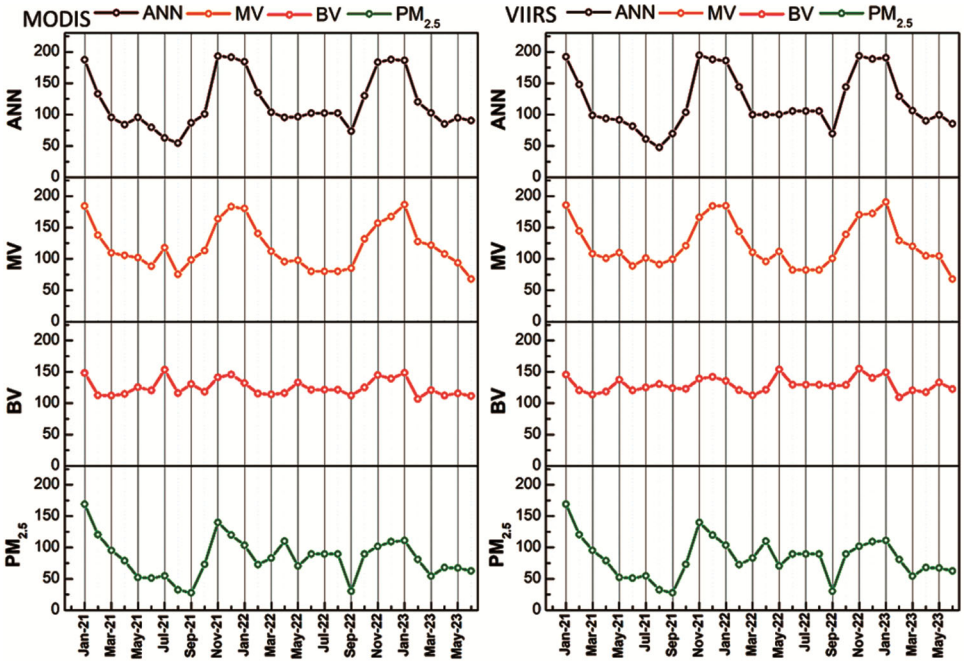

Fig. 9 — Monthly temporal variation of Observed and estimated $PM_{2.5}$ during Jan 2021 to June 2023.

