Indian Journal of Pure & Applied Physics
Vol. 62, September 2024, pp. 764-778
DOI: 10.56042/ijpap.v62i9.11443

A Decadal Study of $PM_{2.5}$ Concentrations over Delhi using MERRA-2 and Ground Measurements: Predictive Insights via Machine Learning

Sumit Singh[a,b],Vikash Singh[c*], Ajay Kumar[a,b], Amarendra Singh[d], Atul Kumar Srivastava[b] & Virendra Pathak[a]

[a]Department of Civil Engineering, Institute of Engineering and Technology, Lucknow, UP 226 021, India
[b]Indian Institute of Tropical Meteorology, Ministry of Earth Sciences, New Delhi 110 060, India
[c]Department of Civil Engineering, Integral University, Lucknow, UP 226 026, India
[d]Centre for Atmospheric Sciences, Indian Institute of Technology, Hauz Khas, New Delhi 110 016, India

Received 28 May 2024; accepted 1 August 2024

This study investigates the spatial and temporal variations of $PM_{2.5}$ concentrations in Delhi from 2014 to 2023, utilizing ground-based measurements from the Central Pollution Control Board (CPCB) and MERRA-2 reanalysis data. The analysis reveals strong positive correlations ($r > 0.90$) across all districts, highlighting city-wide factors influencing $PM_{2.5}$ levels, such as vehicular emissions, industrial activities, and regional weather patterns. Seasonal patterns show $PM_{2.5}$ concentrations peaking during winter, attributed to lower temperatures, reduced wind speeds, and increased emissions from heating sources.To enhance the accuracy of $PM_{2.5}$ predictions, various machine learning (ML) models were employed, including Extra Trees Regressor, Random Forest Regressor, Light Gradient Boosting Machine (LGBM) Regressor, and a Stacking Regressor. These models utilized MERRA-2 sub-parameters like Dust, Organic Carbon, Black Carbon, Sea Salt, and Sulfate. The Stacking Regressor demonstrated the best performance, achieving an R^2 value of 0.67 and a significant improvement in correlation with CPCB measurements ($r = 0.86$). The ML models significantly improved the prediction accuracy of $PM_{2.5}$ concentrations compared to the original MERRA-2 data, reducing the Mean Bias from -39.4 $\mu g/m^3$ to around 10.4$\mu g/m^3$ and the Root Mean Squared Error (RMSE) from 71.1 $\mu g/m^3$ to below 40 $\mu g/m^3$. Additionally, the Fraction of predictions within a factor of 2 increased from 0.61 for MERRA-2 to over 0.89 for all ML models.These findings underscore the effectiveness of integrating machine learning models with MERRA-2 sub-parameters to accurately estimate $PM_{2.5}$ concentrations. This approach provides more reliable predictions of air quality, essential for developing targeted and effective air quality management strategies in Delhi.

Keywords: $PM_{2.5}$ concentrations; Delhi; Machine learning models; Air pollution; MERRA-2

1 Introduction

Air pollution, especially fine particulate matter ($PM_{2.5}$), is a critical environmental and public health concern worldwide. $PM_{2.5}$ refers to particulate matter with a diameter of less than 2.5 micrometers, which is small enough to penetrate deep into the lungs and enter the bloodstream**.** Prolonged exposure to $PM_{2.5}$ has been linked to severe health problems, including respiratory and cardiovascular diseases, lung cancer, and premature death[1-4]. Understanding the temporal and spatial distribution of $PM_{2.5}$ is essential for developing effective air quality management strategies to mitigate these health risks.

Delhi, the capital territory of India, is one of the most polluted cities globally.The city consistently experiences $PM_{2.5}$ levels far exceeding the safe limits established by the World Health Organization (WHO)[5,6]. During winter, air quality in Delhi deteriorates drastically due to a combination of local emissions from vehicular traffic, industrial activities, construction dust, and regional influences such as agricultural burning in neighboring states and long-range transport of pollutants[7-9]. This exacerbation during winter highlights the importance of analyzing the temporal variability of $PM_{2.5}$ to understand seasonal patterns and develop season-specific mitigation measures. Despite the availability of ground-based measurements from the Central Pollution Control Board (CPCB), which provide crucial data on $PM_{2.5}$ levels, these measurements are limited by their spatial coverage and temporal continuity. CPCB stations, while accurate, are often sparse and unable to cover the entire geographical expanse of Delhi comprehensively. To overcome these limitations, reanalysis datasets such as the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) offer an alternative by providing high-resolution, continuous

*Corresponding author: (E-mail: vikashsinghiu96@gmail.com)

estimates of $PM_{2.5}$ and its chemical constituents, including dust, organic carbon, black carbon, sulfate, and sea salt[10,11]. However, while MERRA-2 reanalysis data is invaluable, discrepancies often exist when compared with CPCB ground-based measurements due to various factors, including local emission sources and atmospheric conditions that the reanalysis data might not fully capture. These discrepancies necessitate the refinement of reanalysis data to enhance its correlation with ground-based observations for more accurate air quality assessments.

To address this challenge, this study employs advanced machine learning techniques to improve the correlation between MERRA-2 $PM_{2.5}$ estimates and CPCB ground-based measurements. By training machine learning models on historical data from 2014 to 2022, we aim to predict $PM_{2.5}$ concentrations for the year 2023 with enhanced accuracy. The machine learning models used in this study include Extra Trees Regressor, Random Forest Regressor, Light Gradient Boosting Machine, and Stacked Regressor. These models use MERRA-2 sub-components (dust, organic carbon, black carbon, sulfate, and sea salt) as input parameters to predict $PM_{2.5}$ concentrations more accurately.

Furthermore, to comprehensively understand the sources and transport mechanisms of $PM_{2.5}$ in Delhi, the study integrates the Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) model with National Centers for Environmental Prediction (NCEP) reanalysis data[12]. The HYSPLIT model performs trajectory-based Concentration Weighted Trajectory (CWT) analysis, examining both forward and backward trajectories[13]. This dual approach helps identify potential source regions and pathways of $PM_{2.5}$ transport, shedding light on both local and regional contributions to $PM_{2.5}$ levels in Delhi.The combined use of ground-based measurements, reanalysis data, machine learning predictions, and trajectory analysis addresses significant research gaps in understanding $PM_{2.5}$ pollution in Delhi. This comprehensive methodology offers a robust framework for analyzing the spatiotemporal patterns of $PM_{2.5}$, identifying key sources, and developing effective mitigation strategies. The integration of different data sources and analytical techniques not only improves the accuracy of $PM_{2.5}$ estimates but also provides actionable insights for policymakers to formulate targeted interventions to improve air quality.

In conclusion, this study aims to enhance the understanding of $PM_{2.5}$ pollution in Delhi for the last decade (2014-2023) through an innovative approach that combines reanalysis data, machine learning, and trajectory analysis. By addressing the limitations of existing data and methodologies, this research provides a detailed and accurate assessment of $PM_{2.5}$ levels and their sources.

The methodology used in this study can be applied to other cities with similar air quality issues. Necessary modifications include accounting for local emission sources, adapting to specific meteorological conditions, and ensuring the availability of comparable datasets. For instance, cities with significant industrial emissions might require additional parameters in the machine learning models to accurately predict $PM_{2.5}$ concentrations. To apply this methodology to other urban areas, modifications may include adjusting the model input parameters to reflect local emission sources and environmental conditions.

The findings will support the development of effective policies and strategies to mitigate air pollution and protect public health in Delhi, contributing to broader efforts to improve air quality in urban areas worldwide.

2 Materials and Methodology

2.1 Site Description

Delhi, the capital territory of India, is a sprawling metropolitan area situated in the northern part of the country. It spans 1,484 square kilometers and is bordered by the states of Haryana on three sides and Uttar Pradesh to the east[14].The city lies between latitudes 28.4°N and 28.9°N and longitudes 76.8°E and 77.3°E. With a population exceeding 30 million, Delhi is one of the most densely populated cities in the world. This high population density, combined with rapid urbanization and industrialization, contributes significantly to its air quality challenges. The city experiences diverse meteorological conditions, including hot summers, a monsoon season, and cold winters. These seasonal changes, especially during winter, often result in temperature inversions and lower wind speeds, which trap pollutants near the surface, leading to high concentrations of particulate matter ($PM_{2.5}$).

Delhi's air quality is further impacted by pollution from various sources such as vehicular emissions, industrial activities, construction dust, and biomass

burning. Additionally, transboundary pollution from neighboring states, particularly during crop residue burning seasons, exacerbates the problem. Understanding and analyzing $PM_{2.5}$ concentrations in Delhi is crucial for assessing health impacts on its population and developing effective air quality management strategies. The study area's map (Fig. 1) provides a detailed view of the spatial distribution of Continuous Ambient Air Quality Monitoring Stations (CAAQMS) across Delhi, illustrating average $PM_{2.5}$ concentrations from 2014 to 2023.

Table 1 presents a detailed overview of the monitoring stations distributed across different districts of Delhi, along with their respective average $PM_{2.5}$ concentrations measured in µg/m^3 from 2014 to 2023.Forty CAAQMS stations, marked with blue dots, are strategically placed across Delhi's 11 districts. $PM_{2.5}$ concentrations are categorized as 80-100 µg/m^3, 100-120 µg/m^3, 120-140 µg/m^3, and >140 µg/m^3. High concentrations are observed in central and northern regions such as North Delhi, West Delhi, and Central Delhi, while moderate levels are found in South West and East Delhi. Peripheral areas like parts of South East and South West Delhi show lower concentrations.

