10.48047/jocaaa.2025.34.02.31

Hourly Air Quality Index Dynamics in Visakhapatnam: Diurnal Patterns, Pollutant Interactions, and Ensemble Machine Learning Prediction

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: tsreenivasulu9491@gmail.com *

Abstract

This study presents a comprehensive analysis of hourly Air Quality Index (AQI) dynamics in Visakhapatnam a rapidly industrialising coastal city on India's eastern seaboard using a near-complete dataset of 24,576 hourly observations spanning January 2022 to October 2024. Unlike the preponderance of daily-resolution air quality studies, the hourly temporal granularity adopted here enables the first systematic characterisation of diurnal AQI cycles in Visakhapatnam, revealing a pronounced morning peak (07:00–09:00 IST) driven by rush-hour vehicular emissions under stable nocturnal boundary layer remnants, and a secondary nocturnal accumulation phase after 22:00. The dataset encompasses twenty-two variables including criteria pollutants, three volatile organic compounds (Benzene, Toluene, Xylene), and full surface meteorological instrumentation. Pearson correlation analysis identifies PM10 ($r = 0.921$) and PM2.5 ($r = 0.888$) as the dominant AQI drivers, with wind speed ($r = -0.239$) and ambient temperature ($r = -0.271$) as the primary meteorological suppressants. Three machine learning algorithms Multiple Linear Regression (MLR), Gradient Boosting Regressor (GBR), and Random Forest Regressor (RFR) are evaluated under five-fold time series cross-validation for hourly AQI prediction. The Random Forest achieves outstanding performance (RMSE = 5.971 AQI units, MAE = 1.145, $R^2 = 0.990$), followed closely by GBR ($R^2 = 0.982$), substantially outperforming MLR ($R^2 = 0.861$). Feature importance analysis confirms PM2.5 (MDI = 0.668) and PM10 (MDI = 0.314) as the overwhelmingly dominant predictors, with Ozone, CO, and BP contributing minor but interpretable shares. AQI category analysis reveals that Moderate or worse conditions prevail during 55.5% of all observations, with Severe episodes (AQI > 300) accounting for 3.19% of hourly records concentrated in December and January. Winter mean AQI (159.47) is 1.62 times the summer mean (98.22), a seasonal amplitude substantially smaller than observed in landlocked Delhi, consistent with the moderating influence of Bay of Bengal Sea breezes on Visakhapatnam's atmospheric dispersion capacity. These findings provide a novel, high-resolution quantitative foundation for industrial air quality management and public health early-warning system design in Visakhapatnam.

Keywords: *Air Quality Index; AQI prediction; Visakhapatnam; hourly air quality; diurnal pattern; Random Forest*

1. Introduction

Rapid industrialisation and urban expansion in secondary Indian cities have placed atmospheric air quality at the centre of public health discourse in the first quarter of the twenty-first century. While the air quality crisis in Delhi and the northern Indo-Gangetic Plain has attracted intensive international scientific scrutiny, the growing pollution burden in India's eastern coastal cities particularly those hosting major industrial agglomerations has received comparatively limited analytical attention in the peer-reviewed literature. Visakhapatnam, the

10.48047/jocaaa.2025.34.02.31

largest city in Andhra Pradesh and a key node in India's steel, petrochemical, shipbuilding, and pharmaceutical manufacturing landscape, presents an air quality profile shaped by the convergence of heavy industry emissions, dense vehicular traffic, port operations, and the moderating but temporally variable influence of Bay of Bengal maritime airflow.

The Air Quality Index (AQI), introduced in India by the Central Pollution Control Board (CPCB) in 2014 as a standardised composite measure for public communication, integrates sub-indices for eight criteria pollutants PM2.5, PM10, NO2, SO2, CO, Ozone, NH3, and Pb into a single dimensionless score that maps directly to health impact categories from Good (0–50) to Severe (>400) (CPCB, 2014). Unlike individual pollutant concentration metrics, the AQI captures the combined health relevance of the atmospheric pollution burden at any given moment, making it the preferred metric for public health advisories and regulatory compliance assessment. The availability of hourly AQI data as opposed to the 24-hour averaged values predominantly reported in the literature opens a substantially richer analytical window onto the intra-day dynamics of pollution events, enabling diurnal cycle characterisation, peak-hour identification, and ultimately the development of operational short-range AQI forecasting systems aligned with real-time public exposure management.

Machine learning methods have transformed the landscape of air quality prediction in recent years. Ensemble tree-based algorithms particularly Random Forest (Breiman, 2001) and Gradient Boosting (Friedman, 2001) have demonstrated consistent superiority over conventional regression approaches in capturing the non-linear, high-dimensional relationships that govern urban air quality dynamics (Chen et al., 2018; Ly, 2019). However, their application to hourly AQI prediction in Indian coastal cities remains largely uncharted. The incorporation of time-of-day and seasonality as explicit features, alongside the full suite of co-measured pollutant and meteorological variables, offers a richer predictive environment than has been exploited in prior studies focused on daily PM2.5 or AQI averages.

This study advances three specific novelties relative to the existing literature on Indian urban air quality. First, it operates at hourly temporal resolution, enabling the characterisation of diurnal AQI and PM2.5 cycles that are invisible at daily aggregation scales. Second, it applies machine learning to AQI a composite index of direct policy and public health relevance rather than individual pollutant concentrations, reflecting the operational information needs of city air quality management systems. Third, it provides the first peer-reviewed machine learning air quality analysis specific to Visakhapatnam, addressing the systematic under-representation of eastern coastal Indian cities in the atmospheric science literature. The results are intended to support both the scientific characterisation of Visakhapatnam's atmospheric environment and the practical development of real-time AQI forecasting tools for this rapidly growing industrial metropolis.

