Scientific Reports

https://doi.org/10.1038/s41598-026-37934-9

Article in Press

Estimation of surface $PM_{2.5}$ over the Indo-Gangetic Basin using MERRA-2 reanalysis and machine learning

Received: 10 October 2025

Accepted: 28 January 2026

Published online: 25 March 2026

Cite this article as: Singh V., Singh S., Sharma N. *et al.* Estimation of surface $PM_{2.5}$ over the Indo-Gangetic Basin using MERRA-2 reanalysis and machine learning. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-37934-9

Vivek Singh, Sumit Singh, Nabin Sharma, Amarendra Singh, Aman Srivastava, Atul Kumar Srivastava, Deewan Singh Bisht, Kalpana Patel, Neeti Singh, Mansour Almazroui & Arti Choudhary

Estimation of Surface $PM_{2.5}$ over the Indo-Gangetic Basin using MERRA-2 Reanalysis and Machine Learning

Vivek Singh[1]*, Sumit Singh[1,2], Nabin Sharma[3], Amarendra Singh[4], Aman Srivastava[5] Atul Kumar Srivastava[1], Deewan Singh Bisht[1], Kalpana Patel[3], Neeti Singh[6], Mansour Almazroui[7,8], Arti Choudhary[9]*

[1]*Indian Institute of Tropical Meteorology (Delhi Branch), Prof Ramnath Vij Marg, MoES, New Delhi, 110060, India*
[2]*Civil Engineering Department, Institute of Engineering and Technology, Sitapur Road, Lucknow, Uttar Pradesh, 226021, India*
[3]*Department of Physics, SRM Institute of Science and Technology, Delhi-NCR Campus, Modinagar, Ghaziabad, 201204, India*
[4]*Centre for Atmospheric Sciences, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, 110016, India*
[5]*Department of Civil Engineering, Indian Institute of Technology Roorkee, Haridwar, 247667, India*
[6]*India Meteorological Department, Ministry of Earth Sciences, Lodi road, New Delhi, 110003, India*
[7]*Centre for Excellence for Climate Change Research/Department of Meteorology, King Abdulaziz University, Jeddah, Saudi Arabia*
[8]*Climatic Research Unit, School of Environmental Sciences. University of East Anglia, Norwich, NR4 7TJ, UK*
[9]*Department of Biosciences, School of Physical & Biological Sciences, Faculty of Science, Technology & Architecture, Manipal University Jaipur, Jaipur, 303007, India*

*Corresponding Authors:

(1) Dr. Vivek Singh
Indian Institute of Tropical Meteorology, New Delhi Branch,
Ministry of Earth Sciences,
Prof. Ram Nath Vij Marg, Double Storey, NPL Colony,
New Rajendra Nagar, New Delhi-110060, India
E-mail: vivek.singh@tropmet.res.in; vivek.sgh1@gmail.com

(2) Dr. Arti Choudhary
Department of Biosciences, School of Physical & Biological Sciences,
Faculty of Science, Technology & Architecture,

Manipal University Jaipur, Jaipur, 303007, India
E-mail: arti.choudhary@jaipur.manipal.edu

Highlights:

- Ensemble machine learning framework for $PM_{2.5}$ estimation in the IGB.
- Stacking model successfully upgraded MERRA-2 data with R^2 improved from 0.79-0.82, and RMSE reduced from 27-31 µg m^{-3}.
- The upgraded data successfully captured pollution peaks in the post-monsoon and winter seasons, along with the dip in the monsoon season.
- Trajectory clustering and CWT analysis identified dominant source regions.

Abstract

Fine particulate matter ($PM_{2.5}$) is a significant air pollutant in the Indo Gangetic Basin (IGB), where levels frequently exceed national and WHO air quality standards. Ground observations from 183 CPCB automatic stations, along with MERRA2 reanalysis products and meteorological variables, were utilized in this study to analyse $PM_{2.5}$ over ten years (2014-2023). A machine learning (ML) framework was developed using Random Forest, Extra Trees, LightGBM, and a stacking ensemble model to improve surface $PM_{2.5}$ estimation in four major IGB cities: Delhi, Kanpur, Lucknow, and Patna. It is found that the raw MERRA2 estimates systematically underestimated $PM_{2.5}$, with R^2 values of only 0.28-0.42 and RMSE as high as 82 µg m^{-3}. By contrast, the stacking ensemble achieved R^2 values of 0.79–0.82, FAC2 above 0.94, RMSE reduced to 27-31 µg m^{-3}, and near-zero bias (1.7-2.3 µg m^{-3}). The model successfully reproduced extreme winter pollution episodes as well as monsoon conditions, highlighting the critical role of meteorological parameters such as boundary layer height, wind speed, and precipitation in regulating $PM_{2.5}$ variability. Trajectory clustering and concentration-weighted trajectory (CWT) analysis showed that north-westerly transport contributes 55–65% of wintertime $PM_{2.5}$ in Delhi, Kanpur, and Lucknow, while Patna is affected by both regional inflows and local sources. Major contributing regions include Punjab, Haryana, Rajasthan, and the Nepal plains, associated with crop residue burning and dust transport. By integrating ground observations, reanalysis data, meteorological predictors, and atmospheric transport analysis, this study provides a robust framework for improving $PM_{2.5}$ prediction and identifying dominant pollution sources in the IGB. The results provide scientific evidence for designing both regional and city-

specific mitigation strategies to reduce exposure in one of the world's most polluted and densely populated regions.

Keywords: Machine learning, boundary layer height, Indo Gangetic Basin, Trajectory clustering, concentration-weighted trajectory.

1. Introduction

Fine particulate matter ($PM_{2.5}$) stands out as a significant air pollutant that impacts human health, ecosystems, and climate. Particles with aerodynamic diameters less than 2.5 micrometres possess the ability to penetrate the lungs and subsequently enter the bloodstream. This can lead to significant health issues, including severe respiratory and cardiovascular diseases, neurological complications, and increased risk of early mortality. In recent years, numerous areas around the world have faced increased $PM_{2.5}$ levels, with South Asia, especially the Indo-Gangetic Basin (IGB) of India, standing out as one of the most impacted regions [1,2]. The Indo-Gangetic Basin is a transboundary region extending across northern India and parts of Pakistan, Nepal, and Bangladesh. Of all the cities in the IGB, Delhi, Lucknow, Kanpur, and Patna feature as densely populated and industrialized urban areas. The $PM_{2.5}$ concentrations are evaluated with respect to the National Ambient Air Quality Standards (NAAQS, India) and the World Health Organization (WHO) air quality guidelines [3,4].

The IGB encounters complex air quality issues, majorly resulting from human activities. Key sources of $PM_{2.5}$ in this area consist of emissions from vehicles, industries, burning of coal, combustion of biomass for

household cooking, dust from roads and construction activities, as well as extensive burning of agricultural residues during the post-monsoon season [5,6]. In addition to these emissions, meteorological factors play a crucial role in dispersing the pollutants. During the winter and post-monsoon seasons, low wind speeds, shallow planetary boundary layers, high humidity, and frequent temperature inversions create conditions that trap pollutants near the surface, making it difficult for them to disperse [7,8]. The seasonal conditions contribute to higher $PM_{2.5}$ levels, frequently causing extended periods of poor air quality that impact millions of individuals [9]. Accurate and spatially comprehensive data are crucial for assessing $PM_{2.5}$ pollution and informing effective mitigation strategies. The Central Pollution Control Board (CPCB) manages a system of ground-based Continuous Ambient Air Quality Monitoring Stations (CAAQMS) (https://app.cpcbccr.com/ccr/) that deliver detailed data on $PM_{2.5}$ levels in different cities throughout India. Nonetheless, the distribution of these stations is not uniform, resulting in inadequate coverage in semi-urban and rural areas, which leaves significant portions of the IGB lacking representation [10,11]. Consequently, satellite observations and atmospheric reanalysis datasets are frequently used to enhance ground-based measurements. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), developed by NASA's Global Modelling and Assimilation Office, is extensively utilised because of its extensive temporal coverage, global consistency, and the incorporation of various aerosol types. Although MERRA-2 provides important information regarding regional aerosol dynamics, its limited spatial resolution and reliance on model assumptions often result in an under-representation of near-surface $PM_{2.5}$ (diagnosed from the lowest model layer, centered at ~30–35 m above ground level), particularly during high pollution periods [12]. Machine learning (ML) techniques are highly effective in improving the accuracy of $PM_{2.5}$ estimation from MERRA-2 data. ML models can capture complex and nonlinear relationships between multiple input variables and observed $PM_{2.5}$ concentrations. Previous studies using ML approaches have focused primarily on MERRA-2 aerosol components such as dust,

black carbon, organic carbon, sulphate, and sea salt to estimate $PM_{2.5}$ levels [13,14]. However, many previous studies excluded key meteorological parameters that govern pollutant transport, dispersion, and transformation. In the present study, meteorological variables such as air temperature, relative humidity, wind speed, surface pressure, and planetary boundary layer height are explicitly incorporated to better capture these processes. Including these meteorological variables improves the model's ability to represent seasonal variability and meteorologically driven pollution episodes [15]. Accurate surface $PM_{2.5}$ prediction also requires an understanding of the relative roles of local emissions and regional to long-range transport processes in shaping air quality over the Indo-Gangetic Basin (IGB). Identifying the origin and pathways of polluted air masses is therefore essential for interpreting observed $PM_{2.5}$ variability and for informing effective emission-control and cross-boundary mitigation strategies [16]. Accordingly, this study integrates air-mass transport analysis using the Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) model, with a detailed methodological implementation described in Section 2. The use of Global Data Assimilation System (GDAS) meteorological inputs at 1° × 1° resolution provides improved representation of transport pathways compared to coarser datasets commonly employed in earlier studies, thereby enabling more reliable identification of potential source regions and dominant transport patterns [17].

This model uniquely integrates ground-based $PM_{2.5}$ measurements, MERRA-2 reanalysis, meteorological predictors, and trajectory-based source apportionment within a machine learning framework, providing an innovative approach to accurately predict $PM_{2.5}$ variability and identify dominant pollution sources across the Indo-Gangetic Basin. The findings directly inform both regional and city-specific air quality mitigation strategies.

2. Methodology and Data Sources

2.1 Study Area

The Indo-Gangetic Basin (IGB) is a major transboundary region in South Asia characterized
by high population density, intensive agricultural activity, and extensive urban–industrial development [18,19]. The region extending across northern India and parts of Pakistan, Nepal, and Bangladesh, the basin supports more than 400 million people and represents a key economic and agricultural corridor. The region is bounded by the Himalayas to the north and the Deccan Plateau to the south, a topographic configuration that restricts atmospheric ventilation and favors the accumulation of pollutants emitted from vehicular traffic, industrial activities, biomass combustion, and agricultural residue burning [20,21]. Seasonal meteorological conditions, including shallow planetary boundary layers during winter, frequent temperature inversions, and weak wind speeds, influence pollutant dispersion and contribute to elevated $PM_{2.5}$ concentrations across the basin [19].

