SPATIO-TEMPORAL DYNAMICS OF EXTREME PM2.5 IN INDONESIA: A WEATHER-BASED HYBRID MODELING APPROACH

Dwi Indriyati[1,2], Mochamad Aryono Adhi[2*], and Adi Mulsandi[3]

[1]Agency for Meteorology Climatology and Geophysics, Jakarta 10720, Indonesia

[2]Physics Study Program, Universitas Negeri Semarang, Semarang 50229, Indonesia

[3] Meteorology Study Program, School of Meteorology, Climatology, and Geophysics (STMKG), Tangerang Kota 15119, Indonesia

***E-mail:** *aryono_adhi@yahoo.com*

Received: May 5, 2025 Reviewed: July 7, 2025 Accepted: September 23, 2025

ABSTRACT

Air pollution due to fine particulate matter (PM2.5) has emerged as a critical public health concern in tropical megacities, where rapid urbanization outpaces environmental regulation. Indonesia, with limited air quality monitoring and complex meteorology, remains particularly vulnerable to extreme pollution episodes. This study investigates the spatio-temporal dynamics of extreme PM2.5 concentrations in three Indonesian cities: Kemayoran, Semarang, and Malang. These cities were selected for their contrasting urban morphologies, topographic settings, and meteorological regimes. Using hourly PM2.5 observations (2022–2024) combined with ERA5 reanalysis data, we examine how atmospheric conditions govern pollution variability across both space and time. A hybrid modeling framework was implemented, integrating multiple linear regression with nonlinear algorithms (Random Forest and XGBoost) to predict hourly PM2.5 from meteorological variables including temperature, dew point, wind speed, precipitation, and surface pressure. Machine learning models outperformed linear methods, with Random Forest offering a strong balance between accuracy and interpretability. Wind speed emerged as the most consistent predictor, followed by dew point and precipitation, with notable spatial and seasonal variation. Extreme PM2.5 episodes,defined as hourly concentrations above the 95th percentile, were most frequent during dry and transitional seasons under stagnant, humid, and low-rainfall conditions. Kemayoran exhibited the highest concentrations, while Malang, despite lower emissions, showed susceptibility due to weak ventilation. These results underscore the central role of meteorological stagnation in driving pollution extremes and demonstrate the utility of hybrid models for developing localized early-warning systems in tropical urban contexts.

Keywords: PM2.5, air pollution, Tropical climate, Spatio-temporal analysis, Meteorological drivers, Hybrid modelling.

1. Introduction

Urban air pollution, particularly fine particulate matter (PM2.5), has emerged as a critical global public health issue. Numerous studies have linked elevated PM2.5 exposure to respiratory and cardiovascular diseases, premature mortality, and decreased life expectancy [1]–[6]. While PM2.5 is a pervasive problem in both developed and developing regions, its impacts are especially severe in rapidly urbanizing parts of the Global South, where environmental governance and monitoring remain limited.

In the context of tropical countries such as Indonesia, the challenge of managing PM2.5 pollution is intensified by complex meteorological dynamics and heterogeneous urban structures. Unlike temperate regions where seasonal patterns are more predictable and emission profiles relatively stable, tropical urban environments are characterized by high atmospheric variability, localized convection, mesoscale circulations, and frequent transitions between wet and dry regimes [7]–[10]. These dynamics interact with anthropogenic emissions in ways that complicate pollution forecasting and mitigation strategies.

Several studies have documented the diurnal and seasonal behaviour of PM2.5 in urban settings. For example, bimodal daily peaks, typically during early morning and late evening, have been attributed to traffic emissions and boundary-layer suppression[11]–[14]. Likewise, dry-season pollution maxima are often associated with stagnant air masses and reduced precipitation [15], [16]. Meteorological parameters such as wind speed and rainfall have been widely studied for their roles in pollutant dispersion and scavenging [17]–[19]. However, in tropical settings, the relationship between weather and PM2.5 is less consistent. Spatial heterogeneity in topography and urban form, as well as convective disturbances, introduce lags and nonlinearities that weaken the predictive strength of single variables [17], [18], [20].

Despite this growing body of work, key limitations remain. First, the majority of studies are confined to individual cities, which limits their generalizability. Second, there is insufficient examination of how contrasting urban typologies, such as coastal vs. inland, lowland vs. upland, modulate the expression of canonical PM2.5 patterns. For example, sea-breeze systems may promote ventilation in coastal cities, but their inland penetration can be obstructed by dense urban canopies [21], [22]. Similarly, cities at higher elevations may have lower emissions yet remain vulnerable to stagnation events due to weak vertical mixing [23]–[25].

These limitations reveal a critical research gap: the need for comparative, spatio-temporal assessments that evaluate how PM2.5 dynamics vary across diverse urban and topographic contexts in tropical regions. Existing models often underperform in such environments due to their reliance on linear assumptions and their lack of localized meteorological nuance. Moreover, the scarcity of data from under-monitored tropical cities hampers the development of robust early-warning systems.

To address these issues, this study investigates the spatio-temporal dynamics of PM2.5 in three Indonesian urban centers, Kemayoran (Jakarta), Semarang, and Malang, during the 2022–2024 period. These cities were selected to represent distinct urban morphologies and meteorological regimes. Specifically, the research aims to: (1) characterize diurnal and seasonal pollution patterns; (2) assess meteorological correlations and lag effects; (3) evaluate the predictive performance of multiple modeling approaches, including Multiple Linear Regression, Random Forest, and XGBoost; and (4) identify the meteorological signatures of extreme pollution episodes (above the 95th percentile).

By integrating observational analyses with hybrid machine learning models, the study offers both theoretical and practical contributions. It enhances understanding of the meteorological mechanisms shaping air quality in tropical cities and provides empirical foundations for localized forecasting frameworks. These insights are especially valuable for regions like Southeast Asia, where climate variability and urban expansion intersect to produce heightened vulnerability to pollution extremes.

2. Methods

Study Area and Data Collection. The study focuses on three major urban centres in Indonesia: Jakarta (Kemayoran station), Semarang, and Malang (Figure 1). These cities are located within the tropical monsoonal climate regime of Java Island, yet they exhibit significant contrasts in population density, topographic setting, and meteorological characteristics. According to the DKI Jakarta in Figures 2023 report published by Statistics Indonesia (BPS Jakarta, https://jakarta.bps.go.id), Jakarta has a population of approximately 10.7 million people, with a registered motor vehicle fleet exceeding 23 million units. This results in a vehicle density of over 4,600 units per square kilometre. The city is widely recognized as one of the most polluted capitals in Southeast Asia, with annual average PM2.5 levels ranging between 39 and 53 µg/m³ based on global air quality summaries from IQAir (https://www.iqair.com).

In comparison, according to the Semarang in Figures 2023 report, the city has a total population of around 1.7 million and a lower degree of motorization, typical of medium-sized coastal cities (BPS Semarang, https://semarangkota.bps.go.id). Air quality measurements in Semarang indicate seasonal PM2.5 levels typically ranging between 25 and 35 µg/m³, with observable peaks during the dry season. Meanwhile, Malang in Figures 2023 report states a population of approximately 900,000. The city is characterized by a topographic elevation between 450–550 meters above sea level, cooler average temperatures, and reduced traffic density (BPS Malang, https://malangkota.bps.go.id). Despite its relatively lower emission baseline, Malang remains vulnerable to stagnant air conditions due to topographic enclosure, resulting in occasional PM2.5 concentrations ranging from 15 to 25 µg/m³.

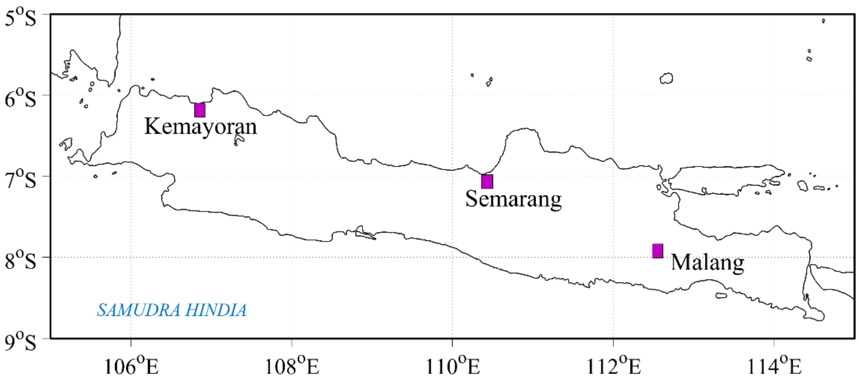

Figure 1. The three study sites: Kemayoran (Jakarta), Semarang, and Malang, are situated on the island of Java, Indonesia. All sites are located within a tropical monsoonal climate regime but they differ in terms of urban density, geographic setting, and elevation.

This combination of demographic, geographic, and climatic differences allows for a comparative analysis of PM2.5 behaviour across diverse urban typologies within a unified tropical climate regime, enhancing the relevance of the findings for other Southeast Asian cities.

