International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Spatio-Temporal Dynamics of PM2.5 Pollution in Delhi: A Multi-Pollutant, Meteorological, and Machine Learning Analysis at Dwarka Sector Monitoring Station

P. Srivyshnavi[1], Sreenivasulu T[2] ,P Shankaraiah [3*], M Darshan Teja [4]
[1]Assistant Professor , Dept. of CS & E, S.P.M.V.V. Engineering College, Tirupati, India.
[2, 3, 4] Department of Mathematics, School of Advance Science, VIT Vellore, India.
Corresponding author: stats4shankar@gmail.com *

Abstract

Fine particulate matter (PM2.5) remains among the most pressing environmental health challenges confronting rapidly urbanising megacities in South Asia. This study presents a comprehensive quantitative analysis of PM2.5 concentrations recorded at the Dwarka Sector Continuous Ambient Air Quality Monitoring (CAAQM) station, Delhi, covering the period January 2019 to December 2023 a complete five-year daily record of 1,826 observations spanning nineteen pollutant and meteorological variables. Employing descriptive statistics, Pearson correlation analysis, and a comparative evaluation of three machine learning models Multiple Linear Regression (MLR), Random Forest Regressor (RFR), and Gradient Boosting Regressor (GBR) under five-fold time series cross-validation, the study characterises the temporal structure, correlate profile, and predictive modelling potential of PM2.5 at this station. The five-year mean PM2.5 concentration of 104.53 µg/m³ substantially exceeds both India's National Ambient Air Quality Standard (NAAQS: 60 µg/m³) and the World Health Organization annual guideline (5 µg/m³). Pronounced seasonal heterogeneity is observed, with winter concentrations (mean: 177.35 µg/m³) approximately 4.5 times those recorded during summer (39.16 µg/m³). Strong positive correlations with PM10 ($r = 0.845$), CO ($r = 0.739$), and NOx ($r = 0.730$), alongside significant negative correlations with ambient temperature ($r = -0.612$), solar radiation ($r = -0.575$), and wind speed ($r = -0.574$), confirm the dominant role of boundary layer dynamics and shared combustion emission sources. The Gradient Boosting Regressor achieves superior predictive performance (RMSE = 35.10 µg/m³, MAE = 22.38 µg/m³, $R^2 = 0.841$) compared to Random Forest ($R^2 = 0.815$) and Linear Regression ($R^2 = 0.772$) under time series cross-validation. PM10 emerges as the dominant predictive feature (MDI importance = 0.731), followed by ambient temperature (0.084) and relative humidity (0.068). These findings carry significant implications for air quality management policy, seasonal early-warning system design, and the evaluation of India's National Clean Air Programme emission reduction targets.

Keywords: *PM2.5 forecasting; urban air quality; Delhi; Dwarka Sector; Gradient Boosting; Random Forest; machine learning; time series*

1. Introduction

Ambient air pollution is consistently identified as the leading environmental contributor to the global burden of disease, responsible for an estimated 6.7 million premature deaths annually when outdoor and indoor sources are combined (Murray et al., 2020). Within the spectrum of atmospheric pollutants, fine particulate matter with an aerodynamic diameter of 2.5 micrometres or less commonly denoted PM2.5 occupies a position of particular public health significance. Owing to its small size, PM2.5 bypasses the upper respiratory tract's mucociliary defence mechanisms and penetrates deep into the pulmonary alveoli, from where

it may translocate into the systemic circulation, triggering inflammatory cascades implicated in cardiovascular disease, stroke, chronic obstructive pulmonary disease, lung cancer, and neurodegenerative conditions (Brook et al., 2010; Pope et al., 2002).

India bears a disproportionate share of the global PM2.5 burden. According to the State of Global Air 2023 report, fourteen of the twenty most polluted cities worldwide by annual PM2.5 concentration are located in India, with the Indo-Gangetic Plain constituting the country's most severely affected region (Health Effects Institute, 2023). Delhi, the National Capital Territory, consistently registers among the worst annual averages in the world. Multiple anthropogenic source categories converge in this megacity: vehicular exhaust from one of the world's largest registered vehicle fleets, industrial and power generation emissions, construction dust, solid waste burning, and the seasonal long-range transport of agricultural residue burning aerosols from Punjab and Haryana during October and November (Guttikunda & Gurjar, 2012; Sharma et al., 2016).

Dwarka Sector, located in south-western Delhi, is a densely populated planned residential district with an estimated resident population exceeding 1.5 million. The neighbourhood is traversed by the high-traffic Urban Extension Road corridor and lies beneath the flight approach path to Indira Gandhi International Airport, rendering it a site of multiple overlapping emission influences. The Central Pollution Control Board (CPCB) maintains a Continuous Ambient Air Quality Monitoring station at Dwarka Sector under the national CAAQM network, providing real-time measurements of criteria pollutants, volatile organic compounds, and surface meteorological parameters. Despite the station's five-year continuous operational record, no peer-reviewed integrated analysis of its full multi-year dataset has previously been published.

Machine learning methods have gained substantial traction in air quality modelling over the past decade, offering the capacity to capture non-linear relationships and interaction effects that elude conventional linear regression frameworks. Ensemble tree-based algorithms particularly Random Forest and Gradient Boosting have demonstrated consistently strong performance in PM2.5 prediction tasks across diverse urban environments, frequently outperforming physics-based chemical transport models when applied to local-scale monitoring data (Chen et al., 2018; Breiman, 2001). Concurrently, a growing recognition of the importance of methodologically rigorous validation specifically, the requirement for time-ordered cross-validation to prevent data leakage in autocorrelated environmental time series has sharpened the standards applied to reported model performance metrics (Roberts et al., 2017).