Four urban centers within the IGB were selected for detailed analysis: Delhi, Kanpur, Lucknow, and Patna. These cities represent distinct sub-regional environments within the basin and are consistently identified among the most polluted urban areas in India [22]. Delhi, the national capital, is a megacity influenced by dense traffic emissions, industrial activities, and episodic crop-residue burning. Kanpur, located in the central IGB, is a major industrial center with substantial manufacturing and coal-based activities [23]. Lucknow, situated in the central-eastern part of the basin, is a rapidly growing urban area characterized by residential, commercial, and traffic-related emissions. Patna, located in the eastern IGB, represents an urban environment influenced by both local emissions and regional transport processes, including biomass burning [18,21,24]. Together, these cities capture a range of emission characteristics and urban settings within the Indo-Gangetic Basin.

Fig. 1(a) shows the geographical extent of the Indo-Gangetic Basin and the spatial distribution of Continuous Ambient Air Quality Monitoring Stations (CAAQMS) operated by the Central Pollution Control Board (CPCB). The monitoring stations are primarily concentrated in urban and peri-urban

regions, including Delhi, Kanpur, Lucknow, and Patna, while large rural areas of the basin remain sparsely monitored. Such uneven spatial coverage has been shown to limit the representativeness of ground-based $PM_{2.5}$ observations at the basin scale [25]. Fig. 1(b) illustrates the temporal expansion of the CAAQMS network across the IGB from 2014 to 2023, reflecting a substantial increase in monitoring capacity following national air-quality initiatives [26].

2.2 Datasets

2.2.1 MERRA-2 Reanalysis Data

The present study utilizes the Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), created by NASA's Global Modelling and Assimilation Office (GMAO). MERRA-2 offers a comprehensive global atmospheric reanalysis spanning from 1980 to the present, with a spatial resolution of 0.5° latitude by 0.625° longitude. This data is produced using the GEOS-5.12.4 assimilation system, which incorporates a diverse array of contemporary observations, such as hyperspectral radiances, microwave sensors, GPS-Radio Occultation profiles, and NASA's ozone measurements. The datasets are available in netCDF-4 formats and can be accessed through NASA's GES DISC [27–30]. This study utilized two complementary data products: the Aerosol Diagnostics dataset (M2T1NXAER: tavg1_2d_aer_Nx) and the Flux Diagnostics dataset (M2T1NXFLX: tavg1_2d_flx_Nx), which are single-level, 2D datasets including surface-level aerosol and meteorological information.

The aerosol diagnostics product provides hourly, time-averaged, single-level fields for key fine particulate species, such as dust (DUSMASS25), organic carbon (OCSMASS), black carbon (BCSMASS), sea salt (SSSMASS25), and sulphate (SO4SMASS) (GMAO, 2015a). In accordance with established methods, sulphate concentrations were adjusted by a factor of 1.375 to compensate for neutralised ammonium sulphate [31] . Surface-level $PM_{2.5}$ concentrations were then estimated as the sum of these species, converted from kg m^{-3} to µg m^{-3} by multiplying with 10^9.

$$PM_{2.5} = DUSMASS25 + OCSMASS + BCSMASS + SSSMASS25 + SO4SMASS \times 1.375$$

The flux diagnostics dataset provides hourly, time-averaged surface meteorological variables that strongly influence aerosol transport and removal processes. These include total precipitation, bias-corrected precipitation, surface air temperature, specific humidity, wind speed and components, latent heat flux, planetary boundary layer height, surface layer height, and surface pressure (GMAO, 2015b). The combination of both datasets from MERRA-2 provides a comprehensive, gap-free dataset for evaluating aerosol behavior and meteorological influences across the Indo-Gangetic Basin.

2.2.2 Central Pollution Control Board (CPCB) Ground-based $PM_{2.5}$

The study uses surface $PM_{2.5}$ observations from India's CAAQMS operated by the CPCB and State Pollution Control Boards (SPCB). CAAQMS provides near–real-time particulate measurements from regulatory-grade analyzers (predominantly beta attenuation monitors), reported in µg m^{-3} with timestamps in Indian Standard Time (IST). Data are disseminated through CPCB's real-time portal and the Central Control Room (CCR) services (https://airquality.cpcb.gov.in/ccr), following CPCB's QA/QC protocols, which include routine instrument calibration, flow audits, span checks, and removal of invalid or flagged data (e.g., negative or calibration values). For this work, we compiled $PM_{2.5}$ data from 183 CAAQMS stations across the IGB, covering the states of Punjab, Haryana, Delhi, Uttar Pradesh, Bihar, and West Bengal, for the period 2014–2023.

The native high-frequency streams were first processed into hourly values and subsequently aggregated into daily averages to enable consistent temporal analysis. For the ML applications, four representative urban locations: Delhi, Kanpur, Lucknow, and Patna were selected. At each of these sites, daily $PM_{2.5}$ values were derived by averaging across all available monitoring stations within the city boundaries to provide a robust city-level estimate, reducing the influence of individual station biases or outages. CPCB's National Ambient Air Quality Standards (NAAQS)

conventions for units and averaging times were followed throughout. While the CAAQMS network offers dense, policy-grade surface constraints for the IGB, known issues such as data gaps, outages, and siting heterogeneity were addressed through completeness filters (at least 75% valid daily observations per year) and robust aggregation prior to further monthly, seasonal, and modelling analyses.

2.2.3. Global Data Assimilation System (GDAS) Data

The Global Data Assimilation System (GDAS), developed by the National Centers for Environmental Prediction (NCEP), assimilates meteorological observations from ground stations, radiosondes, aircraft, and satellites to produce global analyses and short-term forecasts [32]. GDAS outputs are available at different spatial and temporal resolutions, with the GDAS1 product providing fields at a horizontal resolution of 1° × 1° and 3-hourly intervals. Each day, four assimilation cycles (00, 06, 12, and 18 UTC) are performed, generating global meteorological parameters such as three-dimensional wind components, temperature, pressure, and humidity across ~23 pressure levels extending from the surface to about 20 hPa [33]. For HYSPLIT trajectory simulations, meteorological inputs were obtained from the GDAS1 archive in the form of weekly gridded binary files, which contain the required fields at 3-hourly time steps (https://www.ready.noaa.gov/data/archives/gdas1/). These files are widely used in atmospheric transport modelling because they provide continuous, regularly updated datasets suitable for both regional and long-range transport studies. In this study, the GDAS1 7-day files from 2014 to 2023 served as the primary meteorological driver for computing air-mass back trajectories and subsequent transport analyses over the Indo-Gangetic Basin.

2.3 Methodology

2.3.1 Spatio-temporal Analysis

The spatio-temporal distribution of $PM_{2.5}$ over the Indo-Gangetic Basin (IGB) was analyzed using both satellite reanalysis and ground-based observations to capture monthly and seasonal variations. For the MERRA-

2 dataset, the estimated surface-level $PM_{2.5}$ was processed into daily values and subsequently aggregated into monthly means to capture consistent temporal patterns. The processed data were then mapped to visualize large-scale gradients across the IGB. In parallel, ground-based $PM_{2.5}$ concentrations from 183 CAAQMS stations of CPCB were utilized to represent observed air quality. To overcome the spatial sparsity of station locations, an ordinary kriging interpolation technique was applied in ArcGIS to produce continuous spatial surfaces of $PM_{2.5}$ concentrations for each month. This allowed for the generation of comparable gridded fields between satellite-derived and ground-observed datasets.

The combination of these two approaches enabled both temporal trend analysis and spatial pattern identification (regional hotspots). Notably, the methodology allows to highlight areas of persistent high concentrations, such as urban clusters, while also facilitating a systematic comparison between MERRA-2-derived $PM_{2.5}$ and CPCB observations. These spatio-temporal maps provided the foundation for subsequent machine learning modeling by revealing the heterogeneity and seasonal drivers of $PM_{2.5}$ across the basin.

2.3.2 HYSPLIT-Based Trajectory, Clustering, and CWT Analysis

To explore the regional and long-range transport of particulate pollution HYSPLIT framework

(http://www.arl.noaa.gov/HYSPLIT.php) was used. Analyses were conducted for four representative IGB cities: Delhi, Kanpur, Lucknow, and Patna, which are recognized pollution hotspots and serve as receptors of both locally emitted and transported aerosols from long distances. At each site, three interlinked trajectory-based investigations were performed.

Five-day backward trajectories: To identify the dominant inflow directions and origins of air masses reaching each receptor site, 5-day backward trajectories were computed at a starting height of 1000 m above ground level, representing the mixed-layer transport of pollutants [34]. Trajectories were simulated using PySPLIT, a Python-based library that automates HYSPLIT execution and facilitates reproducible visualization [35].

Cluster analysis: Given the large number of trajectory endpoints, cluster analysis was applied to group air mass trajectories with similar transport characteristics [32,36]. This step reduces redundancy and highlights dominant transport regimes affecting each city, such as westerly dust intrusions during pre-monsoon months or easterly inflows during the monsoon. The clustering was carried out using the MeteoInfo (http://www.meteothink.org/) software package, which allows trajectory distance calculation and objective clustering of flow patterns.

Concentration Weighted Trajectory (CWT) analysis: To link air mass origins with observed surface pollution, a receptor-based CWT analysis was implemented. In this method, each grid cell along a trajectory pathway is weighted by the corresponding receptor concentration, in this case CPCB city-level averaged $PM_{2.5}$ concentrations at each of the four sites. This approach enables the identification of potential geographical source regions contributing to elevated PM2.5 levels at the four receptor sites [34,37,38]. CWT computations and spatial mapping were also performed in MeteoInfo, which integrates trajectory datasets with receptor observations.

2.3.3 Machine Learning Modeling

We developed a supervised machine learning framework in Python (v3.10) to predict daily mean $PM_{2.5}$ concentrations measured by CPCB's Continuous Ambient Air Quality Monitoring System (CAAQMS) across four major Indo-Gangetic Basin cities (Delhi, Kanpur, Lucknow, and Patna). For each location, station-level observations were aggregated to generate a single daily $PM_{2.5}$ value (target), while predictor variables were obtained from the MERRA-2 reanalysis products, namely aerosol diagnostics (M2T1NXAER) and flux diagnostics (M2T1NXFLX). The predictor suite encompassed both meteorological variables (surface-layer height, planetary boundary layer height, total precipitation, latent heat flux, air temperature, specific humidity, wind speed, wind direction, and surface pressure) and aerosol components (dust, black carbon, organic carbon, sea salt, and sulphate), thereby integrating both physical and chemical drivers

of air quality. Prior to model development, CPCB ground data and MERRA-2 predictors were temporally aligned at a daily resolution. CPCB $PM_{2.5}$ observations were aggregated spatially across monitoring stations within each city and temporally from hourly to daily resolution to ensure consistency with the coarse spatial representation of the MERRA-2 reanalysis. Although this aggregation smooths hyperlocal emission signals and short-term variability, it reduces representation error when comparing point-based observations with grid-averaged model outputs. Consequently, the results reflect city-scale background $PM_{2.5}$ concentrations rather than neighborhood-level or diurnal pollution dynamics. The daily MERRA-2 surface-level $PM_{2.5}$ dataset was partitioned into training (75%) and testing (25%) subsets for model development and evaluation. Model construction was carried out using the PyCaret regression framework (v3.3.2), which enabled automated preprocessing, standardized model configuration, and consistent performance evaluation within a unified workflow. To enhance robustness and limit overfitting, all models were evaluated using a 10-fold cross-validation strategy based on a K-fold approach, and cross-validation statistics were used for model comparison [39–41].