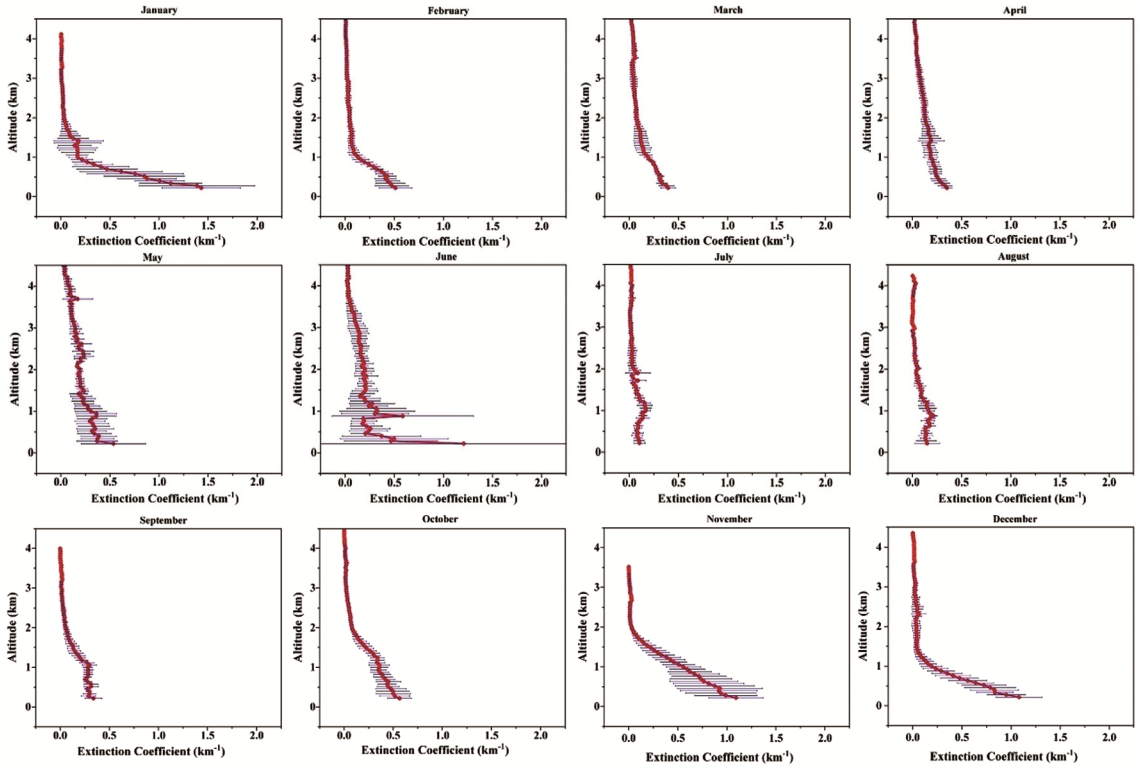

Fig. 10 — Aerosol vertical profile of Extinction coefficient for Lucknow region during 2017 to 2023 (The red line shows mean monthly extinction coefficient and blue lines SD).

Conversely, a contrasting upward trend was observed from July to December. As the year progressed into the latter half, the monthly average $PM_{2.5}$ concentrations in Lucknow increased steadily, indicating a rise in air pollution levels. The highest monthly maximum value occurred on November 2017 at 228.17 μg/m^3, while the minimum was on August 2022 at 21.58 μg/m^3.

3.6 Aerosol vertical profile

The analysis of the longitudinal dataset revealed distinctive patterns in the behaviour of aerosol extinction coefficients across altitudinal layers as shown in (Fig. 10). A consistent feature in the monthly-averaged profiles was the presence of elevated aerosol extinction coefficients in the mid-troposphere, ranging from 1 to 5 kilometres. The monthly-averaged values for this altitude range exhibited notable variations, with average extinction coefficient values ranging between 0.017 km^{-1} and 1.5km^{-1}). Moreover, the examination of the dataset highlighted secondary peaks in aerosol extinction coefficients at higher altitudes. These observations indicate the presence of distinct high-altitude aerosol layers, suggestive of convectively lifted aerosols originating from remote sources and subsequently transported by horizontal upper air currents. Further insight into the temporal variability was obtained by considering the maximum and minimum monthly values of aerosol extinction coefficients. In the mid-troposphere range of 1 to 3 km, the highest aerosol extinction coefficient values were recorded during January, June, and December, reaching values of 1.5, 1.45, and 1.25 km^{-1}, respectively. Therefore, the comprehensive analysis of aerosol extinction coefficients revealed distinctive altitude-dependent patterns in aerosol distribution. The mid-tropospheric elevation demonstrated elevated aerosol concentrations, while secondary peaks at higher altitudes indicated convectively transported aerosol layers. The identification of maximum and minimum monthly values contributes to our understanding of the temporal variability of aerosol distribution in the atmosphere, enhancing our insight into their potential impacts on atmospheric processes and radiative transfer.

4 Conclusion

This study aimed to employ a hybrid methodology, combining bi-variate linear regression (LR), multi-variate linear regression (MLR), and machine learning techniques using artificial neural networks

(ANN).The objective was to estimate the $PM_{2.5}$ concentrations in the highly urbanized city of INDIA for the period from 2021 to 2023. Furthermore, the integration of satellite Aerosol Optical Depth (AOD) with meteorological parameters was employed in the development of models. The results have demonstrated a positive correlation between the models' estimations and the target variable. Hence, the multiple linear regression (MLR) yielded satisfactory results but had limitations, as it performed well only at a few monitoring stations. On the other hand, the artificial neural network (ANN) model was meticulously developed and trained using input datasets to estimate daily averaged $PM_{2.5}$ concentrations across all study sites. The outcomes from the ANN model showed strong consistency and a significant correlation with the observed $PM_{2.5}$ levels. Combining the MLR and ANN approaches resulted in highly meaningful estimations concerning ground-level measurements. The key findings and implications of this integrated approach are summarized as follows.