2.2 Data Description

2.2.1 CPCB $PM_{2.5}$ Concentrations

The Central Pollution Control Board (CPCB) of India operates a network of Continuous Ambient Air Quality Monitoring Stations (CAAQMS) across Delhi to monitor various air pollutants, including $PM_{2.5}$. These stations employ the beta attenuation method for measuring $PM_{2.5}$ concentrations, a widely accepted technique due to its accuracy and reliability[15,16]. The CAAQMS stations provide high-resolution data with a temporal resolution of 15 minutes, capturing the dynamic changes in air quality throughout the day. However, for this study, we have utilized the 24-hour average $PM_{2.5}$ concentration data to ensure consistency and comparability with other datasets. This aggregated data helps in understanding the daily

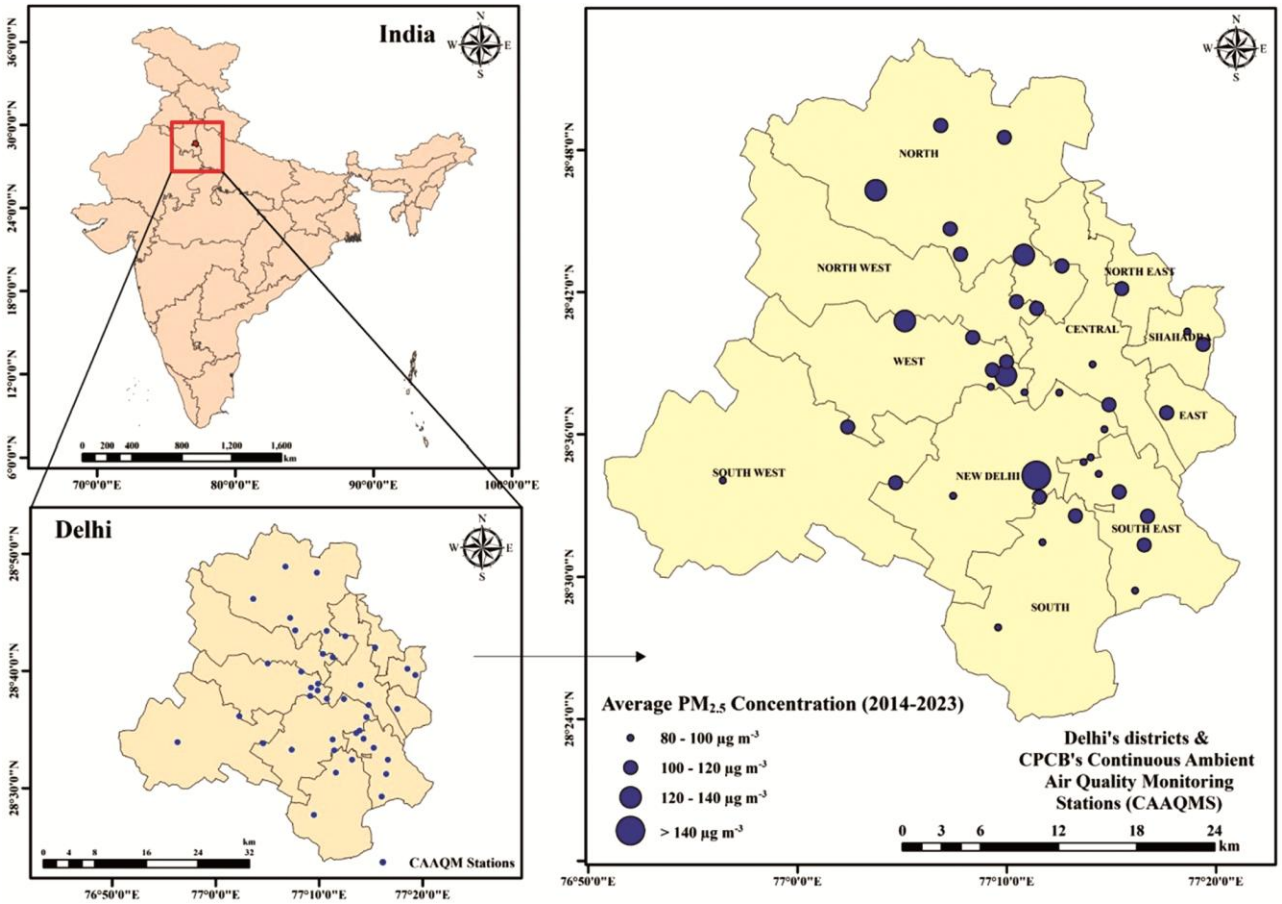

Fig. 1 — Map of Delhi, India, showing the distribution of Continuous Ambient Air Quality Monitoring Stations (CAAQMS) and the averaged $PM_{2.5}$ concentrations (2014-2023)

Table 1 — List of monitoring stations in various districts of Delhi along with the averaged $PM_{2.5}$ concentrations

District	No. of Monitoring Stations	Name of Monitoring Stations	Average $PM_{2.5}$ Concentration ($\mu g\ m^{-3}$)
Central	3	ITO, Sonia Vihar, Chandni Chowk	108.9 ± 84.2
East	1	Patparganj	105.3 ± 87.1
New Delhi	7	IGI Airport, Lodhi Road-IMD, R K Puram, Mandir Marg, Dwarka-Sector 8, Major Dhyan Chand National Stadium, New Moti Bagh	99.4 ± 79.2
North	7	Burari Crossing, DTU, Alipur, Bawana, Jahangirpuri, Narela, Rohini	117.8 ± 93.1
North West	3	Ashok Vihar, Mundka, Wazirpur	117.3 ± 96.5
Shahadra	2	IHBAS Dilshad Garden, Vivek Vihar	102.6 ± 81.8
South	3	Sirifort, Aya Nagar, Sri Aurobindo Marg	92.2 ± 79.3
South East	6	CRRI Mathura Road, Dr. Karni Singh Shooting Range, Jawaharlal Nehru Stadium, Nehru Nagar, Okhla Phase-2, Lodhi Road-IITM	101.6 ± 86.1
South West	2	NSIT Dwarka, Najafgarh	98.1 ± 68.4
West	6	North Campus DU, Punjabi Bagh, Shadipur, Anand Vihar, Pusa-IMD, Pusa-DPCC	107.4 ± 86.5
North East	-	-	-

variations and long-term trends of $PM_{2.5}$ levels in Delhi. The CPCB data serves as the ground truth for training and validating the machine learning models used in this study.

2.2.2 MERRA-2 Reanalysis $PM_{2.5}$

The Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2), produced by NASA's Global Modeling and Assimilation Office (GMAO), provides comprehensive reanalysis data, including $PM_{2.5}$ concentrations and its chemical constituents such as dust (DUSMASS25), organic carbon (OCSMASS), black carbon (BCSMASS), sulfate (SO4SMASS), and sea salt (SSSMASS25)[10]. These components are combined to estimate total $PM_{2.5}$ concentrations[10], offering extensive spatial coverage with a resolution of 0.5° x 0.625° and high temporal resolution with hourly data using Equation (1).

$$PM2.5 = DUSMASS25 + OCSMASS + BCSMASS + SSSMASS25 + SO4SMASS * \left(\frac{132.14}{96.06}\right) \quad \ldots (1)$$

This detailed, gridded data, available from 1980 to the present, is valuable for understanding broader trends and influences on air quality, capturing both natural and anthropogenic sources. However, discrepancies between MERRA-2 estimates and ground-based measurements necessitate the refinement of this data through machine learning techniques to improve its accuracy and reliability. For analyzing $PM_{2.5}$, we utilized MERRA-2 by summing contributions from dust, carbonaceous aerosols, sea salt, and sulfate, and aggregated the hourly data to daily and monthly averages for our analysis. Despite its comprehensive nature, enhancing MERRA-2 data through machine learning is crucial for achieving more accurate and reliable air quality predictions.

The integration of MERRA-2 data with ground-based measurements presented challenges such as differences in spatial and temporal resolution. These were addressed by employing data preprocessing techniques including spatial averaging and temporal alignment to ensure consistency. Machine learning models further helped bridge the gaps by learning the relationships between the datasets.

2.2.3 HYSPLIT NCEP Reanalysis Data and Trajectory model

The Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) model, developed by the National Oceanic and Atmospheric Administration (NOAA), is used for trajectory analysis in this study[12].The model utilizes meteorological data from the National Centers for Environmental Prediction (NCEP) reanalysis to simulate air parcel trajectories and analyze the transport and dispersion of pollutants. Both forward and backward trajectory analyses are performed to understand the movement of air masses affecting Delhi's air quality. Forward trajectories help in identifying the dispersion patterns of pollutants originating from Delhi, while backward trajectories trace the origins of air masses reaching Delhi, highlighting potential upwind pollution sources. This comprehensive trajectory analysis, combined with the Concentration Weighted Trajectory (CWT) approach, provides insights into both local and regional contributions to $PM_{2.5}$ levels in Delhi.The HYSPLIT model used meteorological data from NCEP reanalysis having a spatial resolution of 2.5 degrees and a temporal resolution of 6 hours. Assumptions included constant emission rates and homogeneous

mixing within the boundary layer. These settings ensured a balance between computational efficiency and accuracy of the trajectory analysis.