Three algorithms were utilized: Random Forest Regressor, Extra Trees Regressor, and Light Gradient Boosting Machine (LGBM), in addition to a stacking ensemble performed within the PyCaret framework. Random Forest and Extra Trees are ensemble methods that utilize bagging to construct numerous decision trees from random subsets of data and features, which helps to decrease variance and enhance stability. LGBM is an efficient gradient boosting technique designed for rapid processing and managing large datasets, recognized for its proficiency in capturing intricate nonlinear relationships. The stacking ensemble integrated these learners, utilizing Gradient Boosting, Random Forest, and Extra Trees as foundational models, while a Linear Regression performed as the meta-learner, effectively harnessing their complementary strengths. The evaluation of model performance involved several metrics, such as the coefficient of determination (R^2), root mean square error (RMSE), mean

absolute error (MAE), and mean bias error. The results were compared with the raw MERRA-2 aerosol-derived PM2.5 proxy to assess the enhanced predictive capability of machine learning. Additionally, the clarity of the model was enhanced by analysing feature importance and utilizing SHAP (Shapley Additive exPlanations) values. This analysis indicated that precipitation, planetary boundary layer height, and wind speed were the primary meteorological predictors, while aerosol sub-species like sulphate and dust also had a significant impact.

3. Results and Discussions

3.1 Comparative Assessment of CPCB and MERRA-2 $PM_{2.5}$ Reanalysis

An analysis was conducted on the time series of PM2.5 concentrations spanning from 2014 to 2023 for four prominent cities located in the Indo-Gangetic Basin (IGB): Delhi, Lucknow, Kanpur, and Patna. Fig.2 illustrates the comparison between CPCB ground-based measurements and MERRA-2 reanalysis estimates. In every panel, the CPCB series (blue) illustrates daily $PM_{2.5}$ values, whereas the MERRA-2 series (red) offers the corresponding reanalysis estimates. In various cities, observations from the CPCB show higher daily variations and more significant peaks than those from MERRA-2, highlighting how ground-based monitors react to localized emissions and short-term pollution events. For example, the maximum daily concentrations recorded by CPCB were 678.67 µg/m³ in Lucknow, 499.46 µg/m³ in Kanpur, 496.14 µg/m³ in Delhi, and 495.02 µg/m³ in Patna. In contrast, the highest estimates from MERRA-2 were significantly lower, with 416.84 µg/m³ in Kanpur, 399.79 µg/m³ in both Lucknow and Patna, and 351.04 µg/m³ in Delhi. Previous studies show that MERRA-2 underestimates environmental parameters over South Asia, with discrepancies ranging from sixteen percent to over forty percent [27,42,43]

Despite magnitude differences, both datasets display consistent seasonal cycles, with concentrations peaking during post-monsoon and winter months and reaching minima during the monsoon [44,45]. The observed peaks significantly surpass the WHO 2021 Air Quality Guideline for 24-

hour $PM_{2.5}$ (15 µg/m³) by over thirty to forty times, and exceed the Indian NAAQS limit (60 µg/m³) by six to eleven times during the study period. This highlights the serious and persistent issue of particulate pollution in the IGB, emphasizing the necessity of combining detailed ground observations with global reanalysis datasets to ensure effective air quality management. [46].

3.2 Monthly and Seasonal Spatial–Temporal Patterns of $PM_{2.5}$

The analysis of $PM_{2.5}$ concentrations across the Indo-Gangetic Basin (IGB) on a monthly and seasonal basis, utilizing CPCB ground observations and MERRA-2 reanalysis estimates, reveals distinct seasonal variability with significant peaks during the colder months, as illustrated in Fig.3 and Fig.4, respectively. The CPCB data indicate that the peak concentrations are observed from November to January, typically varying between 150 and 180 µg/m³, especially in the central and western regions of the IGB (Singh et al., 2023). MERRA-2 exhibits similar spatial patterns, but at lower magnitudes ranging from approximately 100 to 120 µg/m³. This observation aligns with established underestimations associated with reanalysis products, attributed to their coarse spatial resolution and the effects of data assimilation smoothing [27].

In the pre-monsoon season spanning March to May, there is a significant reduction in $PM_{2.5}$ levels. Observations from the CPCB report values ranging from approximately 70 to 120 µg/m³, while MERRA-2 data suggest levels between 60 and 90 µg/m³. The reduction is influenced by more intense winds and increased convective mixing, which improve the dispersion of pollutants. However, dust transport and localised human activities can still raise concentrations in specific urban hotspots [48,49]. The monsoon period from June to September showcases the most favourable air quality conditions throughout the IGB, with CPCB levels typically falling below 60 µg/m³ and MERRA-2 values under 50 µg/m³. Heavy and regular precipitation during this timeframe effectively eliminates particulate matter via wet scavenging, while more profound planetary boundary layers enhance the dilution of surface pollutants [1]. After the monsoon season, there is a significant increase in $PM_{2.5}$ levels during

October and November. Data from the CPCB indicate concentrations between 100 and 160 µg/m³, while MERRA-2 reports values ranging from 90 to 120 µg/m³. The significant burning of agricultural waste, the onset of cooler temperatures, and reduced atmospheric dispersion capabilities are the primary factors contributing to this increase in the post-monsoon season [49,50]. Stable atmospheric conditions during this period, combined with frequent nocturnal temperature inversions, trap pollutants close to the surface, while prevailing northwesterly winds transport biomass-burning emissions from Punjab and Haryana towards the central and eastern IGB [51–53]. These transported plumes contribute to regional haze and enhanced secondary aerosol formation through photochemical oxidation and heterogeneous reactions [54,55].

3.3 Prediction of $PM_{2.5}$ Using Machine Learning

The comparative evaluation of the Stacking ensemble model against MERRA-2 reanalysis data for $PM_{2.5}$ estimation across four study locations validates clear advantages of the machine learning approach when compared with CPCB ground-based observations, as shown in Table 1. Across all locations, MERRA-2 consistently underestimates $PM_{2.5}$ concentrations, particularly during peak winter episodes, with Fraction of predictions within a factor of 2 (FAC2) values between 0.546 and 0.634, R^2 ranging from 0.28 to 0.42, and RMSE values between 64.5 and 82.1 µg m^{-3}. Negative mean bias (MB) values (-38.7 to -49.2 µg m^{-3}) indicate systematic underprediction, a limitation associated with its coarse spatial resolution of MERRA-2 [28]. In contrast, the Stacking model achieves markedly better agreement with observations, with FAC_2 values from 0.948 to 0.971, R^2 from 0.79 to 0.82, and substantially reduced Root Mean Square Error (RMSE) (27.4–31.5 µg m^{-3}). The model's near-zero bias (MB = 1.74–2.36 µg m^{-3}) and lower Mean Absolute Percentage Error (MAPE) (0.221–0.249) confirm its ability to reproduce both high and low concentration regimes accurately [54].

The stacking ensemble model shows a clear improvement over the individual machine learning models and the raw MERRA-2 estimates. It effectively captures both high $PM_{2.5}$ pollution events during winter and low

concentrations during the monsoon season, demonstrating consistent performance across different pollution regimes. It captures extreme winter peaks exceeding 300 µg m^{-3} in Delhi and Patna, above 280 µg m^{-3} in Kanpur, and above 250 µg m^{-3} in Lucknow, as well as clean monsoon conditions below 50 µg m^{-3}across all cities. On the other hand, MERRA-2 fails to replicate these seasonal extremes [56]. Scatter plots for all locations (Fig.5, Fig.6, Fig.7, and Fig.8) show Stacking predictions clustering tightly along the 1:1 line, with fewer outliers and reduced dispersion compared to Extra Trees, Random Forest, and LightGBM. Residual analyses (Fig.S1-S4), reinforce these findings; while MERRA-2 residuals skew negatively during high events, stacking residuals remain symmetric and homoscedastic, suggesting uniform performance across the full $PM_{2.5}$ range. Variable importance analysis using SHAP (Fig.S5-S8) reveals that black carbon, organic carbon, and sulphate are the most essential positive predictors across all cities, with black carbon particularly dominant in Delhi and Patna due to heavy vehicular emissions, biomass burning, and industrial activity [57,58]. Overall, the findings verify that the Notably, the stacking ensemble markedly reduces the negative mean bias observed in the raw MERRA-2 $PM_{2.5}$ across all cities, bringing bias values close to zero while simultaneously improving RMSE, FAC2, and agreement metrics, thereby demonstrating effective correction of the systematic underestimation inherent in the reanalysis product. The machine learning framework, including the integration of MERRA-2 aerosols with meteorological predictors, the use of stacking ensembles, and demonstrated performance gains over individual models and reanalysis data. A clear comparison with earlier $PM_{2.5}$ studies should be included to show how the ML framework advances prediction accuracy and reliability.

3.4 Trajectory Cluster and CWT-Based Source Apportionment of $PM_{2.5}$

The five-day backward trajectory cluster analysis (Fig.9) revealed distinct atmospheric transport regimes influencing $PM_{2.5}$ across the Indo-Gangetic Plain (IGP). In Delhi, the dominant cluster (56.5%) originated from the

northwest, covering the Thar Desert and Pakistan. This pattern is consistent with wintertime northwest transport pathways previously identified over northern India using similar trajectory clustering and CWT methods [59]. Secondary pathways from Central Asia (24.4 %) and the Arabian Sea (19.1 %) point to both long-range continental and marine-influenced contributions, similar to seasonal inflow patterns reported for Ghaziabad [60]. Lucknow and Kanpur displayed similar northwesterly dominance (62.8 % and 64.7 % respectively), supplemented by flows from west central India and occasional Arabian Sea intrusion. In contrast, Patna displayed a more complex transport profile with northern (43.5 %), northwestern (30.9 %), and westerly southwesterly (25.6 %) clusters, suggesting multiple source importances, in agreement with mixed pathway observations for eastern IGP sites [61]. The Concentration Weighted Trajectory (CWT) analysis (Fig.10) mapped elevated $PM_{2.5}$ contributions to regions corresponding with the dominant clusters. For Delhi, CWT hotspots (>140 µg/m) were concentrated over the Thar Desert, Punjab Haryana plains, and southern Pakistan, areas associated with dust storms and post-harvest biomass burning, as also reported for peak burning episodes in the IGP [54]. Lucknow and Kanpur showed overlapping CWT peaks across Punjab, Haryana, and Rajasthan, consistent with seasonal agricultural residue burning patterns recognized in the post-monsoon period [62]. In Patna, peak CWT contributions were identified over Bihar, eastern Uttar Pradesh, and the Nepal plains, which are known biomass burning and brick kiln hotspots, in agreement with localized emission source profiles observed in eastern IGP air quality assessments [63,64].