This study addresses three principal research gaps: the absence of a multi-year integrated analysis specific to Dwarka Sector; the scarcity of comparative machine learning evaluations using proper time series cross-validation in the Indian urban air quality literature; and the limited availability of physically interpretable feature importance analyses linking model predictions to atmospheric mechanisms. The study characterises temporal PM2.5 dynamics, quantifies the pollutant–meteorological correlate structure, compares three machine learning models under rigorous validation, and derives actionable insights relevant to air quality management and public health policy in Delhi.

2. Literature Review

2.1 PM2.5 Sources and Atmospheric Dynamics in Delhi

The source attribution of PM2.5 in Delhi has been examined through multiple analytical lenses. Guttikunda and Gurjar (2012) employed emission inventory and receptor modelling

approaches to identify vehicular transport, domestic biomass combustion, and road dust resuspension as the three largest contributors to Delhi's annual PM2.5 loading, collectively accounting for approximately 65% of total mass concentration. Their work established the foundational source profile against which subsequent studies have been calibrated. Sharma et al. (2016) demonstrated through satellite-based fire radiative power data and HYSPLIT trajectory analysis that agricultural stubble burning in Punjab and Haryana contributes between 20% and 40% of daily PM2.5 mass at Delhi monitoring sites during the critical October–November window, with episodic contributions exceeding 50% on individual high-pollution days.

Meteorological controls on near-surface PM2.5 have received detailed treatment in the literature. Seinfeld and Pandis (2016) provide the canonical theoretical treatment of how planetary boundary layer (PBL) dynamics govern the vertical dilution of surface-level emissions; reduced mixing layer heights during nocturnal and winter conditions confine pollutants within a shallow volume, dramatically amplifying surface concentrations. Kumar et al. (2021) applied ERA5 reanalysis data to quantify the contribution of PBL height variability to seasonal PM2.5 fluctuations across the Indo-Gangetic Plain, finding that PBL height accounted for 38–45% of explained variance in winter. Analysing the role of synoptic weather patterns, Dey and Di Girolamo (2010) linked anticyclonic circulation anomalies over north-western India to elevated aerosol optical depth retrievals from MISR satellite data, providing large-scale observational support for the mechanism identified in station-level correlation studies.

2.2 Machine Learning Applications in Air Quality Modelling

The application of supervised machine learning to PM2.5 prediction has expanded substantially since approximately 2015, driven by the growing availability of high-frequency monitoring data and advances in ensemble learning theory. Breiman's (2001) seminal paper establishing the Random Forest framework provided the algorithmic foundation; Friedman's (2001) development of Gradient Boosting Machines offered a complementary boosting-based approach. Chen and Guestrin (2016) subsequently introduced XGBoost, an optimised gradient boosting implementation that has achieved state-of-the-art results across numerous environmental prediction benchmarks.

In the specific context of PM2.5 prediction, Chen et al. (2018) conducted a systematic comparison of Random Forest, Support Vector Regression, Artificial Neural Networks, and gradient boosting methods across multiple Chinese monitoring stations, finding that ensemble tree methods consistently achieved R^2 values of 0.75–0.89 when trained on co-pollutant and meteorological features. Masood and Ahmad (2021) applied Long Short-Term Memory (LSTM) networks to Delhi's monitoring data and achieved mean absolute percentage errors below 12% for 24-hour-ahead predictions, highlighting the additional gain available from explicitly modelling temporal autocorrelation. Brokamp et al. (2018) demonstrated that Random Forest, when properly validated using spatial and temporal hold-out designs, provides reliable PM2.5 predictions suitable for epidemiological exposure assessment an important practical application motivating the modelling exercise in the present study.

A persistent methodological concern in the machine learning air quality literature is the use of random train–test splits, which allow information from future time periods to contaminate training data a form of data leakage that inflates reported performance metrics. Roberts et al. (2017) provided a rigorous statistical treatment of this issue, demonstrating that spatially and temporally blocked cross-validation yields more conservative and reliable performance estimates than random k-fold procedures. The present study adheres strictly to

time-ordered cross-validation throughout, ensuring reported metrics genuinely reflect out-of-sample predictive skill.

2.3 Feature Importance and Interpretability

Beyond predictive accuracy, the interpretability of machine learning models has emerged as a key requirement in environmental policy contexts. Lundberg and Lee (2017) introduced SHAP (SHapley Additive exPlanations), a game-theoretic approach to attributing model predictions to individual features consistently and locally. Applied to air quality models, SHAP analysis has revealed that temperature, wind speed, and boundary layer proxies consistently rank among the highest-importance predictors in diverse urban environments (Zhang et al., 2022). The Random Forest's Mean Decrease in Impurity (MDI) importance measure, while subject to known biases toward high-cardinality continuous features, provides a computationally accessible and broadly consistent ranking of predictive contributors when all features are on continuous scales as is the case in the present dataset (Strobl et al., 2007).