The monitoring instruments employed in this study adhere to national and international air quality monitoring standards, with regular calibration and quality assurance protocols in place to ensure data reliability. Quality control procedures included automated system diagnostics, manual verification of data anomalies, and exclusion of periods flagged for instrument maintenance or operational irregularities.

The resulting dataset comprises continuous hourly PM2.5 records across the three sites, forming a high-resolution basis for examining diurnal, seasonal, and meteorologically driven variations in particulate pollution.

Meteorological data corresponding to the study period were obtained from the ERA5 reanalysis dataset, produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) through the Copernicus Climate Change Service. ERA5 provides hourly estimates of a wide range of atmospheric, land, and oceanic parameters at a horizontal resolution of approximately 31 km, based on a consistent assimilation of observations into a modern numerical weather prediction model.

For this study, hourly data on 2-m air temperature, 2-m dew point temperature, 10-m u- and v- components of wind, surface pressure, and total precipitation were extracted for the grid points nearest to each monitoring station. The ERA5 data were accessed through the Copernicus Climate Data Store (https://cds.climate.copernicus.eu/), and temporal alignment with the PM2.5 observations was performed to ensure consistency in the subsequent analyses.

The use of ERA5 reanalysis enables a physically consistent representation of meteorological conditions across the three study sites, facilitating the examination of atmospheric drivers influencing PM2.5 variability.

Data Pre-processing and Quality Control. Prior to analysis, both PM2.5 and meteorological datasets underwent a series of pre-processing steps designed to ensure consistency and analytical integrity. For each monitoring site, the completeness of the PM2.5 record was first assessed by calculating the proportion of available hourly data relative to the total expected number of observations (26,304 hours for the three-year period). Summary statistics of data issues, including missing values, negative concentrations, and unrealistically high readings, are presented in Table 1. The proportion of problematic records ranged from 3.65% in Kemayoran to 4.46% in Malang. PM2.5 records were subsequently screened for quality. Implausible values, such as negative concentrations, values exceeding 500 μg/m³, and placeholder entries (e.g., 9999.9), were flagged as invalid. These problematic values, along with missing data, were imputed using linear time-based interpolation applied within local temporal windows. Given that these proportions fall below the commonly accepted threshold of 5–10% for acceptable missingness in environmental time series [26]–[28], linear interpolation was deemed suitable for preserving short-term variability without distorting broader trends. After interpolation, only days with at least 75% valid hourly observations were retained for subsequent analyses to further ensure data robustness.

Meteorological variables from ERA5 were temporally aligned with the PM2.5 observation timestamps. Given the slight time-zone differences across the study sites (Western Indonesian Time, UTC+7), local time adjustments were applied to ensure precise temporal matching. Missing values in the meteorological dataset were negligible due to the assimilated nature of reanalysis products and therefore did not require imputation.

To prepare for statistical and machine learning modelling, all predictor variables were standardized by mean-centering and scaling to unit variance. This pre-processing step facilitates numerical stability during model training and enhances the interpretability of model coefficients and feature importance metrics.

Table 1. Summary of PM2.5 data quality issues identified prior to interpolation. Percentages represent the proportion of problematic records relative to the total number of expected hourly observations (26,304 per site)

City	Missing	Negative	>500 μg/m³	Total Flagged	% of Total
Kemayoran	470	19	471	960	3.65%
Semarang	1,095	8	64	1,167	4.44%
Malang	672	69	432	1,173	4.46%

Descriptive statistical analyses. Descriptive statistical analyses were conducted to characterize the temporal variability of PM2.5 concentrations across the three study sites. Key metrics such as mean, median, standard deviation, interquartile range (IQR), and maximum concentrations were computed on hourly, daily, and seasonal bases. Diurnal cycles were visualized using composite mean plots, while seasonal variations were examined through time series decomposition and comparative boxplot visualizations to identify recurrent patterns and anomalies.

To evaluate the statistical association between PM2.5 concentrations and meteorological variables, both Pearson's product-moment correlation coefficient (r) and Spearman's rank-order correlation coefficient (ρ) were calculated using the complete hourly time series data for each site. Pearson correlation was applied to capture linear relationships and is defined mathematically as:

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \cdot \sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (1)$$

where x_i and y_i represent the individual paired observations of PM2.5 and the meteorological variable, and $\bar{x}, \bar{y}$ are their respective means.

Spearman correlation, on the other hand, was used to detect monotonic but potentially nonlinear associations. It is computed based on the ranks of the values rather than their raw magnitudes, as follows:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (2)$$

where d_i is the difference between the ranks of paired observations and n is the number of valid pairs.

Statistical significance of the correlation coefficients was evaluated using p-values derived from two-tailed hypothesis testing. A threshold of $p < 0.05$ was used to determine statistical significance, indicating less than 5% probability that the observed correlation occurred by chance.

In addition to contemporaneous relationships, lagged correlation analyses were performed to capture potential delayed effects of meteorological variables on PM2.5 concentrations. Hourly lag intervals of up to 24 hours were considered for each meteorological parameter, and Pearson correlations were recalculated at each lag step. This approach enabled identification of short-term temporal dependencies, particularly relevant for variables like precipitation, whose impacts on pollutant concentrations may not be immediate.

Lastly, extreme PM2.5 events were defined as hourly concentrations exceeding the 95th percentile of the distribution at each site. These events were isolated and examined in relation to concurrent meteorological conditions to identify atmospheric patterns associated with high-pollution episodes. The analysis aimed to reveal compound effects beyond average behavior, offering insights into episodic risk conditions relevant for early-warning systems.

Hybrid Modelling Framework. To model the relationship between meteorological variables and PM2.5 concentrations, a hybrid statistical–physical framework was developed, integrating conventional regression analysis with machine learning techniques.

As a baseline, multiple linear regression (MLR) models were constructed to quantify the linear associations between standardized meteorological predictors and hourly PM2.5 concentrations at each site. Predictor variables included 2-m air temperature, 2-m dew point temperature, 10-m wind speed and direction (converted from u- and v- components), surface pressure, and total precipitation. Multicollinearity among predictors was assessed using the variance inflation factor (VIF), and stepwise selection procedures based on the Akaike Information Criterion (AIC) were applied to optimize model parsimony.

In parallel, ensemble machine learning models were developed, specifically Random Forest (RF) regression and Extreme Gradient Boosting (XGBoost), to capture potential nonlinear and interaction effects between meteorological variables and PM2.5 concentrations [29], [30], [39], [40], [31]–[38]. Hyperparameter tuning was performed using grid search combined with five-fold cross-validation to minimize overfitting and enhance generalization performance.

Model performance was evaluated using standard metrics, including the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE), computed on an independent test subset comprising 30% of the data. Feature importance scores derived from the ensemble models were analysed to identify the dominant meteorological drivers influencing PM2.5 variability across different temporal scales.

This hybrid approach allows for a robust exploration of both linear and nonlinear relationships between atmospheric conditions and particulate pollution, providing a more comprehensive understanding of PM2.5 dynamics in tropical urban environments.

3. Result

This section examines the spatio-temporal variability of PM2.5 concentrations across the three study sites—Kemayoran, Semarang, and Malang—focusing on diurnal and seasonal patterns observed during the 2022–2024 period. The analysis aims to identify consistent temporal features, such as peak hours and seasonal maxima, while also exploring how these patterns differ spatially across urban typologies and local meteorological regimes. By comparing daily cycles and seasonal fluctuations among cities with distinct geographic and atmospheric characteristics, this subsection provides a baseline for understanding how time-varying pollution dynamics interact with site-specific conditions. Figure 2 presents the average diurnal variation of PM2.5 concentrations across the three study sites, following local time from 00:00 to 23:00.

Spatio-Temporal Characteristics of PM2.5. All locations exhibit a distinct bimodal pattern, with higher concentrations typically occurring during the early morning hours (approximately 06:00–09:00) and again in the late evening to early nighttime (20:00–01:00). The morning peak corresponds with increased vehicular emissions during rush hour combined with a shallow nocturnal boundary layer that inhibits vertical mixing. As the sun rises and surface heating intensifies, convective turbulence promotes vertical dispersion, leading to a pronounced decline in PM2.5 concentrations by midday. This midday minimum, often observed between 11:00 and 15:00, is a common feature in tropical urban environments where solar radiation is intense and promotes mixing. The second peak, typically observed during late evening, arises from a combination of reduced boundary layer height, lower wind speeds, and continued human activity such as cooking, residential heating, or biomass burning. Among the three cities, Kemayoran shows the highest overall concentrations and most pronounced diurnal contrast, reflecting its dense urban structure and higher emission load. Malang, by contrast, maintains relatively lower concentrations throughout the day, likely due to its elevated inland location and stronger afternoon dispersion conditions.

The diurnal PM2.5 patterns observed in this study—characterized by peaks during early morning and late evening—are consistent with those documented in various urban air quality studies, where anthropogenic emissions and suppressed boundary layer mixing are identified as key drivers [11]–[13]. Nevertheless, while these temporal structures appear recurrent across urban areas, our findings reveal notable spatial heterogeneity in their manifestation, shaped by differences in topography, urban density, and local meteorological regimes. Such intra-regional variation is often overlooked in prior studies, which tend to generalize PM2.5 behavior without sufficient attention to geographic specificity. By contrasting three distinct urban settings, this study highlights the importance of contextual factors in modulating pollution dynamics and offers empirical insights that broaden the applicability of diurnal PM2.5 analyses to a wider range of climatic and urban typologies.