The strong agreement between trajectory-derived transport pathways and CWT highlights source areas, reinforcing that northwest regional and long-range transport is the principal component of wintertime $PM_{2.5}$ episodes in Delhi, Lucknow, and Kanpur [59,65]. Conversely, Patna's mixed source profile with significant contributions from eastern and local sources highlights the importance of city-specific mitigation strategies, such as

biomass burning control and industrial emission reduction, consistent with recommendations for similar eastern IGP urban centers [61].

Conclusion

This study provides an integrated evaluation of $PM_{2.5}$ variability in the Indo Gangetic Basin using ten years of CPCB ground measurements, MERRA2 reanalysis, machine learning models, and atmospheric trajectory analysis. Results show that raw MERRA2 consistently underestimated $PM_{2.5}$ concentrations, with R^2 values only 0.28-0.42, RMSE ranging from 64.5 to 82.1 μg m^{-3}, and negative bias between -38.7 and -49.2 μg m^{-3}. In contrast, the stacking ensemble model demonstrated significant improvements, achieving R^2 between 0.79 and 0.82, FAC2 values of 0.948-0.971, and reduced RMSE to 27.4-31.5 μg m^{-3} with near-zero bias (1.7-2.3 μg m^{-3}). These results confirm the strong predictive skill of machine learning in reproducing both extreme pollution episodes and clean monsoon conditions across all four cities. Trajectory clustering and concentration-weighted trajectory analysis further identified north-westerly air masses as the dominant transport pathway contributing more than 55–65% of wintertime $PM_{2.5}$ in Delhi, Kanpur, and Lucknow, while Patna exhibited a more complex profile with mixed contributions from the north, northwest, and local sources. The CWT analysis highlighted hotspot source regions over Punjab, Haryana, Rajasthan, and the Nepal plains, consistent with large-scale crop residue burning and dust transport. The quantitative improvements achieved by the machine learning framework and the clear identification of dominant transport pathways underscore the dual necessity of regional-scale interventions and local mitigation. The integrated framework presented here is scalable and transferable to other regions with sparse monitoring and offers a pathway for real-time forecasting and cross-boundary air pollution management. For city-level air quality management and targeted interventions, ground-based CAAQMS observations and city-specific source apportionment studies remain more appropriate. The limitation of the study is MERRA-2's coarse spatial resolution, which hinders the use of machine learning-enhanced

$PM_{2.5}$ estimates for localized decision-making. Although the approach improves regional $PM_{2.5}$ representation in the Indo-Gangetic Basin, future efforts should focus on integrating higher-resolution satellite data, land-use and emission proxies, and ground-based observations to enhance local air quality monitoring and policy effectiveness.

Acknowledgements

The authors acknowledge the Central Pollution Control Board (CPCB), India, for providing long-term surface $PM_{2.5}$ measurements from its monitoring network, which formed the observational backbone of this study. Authors express their gratitude to Manipal University Jaipur for providing Open access funding for the current publication. We also thank the NASA Global Modeling and Assimilation Office (GMAO) for the provision of the MERRA-2 reanalysis products and the meteorological variables used in this analysis. The computational facilities and research infrastructure provided by the authors' host institutions are duly acknowledged. The integration of multi-source datasets, combined with advanced machine learning frameworks, was made possible through these resources. The authors gratefully acknowledge the Ministry of Earth Sciences (MoES), Government of India, New Delhi for their guidance, support, and collaborative framework that facilitated this research. The authors also acknowledge the scientific discussions and constructive feedback from colleagues that helped refine the methodology and strengthen the interpretations presented in this work. We also express our appreciation for the '**PyCaret'** machine learning framework, an open-source, low-code Python library that streamlines end-to-end ML workflows by automating data preparation, model training, comparison, and deployment.

Funding Declaration

Authors express their gratitude to Manipal University Jaipur, India for providing Open access funding for the current publication.

Declaration of interests

The authors declare that they have no competing financial or personal interests that could have influenced the work reported in this study.

Author statement

The views and conclusions presented in this article are solely those of the authors and do not necessarily represent the perspectives of their affiliated organizations. This work is entirely original, has not been submitted elsewhere, and the copyright of this article is exclusively held by the *Scientific Reports* Journal.

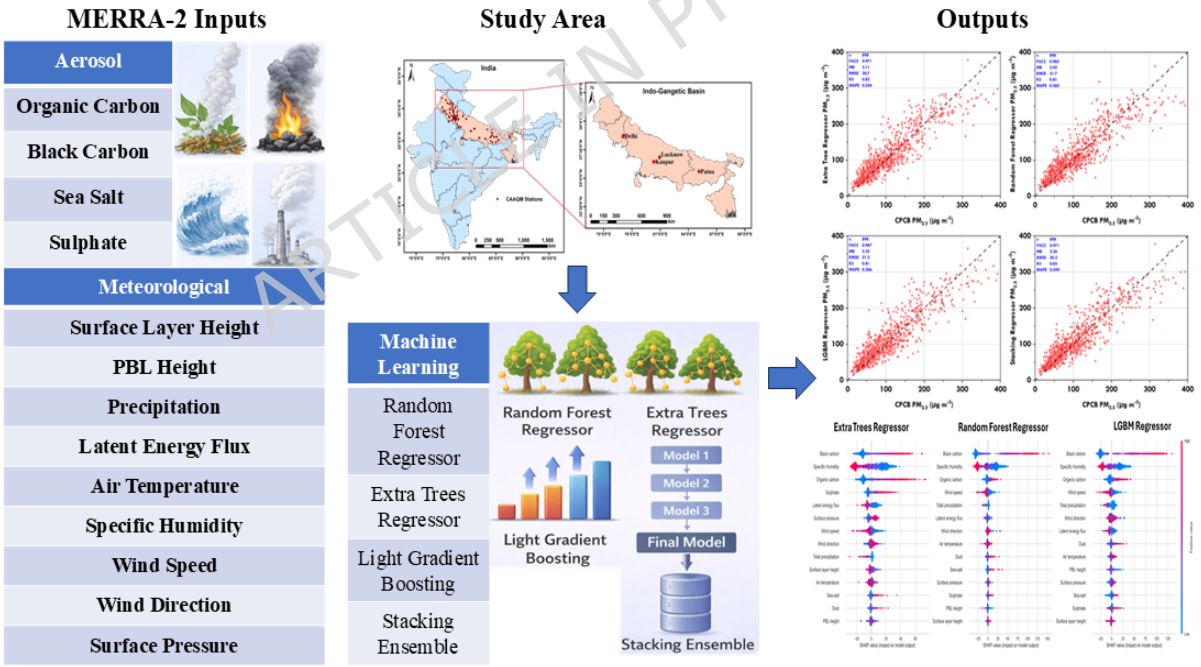
Data Availability

The datasets and codes generated and/or analysed during the current study are not publicly available due to institutional/data policy restrictions, but are available from the corresponding author on reasonable request.

CRediT Taxonomy

VS: Conceptualization; Formal analysis; Visualization; Software, Validation, Writing - original draft; and Writing - review &editing. **SS:** Visualization, Software; **NS:** Visualization, Software, **AS:** Data curation; Formal analysis; Visualization, **AS:** Data Curation; Visualization; **AKS:** Supervision; Validation, **DSB:** Data Curation; Visualization, **KP:** Visualization; Validation, **NS:** Software; Validation, **MA:** Supervision, **AC:** Supervision, Software, Validation.

Graphical Abstract



Figures

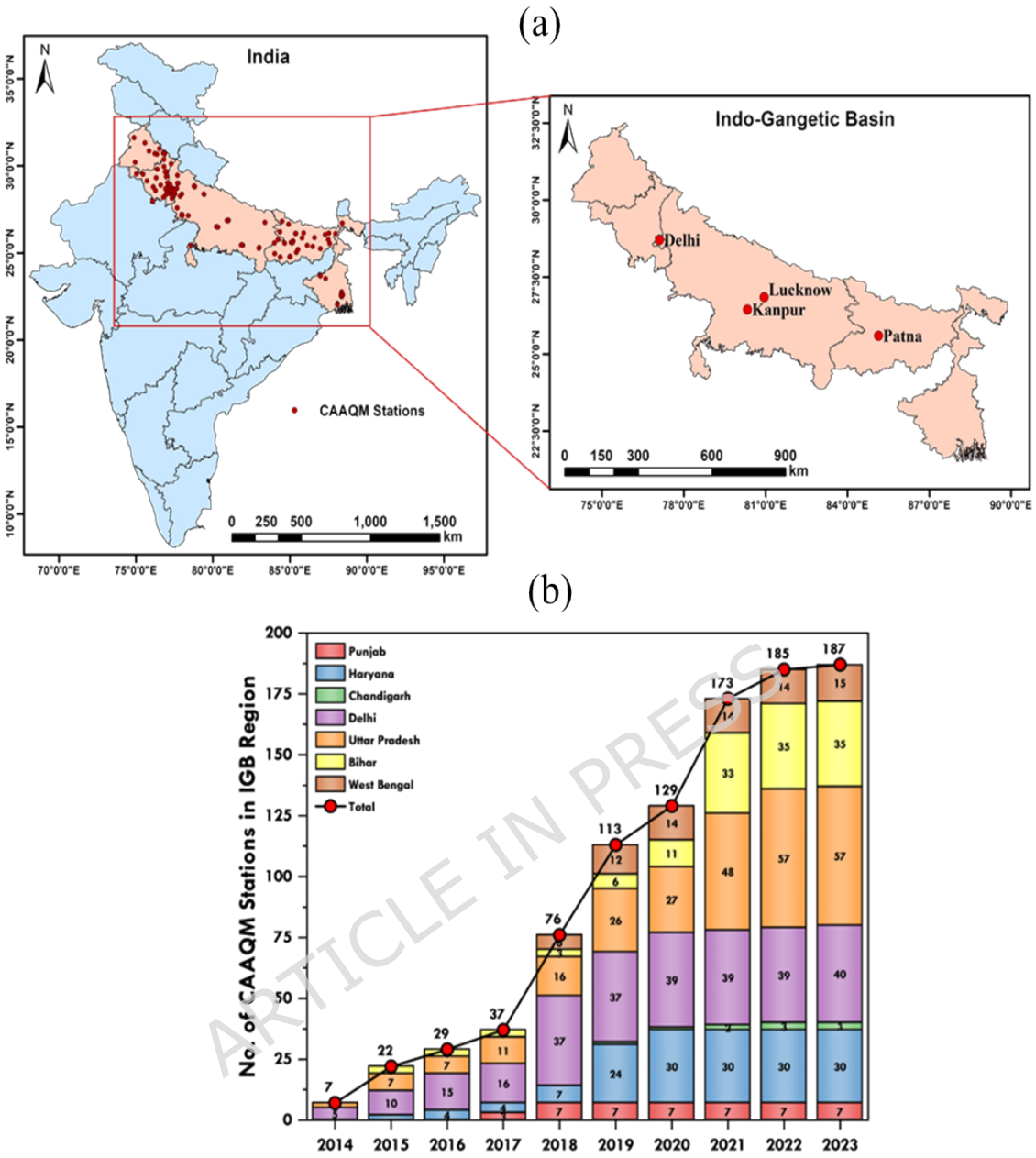


Fig.1. *(*a) Spatial distribution of CAAQMS stations across India and the Indo-Gangetic Basin (IGB), highlighting major cities including Delhi, Lucknow, Kanpur, and Patna. (b) Growth of the CAAQMS network from 2014 to 2023 across different IGB states.