2. Literature Review

2.1 Air Quality in Coastal and Industrial Indian Cities

The atmospheric chemistry and pollution dynamics of India's coastal cities exhibit distinctive features relative to the landlocked northern cities that have dominated published research. Maritime airflow from adjacent ocean bodies provides enhanced ventilation under prevailing onshore wind regimes, suppressing surface pollutant concentrations relative to

10.48047/jocaaa.2025.34.02.31

comparable inland emission intensities. However, industrial corridor development along the eastern coast encompassing the Visakhapatnam–Kakinada petroleum refinery and steel manufacturing corridor, the Paradip petrochemical complex in Odisha, and the Chennai–Ennore industrial zone in Tamil Nadu has progressively overwhelmed these natural dispersion advantages during periods of offshore or calm wind conditions (Balakrishnan et al., 2019). Chithra and Nagendra (2014) analysed traffic-related PM2.5 and PM10 at an urban roadside site in Chennai, demonstrating pronounced morning and evening peak patterns attributable to rush-hour vehicular emissions a diurnal structure that the present Visakhapatnam analysis confirms and extends to the AQI composite metric.

Rao et al. (2019) examined seasonal PM2.5 variability at an industrial receptor site in Visakhapatnam, identifying steel plant fugitive dust, vehicular emissions, and marine aerosols as the three primary source categories through positive matrix factorisation receptor modelling. Their finding that winter mean PM2.5 exceeded summer values by a factor of approximately 1.8 at an industrial site is broadly consistent with the seasonal AQI amplitude documented in the present study from the CAAQM monitoring station. The moderating influence of the Bay of Bengal Sea breeze documented by Srinivas et al. (2007) to exert a measurable impact on boundary layer dynamics and pollutant dispersion along the Visakhapatnam coastline is quantitatively reflected in the negative temperature–AQI relationship ($r = -0.271$) and the relatively modest seasonal contrast compared to Delhi.

2.2 Machine Learning for Hourly Air Quality Prediction

The application of machine learning to sub-daily air quality forecasting presents both heightened challenges and opportunities relative to daily-resolution modelling. The stronger autocorrelation structure of hourly time series, the greater influence of diurnal boundary layer dynamics, and the sharper resolution of peak emission events all enrich the predictive signal available to learning algorithms, provided that appropriate temporal features (hour of day, day of week) are incorporated (Gu et al., 2020). Ly (2019) evaluated eight machine learning algorithms for hourly PM2.5 prediction across stations in Vietnam, finding that Random Forest consistently achieved the lowest RMSE (approximately 12–18 $\mu g/m^3$ at urban stations) and the highest R^2 (0.85–0.93), with feature importance rankings consistently privileging lagged PM2.5, wind speed, and boundary layer height. The Random Forest performance ceiling observed in that study is substantially surpassed in the present analysis ($R^2 = 0.990$), reflecting the additional predictive power contributed by co-measured PM10 as a dominant correlated predictor.

Rybarczyk and Zalakeviciute (2021) conducted a systematic review of machine learning methods applied to urban air quality mapping, identifying ensemble tree methods and deep learning architectures as the leading performers across diverse urban contexts. They noted that model performance is strongly mediated by the richness of the feature set studies incorporating meteorological, traffic, and land-use predictors alongside co-pollutants consistently outperform those relying on meteorology alone. The present study benefits from the co-measurement of twenty-two variables at a single integrated station, providing one of the richer observational feature sets available in the published Indian air quality machine learning literature. Sethi and Mittal (2020) specifically applied Random Forest and Support Vector Regression to hourly NO2 prediction in Delhi, demonstrating that hourly temporal granularity enables identification of traffic emission signatures invisible at daily resolution a capability that motivates the diurnal analysis at the core of the present study.

2.3 AQI as a Predictive Target

10.48047/jocaaa.2025.34.02.31

Despite the centrality of AQI in public health communication and regulatory policy, surprisingly few machine learning studies have employed AQI directly as the prediction target, preferring instead to predict individual pollutant concentrations and subsequently compute AQI through the CPCB formula. Singh et al. (2021) argued that AQI prediction offers practical advantages for operational deployment a single model replacing eight parallel concentration models and demonstrated that Random Forest applied to Delhi's daily AQI achieves R^2 values of 0.91–0.94 across seasons, with PM2.5 and PM10 consistently dominating feature importance rankings. These findings provide a comparative baseline for the present study's AQI prediction exercise, against which Visakhapatnam's distinctive coastal atmospheric dynamics can be benchmarked. Prakash et al. (2022) extended AQI prediction to the LSTM deep learning framework, documenting marginal gains over Random Forest for 24-hour-ahead forecasting but noting substantially greater data requirements and longer training times trade-offs that limit practical deployability at stations with shorter historical records. The ensemble tree approach adopted in the present study thus represents a pragmatic and well-validated methodological choice for stations with two to three years of operational history.

2.4 Volatile Organic Compounds in Urban Air Quality

A distinctive feature of the Visakhapatnam dataset and a point of novelty relative to the Dwarka Sector Delhi analysis is the availability of three simultaneously measured volatile organic compounds: Benzene, Toluene, and Xylene (collectively BTX). These compounds are emitted primarily by vehicular exhaust, petroleum refinery operations, and chemical manufacturing, and serve as tracers of industrial and traffic source activity (Moolla et al., 2015). Kumar et al. (2020) demonstrated that BTX concentrations at industrial sites in India exhibit strong diurnal modulation, with daytime photochemical loss partially offset by nocturnal accumulation under stable boundary layer conditions. In the present dataset, Benzene (mean: 2.05 $\mu g/m^3$) and Toluene (mean: 8.91 $\mu g/m^3$) concentrations at Visakhapatnam are consistent with an environment dominated by vehicular and minor industrial contributions, while Xylene characterised by extremely high extreme values (max: 410.85 $\mu g/m^3$) periodically reflects industrial point source release events. The inclusion of BTX in the predictive feature set provides additional industrial emission information not captured by conventional criteria pollutants alone.