2.3 Data Preprocessing, Machine learning model and validation

Outliers in the $PM_{2.5}$ concentration data were identified using the interquartile range (IQR) method. Data points beyond 1.5 times the IQR were considered outliers. These were addressed using robust statistical techniques such as median imputation to minimize their impact on the model performance. Additionally, data augmentation techniques were used to balance the dataset, ensuring the models accurately learned from rare high-concentration events.

To enhance the accuracy of $PM_{2.5}$ concentration estimates, we employed several machine learning models trained on historical data from 2014 to 2022. The models used in this study include Extra Trees Regressor, Random Forest Regressor, Light Gradient Boosting Machine, and a Stacked Regress or which are less sensitive to outliers. The input parameters for these models are the sub-components of $PM_{2.5}$ from the MERRA-2 reanalysis data (dust, organic carbon, black carbon, sulfate, and sea salt), while the target variable is the actual $PM_{2.5}$ concentrations recorded by CPCB[17].

The dataset was split into an 80:20 ratio for training and testing, respectively, to ensure the robustness of the model predictions. We employed k-fold cross-validation (with k=10) to validate the machine learning models. This technique involves partitioning the data into k subsets and using k-1 subsets for training while the remaining subset is used for testing. This process is repeated k times, ensuring each subset is used for testing once. The results are then averaged to provide an overall performance metric, ensuring robustness and reducing the risk of overfitting. The models were evaluated based on several performance metrics which helped in assessing the models' ability to predict $PM_{2.5}$ concentrations accurately.The integration of machine learning models with reanalysis data addresses the inherent discrepancies between MERRA-2 estimates and ground-based measurements, providing more reliable $PM_{2.5}$ concentration estimates[18].

3 Results and Discussion

3.1 Analysis of Spatial Correlation in $PM_{2.5}$ Concentrations Across Delhi

The heatmap presenting the correlation matrix offers an in-depth view of the relationship between daily $PM_{2.5}$ concentrations averaged over different districts of Delhi from 2014 to 2023 (Fig. 2). The color-coded matrix and Pearson correlation coefficients provide a detailed picture of how $PM_{2.5}$ levels in various districts interrelate. Each cell in the matrix represents the Pearson correlation coefficient (r) between the daily $PM_{2.5}$ concentrations of two

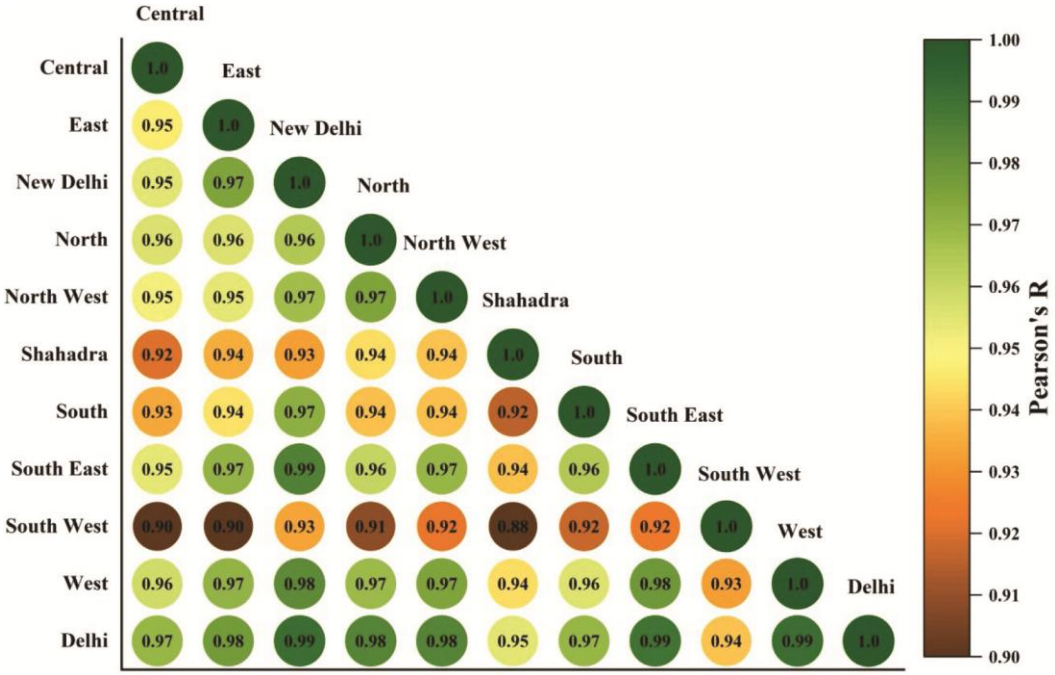

Fig. 2 — Heatmap showing the correlation matrix of daily $PM_{2.5}$ concentrations averaged across different districts of Delhi

districts, with values ranging from 0.88 to 1.0, indicating strong positive correlations.

All districts exhibit very high correlation coefficients (r > 0.90), suggesting that $PM_{2.5}$ concentrations in Delhi are influenced by common factors city-wide, such as vehicular emissions, industrial activities, construction dust, and regional weather patterns. Central and New Delhi show extremely high correlations with each other and other districts (r values ranging from 0.95 to 0.99), likely due to their central location and the concentration of administrative and commercial activities contributing to similar pollution sources. Peripheral districts like South West and Shahdara show slightly lower r values (around 0.88 to 0.92) compared to central districts, possibly due to localized differences in pollution sources, meteorological conditions, or variations in urban density and land use[19]. The row and column representing "Delhi" as a whole show correlation coefficient with individual districts ranging from 0.97 to 0.99, reinforcing the idea that $PM_{2.5}$ levels in individual districts are highly representative of overall air quality trends in the city.The high degree of correlation across districts implies that air quality management policies need to be holistic and city-wide. Strategies like reducing vehicular emissions, controlling industrial pollution, and managing construction dust will likely benefit the entire city due to the interconnected nature of $PM_{2.5}$ levels. The strong correlations suggest that improvements in one district could positively impact neighboring districts, encouraging collaborative efforts across district boundaries.

3.2 Time Series Analysis of $PM_{2.5}$ Concentrations

The time series analysis of $PM_{2.5}$ concentrations in Delhi from 2014 to 2023, compares CPCB ground-based measurements and MERRA-2 reanalysis estimates (Fig. 3). Both panels display daily $PM_{2.5}$ concentrations (blue dots) on the left y-axis and monthly averaged $PM_{2.5}$ concentrations (red line) on the right y-axis. The CPCB data exhibits higher daily variability and peaks compared to MERRA-2, reflecting the sensitivity of ground-based monitors to local emissions and short-term pollution events.

In Fig. 3(a), daily $PM_{2.5}$ levels frequently exceed 400 µg/m^3 during winter months, with some peaks approaching 500 µg/m^3. Conversely, Fig. 3(b) shows MERRA-2 daily $PM_{2.5}$ levels generally staying below 250 µg/m^3. Both datasets reveal a clear seasonal pattern, with $PM_{2.5}$ concentrations peaking during the winter months (November to January), attributed to lower temperatures, reduced wind speeds, and increased emissions from heating sources[20]. The monthly averaged $PM_{2.5}$ concentrations are higher in the CPCB data, often surpassing 200 µg/m^3, whereas MERRA-2 monthly averages typically peak around 150 µg/m^3.

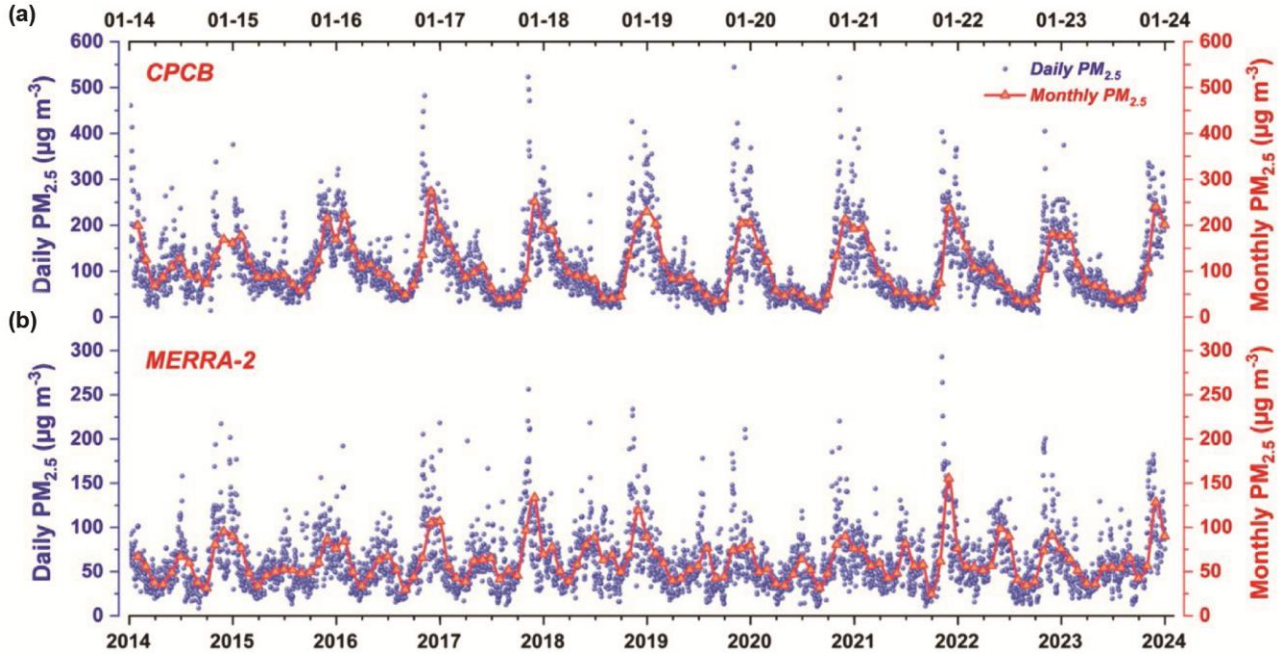

Fig. 3 — Time series analysis of $PM_{2.5}$ concentrations in Delhi comparing (a) ground-based CPCB measurements and (b) MERRA-2 reanalysis estimates

Over the study period, both datasets show an increasing trend in $PM_{2.5}$ concentrations, highlighting the persistent air quality challenges in Delhi. This comparison underscores the reliability of MERRA-2 for understanding broad temporal trends and the necessity of ground-based measurements for capturing acute pollution episodes.