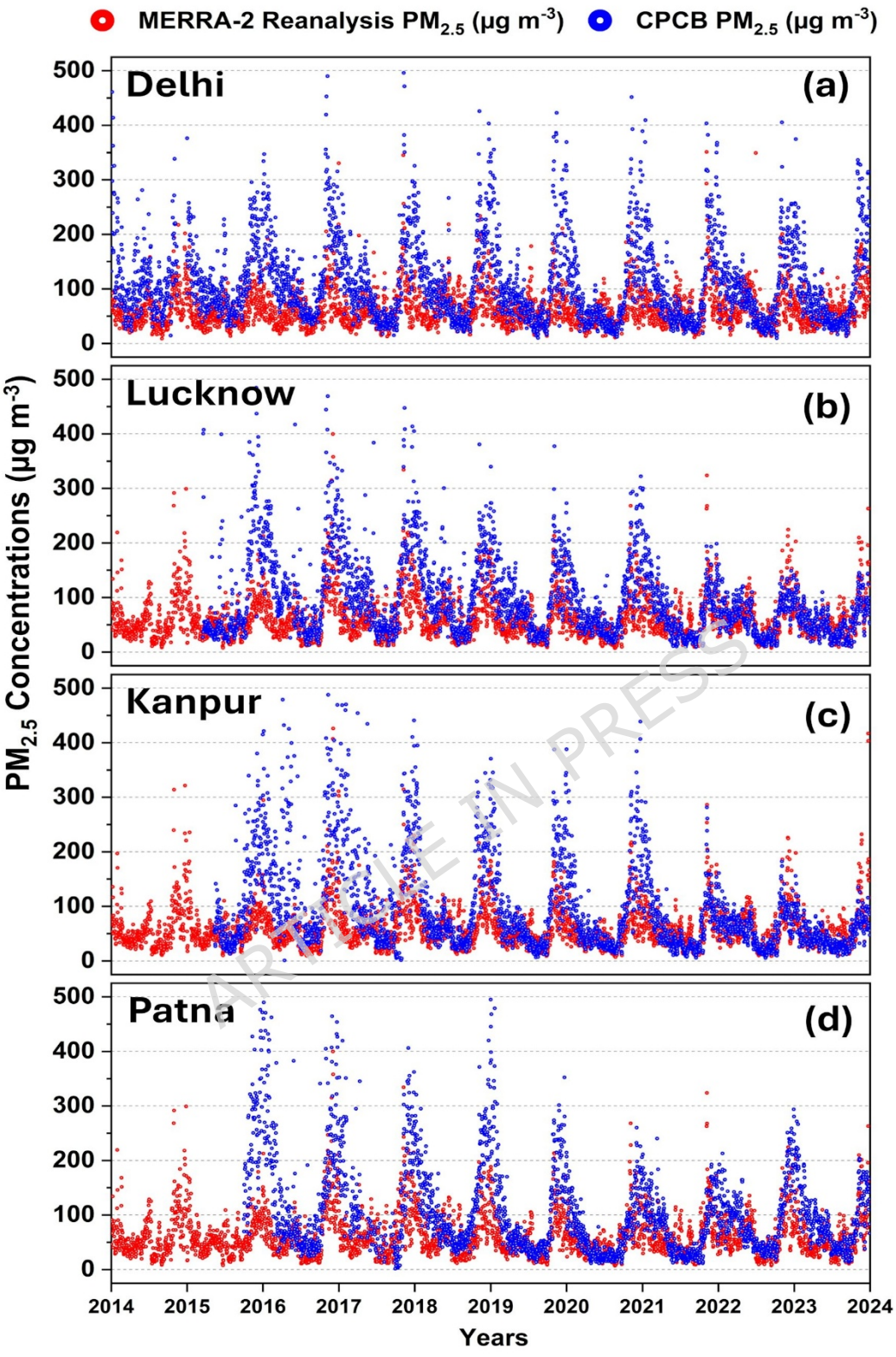


Fig.2. Time series of daily mean $PM_{2.5}$ concentrations from CPCB ground-based monitoring (blue) and MERRA-2 reanalysis (red) for four Indo-Gangetic Basin cities during 2014–2023: (a) Delhi, (b) Lucknow, (c) Kanpur, and (d) Patna.

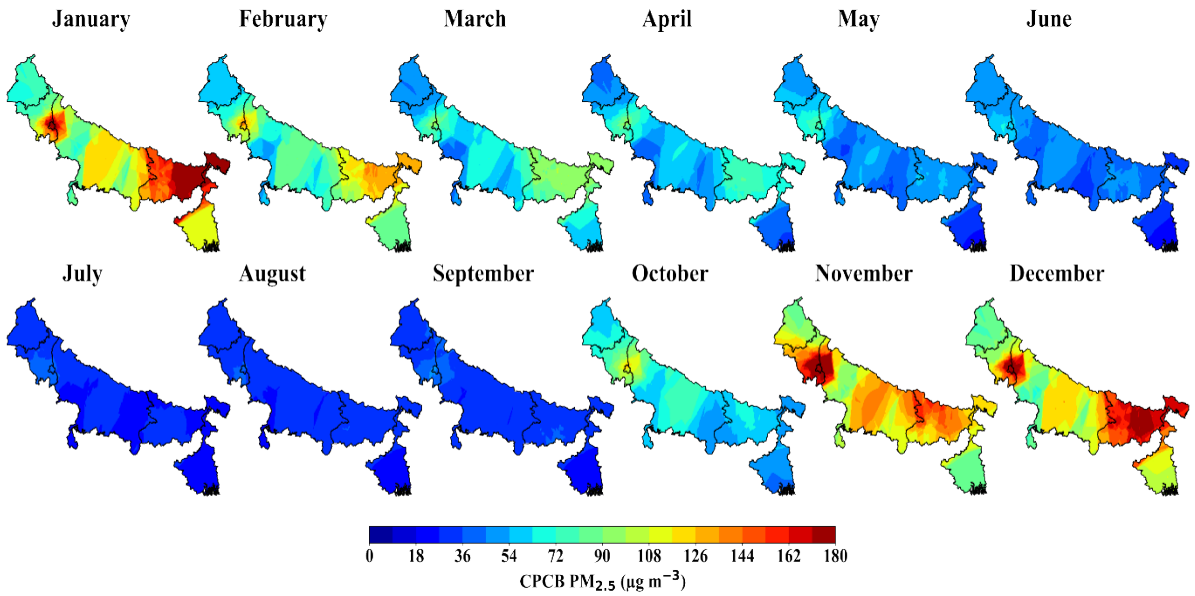


Fig.3. Ten-year (2014–2023) average $PM_{2.5}$ concentrations from CPCB ground observations across IGB.

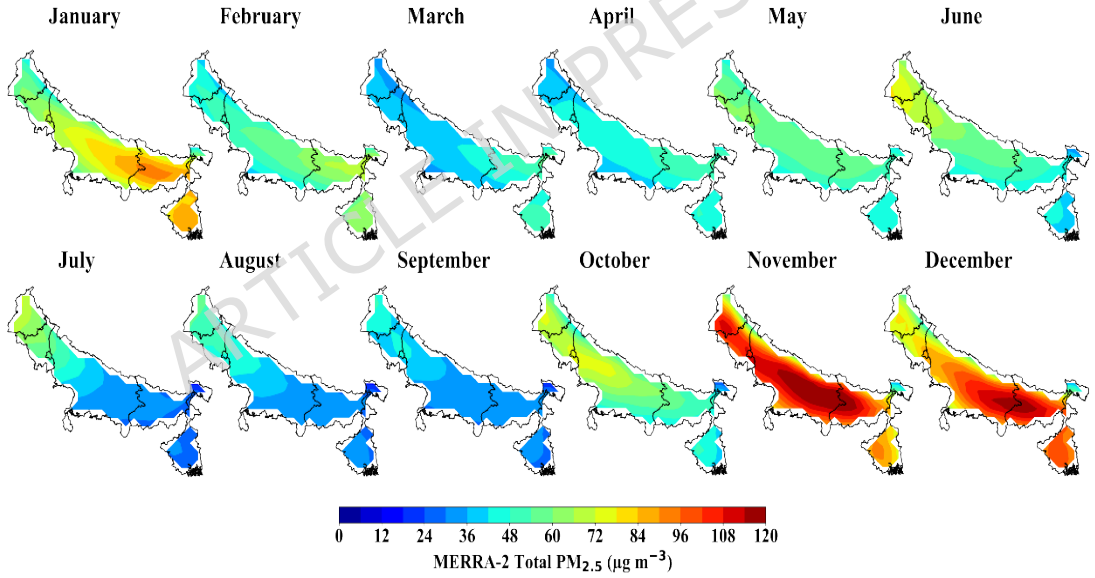


Fig.4. Ten-year (2014–2023) average spatial distribution of $PM_{2.5}$ over the Indo-Gangetic Basin derived from MERRA-2.

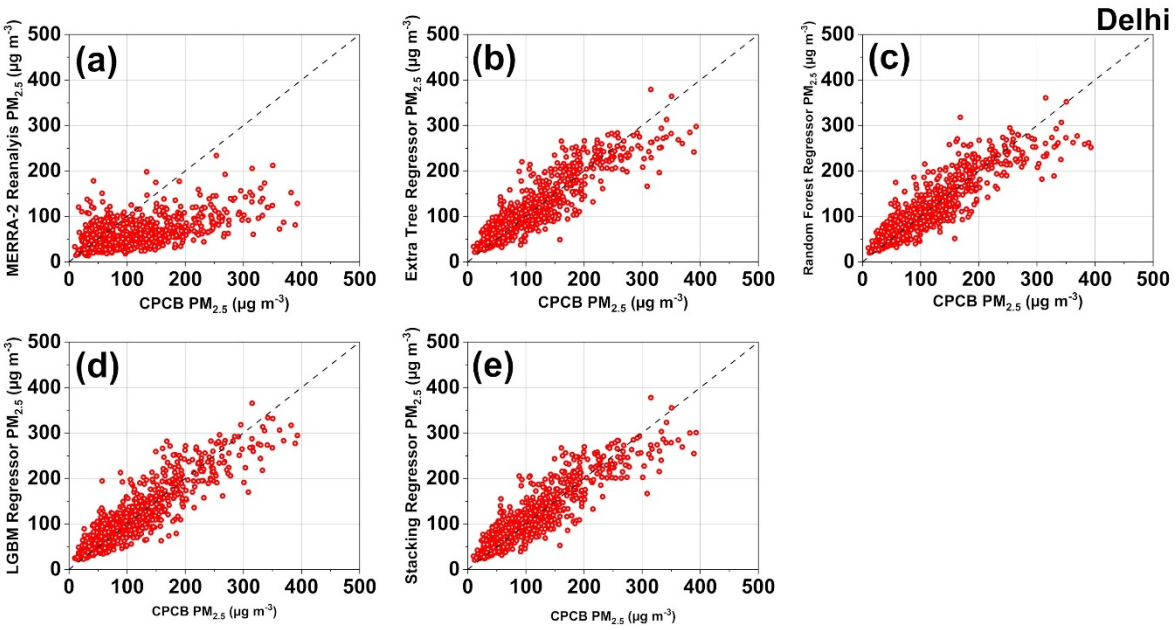


Fig.5. Scatter plots comparing CPCB observed $PM_{2.5}$ concentrations with (a) MERRA-2 reanalysis $PM_{2.5}$ and predictions from (b) Extra Trees Regressor, (c) Random Forest Regressor, (d) Light Gradient Boosting Machine (LGBM), and (e) the stacking ensemble for Delhi.