The observed bimodal diurnal pattern of PM2.5 concentrations in Malang and Semarang, characterized by morning and evening peaks, aligns with findings from other tropical urban areas. For instance, [14] reported similar bimodal patterns in Padang, Manado, Palu, and Pangkalpinang, attributing these peaks to morning traffic emissions and evening atmospheric stability. Conversely, Jakarta exhibited a unimodal pattern with a single peak at night until morning, likely due to its unique urban dynamics and meteorological conditions. These variations underscore the influence of local factors such as traffic density, boundary layer dynamics, and urban morphology on diurnal PM2.5 patterns [14].

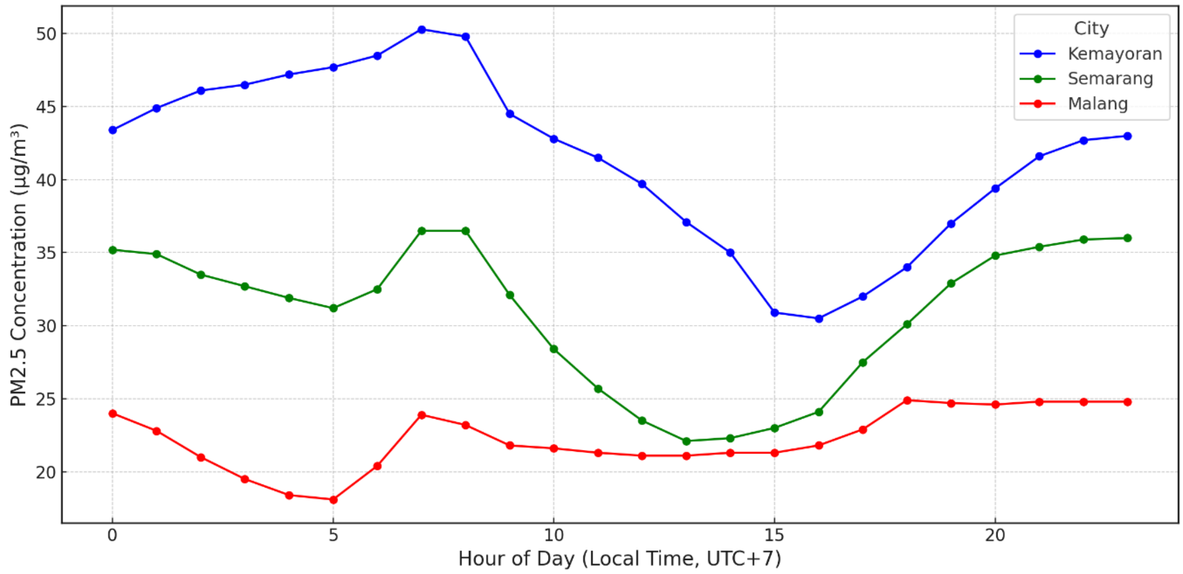

Figure 2. Average diurnal variation of PM2.5 concentrations in Kemayoran (Jakarta), Semarang, and Malang over the 2022–2024 period, based on local time (UTC+7). The figure follows the standard diurnal cycle convention from 00:00 to 23:00.

Figure 3 illustrates the seasonal variation of PM2.5 concentrations in the three study locations, grouped according to standard climatological seasons in the tropics: DJF (wet season), MAM (transition I), JJA (dry season), and SON (transition II). Across all cities, the lowest PM2.5 concentrations are observed during the DJF period, which coincides with increased rainfall frequency and intensity. This condition promotes wet scavenging and facilitates the removal of aerosols from the boundary layer, aligning with expectations in tropical climates.

The seasonal arc of PM2.5 concentrations across the three cities offers an initial impression of symmetry, lower in the wet season, higher in the dry, but that symmetry begins to fracture upon closer inspection. In Semarang, the absence of a strong seasonal swing invites inquiry: is this consistency a sign of atmospheric resilience, or does it mask a more erratic day-to-day variability driven by mesoscale processes? Meanwhile, the sharper seasonal divides in Kemayoran and Malang may reflect not only climatic influences but also divergent urban–atmospheric relationships. The inland location of Kemayoran appears to decouple it from maritime ventilation rhythms that moderate pollution in Semarang, while Malang's elevated inland topography might enhance dry-season entrapment more than dispersion.

The seasonal pattern observed in Kemayoran, with pronounced PM2.5 peaks during the dry season (JJA), is consistent with previous studies in Southeast Asian megacities, where reduced precipitation and persistent boundary layer suppression contribute to pollutant accumulation [15], [16]. These findings are also echoed in other Indonesian urban centers such as Padang and Manado, where [14] reported elevated dry-season concentrations due to stagnant air masses and biomass burning influences.

Conversely, Malang exhibits relatively lower PM2.5 levels year-round, though with a mild peak in the dry season. This stability may be attributed to its elevated topography and frequent convective mixing, which enhances aerosol dispersion. Similar patterns have been noted in other highland cities [23]–[25], where terrain-driven turbulence offsets seasonal stagnation.

Semarang presents an anomalous profile, with PM2.5 levels that do not show strong seasonal modulation. While coastal positioning and flat terrain might suggest enhanced ventilation from sea breeze systems, the muted seasonal swing may reflect compensating effects from persistent urban emissions or irregular convective activity. This is partially aligned with findings from coastal cities, where mesoscale circulations introduce variability that disrupts expected seasonal patterns [21], [22].

These observations reinforce the importance of spatial context in interpreting seasonal air quality patterns. A uniform classification of wet and dry seasons alone is insufficient to explain the pollution dynamics across diverse urban landscapes. Instead, local geography, topographic relief, and mesoscale meteorology must be considered to develop a nuanced understanding and inform location-specific air quality interventions. In addition to meteorological conditions, local emission sources, such as traffic, industry, and residential burning, also play a significant role in shaping PM2.5 levels and should be acknowledged as key contributors, even if not analyzed in detail within this study.

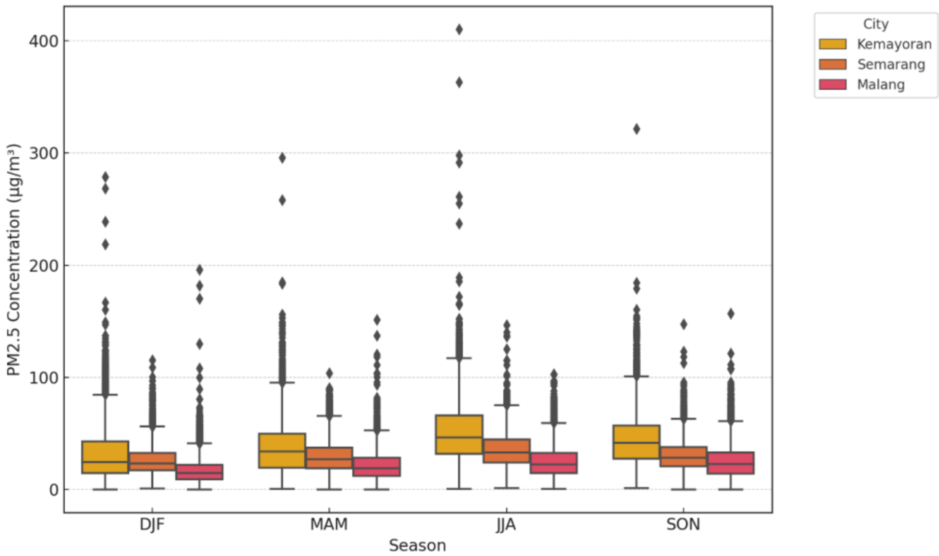

Figure 3. Seasonal variation of PM2.5 concentrations in Kemayoran (Jakarta), Semarang, and Malang, grouped by climatological season: DJF (wet season), MAM (transition I), JJA (dry season), and SON (transition II). Higher concentrations during the dry season reflect weaker atmospheric cleansing and increased pollutant accumulation.

Meteorological Influences on PM2.5. Table 2 summarizes the Pearson and Spearman correlation coefficients between hourly PM2.5 concentrations and five key meteorological variables: air temperature, dew point, wind speed, surface pressure, and precipitation, measured across Kemayoran, Semarang, and Malang. The results reveal some recurring patterns, but also city-specific deviations that merit closer scrutiny.

Wind speed consistently shows a weak to moderate negative correlation with PM2.5 across all three locations (Pearson $r \approx -0.24$ to -0.26, $p < 0.05$). Although the effect size is modest, its consistency is noteworthy, especially given the diverse topographic and urban morphologies involved. In Malang, the inland topography likely amplifies the role of ventilation in regulating surface pollution. Comparable inverse correlations have been reported in both tropical and subtropical cities, such as Bangkok [17] and Ho Chi Minh City [18], where stagnant conditions have been linked with sharp PM build-up.