3. Methodology

3.1 Study Area and Data

The analysis is based on daily averaged ambient monitoring data from the Dwarka Sector CAAQM station, New Delhi (approximate coordinates: 28.59°N, 77.04°E), for the period 1 January 2019 to 31 December 2023, yielding a complete dataset of 1,826 observations. The station is classified as an urban background site and is operated by the CPCB under the National Ambient Air Quality Monitoring Programme. Instruments include Beta Attenuation Mass monitors for PM2.5 and PM10, chemiluminescence analysers for NOx species, pulsed fluorescence analysers for SO2, non-dispersive infrared for CO, and an automatic weather station recording ambient temperature (AT), relative humidity (RH), wind speed (WS), wind direction (WD), rainfall (RF), total cumulative rainfall (TOT-RF), solar radiation (SR), and barometric pressure (BP). Volatile organic compounds Benzene and Toluene are measured by online gas chromatography. Sporadic missing values arising from scheduled instrument maintenance were addressed through linear interpolation constrained to gaps not exceeding three consecutive days, consistent with CPCB data quality guidelines. The resulting dataset is complete with zero residual missing values across all nineteen variables.

Table 1: Descriptive Statistics of Study Variables Dwarka Sector CAAQM Station (n = 1,826; 2019–2023)

Variable	Mean	Std Dev	Min	Max	Unit
PM2.5	104.53	86.38	8.09	600.40	µg/m³
PM10	248.57	136.05	14.63	807.89	µg/m³
NO	39.51	45.73	1.84	360.53	µg/m³
NO2	34.88	19.21	6.05	127.53	µg/m³
NOx	60.86	48.50	7.52	335.49	µg/m³
NH3	45.91	21.85	3.06	194.90	µg/m³
SO2	9.26	8.67	0.38	47.17	µg/m³
CO	1.26	0.69	0.07	5.99	mg/m³
Ozone	27.90	18.30	1.14	109.97	µg/m³
Benzene	2.60	2.24	0.00	13.66	µg/m³
Toluene	14.54	16.28	0.00	132.22	µg/m³
Ambient Temp (AT)	25.52	7.51	6.14	38.58	°C
Rel. Humidity (RH)	59.24	14.59	0.25	97.69	%

Wind Speed (WS)	0.89	0.29	0.32	2.28	m/s
Solar Radiation (SR)	112.71	58.79	10.85	279.53	W/m²
Barometric Pressure (BP)	982.39	7.00	960.08	997.50	hPa

Source: CPCB CAAQM Network, Dwarka Sector, Delhi. All statistics computed from daily arithmetic means.

3.2 Statistical Analysis

Descriptive statistics mean, standard deviation, minimum, and maximum were computed for all nineteen variables. Temporal aggregation was performed at monthly, seasonal, and annual resolutions. Four meteorological seasons were defined following the India Meteorological Department convention adapted for air quality purposes: Winter (December–February), Spring (March–May), Summer (June–August), and Autumn (September–November). Pearson product-moment correlation coefficients were computed across the full variable matrix to quantify pairwise linear associations, with statistical significance assessed at the $p < 0.01$ level. Distributional properties were examined through box-and-whisker plots stratified by season and calendar year.

3.3 Machine Learning Models

Three regression algorithms were implemented and compared. Multiple Linear Regression (MLR) assumes a linear additive relationship between PM2.5 and the predictor vector, estimated by Ordinary Least Squares. It serves as the parametric baseline, enabling quantification of the performance gain attributable to non-linear modelling. The Random Forest Regressor (RFR), introduced by Breiman (2001), constructs an ensemble of 100 decision trees through bootstrap aggregation, averaging predictions to reduce variance. Feature randomisation at each split (max_features = "sqrt") decorrelates individual trees, improving ensemble generalisation. The Gradient Boosting Regressor (GBR), based on Friedman's (2001) framework, constructs trees sequentially, each correcting the residual errors of its predecessors. A learning rate of 0.1 and maximum tree depth of 3 provide regularisation against overfitting. GBR is well suited to tabular environmental datasets with complex non-linear interactions.

All models were trained on seventeen predictor variables: PM10, NO, NO2, NOx, NH3, SO2, CO, Ozone, Benzene, Toluene, AT, RH, WS, WD, RF, SR, and BP. The response variable was daily mean PM2.5. Scikit-learn (version 1.3) was used for all model implementation. A five-fold time series cross-validation (TimeSeriesSplit) protocol was applied, ensuring strict chronological ordering of training and test folds to prevent data leakage. Model performance was evaluated using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2), with scores averaged across folds. Feature importance was derived from the Random Forest's Mean Decrease in Impurity (MDI) scores computed on the full dataset.

4. Results and Discussion

4.1 Temporal Patterns and Descriptive Statistics

The five-year dataset reveals a persistently elevated PM2.5 burden at Dwarka Sector. The overall daily mean of 104.53 µg/m³ (SD = 86.38 µg/m³) exceeds the NAAQS annual standard of 40 µg/m³ by a factor of 2.6 and the WHO 2021 annual guideline of 5 µg/m³ by a factor exceeding 20. The strongly right-skewed distribution (median = 75.08 µg/m³ against mean = 104.53 µg/m³) reflects the frequent occurrence of extreme pollution episodes, with a

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

recorded maximum of 600.40 µg/m³ corresponding to severe wintertime fog-haze events. The annual means demonstrate modest inter-annual variation: 109.19 µg/m³ (2019), 100.14 µg/m³ (2020), 108.47 µg/m³ (2021), 98.37 µg/m³ (2022), and 106.48 µg/m³ (2023). The 2020 decline is attributable to emission reductions enforced during the national COVID-19 lockdown (25 March – 31 May 2020), consistent with Mahato et al. (2020), who documented 40–50% reductions in roadside PM2.5 during the lockdown period. The rapid return to near-2019 levels by 2021 confirms that structural emission drivers vehicular fleets, industrial activity, agricultural burning remained fundamentally unchanged.