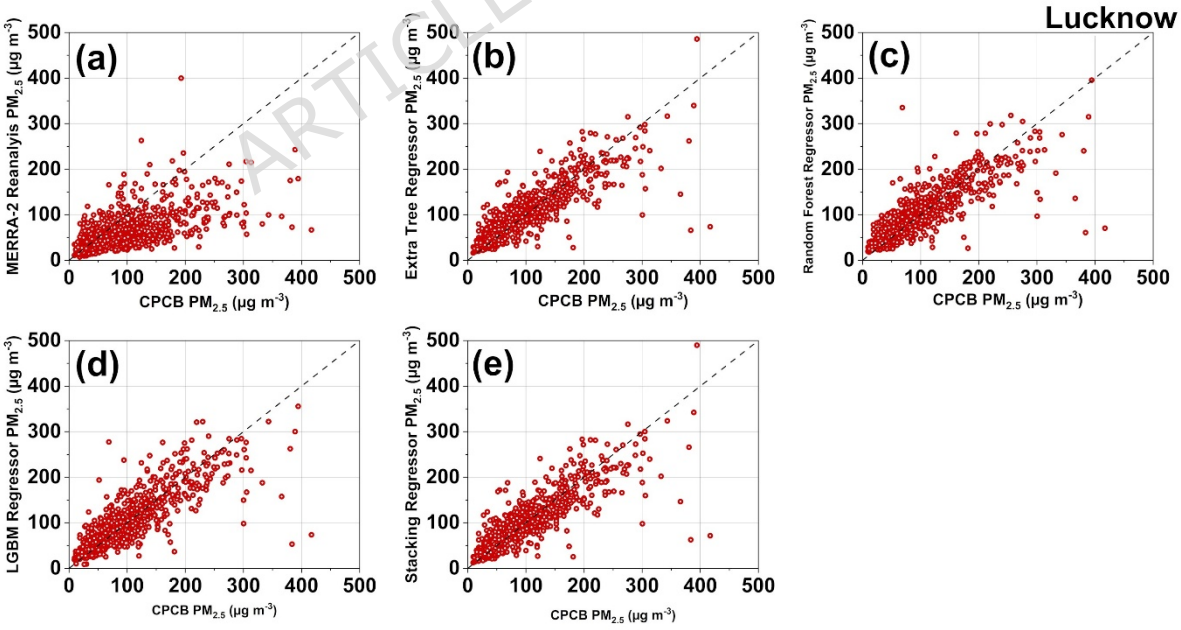


Fig.6. Scatter plots comparing CPCB observed $PM_{2.5}$ concentrations with (a) MERRA-2 reanalysis $PM_{2.5}$ and predictions from (b) Extra Trees

Regressor, (c) Random Forest Regressor, (d) Light Gradient Boosting Machine (LGBM), and (e) the stacking ensemble for Lucknow.

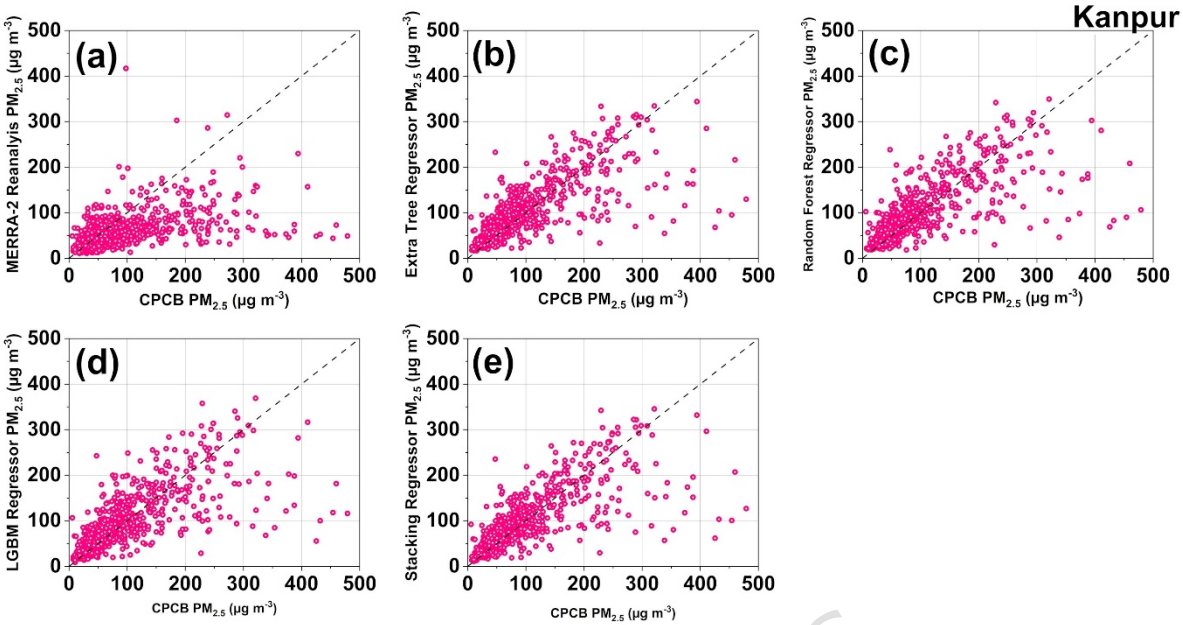


Fig.7. Scatter plots comparing CPCB observed $PM_{2.5}$ concentrations with (a) MERRA-2 reanalysis $PM_{2.5}$ and predictions from (b) Extra Trees Regressor, (c) Random Forest Regressor, (d) Light Gradient Boosting Machine (LGBM), and (e) the stacking ensemble for Kanpur.

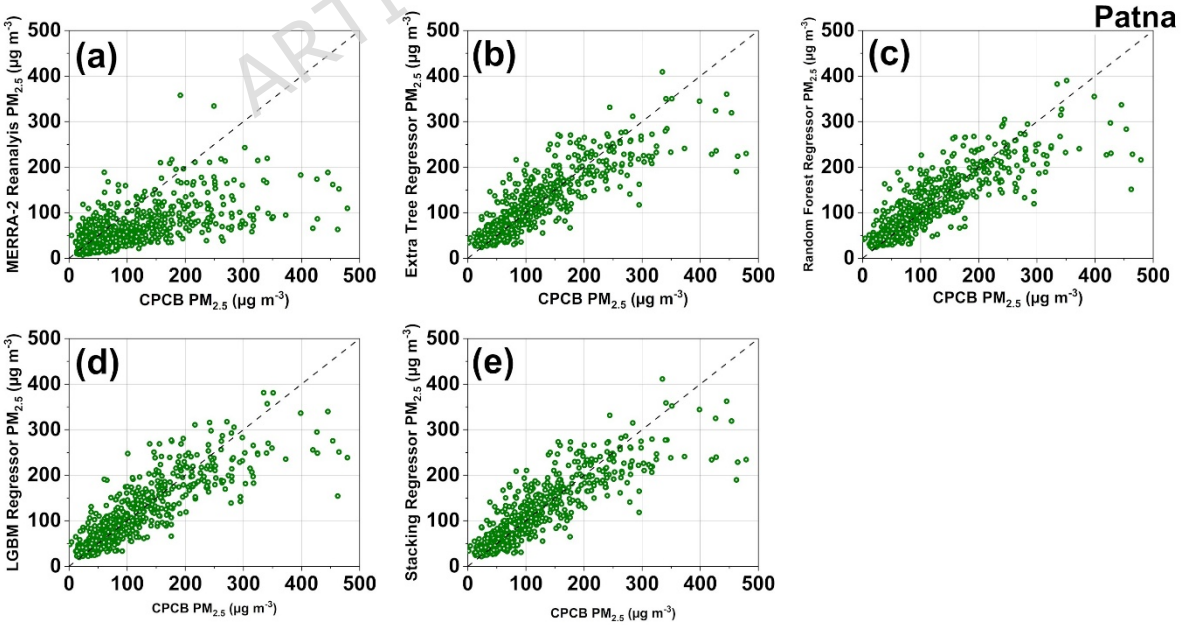


Fig.8. Scatter plots comparing CPCB observed $PM_{2.5}$ concentrations with (a) MERRA-2 reanalysis $PM_{2.5}$ and predictions from (b) Extra Trees Regressor, (c) Random Forest Regressor, (d) Light Gradient Boosting Machine (LGBM), and (e) the stacking ensemble for Delhi.

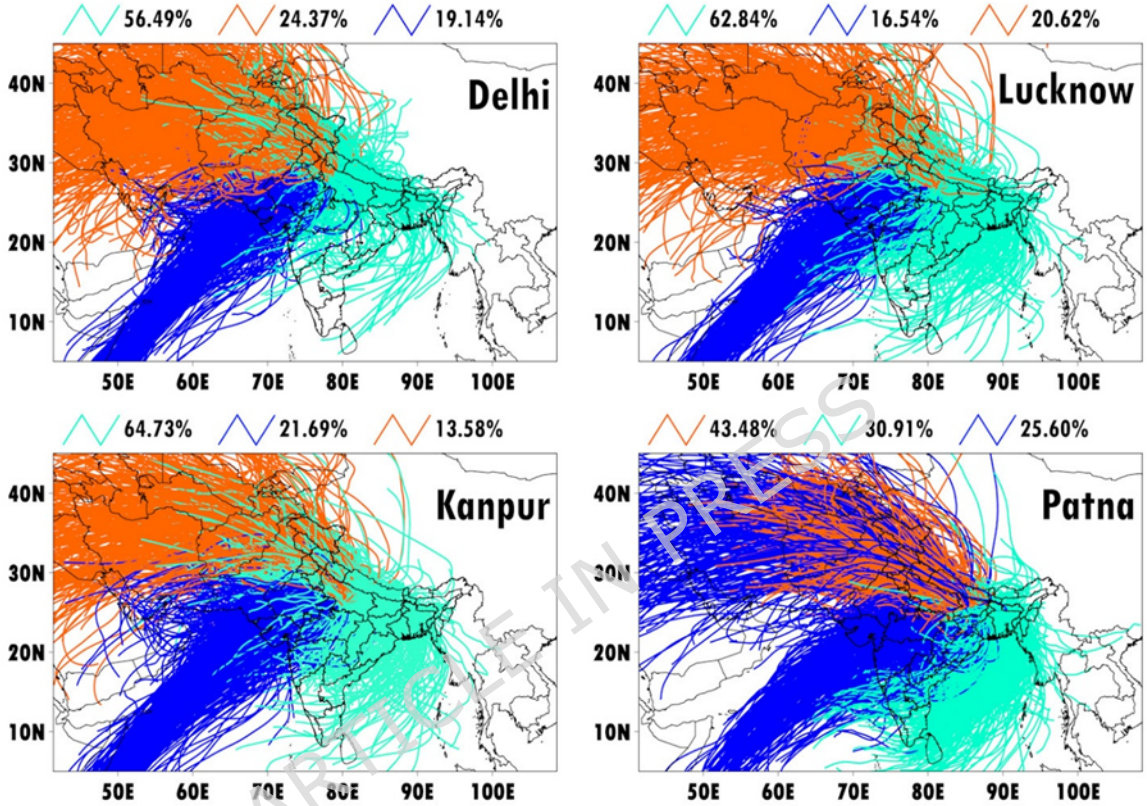


Fig.9. Clustered 5-day air-mass back trajectories for four IGB cities: Delhi, Lucknow, Kanpur, and Patna. Different colours represent trajectory clusters with their percentage contributions, highlighting dominant transport pathways influencing each location.