Temperature and dew point, on the other hand, yield inconsistent correlations. In Semarang and Malang, the relationships with PM2.5 are slightly negative but statistically weak ($r < 0.15$). In Kemayoran, the signal vanishes altogether. This weak coupling challenges assumptions derived from temperate-climate studies, where temperature often emerges as a reliable proxy for boundary-layer depth or photochemical activity. In tropical regions with frequent convective overturning, such assumptions may not hold. Shafi and Scafetta [20] observed similar patterns across Southeast Asia, suggesting that humidity and temperature alone are insufficient to explain short-term PM dynamics under tropical meteorological regimes.

Precipitation shows a marginal but statistically significant negative correlation in Malang ($r = -0.10$, $p < 0.05$), and even weaker in the other cities. This likely reflects the short-lived and spatially patchy nature of convective rainfall in Indonesia. While rainfall is commonly regarded as a natural "cleanser" for the atmosphere, its actual impact on hourly PM2.5 concentrations appears to be diluted unless the precipitation is both intense and prolonged. These findings are consistent with observational studies in northern Thailand, where only multi-hour rainfall episodes showed appreciable PM reduction [17].

Surface pressure correlations hover near zero and lack clear physical interpretability. This is not unusual, as pressure often acts as a background variable and is rarely a direct predictor of air pollution without being linked to large-scale weather regimes.

Altogether, the correlation structure presented here diverges from more deterministic patterns observed in temperate or high-latitude settings. In the urban tropics, pollutant levels appear less directly tied to single meteorological variables and more shaped by the interaction of urban emissions with mesoscale processes like land-sea breezes, convective pulses, and topographic confinement. This underscores the importance of tailored predictive models that go beyond bivariate dependencies, incorporating both spatial dynamics and temporal lags.

Table 2. Pearson and Spearman correlation coefficients between PM2.5 and selected meteorological variables for each study site. Correlations are calculated using hourly data from 2022 to 2024. All values are rounded to three decimal places.

City	**Variable**	**Pearson r**	Pearson p-value	**Spearman r**	Spearman p-value
Kemayoran	Temperature (°C)	0.011	1.69E-160	0.007	1.71E-106
Kemayoran	Dew Point (°C)	0.013	6.01E-85	0.01	3.13E-115
Kemayoran	Wind Speed (m/s)	-0.24	0.00E+00	-0.24	0.00E+00
Kemayoran	Precipitation (mm)	-0.057	4.39E-40	-0.053	4.17E-54
Kemayoran	Surface Pressure (hPa)	0.047	2.31E-112	0.038	1.05E-102
Semarang	Temperature (°C)	-0.034	4.12E-130	-0.027	1.53E-104
Semarang	Dew Point (°C)	0	2.70E-80	0.006	3.23E-95
Semarang	Wind Speed (m/s)	-0.248	2.77E-178	-0.238	3.47E-198
Semarang	Precipitation (mm)	-0.047	4.42E-45	-0.045	1.58E-51
Semarang	Surface Pressure (hPa)	0.063	1.28E-46	0.049	6.80E-36
Malang	Temperature (°C)	-0.145	5.47E-84	-0.122	1.53E-52
Malang	Dew Point (°C)	-0.09	6.18E-32	-0.067	5.97E-30
Malang	Wind Speed (m/s)	-0.26	5.57E-86	-0.233	2.65E-91
Malang	Precipitation (mm)	-0.103	3.50E-39	-0.089	1.65E-41
Malang	Surface Pressure (hPa)	0.053	3.14E-89	0.043	6.89E-86

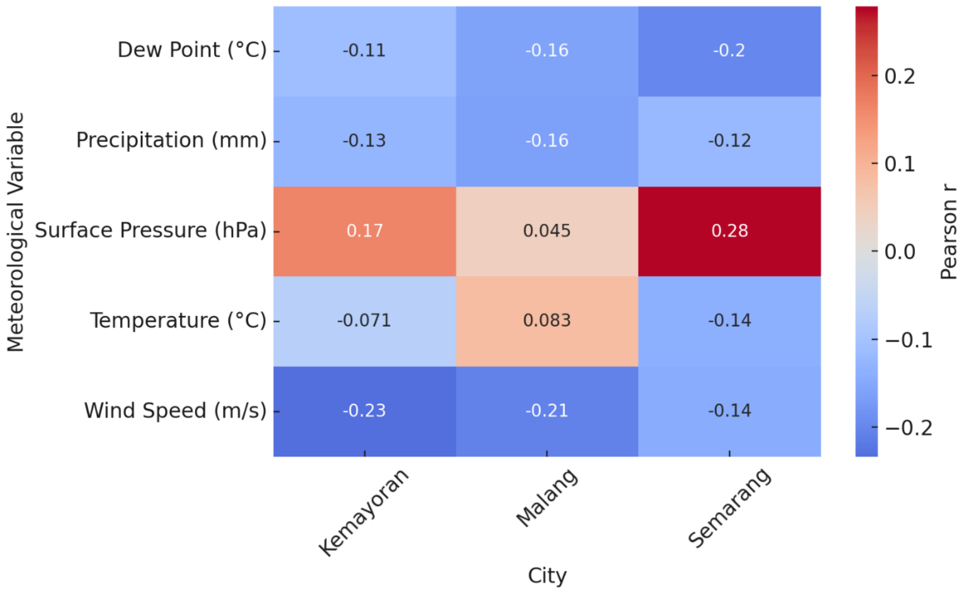

Figure 4. Pearson correlation heatmap between PM2.5 and key meteorological variables across three study sites. Wind speed consistently shows negative correlations, while temperature, dew point, and surface pressure have minimal or inconsistent relationships with PM2.5 levels.

Figure 4 provides a visual summary of the Pearson correlation coefficients between PM2.5 and key meteorological variables. The heatmap highlights the most consistent relationship across all cities: a negative correlation between PM2.5 and wind speed, reinforcing the role of wind in facilitating pollutant dispersion.

In Malang, the negative association with wind speed appears strongest, likely due to its inland and elevated geography where calm conditions can significantly exacerbate pollutant accumulation. Conversely, other meteorological variables such as temperature, dew point, and surface pressure show weak and inconsistent correlations across the three cities, further emphasizing the complex and site-specific nature of PM2.5 variability.

This heatmap complements the numerical values presented in Table 2, offering a rapid and intuitive means to assess which variables may serve as potential predictors in subsequent modeling efforts.

Temporal Lag Analysis of Meteorological Variables. To complement the static correlation analysis, we explored the delayed or cumulative effects of meteorological conditions on PM2.5 concentrations. Simple correlation coefficients, while informative, may overlook such temporal dependencies. For instance, rainfall may not immediately remove particulates but could exert a lagged impact due to scavenging processes, runoff delays, or delayed boundary-layer responses. To capture this dimension, we conducted a lagged Pearson correlation analysis across 24-hour windows for two key meteorological drivers: wind speed and precipitation.

Figure 5 illustrates the lagged correlation matrices for all three cities. Wind speed exhibits a clear and immediate inverse relationship with PM2.5, peaking at lag 0 in all sites. This supports the interpretation that increased wind enhances vertical and horizontal dispersion, thereby lowering near-surface concentrations in real time. The effect is most pronounced in Malang, likely due to its inland topography and restricted ventilation pathways, which makes the atmosphere especially sensitive to even modest wind events. Similar patterns have been documented in northern Thailand [17], where topographic basins amplify the role of local wind conditions, confirming that upland cities may be disproportionately affected by even small shifts in ventilation.

In contrast, precipitation demonstrates a delayed effect. The most significant negative correlations appear between lags 2 to 6 hours in Kemayoran and Malang. This suggests that wet scavenging processes take time to meaningfully reduce airborne particulates, possibly requiring sufficient accumulation or activation of surface drainage and atmospheric mixing. This temporal offset is consistent with findings from urban Vietnam [18], where convective rainfall yielded lagged but appreciable reductions in aerosol levels due to delayed boundary-layer stabilization and washout effects. In Semarang, however, the correlation structure is far less coherent, with weak and inconsistent lagged relationships. This irregularity may reflect the city's exposure to mesoscale meteorological phenomena such as sea breezes and convective outflows, which frequently disrupt local atmospheric stability and modulate precipitation timing and intensity, thus blurring the observable lag structures.