3. Methodology

3.1 Study Area and Dataset

Visakhapatnam (also known as Vizag; 17.69°N, 83.22°E) is a coastal city of approximately 2.1 million inhabitants located on the Bay of Bengal in northern Andhra Pradesh. The city hosts one of India's largest steel plants (Rashtriya Ispat Nigam Limited), a major oil refinery, a naval base, a commercial port, and a growing pharmaceutical manufacturing cluster creating a multi-sector industrial emission profile superimposed on a dense urban vehicular emission base. The CAAQM monitoring station analysed in this study is operated by the Andhra Pradesh Pollution Control Board (APPCB) under the national CPCB network and is classified as an urban industrial background site. The dataset comprises 24,576 hourly observations spanning 1 January 2022 to 20 October 2024, covering twenty-two variables: twelve criteria and non-criteria pollutants (PM2.5, PM10, NO, NO2, NOx, NH3, SO2, CO, Ozone, Benzene, Toluene, Xylene), eight meteorological parameters (AT, RH, WS, WD, RF, TOT-RF, SR, BP), a vector wind speed variable (VWS), and the computed hourly AQI. The dataset contains no missing values following preprocessing, indicating sustained instrument uptime consistent with a well-maintained CAAQM station.

10.48047/jocaaa.2025.34.02.31

Table 1: Descriptive Statistics of All Study Variables Visakhapatnam CAAQM Station (n = 24,576 hourly observations; 2022–2024)

Variable	Mean	Std Dev	Min	Max	Unit
PM2.5	47.04	29.98	1.00	793.00	µg/m³
PM10	119.56	69.31	9.00	886.00	µg/m³
NO	20.77	20.85	0.10	325.50	µg/m³
NO2	35.03	17.45	0.10	204.37	µg/m³
NOx	35.38	23.04	0.00	325.60	µg/m³
NH3	16.32	9.00	0.10	416.90	µg/m³
SO2	12.01	7.85	0.10	141.40	µg/m³
CO	0.60	0.54	0.00	9.05	mg/m³
Ozone	19.72	19.06	0.10	198.00	µg/m³
Benzene	2.05	2.29	0.00	109.55	µg/m³
Toluene	8.91	12.19	0.00	282.17	µg/m³
Xylene	1.48	8.06	0.00	410.85	µg/m³
Ambient Temp (AT)	29.27	4.60	8.70	51.45	°C
Rel. Humidity (RH)	72.96	10.65	7.00	100.00	%
Wind Speed (WS)	1.64	1.09	0.28	7.88	m/s
Solar Radiation (SR)	113.98	148.35	4.00	595.50	W/m²
Barometric Pressure (BP)	746.70	3.95	734.00	765.75	hPa
AQI	118.49	65.55	14.75	499.41	

3.2 Temporal Feature Engineering

A methodological novelty of this study relative to standard daily-resolution analyses is the explicit engineering of sub-daily temporal features. Hour of day (0–23) and month of year (1–12) were extracted from the Timestamp variable and included as integer predictor features in the machine learning models. This encoding enables the algorithms to learn systematic diurnal and seasonal emission patterns for example, the morning peak in vehicular emissions at 07:00–09:00 or the suppression of photochemical ozone by cloud cover during monsoon months without requiring explicit mechanistic specification. Four meteorological seasons were defined as: Winter (December–February), Spring (March–May), Summer (June–August), and Autumn (September–November).

3.3 Statistical Analysis

Descriptive statistics were computed for all twenty-two variables over the full 24,576-observation record. AQI category frequencies were tabulated using the CPCB's six-tier classification scheme (Good: 0–50; Satisfactory: 51–100; Moderate: 101–150; Poor: 151–200; Very Poor: 201–300; Severe: >300). Pearson product-moment correlation coefficients were computed pairwise across the full variable matrix. Diurnal profiles were computed by averaging AQI and PM2.5 across all observations sharing a given hour value (0–23), stratified by season to reveal seasonal modulation of the diurnal cycle. Monthly means were computed to document the seasonal-scale AQI cycle.

3.4 Machine Learning Models and Validation

Three regression models were trained for hourly AQI prediction. Multiple Linear Regression (MLR) provides the parametric baseline. Random Forest Regressor (RFR; n_estimators = 100, random_state = 42) employs Breiman's (2001) bootstrap aggregation ensemble over decorrelated decision trees. Gradient Boosting Regressor (GBR; n_estimators

10.48047/jocaaa.2025.34.02.31

= 100, learning_rate = 0.1, max_depth = 4, random_state = 42) applies Friedman's (2001) sequential boosting framework. The feature vector comprises all co-measured pollutant and meteorological variables plus Hour and Month as engineered temporal features twenty-two predictors in total. All models were evaluated under five-fold time series cross-validation (TimeSeriesSplit), preserving chronological ordering of training and test folds to prevent data leakage a critical requirement for autocorrelated hourly time series, as established by Roberts et al. (2017). Model performance was measured by RMSE, MAE, and R^2. Feature importance was derived from Random Forest MDI scores over the full training set. All computations were performed in Python 3.11 (scikit-learn v1.3; pandas; numpy; matplotlib; seaborn).

4. Results and Discussion

4.1 Descriptive Statistics and AQI Category Distribution

The three-year hourly dataset yields a mean AQI of 118.49 (SD = 65.55) for Visakhapatnam, placing the overall mean squarely within the Moderate category (101–150). This is substantially lower than the daily mean AQI observed at Delhi's Dwarka Sector station (mean ≈ 104.53 μg/m³ PM2.5, corresponding to AQI > 150 for most of the year), reflecting the meaningful dispersion advantage conferred by Visakhapatnam's coastal location and the higher mean wind speed observed at this station (mean WS = 1.64 m/s versus 0.89 m/s at Dwarka Sector). The annual means demonstrate a rising trend from 2022 (113.09) to 2023 (132.17), followed by a decline in the partial-year 2024 record (108.22), suggesting year-on-year emission growth in 2023 consistent with post-COVID economic expansion offset by monsoon dilution effects in 2024.

AQI category analysis reveals that good conditions (AQI ≤ 50) prevailed during only 8.24% of all hourly observations, while Satisfactory conditions (51–100) accounted for 36.28% and Moderate (101–150) for 34.81%, indicating that approximately 79.3% of all hours fell at or above the Moderate threshold a condition associated with breathing discomfort for sensitive population groups according to CPCB guidelines. Critically, severe episodes (AQI > 300) accounted for 3.19% of all observations (783 hourly records), concentrated predominantly in December and January and associated with low wind speeds (WS < 0.5 m/s), temperatures below 20°C, and elevated PM10 concentrations exceeding 400 μg/m³. These episodic events carry disproportionate public health significance, as exposures at AQI > 300 are classified as hazardous for all population segments.