Monthly disaggregation exposes a pronounced bimodal annual cycle. November records the highest mean concentration (231.40 µg/m³), followed by December (212.01 µg/m³) and January (189.21 µg/m³). The autumn secondary peak in October (111.90 µg/m³) coincides with post-monsoon boundary layer re-establishment and the onset of paddy stubble burning in north-western India. Minimum concentrations occur in August (32.20 µg/m³) and July (35.49 µg/m³), driven by efficient wet scavenging and enhanced convective turbulent mixing during the Southwest Monsoon. Seasonally, winter mean PM2.5 (177.35 µg/m³) is 4.5 times the summer mean (39.16 µg/m³), underscoring the overwhelming dominance of boundary layer dynamics in controlling surface concentration levels relative to emission variability alone.

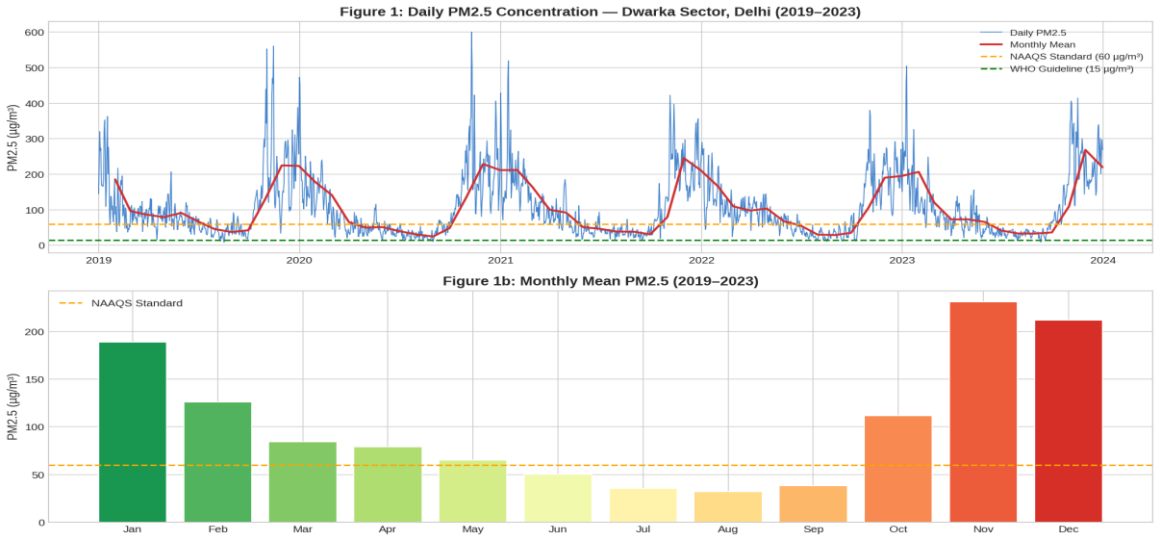

***Figure 1**: Daily PM2.5 time series (2019–2023) with 30-day rolling mean (top) and mean monthly PM2.5 across the study period (bottom). Dashed lines indicate NAAQS (60 µg/m³) and WHO (15 µg/m³) reference values. Dwarka Sector CAAQM Station, Delhi.*

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

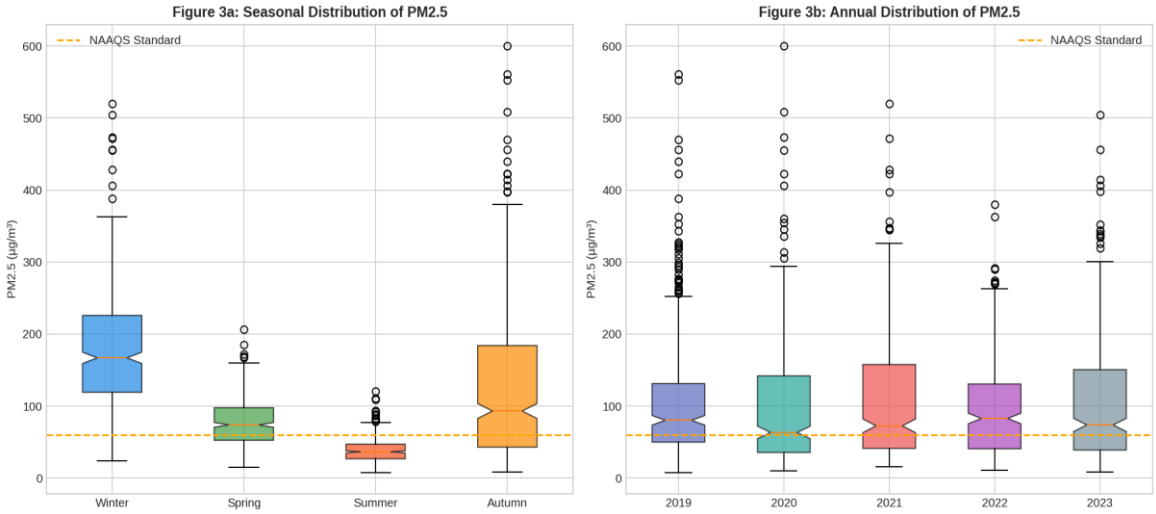

Figure 2: *Box-and-whisker plots of PM2.5 distributions by (a) meteorological season and (b) calendar year. Notches indicate 95% confidence intervals on the median. Orange dashed line: NAAQS 60 µg/m³ standard.*