3.3 Monthly Temporal Variability Analysis

The monthly temporal variability of $PM_{2.5}$ concentrations in Delhi was investigated using a violin box plot that compares CPCB's and MERRA-2 reanalysis $PM_{2.5}$ data (Fig. 4). This visualization highlights both central tendencies and variability across months.

A violin box plot combines features of a box plot with a density plot, offering a comprehensive view of data distribution. The shape and width represent the kernel density estimate, showing the probability density at different values. Within each violin, a traditional box plot indicates the interquartile range (IQR), the median (white dot), and whiskers extending to 1.5 times the IQR. For CPCB data, the plots exhibit distinct seasonal patterns, with broader distributions and higher median values during late post-monsoon and winter months. This reflects increased $PM_{2.5}$ levels due to factors such as lower temperatures, biomass burning, and reduced dispersion conditions[21-22]. Greater variability is observed during winter months compared to summer, likely due to episodic pollution events and fluctuations in meteorological conditions. The highest concentrations are observed in December and January, with median values around 200-250 $\mu g/m^3$ and some daily values exceeding 400 $\mu g/m^3$, highlighting severe pollution episodes. The MERRA-2 data also show seasonal trends, with higher median $PM_{2.5}$ concentrations during winter and lower concentrations in summer. However, the overall concentrations are lower compared to CPCB data, indicating a potential underestimation by the reanalysis model. The violin plots for MERRA-2 data are generally narrower, suggesting less variability in the reanalysis estimates compared to ground measurements, likely due to the smoothing effect of the reanalysis process. Despite lower absolute values, the seasonal patterns and trends observed in MERRA-2 data are consistent with those from CPCB measurements.

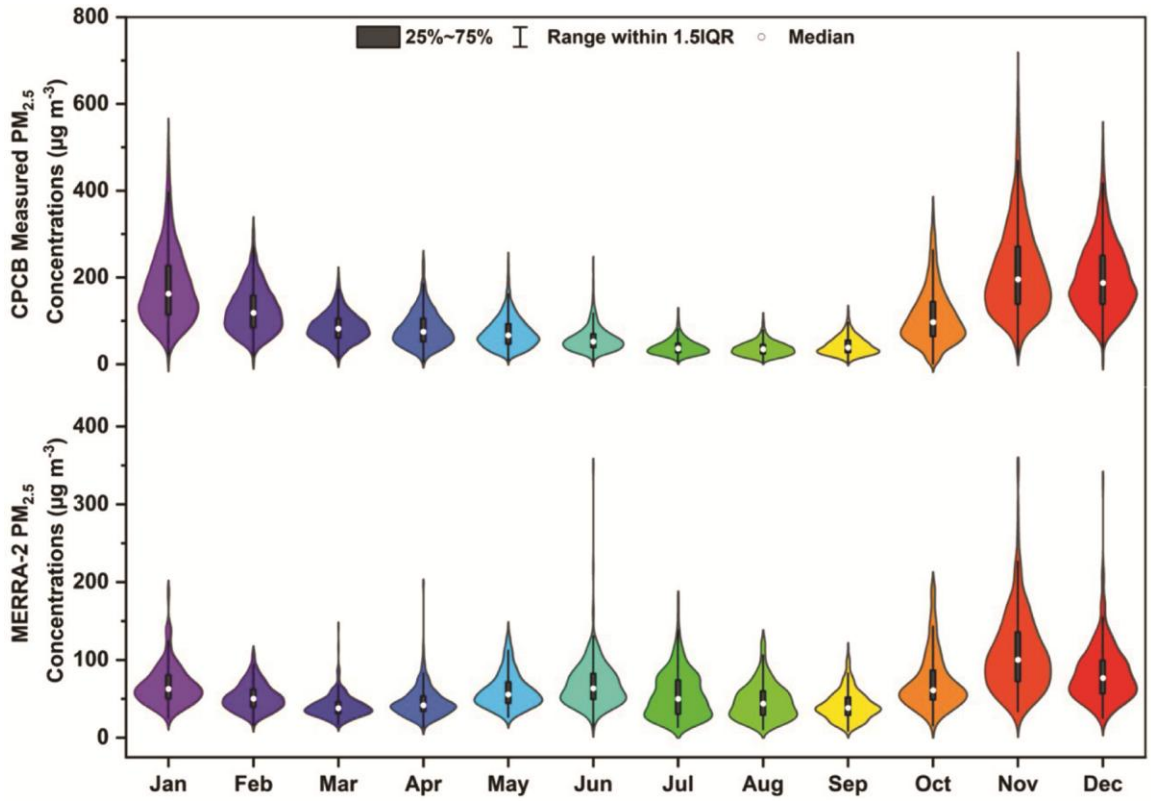

Fig. 4 — Monthly temporal variability of $PM_{2.5}$ concentrations in Delhi using violin box plot comparing CPCB and MERRA-2 reanalysis data

Both datasets confirm significant seasonal variation, with winter months experiencing the highest pollution levels. The general agreement between CPCB and MERRA-2 data in terms of seasonal trends validates the use of MERRA-2 reanalysis for temporal analysis, though ground measurements are crucial for capturing extreme values.

3.4 Seasonal Contribution to $PM_{2.5}$ Concentrations

The seasonal contribution to $PM_{2.5}$ concentrations in Delhi from 2014 to 2023, depicted through a stacked bar chart, reveals significant seasonal variations across different districts (Fig. 5).

Winter (December-February) emerges as the season with the highest contribution to annual $PM_{2.5}$ levels, ranging from 34% in South East Delhi to 40% in East Delhi. This is primarily due to lower temperatures, increased heating, and temperature inversions that trap pollutants near the surface. Biomass burning and heightened vehicular emissions during winter further exacerbate $PM_{2.5}$ levels. Summer (March-June) contributes moderately, between 15% and 21%, with the highest contributions in South East Delhi (21%) and the lowest in Shahadra (15%). During this season, higher temperatures and atmospheric dispersion help dilute pollutants, although dust storms and localized emissions still significantly impact $PM_{2.5}$ concentrations.

Monsoon (July-September) has the lowest contribution, ranging from 8% to 11% across districts, due to the scavenging effect of rainfall that removes particulates from the air. West Delhi shows a slightly higher monsoon contribution (11%), possibly due to local sources or specific meteorological conditions[23]. Post monsoon (October-November) exhibits a significant contribution, between 35% and 38%, driven by agricultural residue burning in neighboring states, increased emissions from vehicles and industries, and festive activities like Diwali. District-wise analysis shows that East Delhi and North West Delhi have the highest winter contributions, indicating severe air quality issues in these areas during winter. South East Delhi's highest summer contribution suggests localized emissions such as construction and traffic. Understanding these seasonal patterns highlights the need for targeted air quality management strategies. During winter, controlling emissions from heating, transportation, and industry, along with enforcing biomass burning regulations, is crucial. Summer strategies should focus on mitigating dust storms and localized emissions. While monsoon requires continuous monitoring despite lower $PM_{2.5}$ levels, post monsoon necessitates interventions to prevent agricultural burning and manage emissions from festive activities. Tailoring policies to each season's challenges can significantly improve air quality in Delhi.

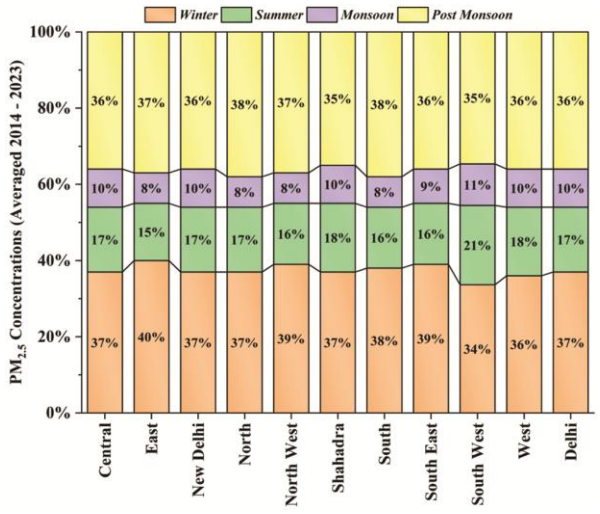

Fig. 5 — Seasonal percentage contribution to $PM_{2.5}$ concentrations across districts in Delhi

3.5 Comparison of $PM_{2.5}$ Measurements: MERRA-2 vs. CPCB Data

The scatter plot reveals the relationship between daily $PM_{2.5}$ concentrations obtained from MERRA-2 reanalysis data and those measured by the Central Pollution Control Board (CPCB) across Delhi (Fig. 6).