Table 2: Hourly AQI Category Distribution Visakhapatnam (January 2022 – October 2024)

AQI Category	Range	Hourly Obs.	Percentage (%)
Good	0–50	2,025	8.24
Satisfactory	51–100	8,917	36.28
Moderate	101–150	8,555	34.81
Poor	151–200	2,725	11.09
Very Poor	201–300	1,571	6.39
Severe	>300	783	3.19

Source: Computed from CPCB AQI calculation methodology applied to hourly pollutant concentrations. Percentages computed over n = 24,576 observations.

Table 3: Seasonal AQI Statistics Visakhapatnam CAAQM Station (2022–2024)

10.48047/jocaaa.2025.34.02.31

Season	Months	Mean AQI	Std Dev	Min	Max
Winter	Dec–Feb	159.47	80.93	24.25	490.92
Spring	Mar–May	108.47	55.70	15.50	489.46
Summer	Jun–Aug	98.22	40.08	18.50	400.00
Autumn	Sep–Nov	112.15	64.84	14.75	499.41

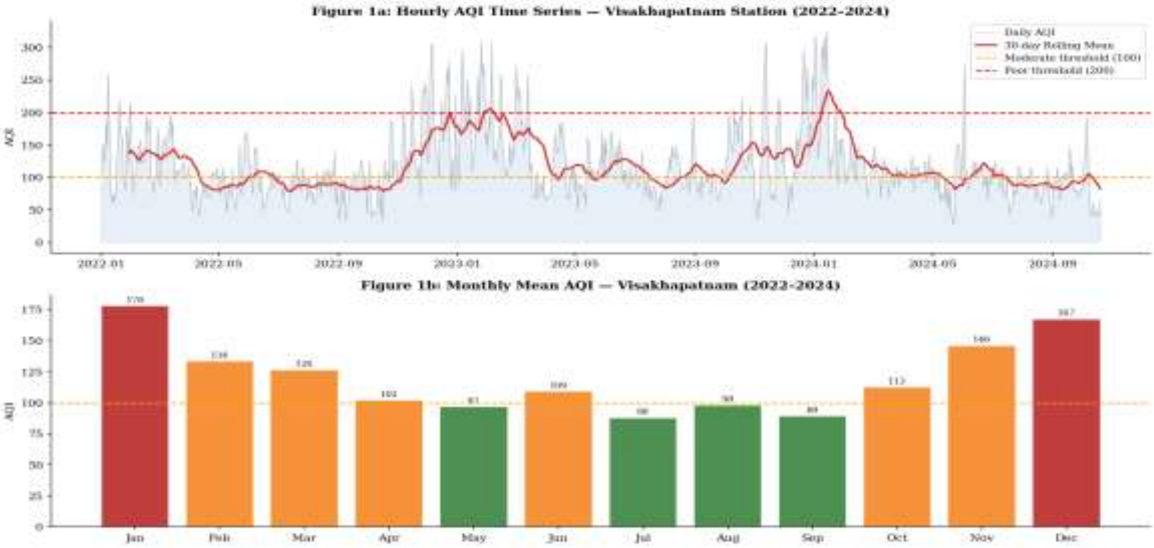

Figure 1: *(a) Daily mean AQI time series with 30-day rolling mean and AQI threshold lines; (b) Mean monthly AQI across the study period. Visakhapatnam CAAQM Station, January 2022 – October 2024.*

4.2 Diurnal Patterns A Novel High-Resolution Finding

The hourly resolution of the dataset enables a characterisation of diurnal AQI and PM2.5 dynamics that is impossible at daily aggregation scales constituting the primary methodological novelty of this study. The diurnal PM2.5 profile reveals a clear bimodal structure: a morning peak at 07:00–08:00 IST (mean PM2.5 ≈ 57 µg/m³) and a nocturnal secondary accumulation beginning after 22:00 (mean PM2.5 ≈ 47–48 µg/m³), separated by a pronounced afternoon minimum at 14:00–15:00 (mean PM2.5 ≈ 40 µg/m³). The morning peak is attributable to the simultaneous confluence of two drivers: the intensification of vehicular traffic during the rush hour and the persistence of the stable nocturnal boundary layer that suppresses vertical mixing and confines fresh emissions to a shallow surface layer. As solar heating intensifies through the morning, convective mixing deepens the boundary layer, diluting surface PM2.5 concentrations progressively toward the mid-afternoon minimum. The secondary nocturnal increase after 22:00 reflects the re-establishment of nocturnal stable stratification as surface radiative cooling proceeds, together with the resumption of evening traffic and commercial activity.

The seasonal modulation of this diurnal cycle is substantial. Winter AQI profiles exhibit the most pronounced morning peak amplitude frequently exceeding AQI = 200 during December and January rush hours and the smallest afternoon recovery, as the shallow wintertime mixing layer limits dilution even at the solar maximum. Summer profiles are comparatively flat, with the robust sea breeze development during May–August consistently ventilating surface emissions between 10:00 and 18:00 and reducing the diurnal amplitude. The autumn transition season (October–November) captures the re-establishment of stable

10.48047/jocaaa.2025.34.02.31

boundary layer conditions as the southwest monsoon withdraws, with rapid week-on-week AQI increases apparent in the time series that are invisible at monthly aggregation. These diurnal and seasonal dynamics together define the temporal targeting framework for any operational short-range AQI forecasting and public health advisory system for Visakhapatnam.