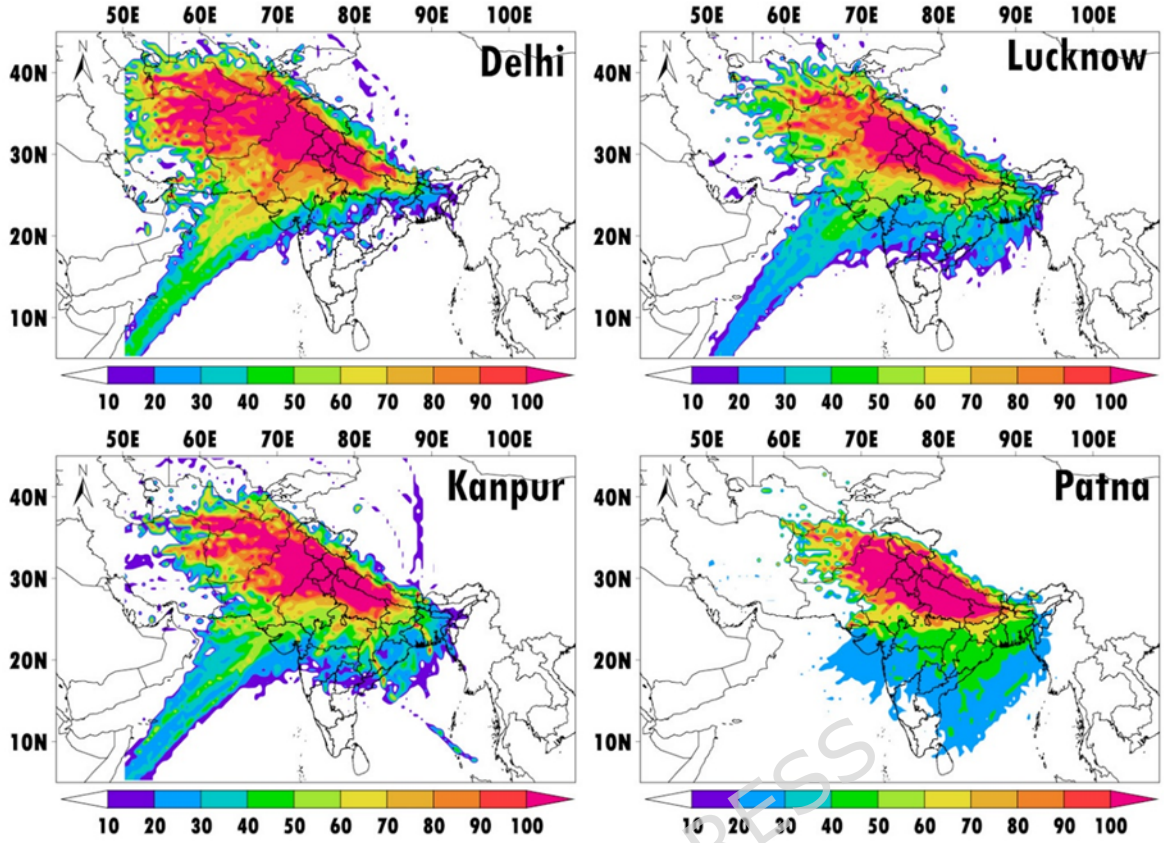


Fig.10. Concentration Weighted Trajectory (CWT) analysis for Delhi, Lucknow, Kanpur, and Patna. The coloured scale indicates potential source regions contributing to high $PM_{2.5}$ levels.

Table 1. Statistical evaluation of model performance for estimated vs ground-level $PM_{2.5}$ concentrations.

City	Model	n	FAC2	MB	MGE	RMSE	r	R^2	P	MAPE
Delhi	MERRA-2 Reanalysis	898	0.557	-47.3	57	76.8	0.56	0.31	3.13E-74	0.499
	Extra Trees	898	0.971	3.11	21.9	30.7	0.91	0.82	0	0.254
	Random Forest	898	0.965	2.42	22.5	31.7	0.90	0.81	1.13e-321	0.26
	LGB Machine	898	0.967	2.35	22.6	31.2	0.90	0.81	0	0.266
	Stacking	**898**	**0.971**	**2.36**	**21.7**	**30.2**	**0.91**	**0.82**	**0**	**0.249**
Lucknow	MERRA-2 Reanalysis	796	0.686	-34.5	45.2	65.3	0.62	0.38	1.02E-84	0.45
	Extra Trees	796	0.938	1.78	23.4	37	0.85	0.73	2.13E-225	0.315
	Random Forest	796	0.921	2.15	25	39.5	0.83	0.69	5.77E-203	0.335
	LGB Machine	796	0.928	1.43	25.1	39.4	0.83	0.69	8.37E-206	0.33
	Stacking	**796**	**0.941**	**0.0263**	**23**	**36.9**	**0.85**	**0.73**	**2.77E-226**	**0.293**
Kanpur	MERRA-2 Reanalysis	768	0.698	-34.5	46.9	77.8	0.49	0.24	2.15E-47	0.47
	Extra Trees	768	0.888	2.26	31	53.5	0.75	0.56	2.66E-137	0.417
	Random Forest	768	0.882	2.36	32.1	55.2	0.73	0.53	8.80E-128	0.43
	LGB Machine	768	0.876	0.206	33.4	56.2	0.72	0.52	7.25E-123	0.436
	Stacking	**768**	**0.888**	**-0.0985**	**30.7**	**53.6**	**0.75**	**0.56**	**4.43E-137**	**0.392**
Patna	MERRA-2 Reanalysis	713	0.581	-41.3	54.2	79.6	0.60	0.36	1.08E-69	0.598
	Extra Trees	713	0.92	0.689	26.9	41.3	0.87	0.76	3.42E-223	0.373
	Random Forest	713	0.92	-0.132	27.8	43.3	0.86	0.74	2.49E-209	0.385
	LGB Machine	713	0.923	-0.911	28	42.6	0.86	0.74	1.84E-213	0.392

	Stacking	**713**	**0.927**	**-0.36**	**26.7**	**41**	**0.87**	**0.76**	**1.99E-225**	**0.362**

References

1. Ojha, N. *et al.* On the widespread enhancement in fine particulate matter across the Indo-Gangetic Plain towards winter. *Scientific Reports* **10**, 5862 (2020).

2. Ali, M. A. *et al.* Long-term PM2.5 exposure in Bangladesh: identification of pollution hotspots, trends, sources and health risk assessment. *Air Quality, Atmosphere & Health* (2025) doi:10.1007/s11869-025-01768-7.

3. Wan Mahiyuddin, W. R., Ismail, R., Mohammad Sham, N., Ahmad, N. I. & Nik Hassan, N. M. N. Cardiovascular and Respiratory Health Effects of Fine Particulate Matters (PM2.5): A Review on Time Series Studies. *Atmosphere* **14**, 856 (2023).

4. Nkansah, F. K., Durosimi Belford, E. J., Hogarh, J. N. & Anim, A. K. Assessment of ambient air quality and health risks from vehicular emissions in urban Ghana: A case study of Winneba. *Journal of Air Pollution and Health* (2025) doi:10.18502/japh.v10i1.18092.

5. Chatterjee, D. *et al.* Source Contributions to Fine Particulate Matter and Attributable Mortality in India and the Surrounding Region. *Environmental Science & Technology* **57**, 10263–10275 (2023).

6. Sharma, N., Dahal, S., Patel, K. & Kumar, S. Study of the Correlation between Angstrom Exponent and Fine Mode Fraction in the Indo-Gangetic Plain Using Ground-Based Remote Sensing AERONET Data. *Journal of the Indian Society of Remote Sensing* **53**, 975–991 (2025).

7. Aslam, M. Y. *et al.* Seasonal characteristics of boundary layer over a high-altitude rural site in Western India: implications on dispersal of particulate matter. *Environmental Science and Pollution Research* **28**, 35266–35277 (2021).

8. Paulot, F., Naik, V. & W. Horowitz, L. Reduction in Near-Surface Wind Speeds With Increasing CO 2 May Worsen Winter Air Quality in the Indo-Gangetic Plain. *Geophysical Research Letters* **49**, (2022).

9. Dwivedi, P., Radha, R. S., Shekhar, H. & Sharma, S. K. The impact assessment of diwali firecrackers emissions on air quality in Delhi, India: a comparative study of eight consecutive years (2017–2024). *Journal of Atmospheric Chemistry* **82**, 9 (2025).

10. Mandal, S. *et al.* Nationwide estimation of daily ambient PM2.5 from 2008 to 2020 at 1 km2 in India using an ensemble approach. *PNAS Nexus* **3**, (2024).

11. Sharma, N., Dave, J. A., Kumar, S., Patel, K. & Singh, A. K. Variability in the concentration of particulate matter in Delhi-NCR: analysis and prediction using machine learning algorithms. *Atmospheric Environment* **360**, 121422 (2025).

12. Wang, S. *et al.* Reconstructing long-term (1980–2022) daily ground

particulate matter concentrations in India (LongPMInd). *Earth System Science Data* **16**, 3565–3577 (2024).

13. Mandal, S. *et al.* Assessing daily PM2.5 at every square kilometer of India over 2008-2020 using a machine learning framework. *ISEE Conference Abstracts* **2022**, (2022).

14. Anand M, D., Sahu, A. & Prakash, J. Assessment of Fine Aerosol in Two Different Climate Regions of India Using MERRA-2 Products, Ground-based Measurements, and Machine Learning. *Aerosol Science and Engineering* (2025) doi:10.1007/s41810-024-00279-9.

15. Masood, A. *et al.* Improving PM2.5 prediction in New Delhi using a hybrid extreme learning machine coupled with snake optimization algorithm. *Scientific Reports* **13**, 21057 (2023).