Each point on the scatter plot represents a paired daily $PM_{2.5}$ concentration value from both datasets.

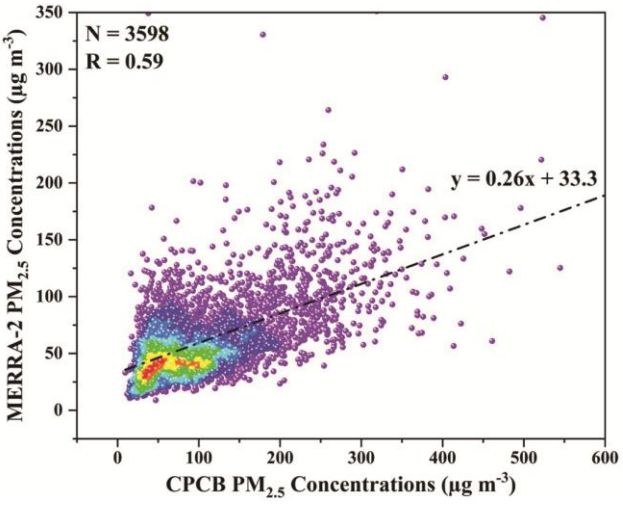

Fig. 6 — Scatter plot showing the relationship between daily $PM_{2.5}$ concentrations from MERRA-2 reanalysis data and CPCB measurements in Delhi

The positive linear relationship, evident from the upward trend in the scatter plot and corroborated by a Pearson correlation coefficient (R) of 0.59, indicates a moderate correlation. The regression line, described by the equation y = 0.26x + 33.3, suggests that for every unit increase in CPCB-measured $PM_{2.5}$, the MERRA-2 data reflects an increase of approximately 0.26 units, starting from a baseline of 33.3 µg/m^3. This relationship underscores the potential of MERRA-2 reanalysis data to approximate ground-based measurements, albeit with some variability, especially at higher concentration levels.

Despite the moderate correlation, significant spread in data points, particularly at elevated $PM_{2.5}$ levels, highlights the variability in how MERRA-2 captures extreme pollution events compared to CPCB observations. The concentration of data points around lower $PM_{2.5}$ values with noticeable scatter as values increase suggests limitations in MERRA-2's accuracy in representing local pollution spikes. This discrepancy could be attributed to the spatial resolution and inherent smoothing of reanalysis data, which may not capture the fine-scale variations detected by ground stations[11]. Outliers indicate instances where MERRA-2 either overestimated or underestimated actual $PM_{2.5}$ concentrations, reflecting challenges in modeling local emission sources and meteorological influences accurately. Despite these limitations, MERRA-2 data is invaluable for providing continuous spatial coverage, essential for historical air quality studies and regions lacking extensive ground monitoring networks. However, integrating reanalysis data with ground-based observations is crucial for enhancing accuracy, emphasizing the complementary nature of both data sources in air quality research.

3.6 Spatial-Temporal Analysis of $PM_{2.5}$ Concentration in Delhi

This section provides a comprehensive analysis of the $PM_{2.5}$ concentration trends in Delhi over the years, examining both yearly and seasonal variations. By utilizing data from multiple monitoring stations across the city, we can observe how pollution control measures and seasonal factors influence air quality (Fig. 7). The spatial maps were generated using Kriging interpolation, a geostatistical method providing best linear unbiased predictions for spatial data. This method considers both the distance and the

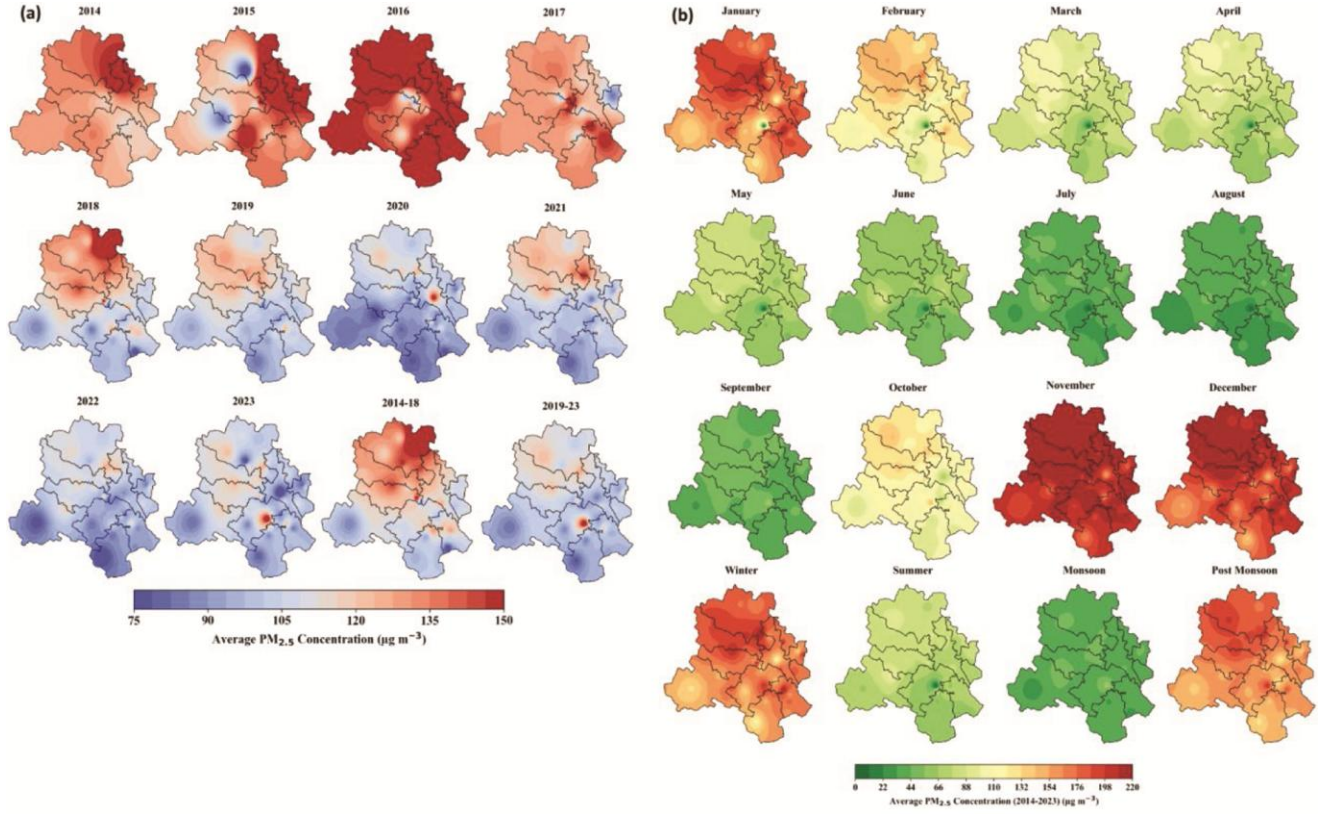

Fig. 7 — Spatial-temporal Analysis of $PM_{2.5}$ concentration in Delhi (a) Yearly variations highlighting trends (b) Monthly and seasonal variations throughout the year

degree of variation between known data points, enhancing the accuracy and reproducibility of the spatial distribution maps.

3.6.1 Yearly Variations

The spatial-temporal analysis of $PM_{2.5}$ concentrations in Delhi from 2014 to 2023 highlights significant trends and the impact of pollution control measures (Fig. 7(a)). Data from 40 Continuous Ambient Air Quality Monitoring Stations (CAAQMS) across Delhi's 11 districts reveal that from 2014 to 2018, $PM_{2.5}$ levels were consistently high, especially in the northern and central regions, with annual averages ranging from 135 to 150 μg/m^3. These levels far exceeded national and international safe limits, indicating severe air quality issues primarily due to vehicular emissions, industrial activities, construction dust, rapid urbanization, and seasonal factors like winter temperature inversions. A notable improvement in $PM_{2.5}$ levels was observed from 2019 onwards, particularly in 2020, due to the COVID-19 lockdown, which significantly reduced human activities. This led to average $PM_{2.5}$ concentrations dropping to 75 to 90 μg/m^3 in several areas. Despite a slight rebound in 2021, levels remained below pre-2019 figures, demonstrating the effectiveness of regulatory measures and increased public awareness. Comparing the periods 2014-2018 and 2019-2023 underscores the success of air quality management strategies such as the Graded Response Action Plan (GRAP), stricter vehicle emission norms, and cleaner technologies[24]. However, some northern regions still showed elevated $PM_{2.5}$ levels, necessitating targeted interventions. Continuous monitoring, timely interventions, and public awareness are crucial for maintaining and improving air quality. Policies promoting electric vehicles, expanding public transportation, and enforcing stringent industrial regulations are essential for long-term success.