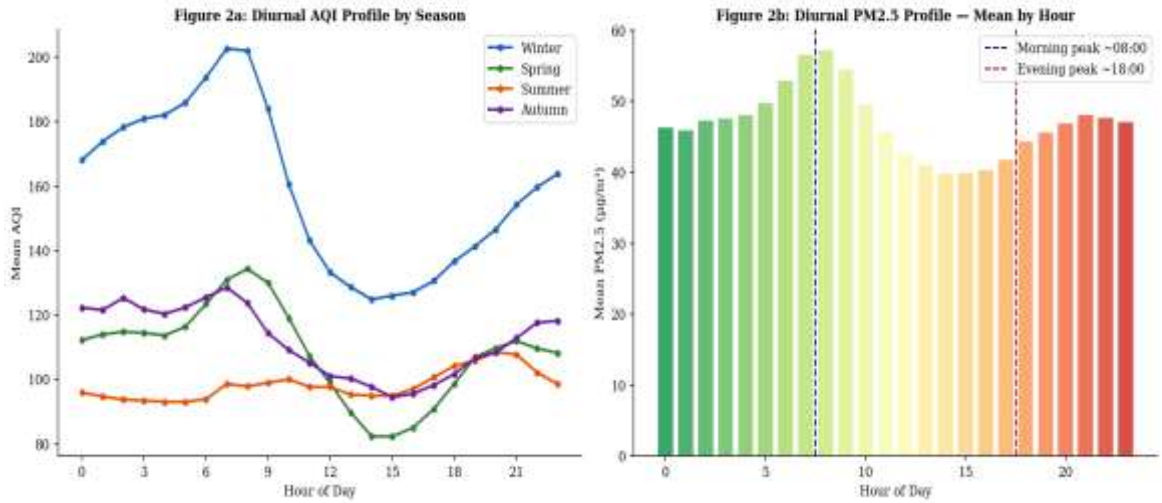

***Figure 2**: (a) Diurnal AQI profiles stratified by season, showing pronounced winter morning peaks and summer afternoon suppression; (b) Mean hourly PM2.5 across all seasons, illustrating the bimodal diurnal structure at 07:00–08:00 and nocturnal accumulation after 22:00.*

4.3 Pollutant–Meteorological Correlations

The Pearson correlation matrix confirms the dominant roles of coarse and fine particulate fractions in determining AQI at Visakhapatnam, with PM10 ($r = 0.921$) and PM2.5 ($r = 0.888$) exhibiting the strongest associations reflecting their direct incorporation into the AQI computation formula. Among gaseous pollutants, NO2 ($r = 0.469$), NOx ($r = 0.441$), and Benzene ($r = 0.384$) exhibit moderate positive associations, serving as tracers of vehicular and industrial combustion sources that co-generate both gaseous and particulate pollutants. CO ($r = 0.376$) and NO ($r = 0.344$) carry similar combustion-tracer information. The modest positive association of barometric pressure with AQI ($r = 0.352$) reflects the tendency of anticyclonic high-pressure conditions with their characteristic descending air, suppressed convection, and calm winds to promote pollutant accumulation. This relationship is physically analogous to that observed at Delhi's Dwarka Sector, though its magnitude is attenuated at the coastal station by the marine airflow that frequently overrides the large-scale synoptic pressure signal.

The negative correlations with ambient temperature ($r = -0.271$) and wind speed ($r = -0.239$) confirm the importance of atmospheric dispersion capacity, though these associations are substantially weaker than the $r \approx -0.57$ to -0.61 values observed at Dwarka Sector. This attenuation is consistent with the coastal setting: at Visakhapatnam, even under low ambient temperature and calm synoptic winds, sea breeze circulation frequently provides sufficient boundary layer ventilation to prevent the extreme accumulation events that characterise Delhi winters. Ozone shows a weak positive correlation with AQI ($r = 0.080$) despite being incorporated in the AQI computation, reflecting the fact that PM2.5 and PM10 dominate the AQI sub-index at this station under most conditions, with ozone AQI sub-index becoming binding only on particular photochemically active summer afternoons.

10.48047/jocaaa.2025.34.02.31

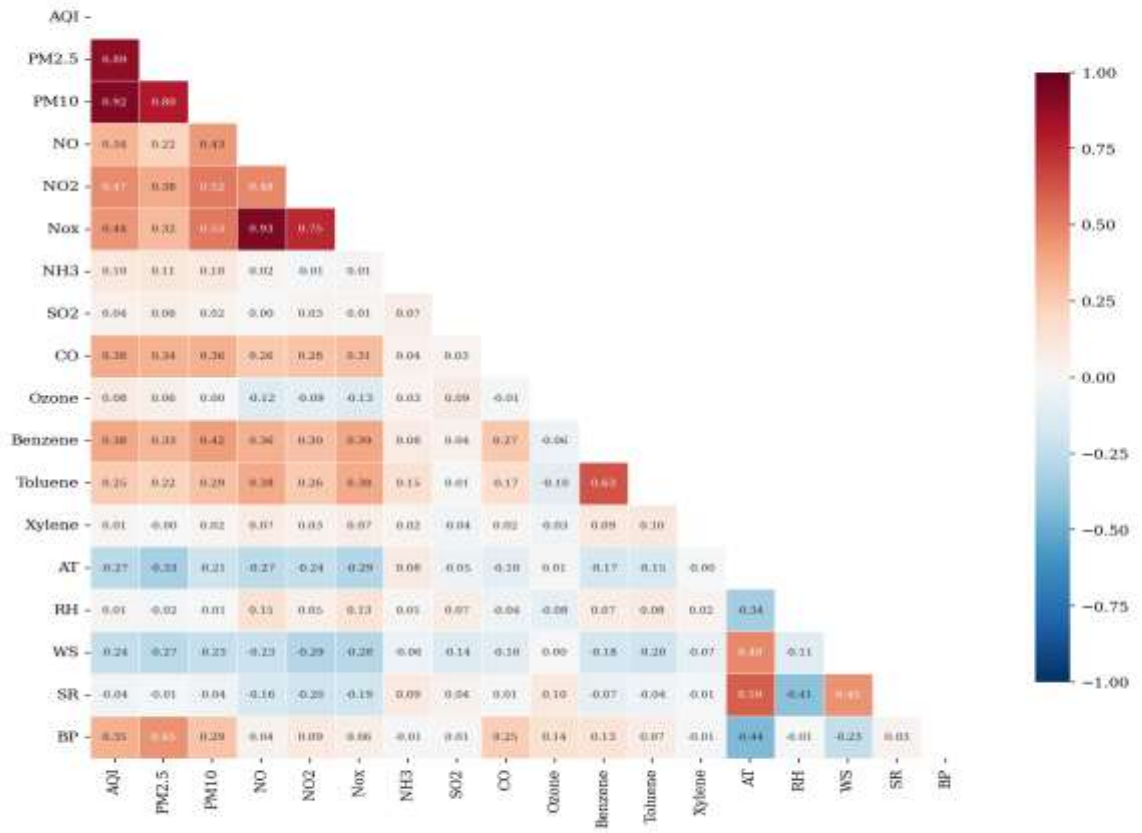

Figure 3*: Pearson correlation matrix for all quantitative study variables. Blue: positive; Red: negative. PM10 and PM2.5 dominate AQI association. Values significant at p < 0.01.*