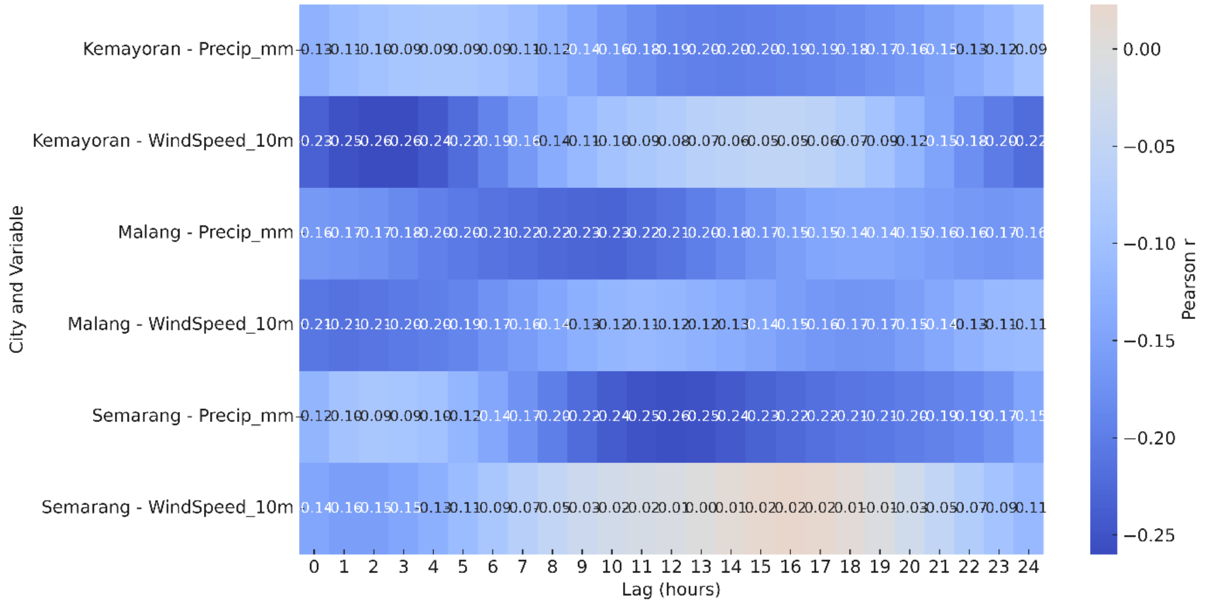

Figure 5. Lagged Pearson correlation heatmap between PM2.5 and two key meteorological variables (wind speed and precipitation) across Kemayoran, Semarang, and Malang. Each row represents a city-variable combination, and each column represents the lag in hours (0–24). Wind speed generally shows peak negative correlations at lag 0, while precipitation effects are more delayed, peaking between 2–6 hours depending on location.

These city-specific lag profiles underscore the challenge of building generalized prediction frameworks for tropical environments. While wind speed exerts its strongest effect with no delay, precipitation's impact is temporally staggered and spatially heterogeneous. This has important implications for modeling: wind can be treated as a contemporaneous input, whereas precipitation might require lagged terms or cumulative indices to fully capture its influence. Previous studies in Southeast Asia [17], [18] have similarly concluded that lagged meteorological responses especially under convective regimes pose a fundamental obstacle to one-size-fits-all modeling strategies.

Nonetheless, for the modeling framework used in this study, we chose not to incorporate lagged meteorological features, primarily to maintain parsimony and minimize the risk of overfitting, especially given the relatively modest correlation magnitudes observed. Furthermore, preliminary tests showed limited additional predictive gain when including lagged variables, particularly for Semarang. As such, our final models rely only on current-hour meteorological predictors, which simplify implementation while retaining core predictive structure.

However, we recognize the potential utility of lag-aware models, especially in more complex or less ventilated environments such as Malang or during specific seasonal episodes. Future work should explore the inclusion of lagged rainfall or wind profiles, possibly via feature engineering or recurrent neural network (RNN) architectures that are designed to capture temporal dependencies over time. Such efforts may help close the performance gap in cities with less predictable PM2.5 dynamics and enhance the responsiveness of air quality warning systems.

Hybrid Model Performance: Model Description and Setup. To evaluate the predictive capacity of meteorological variables on PM2.5 concentrations, a hybrid modeling framework was implemented using both linear and non-linear algorithms. Three models were selected to represent different levels of complexity and interpretability: multiple linear regression (MLR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost).

MLR was employed as a baseline statistical model to assess the linear contribution of meteorological variables to PM2.5 variability. RF, a bagging-based ensemble learning algorithm, was chosen for its ability to model nonlinear interactions and its built-in feature importance metrics [36]–[41]. XGBoost, a boosting-based algorithm, was included due to its reputation for high predictive accuracy and efficient handling of multicollinearity and missing data [36]–[40].

All models were trained using hourly data from 2022 to 2024 for each city separately. Predictor variables included 2-m air temperature, 2-m dew point, 10-m wind speed (computed from u and v components), surface pressure, and precipitation. Prior to modeling, all predictors were standardized (mean-centered and scaled to unit variance) to facilitate model convergence and comparability.

The datasets were split into training (70%) and testing (30%) sets using a stratified temporal approach. Five-fold cross-validation was applied during model tuning to prevent overfitting. Model performance was evaluated using three standard metrics: the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE), computed on the test dataset. Feature importance scores were extracted from the RF and XGBoost models to aid in interpretation of predictor influence.

Model Evaluation Results. Table 3 presents the predictive performance of three models, multiple linear regression (MLR), Random Forest (RF), and XGBoost, trained to estimate hourly pollution levels based on meteorological variables. Across all three cities, non-linear machine learning models substantially outperformed the baseline MLR, as reflected by higher R^2 values and lower RMSE and MAE scores.

In Kemayoran, XGBoost achieved the best overall performance, with an R^2 of 0.534, RMSE of 16.34 µg/m³, and MAE of 11.70 µg/m³. This marked improvement over MLR ($R^2 = 0.353$) underscores the added value of capturing nonlinear interactions and threshold effects in a densely urbanized and emission-intensive environment. Similarly, in Semarang, XGBoost outperformed both RF and MLR, though the overall predictive power remained slightly lower ($R^2 = 0.435$). The relatively moderate performance in Semarang may reflect more variable local meteorological conditions and possibly lower signal-to-noise ratio in emissions data.

In Malang, all models yielded lower predictive accuracy, with the best R^2 from XGBoost reaching only 0.377. This suggests that particulate levels in this upland city are less directly explained by the selected meteorological variables. Factors such as long-range transport, localized biomass burning, or orographic effects may introduce additional complexity not captured in the current input features.

The consistent superiority of machine learning models, particularly XGBoost, across all sites indicates that the relationship between meteorology and PM2.5 is fundamentally nonlinear. Moreover, the variation in performance among cities highlights the importance of considering site-specific dynamics and spatial heterogeneity in air quality modeling frameworks.

Figure 6 visualizes the relationship between observed and predicted PM2.5 concentrations for the three study locations using Random Forest models trained on one-month subsets of hourly data. Across all sites, the scatter points cluster around the 1:1 line, indicating that the model captures the general trend of PM2.5 variability.

The prediction skill appears most consistent in Kemayoran, likely due to its more stable emission patterns and higher pollution baseline, which reduce relative noise. In contrast, scatter in Malang and Semarang is slightly more dispersed, especially at lower PM2.5 concentrations, suggesting a weaker signal-to-noise ratio in the model input or greater variability in local meteorological or emission sources.

While the models do not perfectly match extreme values, the overall alignment supports the use of Random Forest for short-term air quality estimation in tropical urban environments. This plot complements the quantitative metrics in Table 3 by providing a qualitative assessment of model performance across the full range of observed concentrations.

Table 3. Performance metrics (R^2, RMSE, MAE) of PM2.5 prediction models, MLR, Random Forest (RF), and XGBoost, across three cities using hourly meteorological inputs. XGBoost consistently outperforms other models in terms of both error reduction and explanatory power.

City	Model	R^2	RMSE (µg/m³)	MAE (µg/m³)
Kemayoran	MLR	0.353	19.42	13.69
Kemayoran	RF	0.522	16.59	11.85
Kemayoran	XGBoost	0.534	16.34	11.7
Semarang	MLR	0.348	16.28	11.55
Semarang	RF	0.412	15.37	10.8
Semarang	XGBoost	0.435	14.99	10.53
Malang	MLR	0.205	10.75	7.77
Malang	RF	0.35	9.85	7.2
Malang	XGBoost	0.377	9.54	7

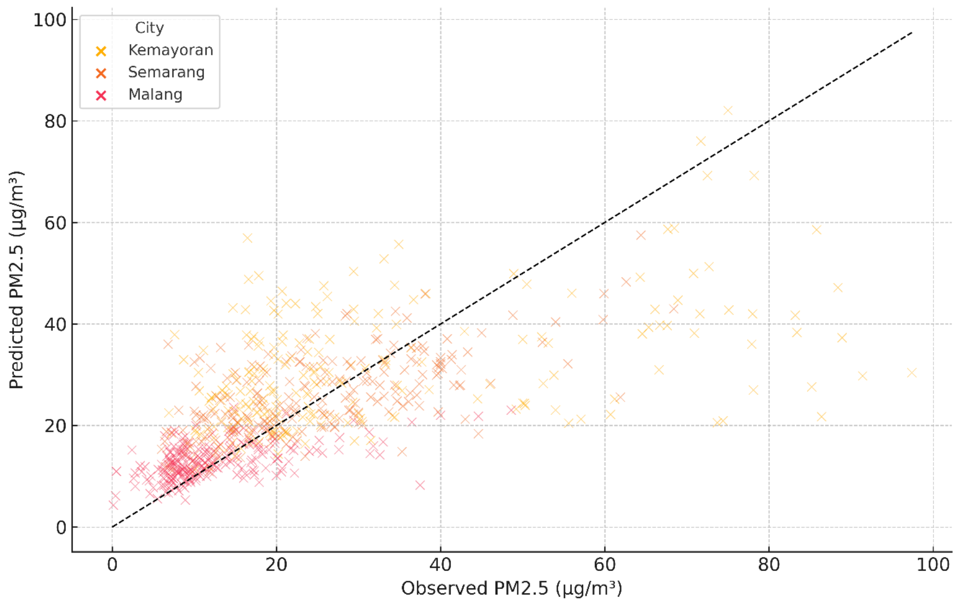

Figure 6. Observed versus predicted PM2.5 concentrations using Random Forest regression for Kemayoran, Semarang, and Malang, based on one-month subsets of hourly data. The dashed line indicates the 1:1 identity line. Predicted values generally follow the observed trend with some dispersion, especially at higher concentrations.