3.6.2 Monthly and Seasonal Variations

The monthly and seasonal spatial-temporal analysis of $PM_{2.5}$ concentrations from 2014 to 2023 provides insights into how levels fluctuate throughout the year (Fig. 7(b)) January and February exhibit the highest $PM_{2.5}$ levels, often exceeding 154 μg/m^3, especially in northern and central Delhi, due to cold weather, temperature inversions, increased emissions from heating sources, and lower wind speeds. March and April see a decrease to 44-110 μg/m^3 due to warmer temperatures and higher wind speeds, while May and June show further reductions to 22-88 μg/m^3 thanks to increased atmospheric mixing. The monsoon season (July to September) records the lowest $PM_{2.5}$ levels, typically below 66 μg/m^3, as rains wash out particulate matter. However, levels rise again in October (88-132μg/m^3) due to post-monsoon agricultural burning. November and December see dramatic increases, often exceeding 176 μg/m^3, driven by lower temperatures and crop residue burning, leading to severe air quality deterioration[21]. Winter (December to February) has the highest $PM_{2.5}$ concentrations due to low temperatures, reduced wind speeds, and increased heating emissions. Summer (March to June) shows significant improvement, with levels between 22-88 μg/m^3, while the monsoon season (July to September) has the lowest levels due to frequent rainfall. The post-monsoon period (October to November) sees a sharp increase in $PM_{2.5}$ due to agricultural burning and cooler temperatures.

3.7 Concentration-Weighted Trajectory (CWT) analysis

The CWT for $PM_{2.5}$ in Delhi, from 2014 to 2023, provides insights into pollution sources and transport pathways across seasons using 5-day backward and forward trajectories (Figs. 8 & 9)

3.7.1 Backward Trajectory Analysis

Using NCEP reanalysis data, backward trajectories reveal the origins of air masses contributing to $PM_{2.5}$ levels in Delhi (Fig. 8). During winter, air masses predominantly come from the northwest regions such as Punjab, Haryana, and parts of Pakistan. These regions contribute to high $PM_{2.5}$ levels, often exceeding 200 μg/m^3, due to industrial emissions, urban pollution, and agricultural burning. Clustering of trajectories shows main pathways from the northwest (27.94%), west (32.93%), north (8.43%), and mixed (30.71%). In summer, air masses are more dispersed, including those from northwest and central India, with $PM_{2.5}$ levels up to 100 μg/m^3, benefiting from better atmospheric dispersion. Key sources include dust storms from the Thar Desert and emissions from industrial and urban areas, with clusters from northwest (31.23%), west (14.67%), central (13.61%), and mixed (40.49%)[23].During the monsoon season, southwest monsoon winds result in the lowest $PM_{2.5}$ levels, below 80 μg/m^3, due to the cleansing effect of rainfall. Air masses originate mainly from the Arabian Sea and Western Ghats, with minimal anthropogenic emissions. Clusters for this period are southwest (20.11%), west (26.96%), south (29.13%), and mixed (23.80%). In the post-monsoon season, $PM_{2.5}$ levels rise again, exceeding 200 μg/m^3,

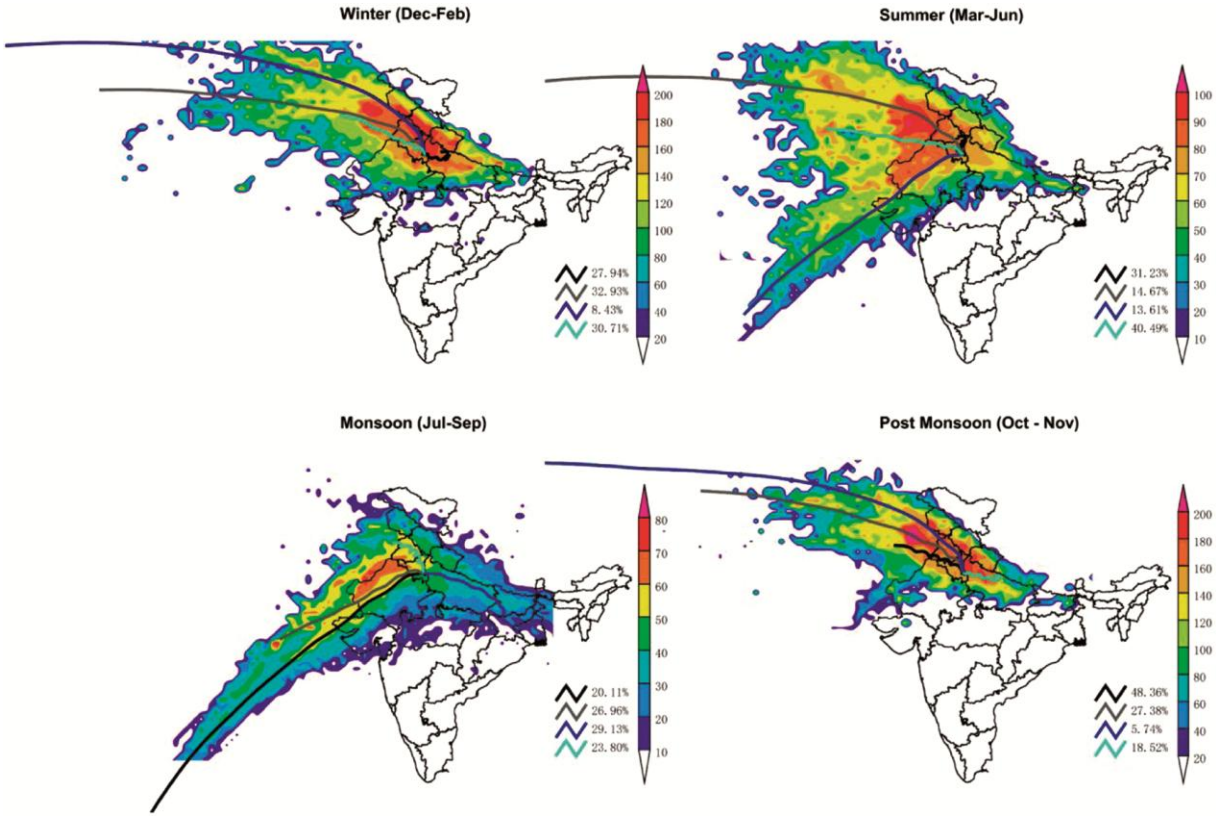

Fig. 8 — Backward Trajectory and CWT analysis showing seasonal origins and cluster pathways of air masses contributing to $PM_{2.5}$ levels in Delhi

with air masses coming from northwest and northern regions, primarily due to agricultural burning in Punjab and Haryana. The clusters are northwest (48.36%), west (27.38%), north (5.74%), and mixed (18.52%).

3.7.2 Forward Trajectory Analysis

Forward trajectories provide insight into how pollutants from Delhi disperse and affect air quality in other regions (Fig. 9).

In winter, pollutants primarily disperse towards the east and southeast, impacting Uttar Pradesh, Bihar, and West Bengal, with $PM_{2.5}$ concentrations often exceeding 200 µg/m^3. The predominant pathways are northwest (31.49%), west (6.65%), north (29.05%), and mixed (32.82%)[24]. During summer, dispersion is more widespread towards the east, northeast, and central India, with lower $PM_{2.5}$ concentrations up to 100 µg/m^3 due to increased atmospheric mixing and higher temperatures. Key pathways are northwest (27.61%), west (28.39%), central (11.24%), and mixed (32.72%). In the monsoon season, pollutants are dispersed mainly towards the northeast and east, driven by monsoon winds, with $PM_{2.5}$ concentrations below 70 µg/m^3 due to heavy rains. Clusters are northeast (44.74%), east (30.54%), southeast (22.37%), and mixed (2.35%). In the post-monsoon season, high $PM_{2.5}$ concentrations again exceed 200 µg/m^3, with significant dispersion towards the east and southeast, influenced by agricultural burning and urban emissions. Key pathways are northeast (52.55%), east (8.87%), southeast (25.94%), and mixed (12.64%).

Both backward and forward CWT analyses highlight the importance of understanding seasonal variations in $PM_{2.5}$ pollution sources and dispersion patterns.