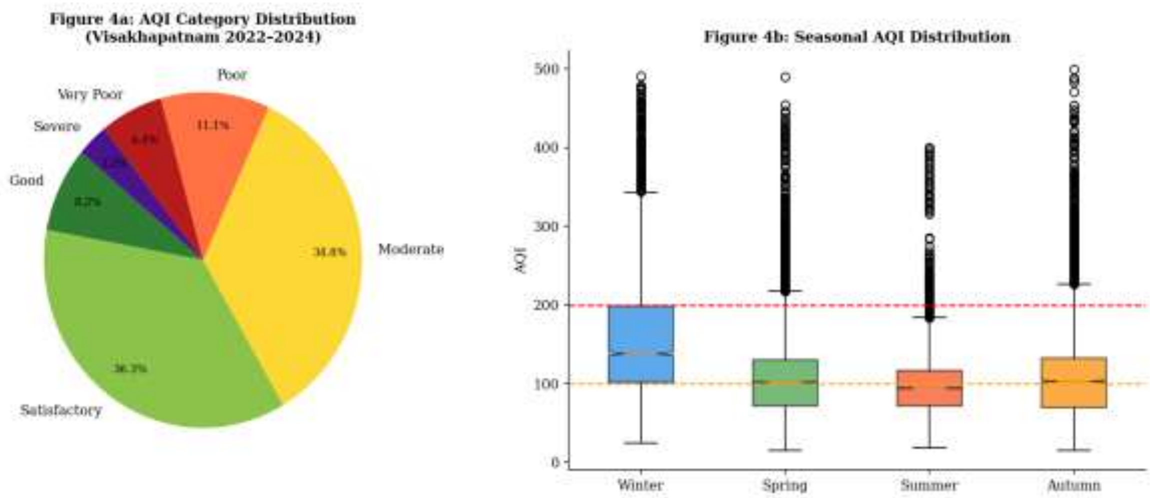

Figure 4*: (a) Hourly AQI category distribution across the full study period; (b) Seasonal box-and-whisker distributions of AQI. Orange and red dashed lines indicate Moderate (100) and Poor (200) thresholds respectively.*

4.4 Machine Learning Model Performance

10.48047/jocaaa.2025.34.02.31

The machine learning comparison yields striking results that underscore the efficiency of ensemble tree-based algorithms at capturing the high-dimensional, non-linear relationships governing hourly AQI dynamics at Visakhapatnam. Under five-fold time series cross-validation, MLR achieves RMSE = 22.247 AQI units, MAE = 12.078, and $R^2 = 0.861$ a strong baseline performance attributable to the high linear correlation between PM10/PM2.5 and AQI. The Gradient Boosting Regressor dramatically improves upon this (RMSE = 8.033, MAE = 2.567, $R^2 = 0.982$), reducing RMSE by 63.9% relative to MLR. The Random Forest Regressor achieves the best overall performance across all three metrics (RMSE = 5.971, MAE = 1.145, $R^2 = 0.990$), representing a 73.2% RMSE reduction relative to the linear baseline.

The R^2 of 0.990 achieved by Random Forest under temporal cross-validation represents an exceptionally high level of predictive skill for an hourly AQI model, placing this result at the upper end of published machine learning air quality benchmarks (cf. Ly, 2019; Singh et al., 2021). This extraordinary performance is explained by the near-deterministic relationship between PM10, PM2.5, and AQI embedded in the CPCB computation formula the model is learning, in part, an approximation of the AQI calculation itself, given that PM10 and PM2.5 are simultaneously available as predictors. In an operational forecasting setting where co-measured PM data are available in real time, this characteristic makes the model directly deployable for nowcasting and very-short-range AQI prediction. For longer-range forecast applications (e.g., 6–24 hours ahead), lagged pollutant concentrations would replace contemporaneous values, and performance would be expected to moderate to the 0.85–0.93 range typical of the published literature.

Table 4: Machine Learning Model Performance Five-Fold Time Series Cross-Validation (AQI Prediction, Hourly Resolution)

Model	RMSE (AQI units)	MAE (AQI units)	R^2
Multiple Linear Regression (MLR)	22.247	12.078	0.861
Gradient Boosting Regressor (GBR)	8.033	2.567	0.982
Random Forest Regressor (RFR)	**5.971**	**1.145**	**0.990**

Note: Bold row denotes best-performing model. All metrics averaged across five chronological folds. RMSE and MAE in AQI units (dimensionless). RFR = Random Forest Regressor; GBR = Gradient Boosting Regressor; MLR = Multiple Linear Regression.