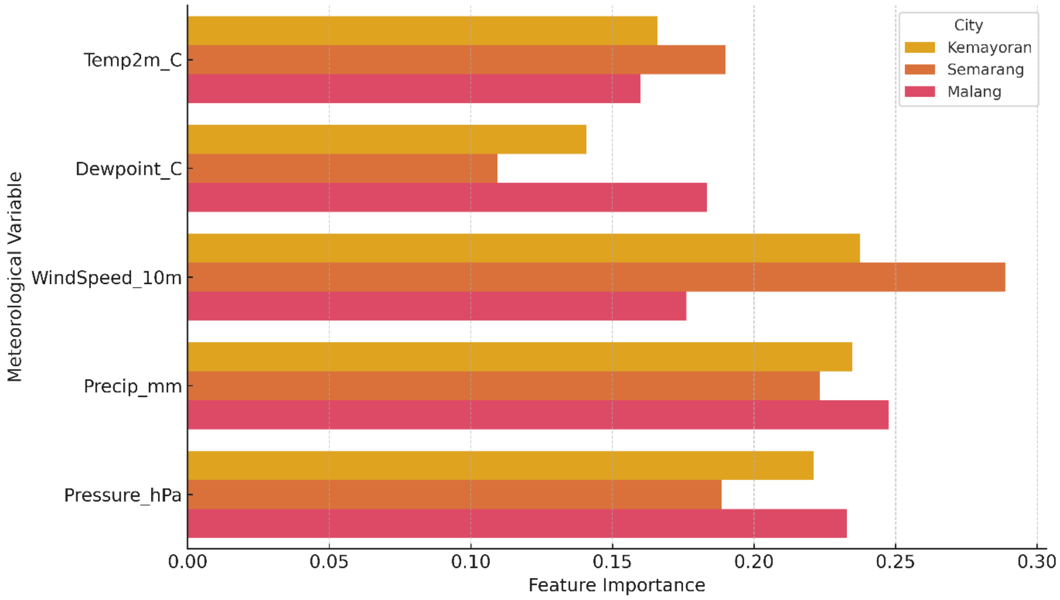

Figure 7. Relative feature importance from Random Forest models trained on one-month subsets of hourly data for Kemayoran, Semarang, and Malang. Wind speed consistently ranks as the most influential predictor, followed by dew point and precipitation. Feature contributions vary by location, reflecting local meteorological dynamics and emission characteristics.

Random Forest was selected for visualization in Figure 6 due to its strong and balanced performance across all three cities, coupled with lower computational overhead compared to more complex models such as XGBoost. While XGBoost offered slightly higher numerical accuracy (Table 3), Random Forest produced consistent and interpretable predictions with lower memory demands, making it ideal for illustrative purposes in a cross-city comparison. Linear regression was excluded from visual evaluation due to its limited explanatory power and poor fit to observed data variability.

Feature Importance Analysis. Figure 7 presents the relative importance of five meteorological variables in predicting hourly PM2.5 concentrations using Random Forest models. Across all three cities, wind speed consistently ranks as the most influential predictor, aligning with its known role in pollutant dispersion and boundary layer ventilation. This reinforces previous findings from both correlation and lag analyses, which also identified wind speed as the dominant meteorological driver of PM2.5 variability.

Dew point and precipitation follow as the next most relevant predictors, suggesting that moisture-related processes such as condensation growth and wet deposition play a secondary but notable role in modulating PM2.5 levels. In contrast, surface pressure and air temperature contributed minimally to model performance, though slight variations between cities indicate the influence of local meteorological dynamics.

The convergence of evidence, from correlation, lag, and machine learning feature importance, clearly underscores the primacy of wind-driven ventilation in shaping urban air pollution in tropical settings. Understanding these site-specific differences is essential for tailoring air quality models and mitigation strategies to regional contexts, especially in data-limited tropical cities.

Interpretation and Implications. The hybrid modeling framework employed in this study reveals both the promise and limitations of using meteorological variables to predict short-term PM2.5 concentrations in tropical urban environments. While non-linear models such as Random Forest and XGBoost substantially outperformed classical linear regression, their predictive power remained moderate, with R^2 values ranging from 0.35 to 0.53 across different cities and model types (Table 3). This outcome affirms a critical insight: meteorological variables, though important, are insufficient as sole predictors of particulate pollution in complex urban settings.

The disparity in model performance among cities further underscores the role of spatial heterogeneity. In Malang, lower accuracy suggests a weaker meteorological signal or a stronger influence from localized and non-meteorological drivers, such as topographic trapping, long-range transport, or biomass burning, none of which are explicitly captured in the current input set. In contrast, Kemayoran demonstrated relatively higher predictability, likely due to its more consistent emission sources and better alignment between meteorological variation and pollutant dynamics. This inter-city variation supports findings from other tropical modeling efforts (e.g.,[42], [43]), which also caution against the direct transferability of model structures across regions with different climatic and anthropogenic baselines.

Visual diagnostics of model output (Figure 6) and feature attribution (Figure 7) reinforce these insights. Wind speed consistently emerged as the most influential variable across all three sites, reflecting its immediate role in enhancing boundary-layer mixing and pollutant dispersion. Yet, the relevance of secondary predictors, such as precipitation and dew point, varied by location, suggesting that site-specific mesoscale processes and microclimates modulate pollution dynamics in nuanced ways. This heterogeneity aligns with evidence from coastal and monsoon-influenced cities, where intermittent convective activity can intermittently override typical predictor-response relationships.

From a practical perspective, these findings point to actionable strategies in urban air quality management. In cities like Kemayoran, where meteorology-pollution coupling is relatively strong, real-time data on wind and humidity could be integrated into early warning systems with reasonable reliability. However, in less predictable environments such as Malang, forecasting tools may require broader input sources, such as satellite-derived aerosol optical depth, emissions inventories, or land-use proxies, to attain operational robustness.

Importantly, the hybrid statistical–physical approach adopted in this study offers a pragmatic solution for data-scarce regions. Its reliance on widely available meteorological inputs and machine learning architecture enables scalable deployment across Southeast Asian cities with limited monitoring infrastructure. The method is also computationally efficient and relatively interpretable, making it suitable not only for scientific forecasting but also for informing regulatory action and public health advisory systems. Recent studies (e.g.,[44]) have similarly demonstrated that combining empirical predictors with machine learning frameworks offers a viable path toward rapid, localized air quality prediction under tropical atmospheric complexity.

Nonetheless, a degree of predictive uncertainty persists, particularly in cities where meteorological control is weak or chaotic. To address this, future modeling should incorporate explicit uncertainty quantification techniques, such as bootstrapped confidence intervals, ensemble spread, or Bayesian posterior estimates. These tools can help decision-makers understand the range of possible outcomes rather than relying on single-point predictions, especially under conditions of meteorological instability or emission anomalies. Without such safeguards, there is a risk of overconfidence in forecasts, especially in high-stakes scenarios involving health alerts or policy interventions.

Characterization of Extreme PM2.5 Events. Table 4 summarizes the composite meteorological profiles during extreme PM2.5 events, defined as hourly concentrations exceeding the 95th percentile, across three Indonesian urban sites from 2022 to 2024. These events offer a window into the meteorological circumstances most conducive to severe air quality degradation in tropical environments.

Kemayoran exhibited the most alarming extremes, with peak concentrations surpassing 410 µg/m³. These episodes coincided with classic stagnation conditions: moderate temperatures (~25°C), low wind speeds (1.63 m/s), and limited precipitation (3.5 mm). Such a profile reflects the suppression of boundary-layer ventilation, a phenomenon well documented in urban meteorology where nocturnal inversion, weak synoptic forcing, and urban canopy drag synergize to trap pollutants near the surface. Similar high-pollution signatures have been observed in coastal megacities like Manila and Bangkok [17], especially during dry-season subsidence and calm-wind regimes.

In Semarang and Malang, maximum concentrations were notably lower, yet the meteorological fingerprints remained consistent. Semarang's composite episodes featured weak winds (~1.42 m/s) and elevated humidity, indicating that even modest convective inhibition in coastal environments can facilitate pollutant accumulation. Malang, despite having the cleanest overall profile, experienced extreme events under even weaker wind conditions (~0.63 m/s), highlighting the city's vulnerability to local stagnation. These findings echo results from Chiang Mai [45], where upland cities with limited anthropogenic sources still encountered sharp pollution spikes during calm atmospheric episodes.