1 The results indicated that the performance of the ANN model surpassed the MLR model and LR in dealing with large datasets, offering higher efficiency, improved accuracy, and reduced error.
2 The accuracy of $PM_{2.5}$ estimation was enhanced by including the meteorological parameters use data with satellite AOD
3 When estimating $PM_{2.5}$ concentrations, if the connection between input and output vectors exhibits non-linear behavior, deep neural networks can effectively approximate $PM_{2.5}$ levels by leveraging the power of nonlinear activation functions.

Incorporating CALIPSO's vertical profile observations significantly advances our understanding of particulate matter distribution, allowing for a more comprehensive assessment of atmospheric aerosol dynamics and their potential implications for air quality and human health.

Acknowledgements

The authors express their gratitude to NASA for providing access to the MODIS, VIIRS, and CALIPSO aerosol products through the Giovanni data portal (http://giovanni.gsfc.nasa.gov/) and to ERA5 for supplying meteorological parameters. We also acknowledge to CPCB for providing $PM_{2.5}$ data.

References

1 Kumar S & Dwivedi S K,*Environ Res*,(2021) 111754.
2 Kumar S, Bharti S K & Kumar N, *J Geol Soc India*, 99 (2023) 666.
3 IQ Air, World air quality report, https://www.iqair.com/world-most-polluted-cities, 2019.
4 Bharti S K, Kumar D, Anand S, Barman S C & Kumar N, *Micron*, 103(2017) 90.
5 Akanksha P, Pandey P & Somvanshi S, *Int J Res Appl Sci Eng Technol*,(2020) 2321.
6 Dockery D W, Pope III C A, Xu X, Spengler J D, Ware J H, Fay M E & Speizer F E, *The New Eng J Med*, 329 (1993) 1753.
7 Kappos A D, Bruckmann P, Eikmann T, Englert N, Heinrich U, Hoppe P, € Koch E, Krause G H, Kreyling W G & Rauchfuss K,*Int J Hyg Environ Health*, 207 (2004) 399.
8 Neuberger M, Schimek M G, Horak F, Moshammer H, Kundi M, Frischer T, Gomiscek B, Puxbaum H & Hauck H, *Atmos Environ*, 38(2004) 3971.
9 Wilson J G, Kingham S, Pearce J & Sturman A P, *Atmos Environ*, 39 (2005) 6444
10 Bravo M A & Bell M L, *J Air Waste Manag Assoc*, 61(2011) 69.
11 Paciorek C J & Liu Y, *Environ Health Perspect*, 117 (2009) 904.
12 Van D A, Martin RV, Brauer M, Kahn R, Levy R, Verduzco C, *et al*., *Environ Health Perspect*, 118 (2010) 847.
13 Lee HJ, Liu Y, Coull B & Schwartz K P, *Epidemiology*, 22(2011) S215.
14 Dee DP, Kall ¨ en E, Simmons A J & Haimberger L, *Bull Amer Meteorol Soc*, 92 (2011) 65.
15 Wu C & Yu J Z, *Atmos Meas Tech*, 11(2018) 1233.
16 Saunders R O, Kahl J D & Ghorai J K, *Atmos Environ*, 91(2014) 146.
17 Elangasinghe M, Dirks K, Singhal N, Costello S, Longley I & Salmond J, *Atmos Environ*, 83 (2014) 99.
18 Sharma V, Ghosh S, Dey S & Singh S, Ann GIS,(2023) 1.
19 Bera B, Bhattacharjee S, Sengupta N & Saha S, *Environ Chall*, 4(2021) 100155.
20 Li T, Shen H, Yuan Q, Zhang X & Zhang L, *Geophys Res Lett*, 44(2017) 11.
21 Ahmad M, Alam K, Tariq S, Anwar S, Nasir J & Mansha M, *Atmos Environ*, 219 (2019) 117050.