3.8 $PM_{2.5}$ Estimation using Machine Learning (ML) approach

To improve the accuracy ofestimation of $PM_{2.5}$ concentrations, a robust approach integrating machine learning techniques was adopted. This section discusses the methodology and results of applying various machine learning models to enhance the correlation between MERRA-2 reanalysis data and CPCB ground measurements (Fig. 10)

3.8.1 Training of ML Models

To address the low correlation between MERRA-2 reanalysis $PM_{2.5}$ data and the actual CPCB ground

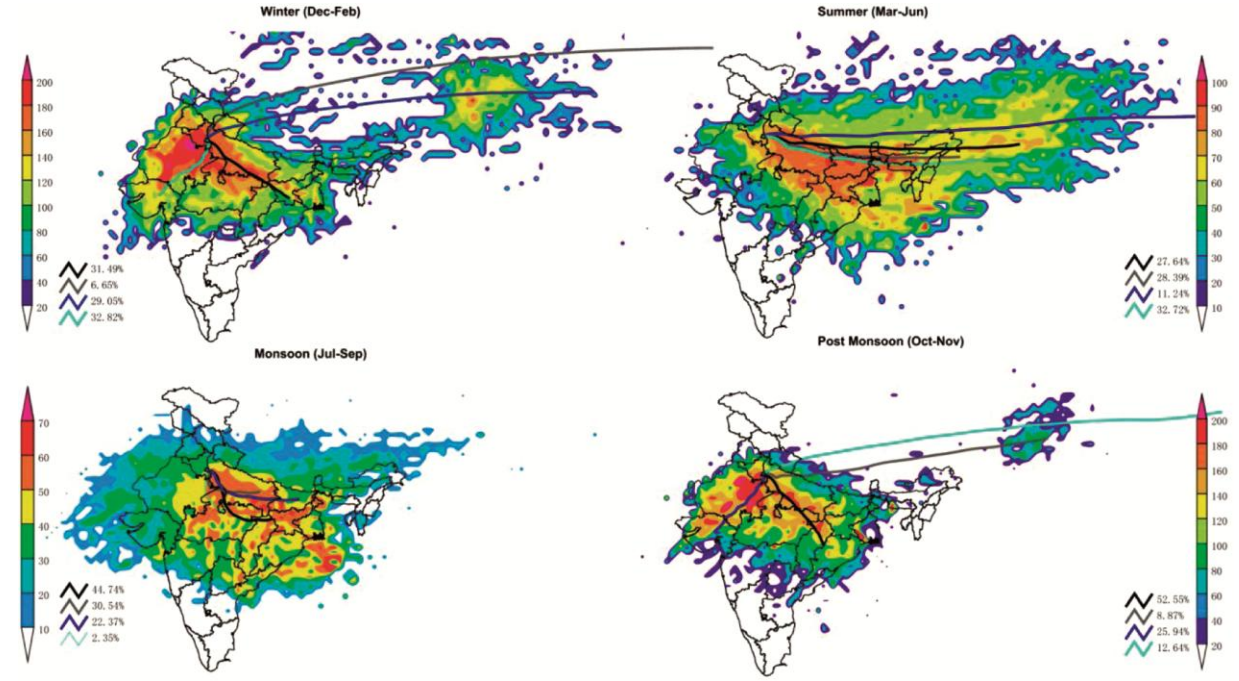

Fig. 9 — Forward Trajectory and CWT analysis illustrating seasonal dispersion patterns and cluster pathways of pollutants from Delhi

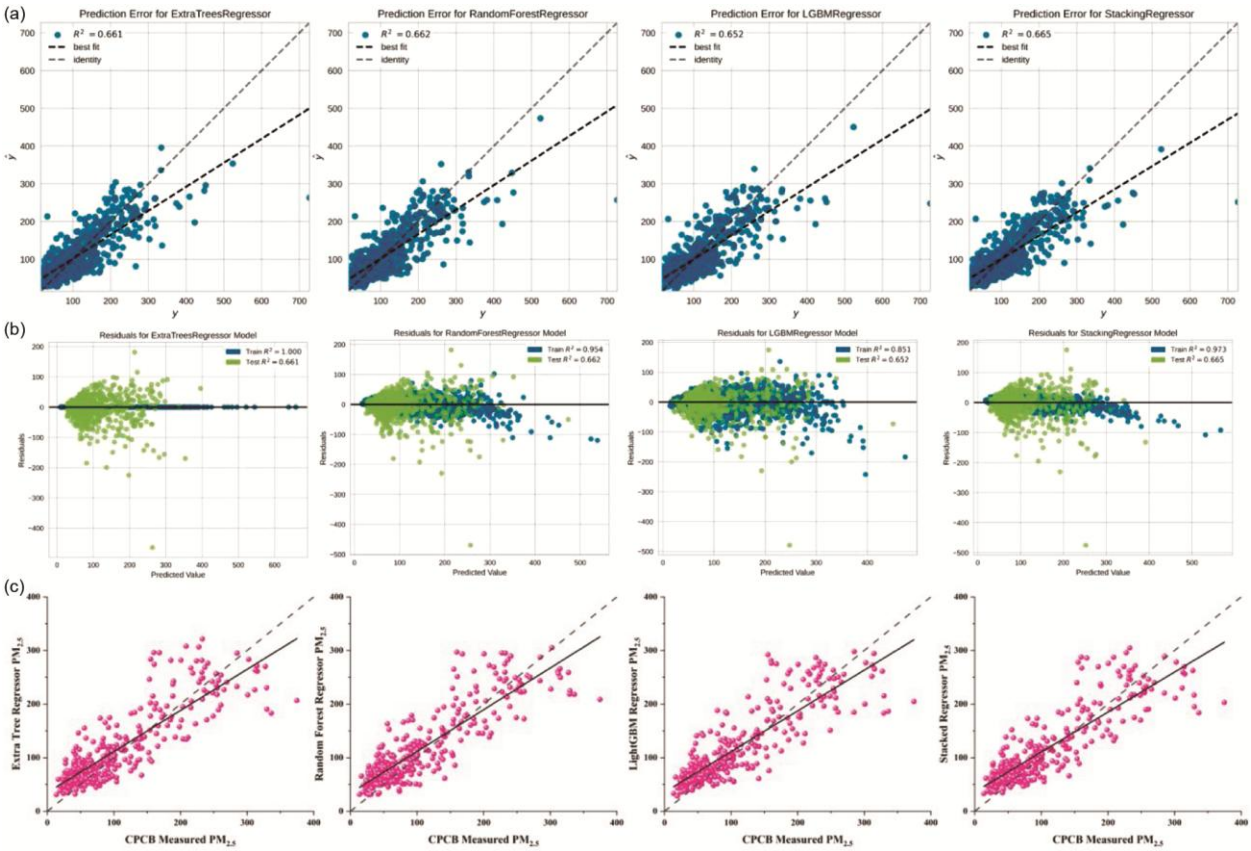

Fig. 10 — $PM_{2.5}$ Estimation using Machine Learning (ML) Approach. (a-b) Prediction error and residual plots during the training stage (c) Scatter plots between machine learning model estimates and CPCB measurements from the trained models

measurements, various machine learning techniques were employed. The MERRA-2 dataset includes sub-parameters of $PM_{2.5}$ such as Dust, Organic Carbon, Black Carbon, Sea Salt and Sulfate. These components are used to calculate the total $PM_{2.5}$ using Equation (1). The goal was to enhance the prediction accuracy of $PM_{2.5}$ concentrations for the year 2023 by training machine learning models with data from 2014 to 2022, split into an 80:20 ratio for training and testing

Four regression models have been trained: Extra Trees Regressor, Random Forest Regressor, Light Gradient Boosting Machine (LGBM) Regressor, and a Stacking Regressor that integrates the predictions of the other models. The prediction error plots display observed versus predicted $PM_{2.5}$ concentrations for each model. The R^2 values for the test set are as follows: Extra Trees (0.661), Random Forest (0.662), LGBM (0.652), and Stacking (0.665), showing a moderate level of predictive accuracy, with the Stacking Regressor performing marginally better. During training, the R^2 values were significantly higher: Extra Trees (1.0), Random Forest (0.954), LGBM (0.851), and Stacking (0.973), indicating strong fitting to the training data (Fig. 10(a)).

Residual plots provide further insights into model performance, showing differences between observed and predicted values (Fig. 10(b)). Extra Trees Regressor residuals are uniformly distributed around zero but spread increases with higher values, indicating reduced accuracy for extreme concentrations. The Random Forest Regressor shows a similar pattern with slightly less variability. LGBM Regressor residuals cluster tightly around zero for lower values but exhibit some overestimation and underestimation at extremes. The Stacking Regressor has the most balanced residual distribution, with reduced spread and bias across the range, suggesting ensemble learning effectively mitigates individual model weaknesses.

However, moderate R^2 values suggest significant variability in $PM_{2.5}$ concentrations remains unexplained due to complex pollution dynamics influenced by local emissions, weather conditions, and regional pollutant transport. The models struggle with high $PM_{2.5}$ concentrations, highlighting challenges in predicting extreme pollution events.

Historical data from 2014 to 2023 reveal distinct seasonal $PM_{2.5}$ patterns in Delhi. Winter months consistently show the highest pollution levels, influenced by lower temperatures, increased heating emissions, and poor dispersion conditions. The monsoon season has the lowest levels, benefiting from wet deposition and better dispersion due to rainfall. These patterns underscore the need for models that adjust to seasonal and meteorological variations to improve predictive accuracy.

3.8.2 Estimation of $PM_{2.5}$ concentrations

The machine learning models demonstrated a significant improvement over the original MERRA-2 reanalysis data[25]. This enhancement is evidenced by the increased correlation with CPCB measured $PM_{2.5}$ concentrations, as shown in the scatter plots (Fig. 10(c)). The machine learning models, including Extra Trees Regressor, Random Forest Regressor, Light Gradient Boosting Machine, and Stacked Regressor, were able to capture the variability in $PM_{2.5}$ concentrations more effectively. For instance, the Correlation Coefficient improved from 0.694 for MERRA-2 to values above 0.86 for all machine learning models, indicating a stronger linear relationship between the predicted and observed values.