Collectively, these results reinforce a central hypothesis in tropical air pollution science: meteorological suppression—especially low wind speed and reduced wet deposition—is a key driver of extreme PM2.5 events. This mechanism explains how hazardous air quality episodes can still emerge in under-monitored or semi-clean cities, even under relatively low emission baselines. However, in highly urbanized environments such as Jakarta, the magnitude and variability of anthropogenic emissions may equal or even surpass the influence of meteorological conditions. In such contexts, emission loads can amplify or compound stagnation effects, making pollution episodes more intense and persistent. Therefore, emission control and meteorological awareness must be treated as complementary and co-dominant priorities, not sequential steps. Recognizing this duality is essential for the development of robust early-warning systems and urban air quality interventions.

From a policy and planning perspective, these findings provide empirical justification for integrating real-time meteorological forecasting, particularly near-surface wind and humidity, into air quality alert systems. Seasonally tuned early-warning frameworks should be prioritized, especially during the dry season (JJA) and the second transitional season (SON), when stagnation conditions are most frequent. Rapid-response interventions such as traffic regulation, outdoor activity restrictions, or industrial curtailment could be deployed on days flagged by such systems.

Finally, future research should explore whether large-scale precursors, such as regional subsidence patterns, monsoon break phases, or MJO inactivity, can serve as mesoscale indicators of upcoming stagnation risk. The integration of meteorological pattern recognition with high-resolution emission inventories may significantly improve the anticipatory capability of early-warning systems in tropical cities.

Figure 8 illustrates the composite meteorological conditions associated with extreme PM2.5 events, defined as concentrations exceeding the 95th percentile, across Kemayoran, Semarang, and Malang during the 2022–2024 period. Despite substantial differences in absolute pollution levels, all cities share a common atmospheric signature during extreme episodes: low wind speed, limited precipitation, and high humidity, which together indicate boundary layer stagnation and weak pollutant dispersion.

Kemayoran recorded the highest PM2.5 concentrations and also exhibited elevated temperature and pressure values, suggesting that pollution extremes may coincide with suppressed convection and stable stratification near the surface. Semarang and Malang, while showing lower absolute PM2.5 values, both experienced similar reductions in wind speed during their respective extreme hours. Malang, in particular, stood out for having the lowest wind speed (~0.63 m/s), indicating its susceptibility to localized accumulation despite its upland location and generally cleaner baseline.

Table 4. Composite mean meteorological conditions during extreme PM2.5 hours (>95th percentile) for three Indonesian cities (2022–2024). Values include temperature, dew point, wind speed, precipitation, and pressure.

City	PM2.5 (μg/m³)	Temperature (°C)	Dew Point (°C)	Wind (m/s)	Precipitation (mm)	Pressure (hPa)
Kemayoran	410.4	25.17	23.23	1.63	3.5	1005.4
Semarang	58.1	25.24	21.17	1.42	2.1	985.6
Malang	37.6	22.12	18.89	0.63	3.42	921.72

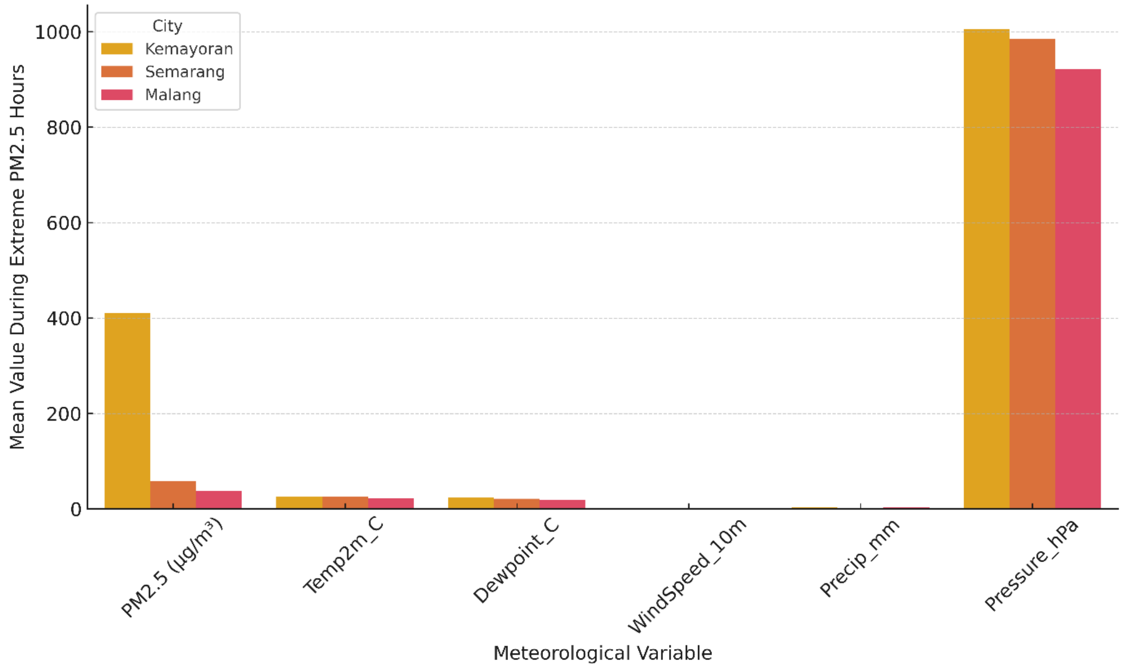

Figure 8. Composite mean values of meteorological variables during extreme PM2.5 events (>95th percentile) in Kemayoran, Semarang, and Malang over the 2022–2024 period. All cities experienced low wind speed and minimal rainfall during extreme episodes, with Kemayoran exhibiting the highest PM2.5 levels and Malang the weakest atmospheric ventilation.

These findings reinforce the conclusion that weak atmospheric mixing, rather than emissions alone, governs the occurrence of extreme pollution in tropical cities. This underscores the importance of integrating near-surface wind fields, precipitation forecasts, and seasonal circulation regimes into air quality early warning systems, especially during transitional and dry months.

Comparative Discussion Across Sites. While spatial variability has been examined in the preceding sections, this subsection provides a synthesized comparison to highlight distinct atmospheric mechanisms influencing PM2.5 levels across Kemayoran, Semarang, and Malang. The assessment reveals both recurring patterns and localized characteristics that shape air quality outcomes in tropical urban environments.

Across all three cities, extreme pollution episodes consistently coincide with stagnant atmospheric setups, marked by low wind speeds, limited rainfall, and elevated humidity. These conditions, repeatedly captured in composite analyses (Table 4) and visualized in Figure 8, reflect a shared climatological susceptibility to poor ventilation.

Yet important differences persist. Kemayoran, a coastal megacity with high emission intensity, exhibits the most severe PM2.5 peaks, exceeding 400 µg/m³, particularly during the dry season (JJA), when vertical mixing is suppressed and large-scale airflow is weak. Despite its coastal position, urban canopy effects and regional subsidence likely diminish the effectiveness of sea-breeze ventilation.

Malang, by contrast, registers the lowest pollution extremes, yet also experiences the weakest wind conditions (~0.63 m/s). This creates a paradox: although emissions are relatively modest, the meteorological environment favors stagnation and accumulation. Thus, while the severity is lower, the risk of local buildup remains nontrivial.

Semarang presents an intermediate profile, geographically, meteorologically, and in PM2.5 levels. Its greater exposure to mesoscale sea-breeze circulations and convective disturbances likely contributes to its higher variability and weaker correlations between meteorological drivers and pollution concentrations, as shown in earlier analyses.

Overall, these findings support a nuanced understanding of tropical air pollution: while common meteorological triggers such as dryness and stagnation are evident, their interaction with topography, emissions, and coastal dynamics determines site-specific outcomes. This reinforces the importance of localized forecasting models and targeted mitigation, as opposed to blanket policy prescriptions. Future modeling frameworks may be strengthened by integrating land use, spatial drivers, or regional transport effects.

From a practical standpoint, these insights could inform adaptive early warning systems tailored to seasonality and meteorological thresholds. For instance, real-time alerts might be issued on dry-season mornings with forecasted low wind speeds—conditions that elevate the risk of air stagnation. Temporary interventions such as time-bound vehicle restrictions or stricter emissions controls may be prioritized on days flagged by predictive systems. Translating these outputs into actionable guidance will be essential to protect urban populations from short-term pollution spikes.

Nonetheless, it is important to recognize the inherent limitations of the current modeling framework. While machine learning approaches like Random Forest and XGBoost substantially outperform linear models, some degree of predictive uncertainty remains, particularly in cities like Malang where meteorological signals are weaker or more erratic. This uncertainty stems not only from unmeasured emissions and land cover changes, but also from the chaotic nature of tropical atmospheric processes. Standard error metrics (RMSE, MAE) offer a general picture of model fit, but may underestimate extreme deviations. Incorporating uncertainty quantification methods—such as prediction intervals or ensemble-based spread, would improve the transparency and operational reliability of air quality forecasts, particularly when used to support real-time decision-making.