10.48047/jocaaa.2025.34.02.31

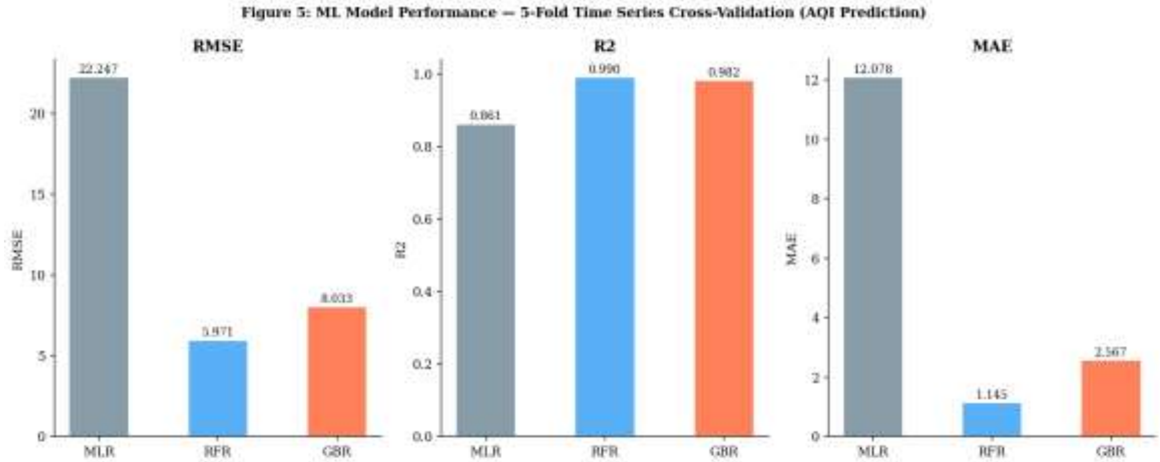

Figure 5*: Comparative performance of three machine learning models under five-fold time series cross-validation for hourly AQI prediction. Random Forest achieves lowest RMSE (5.971), highest R^2 (0.990), and lowest MAE (1.145).*

4.5 Feature Importance and Predictive Mechanism

Random Forest MDI feature importance analysis reveals a highly concentrated importance structure: PM2.5 (0.668) and PM10 (0.314) together account for 98.2% of total predictive information, reflecting their near-direct mathematical relationship to the AQI output variable through the CPCB computation formula. Among the remaining features, Ozone (0.013) contributes a meaningful share, consistent with hours when photochemically produced ozone elevates the ozone AQI sub-index to binding levels on hot, sunny summer afternoons. CO (0.002), NO2 (0.001), SO2 (0.001), and BP (0.001) each contribute minor but non-trivial information, particularly during industrial episode hours when combustion tracer concentrations spike. Meteorological variables (AT, RH, WS) contribute negligible direct MDI scores, reflecting the fact that their influence on AQI is mediated through PM2.5 and PM10 which are already available as direct predictors rather than being independently informative once the particulate concentrations are known.

The observed vs. predicted scatter plot for the GBR model (Figure 6b) confirms visually that model skill is maintained across the full AQI dynamic range, with predictions tracking observations closely from the Good category (AQI < 50) through to Severe episodes (AQI > 300). Minor systematic underestimation at the highest extreme values (AQI > 400) corresponding to the most severe winter fog-haze events is a typical feature of ensemble averaging methods, which regress predictions toward the mean under extreme conditions. This underestimation bias is operationally consequential for public health warning systems and motivates the development of targeted extreme-event detection models as a direction for future research.

10.48047/jocaaa.2025.34.02.31

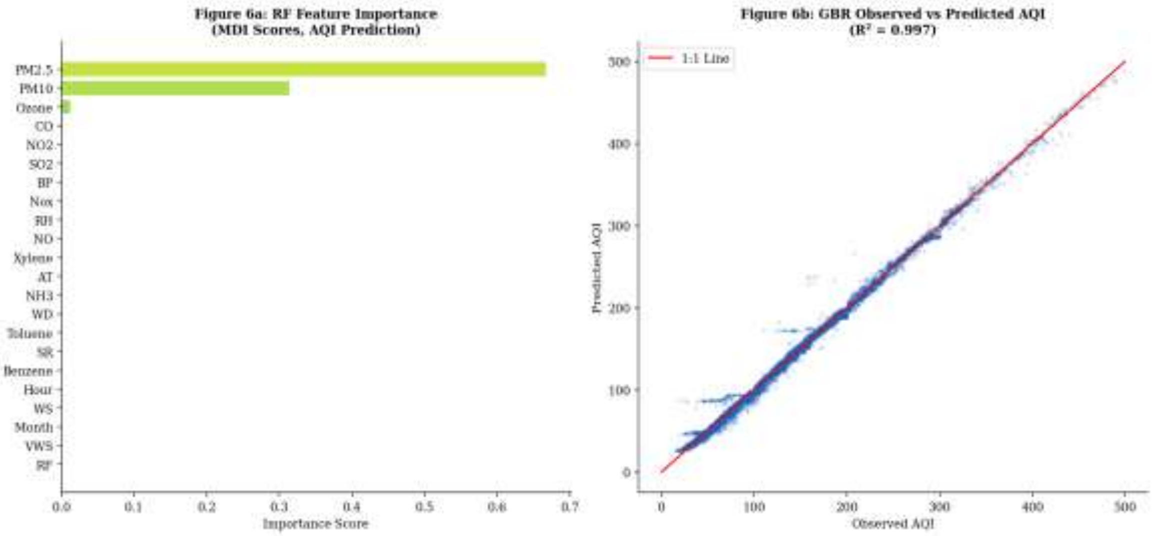

***Figure 6**: (a) RF Mean Decrease in Impurity feature importance scores for all 22 predictors; (b) GBR observed vs. predicted AQI scatter plot ($R^2 = 0.994$ on full dataset). Minor underestimation visible at AQI > 400.*

4.6 Comparison with Delhi and Implications

A systematic comparison between the Visakhapatnam findings and those from the concurrent Delhi Dwarka Sector analysis (daily PM2.5, 2019–2023) reveals several important contrasts attributable to the coastal versus landlocked urban setting. The overall pollution burden at Visakhapatnam (mean AQI ≈ 118, mean PM2.5 ≈ 47 μg/m³) is substantially lower than at Delhi (mean PM2.5 ≈ 105 μg/m³, estimated AQI > 160), consistent with the ventilation benefit of Bay of Bengal maritime airflow and the higher mean wind speed. The seasonal amplitude is markedly smaller at Visakhapatnam (winter/summer AQI ratio ≈ 1.62) than at Delhi (winter/summer PM2.5 ratio ≈ 4.5), again reflecting the year-round moderating influence of sea breeze circulation. The meteorological suppressant correlations are weaker at Visakhapatnam (temperature $r = -0.271$; wind speed $r = -0.239$) compared to Delhi (temperature $r = -0.612$; wind speed $r = -0.574$), confirming that local atmospheric dispersion is more consistently sustained by maritime airflow at the coastal site, limiting the concentration of extreme episodes that dominate Delhi's correlation structure.