4.2 Pollutant and Meteorological Correlations

The Pearson correlation matrix reveals a physically coherent and interpretable structure. Among co-pollutants, PM10 exhibits the strongest positive association with PM2.5 ($r = 0.845$, $p < 0.01$), reflecting shared combustion and resuspension source categories and common meteorological dispersion controls. CO ($r = 0.739$) and NOx ($r = 0.730$) follow closely, consistent with their co-emission from vehicular and industrial combustion sources that also generate fine carbonaceous and nitrate aerosol components. NO ($r = 0.705$) tracks fresh traffic emission, while NO2 ($r = 0.596$) is somewhat weaker owing to its dual origin through direct emission and photochemical conversion of NO. Benzene ($r \sim 0.60$) and Toluene ($r \sim 0.55$) similarly co-vary with PM2.5 as tracers of combustion and evaporative vehicular emission.

Among meteorological variables, barometric pressure exerts the strongest positive association ($r = 0.630$), consistent with anticyclonic high-pressure synoptic patterns that suppress convective mixing and favour pollutant accumulation. Negative correlations with ambient temperature ($r = -0.612$), solar radiation ($r = -0.575$), and wind speed ($r = -0.574$) are mechanistically interrelated: higher surface temperatures and solar radiation intensity drive convective boundary layer growth, diluting surface concentrations; higher wind speeds enhance horizontal advective ventilation. These findings align closely with the meteorological influence documented by Seinfeld and Pandis (2016) and Kumar et al. (2021) for Indian urban sites. Relative humidity shows a modest positive association ($r = 0.253$), reflecting hygroscopic particle growth at elevated RH values and the suppression of turbulent mixing by stable, humid air masses.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

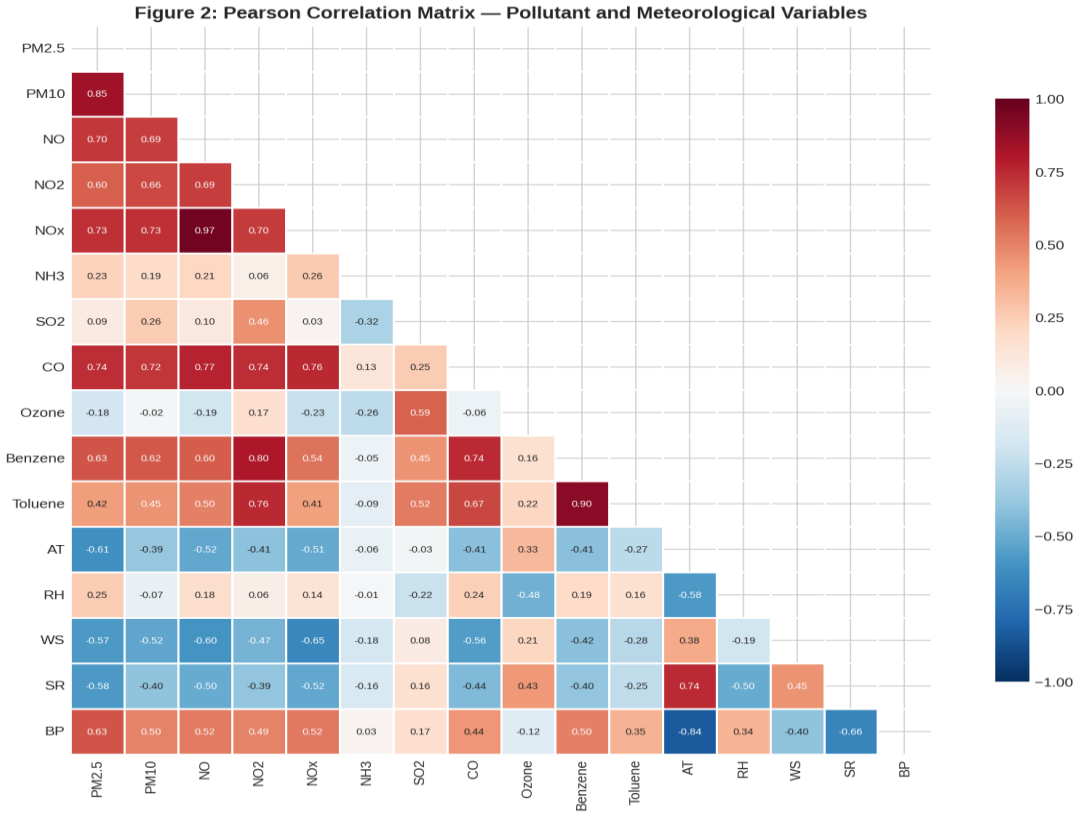

Figure 3: *Lower-triangular Pearson correlation heatmap for all quantitative study variables. Blue shading indicates positive correlation; red shading indicates negative correlation. Values statistically significant at $p < 0.01$.*