Table 2 compares the performance of different estimation models in estimating $PM_{2.5}$ concentrations based on CPCB and MERRA-2 reanalysis data. Additionally, the Fraction of predictions within a factor of 2 (FAC2), which is a critical metric for model accuracy, increased significantly from 0.611 for MERRA-2 to over 0.89 for the machine learning models, with the Random Forest Regressor achieving the highest at 0.907. This improvement suggests that the machine learning models can predict $PM_{2.5}$ concentrations with much greater precision. The Mean Bias (MB) was also substantially reduced, from -39.4 $\mu g/m^3$ for MERRA-2 to around 11 $\mu g/m^3$ for the

Table 2 — Statistical performance matrix of the $PM_{2.5}$ estimating machine learning models

Estimating Models	n	FAC2	MB	MGE	RMSE	r	COE	p-value
Linear MERRA-2 Reanalysis	365	0.611	-39.4	51.9	71.1	0.694	0.148	1.20E-53
Extra Trees Regressor	365	0.899	11.3	30.3	39.9	0.866	0.503	2.18E-111
Random Forest Regressor	365	0.907	11.5	29.2	38	0.881	0.52	8.29E-120
Light Gradient Boosting Machine	365	0.89	11.2	29.5	39.5	0.869	0.516	4.37E-113
Stacked Regressor	365	0.893	10.4	29.8	39	0.871	0.511	3.06E-114

machine learning models, highlighting a significant reduction in systematic error.Furthermore, the Mean Gross Error and Root Mean Squared Error (RMSE) were considerably lower in the machine learning models compared to the MERRA-2 reanalysis. The Mean Gross Error (MGE) decreased from 51.9 $\mu g/m^3$ to approximately 30 $\mu g/m^3$, and the RMSE dropped from 71.1 $\mu g/m^3$ to below 40 $\mu g/m^3$. These reductions underscore the machine learning models' enhanced capability to minimize the deviations between predicted and actual $PM_{2.5}$ concentrations. The Coefficient of Efficiency (COE) also saw a notable improvement, increasing from 0.148 for MERRA-2 to over 0.5 for the machine learning models, demonstrating their superior performance in capturing the data's variability.

In summary, integrating machine learning models with MERRA-2 sub-parameters substantially improves the accuracy of $PM_{2.5}$ estimations. These models offer a more reliable prediction of air quality by effectively correlating with the actual ground measurements from CPCB.

4 Conclusion

The comprehensive study of $PM_{2.5}$ concentrations in Delhi from 2014 to 2023 provides valuable insights into the spatial and temporal patterns of air pollution in the region. By integrating data from the Central Pollution Control Board (CPCB) with MERRA-2 reanalysis data, the study highlights the pervasive influence of common pollution sources across the city's districts and underscores the importance of a holistic approach to air quality management. The key conclusions are as follows:

a) Strong Spatial Correlation: The analysis revealed strong positive correlations ($r > 0.90$) between $PM_{2.5}$ concentrations across all districts in Delhi. This indicates that air pollution in the city is driven by common sources such as vehicular emissions, industrial activities, and construction dust. The interconnected nature of $PM_{2.5}$ levels suggests that improvements in one district could positively impact neighboring districts, emphasizing the need for city-wide air quality management policies.

b) Seasonal Variations: The study found significant seasonal variations in $PM_{2.5}$ concentrations, with peaks during the winter months (November to January) due to lower temperatures, reduced wind speeds, and increased emissions from heating sources. The monsoon season (July to September) recorded the lowest levels, benefiting from the scavenging effect of rainfall. Post-monsoon (October to November) saw increased $PM_{2.5}$ levels due to agricultural residue burning in neighboring states and festive activities.

c) Machine Learning Enhancements: The application of machine learning (ML) models significantly improved the prediction accuracy of $PM_{2.5}$ concentrations. Among the models tested, the Stacking Regressor performed the best, achieving an R^2 value of 0.665 and an improved correlation with CPCB measurements ($r = 0.86$). This demonstrates the potential of ML models to enhance air quality predictions, offering a more accurate tool for understanding and managing air pollution.

d) Policy Implications and Management Strategies: The high degree of correlation across districts highlights the need for comprehensive, city-wide air quality management strategies. Effective measures include reducing vehicular emissions, controlling industrial pollution, and managing construction dust. Seasonal strategies are also crucial, such as controlling emissions from heating and transportation during winter, mitigating dust storms and localized emissions during summer, and preventing agricultural burning and managing emissions from festive activities during the post-monsoon period.

e) Limitation: While the CPCB data provides high accuracy, it is limited by its spatial coverage. Similarly, MERRA-2 reanalysis data, despite its extensive coverage, might not fully capture local emission sources and short-term pollution events. These limitations could introduce potential biases and gaps, necessitating the use of machine learning models to enhance data reliability.

In conclusion, this study emphasizes the importance of integrating advanced machine learning models with traditional monitoring and reanalysis data to improve the accuracy of air quality predictions. The enhanced $PM_{2.5}$ estimations using ML models provide a robust tool for policymakers to develop effective and targeted air quality management strategies. Continuous monitoring, timely interventions, and public awareness are essential to maintaining and improving air quality in Delhi. Future research should focus on further refining these models and exploring additional factors influencing air pollution to achieve even greater accuracy and effectiveness in air quality management.

Acknowledgements

The authors are grateful to the Central Pollution Control Board (CPCB) for providing comprehensive ground measurement data on $PM_{2.5}$ concentrations in Delhi, essential for benchmarking analyses and validating models. Thanks are extended to the Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), and the National Centers for Environmental Prediction (NCEP) for their high-quality reanalysis data, crucial for understanding $PM_{2.5}$ variability and conducting trajectory analyses. Appreciation is also extended to all individuals and teams involved in maintaining and disseminating these datasets, whose contributions were indispensable to our research. The authors express sincere gratitude to the Director of IITM, Pune for his invaluable encouragement and support throughout this study.

References

1 Brook R D, Rajagopalan S, Pope III C A, Brook J R, Bhatnagar A, Diez-Roux AV, Holguin F, Hong Y, Luepker R V, Mittleman M A & Peters A, *Circulation*, 121 (2010) 2331.
2 Coleman N C, Burnett R T, Ezzati M, Marshall J D, Robinson A L & Pope C A, *Environ Health Perspect*, 128 (2020) 107004.
3 Li C Y, Wu C D, Pan W C, Chen Y C & Su H J, *Epidemiology*, 30 (2019) S67.
4 Yang X, Zhang T, Zhang Y, Chen H & Sang S, *Sci Total Environ*, 796 (2021) 148819.
5 Nagpure A S, Gurjar B R & Martel J C, *Atmos Pollut Res*, 5 (2014) 371.
6 Guttikunda S K & Goel R, *Environ Dev*, 6 (2013) 8.
7 Lakshmanan S, Upadhayay A, Kumar N & Bhattacharya S, *Sci Total Environ*, 900 (2023) 165838.
8 Zhang X, Han L, Wei H, Tan X, Zhou W, Li W & Qian Y, *J Clean Prod*, 346 (2022) 130988.
9 Guttikunda S K & Jawahar P, *Atmos Environ*, 92 (2014) 449.
10 Gelaro R, McCarty W, Suárez M J, Todling R, Molod A, Takacs L, Randles C A, Darmenov A, Bosilovich M G, Reichle R & Wargan K, *J Clim*, 30 (2017) 5419.
11 Spandana B, Srinivasa Rao S, Upadhya A R, Kulkarni P & Sreekanth V, *Adv Sp Res*, 67 (2021) 3134.
12 Saha S, Moorthi S, Pan H L, Wu X, Wang J, Nadiga S, Tripp P, Kistler R, Woollen J, Behringer D & Liu H, *Bull Am Meteorol Soc*, 91 (2010) 1015.
13 Dimitriou K, Remoundaki E, Mantas E & Kassomenos P, *Atmos Environ*, 116 (2015) 138.
14 Shahfahad, Bindajam A A, Naikoo M W, Talukdar S, Asif, Mallick J & Rahman A, *Nat Hazards*, Jul (2023) 1.
15 Singh A, Srivastava A K, Varaprasad V, Kumar S, Pathak V & Shukla A K, *Meteorol Atmos Phys*, 133 (2021) 1127.
16 Kumar A, Singh S, Singh A, Srivastava A K & Pathak V, *Indian J Pure Appl Phys*, 62 (2024) 350.
17 Xiao Q, Chang H H, Geng G & Liu Y, *Environ Sci Technol*, 52 (2018) 13260.
18 Mandal S, Madhipatla K K, Guttikunda S, Kloog I, Prabhakaran D & Schwartz J D, *Atmos Environ*, 224 (2020) 117309.
19 Pant P, Habib G, Marshall J D & Peltier R E, *Environ Res*, 156 (2017) 167.
20 Sharma S, Khare M & Kota S H, *Aerosol Air Qual Res*, 22 (2022) 210377.
21 Singh A, Srivastava A K, Pathak V & Shukla A K, *Atmos Environ*, 270 (2022) 118893.
22 Kumar S, Singh A, Srivastava A K, Sahu S K, Hooda R K, Dumka U C & Pathak V, *Urban Clim*, 38 (2021) 100880.
23 Sarkar S, Chauhan A, Kumar R & Singh R P, *Geo Health*, 3 (2019) 67.
24 Ghosh S, Biswas J, Guttikunda S, Roychowdhury S & Nayak M, *J Air Waste Manage Assoc*, 65 (2015) 218.
25 Dhandapani A, Iqbal J & Kumar R N, *Chemosphere*, 340 (2023) 139966.