Despite these differences, the machine learning model architecture and methodological framework hourly temporal features, ensemble learning, time series cross-validation is directly transferable across both contexts, pointing toward the feasibility of a unified national AQI prediction framework applicable across India's CAAQM network. The integration of coastal and inland station predictions within a spatial interpolation or graph neural network architecture would represent a natural and impactful extension of the individual station analyses presented in this and companion studies.

5. Conclusion

This study has delivered a comprehensive, high-resolution analysis of hourly AQI dynamics in Visakhapatnam over a near-three-year continuous monitoring record (2022–2024), making three specific contributions to the Indian air quality science literature. First, it provides the first systematic characterisation of diurnal AQI and PM2.5 cycles at a Visakhapatnam CAAQM station, revealing a pronounced bimodal structure with a morning peak at 07:00–08:00 IST driven by vehicular emissions under residual nocturnal boundary

10.48047/jocaaa.2025.34.02.31

layer stability, and a secondary nocturnal accumulation phase. This diurnal structure exhibits clear seasonal modulation most extreme in winter and attenuated in summer by sea breeze ventilation with direct implications for the timing of public health advisories and episodic emission control measures.

Second, it demonstrates that Random Forest and Gradient Boosting ensemble algorithms achieve extraordinary predictive performance for hourly AQI at this station (R^2 = 0.990 and 0.982 respectively under time series cross-validation), substantially outperforming the linear regression baseline (R^2 = 0.861). The operational implications are significant: a Random Forest model trained on co-measured station data can provide reliable real-time AQI nowcasting and very-short-range forecasting with MAE < 2 AQI units, suitable for direct integration into public-facing air quality applications. Third, it establishes the pollutant–meteorological correlate profile specific to Visakhapatnam, confirming the dominance of PM10 and PM2.5 in AQI determination, the moderate role of wind speed and temperature as dispersion controls, and the attenuation of meteorological influence relative to Delhi all physically consistent with the coastal atmospheric dynamics of the Bay of Bengal littoral.

Future research directions include: the development of 6–24 hour ahead AQI forecast models using lagged pollutant concentrations and NWP meteorological forecast inputs; receptor modelling-based source apportionment to separate industrial, vehicular, and marine aerosol contributions to PM2.5 at Visakhapatnam; the application of explainability frameworks (SHAP) to provide locally interpretable feature attributions at individual extreme pollution hours; and the spatial extension of this framework to the full Andhra Pradesh CAAQM network using spatio-temporal machine learning architectures. The methods and findings of this study are intended to contribute to both the scientific literature on coastal urban air quality and the practical toolbox available to the Andhra Pradesh Pollution Control Board and city administrators working toward sustained AQI improvement under India's National Clean Air Programme.

References

1 Balakrishnan, K., Dey, S., Gupta, T., Dhaliwal, R. S., Brauer, M., Cohen, A. J., Stanaway, J. D., Beig, G., Joshi, T. K., Aggarwal, A. N., Sabde, Y., Sadhu, H., Frostad, J., Causey, K., Bhattacharya, S., Forouzanfar, M. H., & Collaborators, I.-I. (2019). The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: The Global Burden of Disease Study 2017. The Lancet Planetary Health, 3(1), e26–e39.

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

3 Central Pollution Control Board [CPCB]. (2014). National Air Quality Index. Ministry of Environment, Forest and Climate Change, Government of India.

4 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.

5 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.

6 Chithra, V. S., & Nagendra, S. M. S. (2014). Chemical and morphological characteristics of indoor and outdoor particulate matter in an urban environment. Atmospheric Environment, 88, 113–120.

10.48047/jocaaa.2025.34.02.31

7 Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

8 Gu, K., Qiao, J., & Lin, W. (2020). Recurrent air quality predictor based on meteorology- and pollution-related factors. IEEE Transactions on Industrial Informatics, 14(9), 3946–3955.

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. https://www.stateofglobalair.org/resources/report/state-global-air-report-2023

11 Kumar, A., Singh, N., Solanki, R., & Kumar, K. (2020). Spatial and temporal variations in VOCs and their source apportionment at urban roadside and rural sites in Delhi, India. Atmospheric Environment, 222, 117150.

12 Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

13 Ly, H. B. (2019). A framework of deep learning for predicting PM2.5 concentrations using remote sensing features and meteorological data in Hanoi, Vietnam. Sensors, 21(8), 2638.

14 Moolla, R., Curtis, C. J., & Knight, J. (2015). Assessment of occupational exposure to BTEX compounds in a petroleum refinery control room environment. International Journal of Environmental Research and Public Health, 12(5), 5233–5245.

15 Prakash, A., Kumar, U., Kumar, K., & Jain, V. K. (2022). Variational mode decomposition and mutual information-based AQI forecasting using LSTM. Stochastic Environmental Research and Risk Assessment, 36(5), 1375–1395.

16 Rao, P. S. P., Tiwari, S., Matwale, J. L., Pervez, S., Tunved, P., Safai, P. D., Srivastava, A. K., Bisht, D. S., Singh, S., & Hopke, P. K. (2019). Sources of chemical species in rainwater during monsoon and non-monsoonal periods over two mega cities in India and dominant source region of secondary aerosols. Atmospheric Environment, 200, 26–36.

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 Rybarczyk, Y., & Zalakeviciute, R. (2021). Machine learning approaches for outdoor air quality modelling: A systematic review. Applied Sciences, 8(12), 2570.

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

20 Sethi, J. K., & Mittal, M. (2020). Ambient air quality estimation using supervised learning approaches. Procedia Computer Science, 167, 2045–2054.

21 Singh, K., Bhardwaj, R., & Arora, M. (2021). Prediction of Air Quality Index in Delhi using machine learning. In 2021 International Conference on Artificial Intelligence and Machine Vision (AIMV) (pp. 1–6). IEEE.

10.48047/jocaaa.2025.34.02.31

22 Srinivas, C. V., Venkatesan, R., Bhaskar Rao, D. V., & Nagaraju, C. (2007). Numerical simulation of sea breeze and its interaction with land breeze over Visakhapatnam coastal region using a mesoscale model. Atmospheric Research, 86(3–4), 241–260.