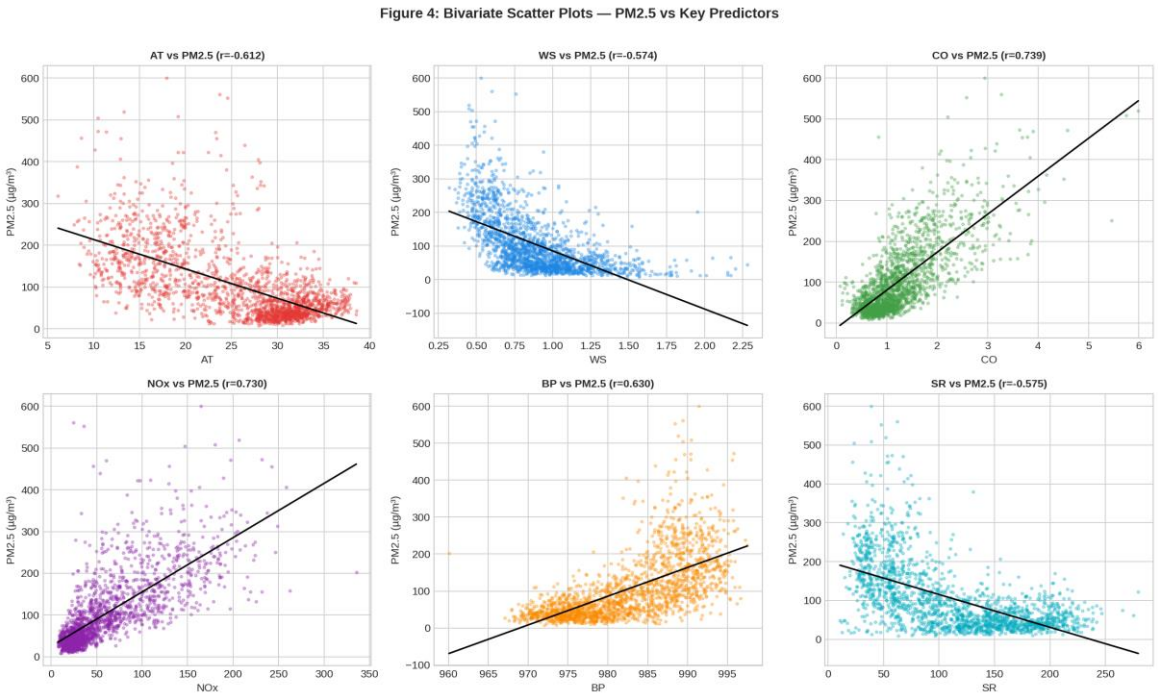

Figure 4*: Bivariate scatter plots of PM2.5 against six key predictors with OLS regression lines and Pearson r values. n = 1,826 daily observations, Dwarka Sector, Delhi (2019–2023).*

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

4.3 Machine Learning Model Performance

Under five-fold time series cross-validation, all three models demonstrate meaningful predictive skill, with the two ensemble methods substantially outperforming the linear baseline. MLR achieves RMSE = 41.60 µg/m³, MAE = 30.93 µg/m³, and R^2 = 0.772, confirming the utility of the predictor set while revealing the inadequacy of a purely linear functional form given the known non-linear interactions among atmospheric variables. The Random Forest improves across all metrics (RMSE = 37.43 µg/m³, MAE = 23.16 µg/m³, R^2 = 0.815), with the 25.1% reduction in MAE relative to MLR being particularly notable, suggesting that RF's capacity to model threshold effects such as the near-zero wind speed conditions that precipitate severe episodes substantially reduces systematic bias on extreme pollution days.

The Gradient Boosting Regressor delivers the strongest overall performance (RMSE = 35.10 µg/m³, MAE = 22.38 µg/m³, R^2 = 0.841), representing a 15.6% reduction in RMSE and a 6.9-percentage-point improvement in R^2 relative to MLR. These gains are consistent with findings from Chen et al. (2018) and Brokamp et al. (2018) in comparable urban monitoring contexts. The GBR's sequential error-correction mechanism enables it to capture persistent residual structure for instance, the non-linear amplification of PM2.5 under simultaneously low wind speed and low temperature conditions that the RF's parallel averaging framework cannot fully exploit. The unexplained variance (approximately 15.9%) is attributable primarily to predictors absent from the local station dataset: planetary boundary layer height, regional emission transport indicators derived from fire radiative power satellites, and fugitive dust resuspension events all of which represent well-established targets for future model enhancement.

Table 2: Machine Learning Model Performance Five-Fold Time Series Cross-Validation

Model	RMSE (µg/m³)	MAE (µg/m³)	R^2
Multiple Linear Regression	41.60	30.93	0.772
Random Forest Regressor	37.43	23.16	0.815
Gradient Boosting Regressor	**35.10**	**22.38**	**0.841**