16. Prakriti *et al.* Deciphering Seasonal Variability and Source Dynamics of Urban Pollutants Over Delhi Under Surface Meteorological Influence Using Ground-Based and Trajectory Modeling Techniques. *Earth Systems and Environment* **9**, 1447–1463 (2025).

17. Shukla, G. & Kumar, A. Chemical composition of aerosols over the Bay of Bengal based on global reanalyses data and on-board ship measurements. *International Journal of Remote Sensing* 1–28 (2025) doi:10.1080/01431161.2025.2577974.

18. Pant, P. *et al.* Characterization of ambient PM2.5 at a pollution hotspot in New Delhi, India and inference of sources. *Atmospheric Environment* **109**, 178–189 (2015).

19. Das, M., Das, A., Ghosh, S., Sarkar, R. & Saha, S. Spatio-temporal concentration of atmospheric particulate matter (PM2.5) during pandemic: A study on most polluted cities of indo-gangetic plain. *Urban Climate* **35**, 100758 (2021).

20. Srimuruganandam, B. & Shiva Nagendra, S. M. Source characterization of PM10 and PM2.5 mass using a chemical mass balance model at urban roadside. *Science of The Total Environment* **433**, 8–19 (2012).

21. Saharan, U. S. *et al.* Hotspot driven air pollution during crop residue burning season in the Indo-Gangetic Plain, India. *Environmental Pollution* **350**, 124013 (2024).

22. Gargava, P. & Rajagopalan, V. Source apportionment studies in six Indian cities—drawing broad inferences for urban PM10 reductions. *Air Quality, Atmosphere & Health* **9**, 471–481 (2016).

23. Sharma, N., Kumar, S. & Patel, K. Variability of the optical and radiative characteristics of aerosols and classification of aerosol types over the Indo-Gangetic Plain during 2008 to 2018. *JOURNAL OF SCIENTIFIC RESEARCH* **69**, 48–58 (2025).

24. Sharma, N., Dahal, S., Chaurasiya, S. K., Kumar, S. & Patel, K. Impact

of climatic factors on volume aerosol size distribution over Northern India. *Journal of Atmospheric and Solar-Terrestrial Physics* **277**, 106633 (2025).

25. Dey, S. *et al.* A Satellite-Based High-Resolution (1-km) Ambient PM2.5 Database for India over Two Decades (2000–2019): Applications for Air Quality Management. *Remote Sensing* **12**, 3872 (2020).

26. Ganguly, T., Selvaraj, K. L. & Guttikunda, S. K. National Clean Air Programme (NCAP) for Indian cities: Review and outlook of clean air action plans. *Atmospheric Environment: X* **8**, 100096 (2020).

27. Randles, C. A. *et al.* The MERRA-2 Aerosol Reanalysis, 1980 Onward. Part I: System Description and Data Assimilation Evaluation. *Journal of Climate* **30**, 6823–6850 (2017).

28. Gelaro, R. *et al.* The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *Journal of Climate* **30**, 5419–5454 (2017).

29. Singh, S. *et al.* Assessment of Surface PM2.5 Concentrations over India using Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) Reanalysis Data. *Pure and Applied Geophysics* (2025) doi:10.1007/s00024-025-03666-6.

30. Buchard, V. *et al.* Evaluation of the surface PM2.5 in Version 1 of the NASA MERRA Aerosol Reanalysis over the United States. *Atmospheric Environment* **125**, 100–111 (2016).

31. Peng, W. & Weng, F. Impacts of Aerosol Scattering and Absorption on FY-4B Geostationary Interferometric Infrared Sounder (GIIRS) Observations. *Journal of Geophysical Research: Atmospheres* **130**, (2025).

32. Su, L., Yuan, Z., Fung, J. C. H. & Lau, A. K. H. A comparison of HYSPLIT backward trajectories generated from two GDAS datasets. *Science of The Total Environment* **506–507**, 527–537 (2015).

33. Luo, Y., Wei, H. & Yang, K. The impact of biomass burning occurred in the Indo-China Peninsula on PM2.5 and its spatiotemporal characteristics over Yunnan Province. *Science of The Total Environment* **908**, 168185 (2024).

34. Dimitriou, K., Remoundaki, E., Mantas, E. & Kassomenos, P. Spatial distribution of source areas of PM2.5 by Concentration Weighted Trajectory (CWT) model applied in PM2.5 concentration and composition data. *Atmospheric Environment* **116**, 138–145 (2015).

35. Warner, M. S. C. Introduction to PySPLIT: A Python Toolkit for NOAA ARL's HYSPLIT Model. *Computing in Science & Engineering* **20**, 47–62 (2018).

36. Cui, L., Song, X. & Zhong, G. Comparative Analysis of Three Methods

for HYSPLIT Atmospheric Trajectories Clustering. *Atmosphere* **12**, 698 (2021).

37. Dimitriou, K. The Dependence of PM Size Distribution from Meteorology and Local-Regional Contributions, in Valencia (Spain) - A CWT Model Approach. *Aerosol and Air Quality Research* **15**, 1979–1989 (2015).

38. Brereton, C. A. & Johnson, M. R. Identifying sources of fugitive emissions in industrial facilities using trajectory statistical methods. *Atmospheric Environment* **51**, 46–55 (2012).

39. Sayeed, A. *et al.* Hourly and Daily PM 2.5 Estimations Using MERRA-2: A Machine Learning Approach. *Earth and Space Science* **9**, (2022).

40. Hu, X. *et al.* Estimating PM 2.5 Concentrations in the Conterminous United States Using the Random Forest Approach. *Environmental Science & Technology* **51**, 6936–6944 (2017).

41. Doris, M. *et al.* Eighteen years of daily PM2.5 predictions (2005–2022) for a region of western Canada: Machine learning and satellite inputs for applications in rural health. *Atmospheric Environment* **355**, 121281 (2025).

42. Navinya, C. D., Vinoj, V. & Pandey, S. K. Evaluation of PM2.5 Surface Concentrations Simulated by NASA's MERRA Version 2 Aerosol Reanalysis over India and its Relation to the Air Quality Index. *Aerosol and Air Quality Research* **20**, 1329–1339 (2020).

43. Dhandapani, A., Iqbal, J. & Kumar, R. N. Application of machine learning (individual vs stacking) models on MERRA-2 data to predict surface PM2.5 concentrations over India. *Chemosphere* **340**, 139966 (2023).

44. Kumar, P. *et al.* New directions: Air pollution challenges for developing megacities like Delhi. *Atmospheric Environment* **122**, 657–661 (2015).

45. Guttikunda, S. K. & Jawahar, P. Atmospheric emissions and pollution from the coal-fired thermal power plants in India. *Atmospheric Environment* **92**, 449–460 (2014).

46. Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D. & Pozzer, A. The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature* **525**, 367–371 (2015).

47. Singh, A., Patel, A., Satish, R., Tripathi, S. N. & Rastogi, N. Wintertime oxidative potential of PM2.5 over a big urban city in the central Indo-Gangetic Plain. *Science of The Total Environment* **905**, 167155 (2023).

48. Kumar, A., Yadav, I. C., Shukla, A. & Devi, N. L. Seasonal variation of PM2.5 in the central Indo-Gangetic Plain (Patna) of India: chemical

characterization and source assessment. *SN Applied Sciences* **2**, 1366 (2020).

49. Singh, S. *et al.* Assessment of Surface PM2.5 Concentrations over India using Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) Reanalysis Data. *Pure and Applied Geophysics* **182**, 1713–1735 (2025).

50. Kumari, S., Verma, N., Lakhani, A. & Kumari, K. M. Severe haze events in the Indo-Gangetic Plain during post-monsoon: Synergetic effect of synoptic meteorology and crop residue burning emission. *Science of The Total Environment* **768**, 145479 (2021).

51. Tripathi, S. N., Yadav, S. & Sharma, K. Air pollution from biomass burning in India. *Environmental Research Letters* **19**, 073007 (2024).

52. Roy, C., Ayantika, D. C., Girach, I. & Chakrabarty, C. Intense Biomass Burning Over Northern India and Its Impact on Air Quality, Chemistry and Climate. in 169–204 (2022). doi:10.1007/978-981-16-7727-4_8.

53. Sharma, N., Kumar, S. & Patel, K. Aerosol type classification and its temporal distribution in Kanpur using ground-based remote sensing. *Journal of Atmospheric and Solar-Terrestrial Physics* **265**, 106366 (2024).

54. Singh, N. *et al.* Aerosol chemistry, transport, and climatic implications during extreme biomass burning emissions over the Indo-Gangetic Plain. *Atmospheric Chemistry and Physics* **18**, 14197–14215 (2018).

55. Mogno, C., Palmer, P. I., Knote, C., Yao, F. & Wallington, T. J. Seasonal distribution and drivers of surface fine particulate matter and organic aerosol over the Indo-Gangetic Plain. *Atmospheric Chemistry and Physics* **21**, 10881–10909 (2021).

56. Buchard, V. *et al.* The MERRA-2 Aerosol Reanalysis, 1980 Onward. Part II: Evaluation and Case Studies. *Journal of Climate* **30**, 6851–6872 (2017).

57. Arif, M., Kumar, R., Kumar, R., Eric, Z. & Gourav, P. Ambient black carbon, PM2.5 and PM10 at Patna: Influence of anthropogenic emissions and brick kilns. *Science of The Total Environment* **624**, 1387–1400 (2018).

58. Kumar, P. *et al.* Seasonal and Spatial Variations in Particulate Matter, Black Carbon and Metals in Delhi, India's Megacity. *Urban Science* **8**, 101 (2024).

59. Ghosh, S., Biswas, J., Guttikunda, S., Roychowdhury, S. & Nayak, M. An investigation of potential regional and local source regions affecting fine particulate matter concentrations in Delhi, India. *Journal of the Air & Waste Management Association* **65**, 218–231 (2015).

60. Gupta, L. *et al.* Assessment of PM10 and PM2.5 over Ghaziabad, an industrial city in the Indo-Gangetic Plain: spatio-temporal variability and associated health effects. *Environmental Monitoring and Assessment* **193**, 735 (2021).

61. Ravindra, K., Singh, T., Mandal, T. K., Sharma, S. K. & Mor, S. Seasonal variations in carbonaceous species of PM2.5 aerosols at an urban location situated in Indo-Gangetic Plain and its relationship with transport pathways, including the potential sources. *Journal of Environmental Management* **303**, 114049 (2022).

62. Sembhi, H. *et al.* Post-monsoon air quality degradation across Northern India: assessing the impact of policy-related shifts in timing and amount of crop residue burnt. *Environmental Research Letters* **15**, 104067 (2020).

63. Rahman, M. M., Begum, B. A., Hopke, P. K., Nahar, K. & Thurston, G. D. Assessing the PM2.5 impact of biomass combustion in megacity Dhaka, Bangladesh. *Environmental Pollution* **264**, 114798 (2020).

64. Emily, U. Bending Agricultural Burning Trajectories in Eastern India. *Graduate School of Cornell University* (2023).

65. Mazzeo, A. *et al.* Impact of residential combustion and transport emissions on air pollution in Santiago during winter. *Atmospheric Environment* **190**, 195–208 (2018).