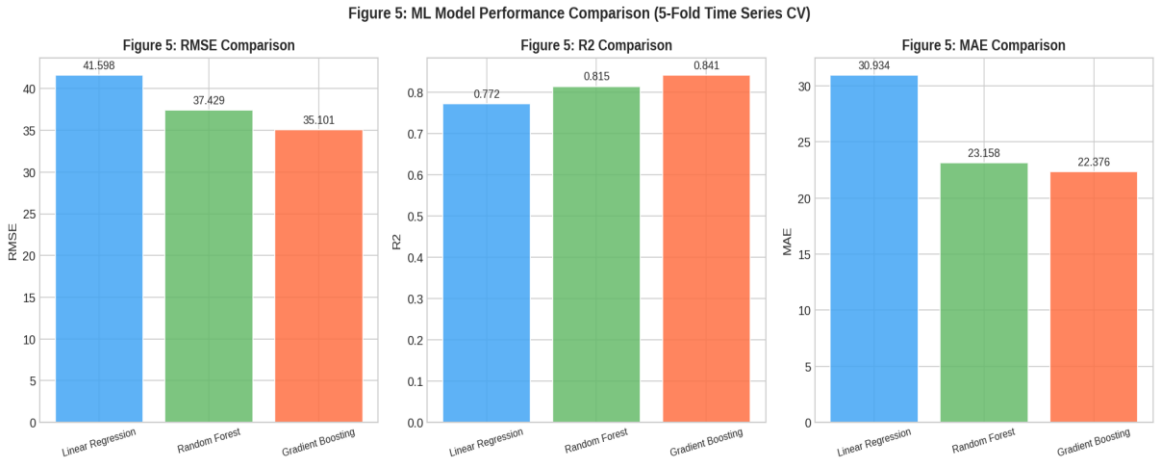

Figure 5: *Comparative performance metrics (RMSE, R^2, MAE) across three machine learning models under five-fold time series cross-validation. Gradient Boosting achieves best performance on all three criteria.*

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

4.4 Feature Importance Analysis

The Random Forest MDI feature importance analysis reveals a strongly hierarchical predictor structure. PM10 dominates with an importance score of 0.731, accounting for nearly three-quarters of the total predictive signal a consequence of its very high bivariate correlation with PM2.5 (r = 0.845) and the shared physicochemical pathways through which both size fractions are generated and transported. Ambient temperature ranks second (0.084), confirming its role as a proxy for convective boundary layer development that governs vertical dilution capacity. Relative humidity (0.068) ranks third, consistent with its dual role in secondary aerosol formation and hygroscopic particle growth, as documented by Seinfeld and Pandis (2016). Solar radiation (0.030) and barometric pressure (0.014) complete the top five predictors, each linked to distinct atmospheric dispersion mechanisms. Gaseous pollutants NO, CO, and NOx contribute relatively modest MDI scores when PM10 is included in the feature set, reflecting their substantial collinearity with PM10 rather than orthogonal predictive information. This finding suggests that in an operational forecasting context lacking co-located PM10 measurements, gaseous combustion tracers could serve as partial proxies for fine particulate loading.

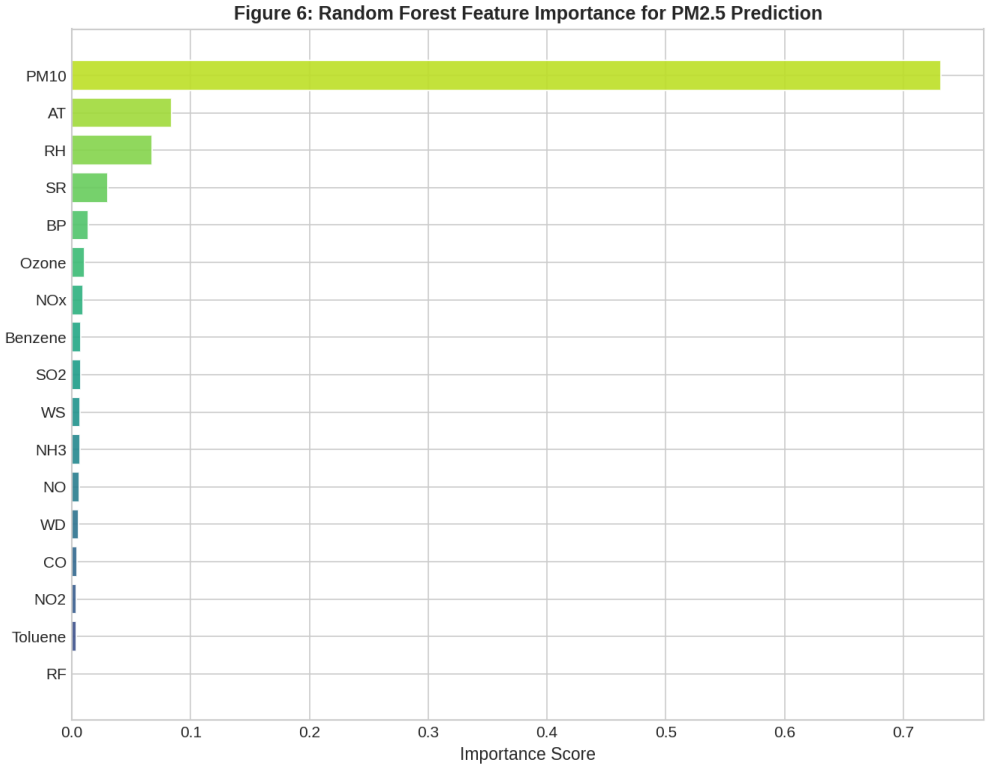

***Figure 6**: Random Forest Mean Decrease in Impurity (MDI) feature importance scores for all 17 predictor variables. Scores normalised to sum to unity. Computed from model trained on complete 1,826-day dataset.*

5. Conclusion

This study has presented a systematic, empirically grounded investigation of PM2.5 pollution dynamics at the Dwarka Sector CAAQM monitoring station, Delhi, over a five-year

continuous daily record from 2019 to 2023. The analysis integrates temporal characterisation, correlation analysis, and comparative machine learning modelling within a methodologically rigorous time series cross-validation framework, yielding findings of both scientific and policy relevance.

The central quantitative finding a five-year mean PM2.5 of 104.53 µg/m³, representing 2.6 times the NAAQS annual standard confirms the chronic and severe nature of fine particulate pollution at this site. The absence of a sustained downward trend across the study period, notwithstanding the transient 2020 reduction attributable to COVID-19 lockdown-induced emission suppression, indicates that the structural drivers of Delhi's air quality crisis remain fundamentally unmitigated. This finding is particularly sobering in the context of the National Clean Air Programme's stated target of a 20–30% PM2.5 reduction by 2024 relative to 2017 baseline levels. The 4.5-fold seasonal amplitude between winter peak and summer trough concentrations underscores that meteorological boundary layer dynamics rather than emission variability alone constitute the dominant modulator of surface PM2.5 accumulation at intra-annual timescales, with important implications for the design and timing of emission control interventions.

The Gradient Boosting Regressor, achieving $R^2 = 0.841$ under rigorous time series cross-validation, represents the most appropriate modelling architecture for daily PM2.5 prediction from this multi-variable observational dataset, outperforming both Random Forest and Multiple Linear Regression across all three-evaluation metrics. The feature importance analysis identifies PM10, ambient temperature, relative humidity, solar radiation, and barometric pressure as the principal predictive contributors, each linking model behaviour to established atmospheric physical mechanisms of boundary layer mixing, hygroscopic growth, and synoptic pressure control. These results provide a replicable methodological template applicable to other CAAQM monitoring stations across India's major cities.

Future research should extend the predictive framework in several directions: incorporation of planetary boundary layer height from ERA5 reanalysis or Doppler wind lidar observations; integration of satellite-derived fire radiative power from the VIIRS instrument as a time-varying regional emission proxy during the October–November crop residue burning season; application of deep learning temporal sequence architectures (LSTM, Temporal Convolutional Networks) to exploit lagged PM2.5 autocorrelation; and spatio-temporal extension across the full Delhi-NCR CAAQM network using graph neural network frameworks. These advances would support the development of an operationally deployable, interpretable air quality forecasting system capable of providing 24–72-hour advance warning of severe pollution episodes a critical public health infrastructure requirement for one of the world's most polluted megacities.

References

1 Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

2 Brook, R. D., Rajagopalan, S., Pope, C. A., Brook, J. R., Bhatnagar, A., Diez-Roux, A. V., Holguin, F., Hong, Y., Luepker, R. V., Mittleman, M. A., Peters, A., Siscovick, D., Smith, S. C., Whitsel, L., & Kaufman, J. D. (2010).

3 Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the American Heart Association. Circulation, 121(21), 2331–2378.

4 Brokamp, C., Jandarov, R., Hossain, M., & Ryan, P. (2018). Predicting daily urban fine particulate matter concentrations using a random forest model. Environmental Science & Technology, 52(7), 4173–4179.

5 Chen, G., Li, S., Knibbs, L. D., Hamm, N. A. S., Cao, W., Li, T., Guo, J., Ren, H., Abramson, M. J., & Guo, Y. (2018). A machine learning method to estimate PM2.5 concentrations across China with remote sensing, meteorological and land use information. Science of the Total Environment, 636, 52–60.

6 Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM.

7 Dey, S., & Di Girolamo, L. (2010). A climatology of aerosol optical and microphysical properties over the Indian subcontinent from 9 years (2000–2008) of Multiangle Imaging Spectroradiometer (MISR) data. Journal of Geophysical Research: Atmospheres, 115(D15).

8 Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

9 Guttikunda, S. K., & Gurjar, B. R. (2012). Role of meteorology, emission sources, and chemistry in the spatio-temporal variability of particulate matter in Delhi, India. Environmental Science and Pollution Research, 19(3), 1036–1048.

10 Health Effects Institute. (2023). State of Global Air 2023. Health Effects Institute.

11 Kumar, A., Sarin, M. M., & Srinivas, B. (2021). Impact of meteorology on PM2.5 concentration and composition over the Indo-Gangetic Plain: Source apportionment and long-range transport. Atmospheric Environment, 254, 118377.

12 Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

13 Mahato, S., Pal, S., & Ghosh, K. G. (2020). Effect of lockdown amid COVID-19 pandemic on air quality of the megacity Delhi, India. Science of the Total Environment, 730, 139086.

14 Masood, A., & Ahmad, K. (2021). A model for particulate matter (PM2.5) prediction for Delhi based on machine learning approaches. Procedia Computer Science, 167, 2101–2110.

15 Murray, C. J. L., Aravkin, A. Y., Zheng, P., Abbafati, C., Abbas, K. M., Abbasi-Kangevari, M., Abd-Allah, F., Abdelalim, A., Abdollahi, M., Abdollahpour, I., & GBD 2019 Risk Factors Collaborators. (2020). Global burden of 87 risk factors in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. The Lancet, 396(10258), 1223–1249.

16 Pope, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., & Thurston, G. D. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA, 287(9), 1132–1141.

17 Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913–929.

18 Seinfeld, J. H., & Pandis, S. N. (2016). Atmospheric chemistry and physics: From air pollution to climate change (3rd ed.). Wiley.

19 Sharma, A. K., Bali, K., & Kumar, R. (2016). Assessment of the impact of crop residue burning in north-western India on air quality of Delhi. International Journal of Research in Chemistry and Environment, 6(2), 52–58.

20 Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics, 8, 25.

21 Zhang, L., Liu, P., Zhao, L., Wang, G., Zhang, W., & Liu, J. (2022). Air quality predictions with a semi-supervised bidirectional LSTM neural network. Atmospheric Pollution Research, 12(3), 328–339.