4. Conclusion

This study employed a hybrid statistical–physical modeling framework to investigate the meteorological controls of PM2.5 variability in three Indonesian cities—Kemayoran, Semarang, and Malang—using high-frequency observations and ERA5-based reanalysis data from 2022 to 2024. Across all sites, extreme PM2.5 episodes were consistently associated with stagnation conditions marked by low wind speeds, limited precipitation, and boundary-layer stability.

Despite this shared pattern, each city exhibited distinct pollution dynamics influenced by local geography, urban morphology, and emission baselines. Kemayoran recorded the most severe concentrations, linked to dry-season subsidence and persistent weak dispersion. Malang, though generally cleaner, proved highly sensitive to minor ventilation changes due to its enclosed topography. Semarang presented an intermediate profile, shaped by intermittent mesoscale sea breeze activity and convective processes.

The use of hybrid modeling—combining multiple linear regression (MLR) with non-linear machine learning techniques such as Random Forest and XGBoost—proved essential in improving the understanding and predictive accuracy of PM2.5 dynamics. Compared to conventional linear methods, machine learning models more effectively captured non-linear interactions and complex dependencies between meteorological drivers and pollution levels. In particular, they revealed the dominant and immediate role of wind speed, the spatially heterogeneous influence of precipitation, and the limited explanatory power of surface pressure and temperature. These insights would have been difficult to extract using linear models alone, underscoring the added value of hybrid approaches in dissecting tropical urban air pollution systems.

While moderate R^2 values (0.35–0.53) suggest room for improvement, especially through the integration of emissions, land-use, or transport variables, the hybrid models nonetheless advanced the diagnostic capacity to interpret PM2.5 variability across cities with contrasting atmospheric regimes. They also demonstrated operational potential for real-time air quality prediction in data-limited tropical environments.

From a practical standpoint, the findings advocate for seasonal air quality forecasting systems that monitor meteorological precursors of pollution events—such as stagnation, moisture buildup, and suppressed convection. Instead of relying on static thresholds, cities should adopt dynamic, location-specific early-warning protocols, particularly during the dry and transitional seasons (JJA and SON), when pollution extremes are most frequent.

Ultimately, this study illustrates that hybrid modeling not only improves prediction accuracy but also deepens conceptual understanding of how meteorology drives pollution behavior in tropical cities. By integrating physical reasoning, data-driven learning, and seasonal climatology, the approach offers a replicable and scalable framework for advancing air quality forecasting and risk management across Southeast Asia and other tropical urban regions.

Suggestion

While this study offers valuable insights into meteorologically driven PM2.5 variability in tropical urban environments, several avenues remain open for future exploration. First, model performance could be enhanced by integrating spatially resolved variables, such as land use, traffic density, or satellite-based aerosol indicators, to better capture local emission heterogeneity. Second, more sophisticated time-series architectures, including Long Short-Term Memory (LSTM) networks and Temporal Convolutional Networks (TCNs), warrant investigation for their ability to model lagged and nonlinear dependencies inherent in pollutant accumulation.

Third, the exclusive reliance on meteorological inputs constitutes a key limitation, as the absence of direct emissions data constrains the model's capacity to disentangle atmospheric transport from source-driven variability. Future studies should consider coupling machine learning frameworks with emission inventories or chemical tracer data to improve both interpretability and explanatory power. Lastly, the proposed hybrid framework should be tested across diverse tropical settings with varying climatological and anthropogenic baselines to evaluate its transferability. The fusion of high-resolution spatial, chemical, and temporal information with robust learning techniques holds considerable promise for advancing predictive air quality systems in data-limited urban regions.

Acknowledgement

The authors would like to express their gratitude to the Meteorological, Climatological, and Geophysical Agency of Indonesia (BMKG) for providing access to high-resolution PM2.5 observation data. The authors also acknowledge the Copernicus Climate Change Service (C3S) for offering open access to ERA5 reanalysis data through the Climate Data Store (CDS) platform. These datasets were essential in enabling the high-frequency and seasonally resolved analysis presented in this study.

References

[1] A. Verma and L. Bhatia, “A Study of Seasonal and Temporal Variances in Ambient Air Quality of Highly Polluted Cities in Rajasthan,” *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, 2024, doi: 10.32628/cseit24104103.

[2] A. Rivera, P. Ponce, O. Mata, A. Molina, and A. Meier, “Local Weather Station Design and Development for Cost-Effective Environmental Monitoring and Real-Time Data Sharing,” Sensors, 2023, doi: 10.3390/s23229060.

[3] I. M. Honório Coelho, E. F. Cartaxo, R. Marques, and T. M. de Freitas, “Respirable Particles, the Influence of the Tropical Climate and the Importance of Monitoring Air Quality in the City of Manaus-Am,” 2024, doi: 10.56238/sevened2024.004-021.

[4] D. Kairúz-Cabrera, V. Hernandez-Rodriguez, O. Schalm, A. Martínez, P. M. Laso, and D. Alejo-Sánchez, “Development of a Unified IoT Platform for Assessing Meteorological and Air Quality Data in a Tropical Environment,” Sensors, 2024, doi: 10.3390/s24092729.

[5] K. G. Vohra et al., “Rapid Rise in Premature Mortality Due to Anthropogenic Air Pollution in Fast-Growing Tropical Cities From 2005 to 2018,” Sci. Adv., 2022, doi: 10.1126/sciadv.abm4435.

[6] J. Zhang et al., “Associations of Long-Term Exposure to Air Pollution, Physical Activity With Blood Pressure and Prevalence of Hypertension: The China Health and Retirement Longitudinal Study,” Front. Public Heal., 2023, doi: 10.3389/fpubh.2023.1137118.

[7] G. V Levina, “Helical Cyclogenesis as an Extreme Threshold Phenomenon in a Rotating Stratified Moist Atmosphere,” Front. Earth Sci., 2024, doi: 10.3389/feart.2023.1296067.

[8] Y. Sun, N. Zhang, S. Miao, F. Kong, Y. Zhang, and N. Li, “Urban Morphological Parameters of the Main Cities in China and Their Application in the WRF Model,” J. Adv. Model. Earth Syst., 2021, doi: 10.1029/2020ms002382.

[9] L. Zhu, L. Bai, G. Chen, Y. Q. Sun, and Z. Meng, “Convection Initiation Associated With Ambient Winds and Local Circulations Over a Tropical Island in South China,” Geophys. Res. Lett., 2021, doi: 10.1029/2021gl094382.

[10] Q. Doan et al., “Urban-induced Modifications to the Diurnal Cycle of Rainfall Over a Tropical City,” Q. J. R. Meteorol. Soc., 2021, doi: 10.1002/qj.3966.

[11] X. Jin et al., “Regional PM2.5 Pollution Confined by Atmospheric Internal Boundaries in the North China Plain: Boundary Layer Structures and Numerical Simulation,” Atmos. Chem. Phys., 2022, doi: 10.5194/acp-22-11409-2022.

[12] N. Ojha et al., “On the Widespread Enhancement in Fine Particulate Matter Across the Indo-Gangetic Plain Towards Winter,” Sci. Rep., 2020, doi: 10.1038/s41598-020-62710-8.

[13] Q. Liu et al., “Reduced Volatility of Aerosols From Surface Emissions to the Top of the Planetary Boundary Layer,” *Atmos. Chem. Phys.*, 2021, doi: 10.5194/acp-21-14749-2021.

[14] N. Cholianawati et al., “Diurnal and DailyVariations of PM2.5 and its Multiple-Wavelet Coherence with Meteorological Variables in Indonesia,” Aerosol Air Qual. Res., vol. 24, no. 3, p. 230158, 2024, doi: 10.4209/aaqr.230158.

[15] M. S. Rahman, S. S. Bhuiyan, Z. Ahmed, N. Saha, and B. A. Begum, “Characterization and source apportionment of elemental species in PM2.5 with especial emphasis on seasonal variation in the capital city ‘Dhaka’, Bangladesh,” Urban Clim., vol. 36, p. 100804, 2021, doi: https://doi.org/10.1016/j.uclim.2021.100804.

[16] N. Amil, M. T. Latif, M. F. Khan, and M. Mohamad, “Seasonal variability of PM2.5 composition and sources in the Klang Valley urban-industrial environment,” Atmos. Chem. Phys., vol. 16, no. 8, pp. 5357–5381, 2016, doi: 10.5194/acp-16-5357-2016.

[17] T. Amnuaylojaroen, P. Kaewkanchanawong, and P. Panpeng, “Distribution and Meteorological Control of PM2.5 and Its Effect on Visibility in Northern Thailand,”

Atmosphere (Basel)., vol. 14, no. 3, 2023, doi: 10.3390/atmos14030538.
[18] G. T. H. Nguyen, L. T. La, H. Hoang-Cong, and A. H. Le, "An exploration of meteorological effects on PM2.5 air quality in several provinces and cities in Vietnam," J. Environ. Sci., vol. 145, pp. 139–151, 2024, doi: https://doi.org/10.1016/j.jes.2023.07.020.
[19] S. D. A. Kusumaningtyas et al., "Estimating PM2.5 Concentration Using Multiple Linear Regression (MLR) and the Random Forest (RF) Approach in Jakarta, Indonesia," in Proceedings of the International Conference on Radioscience, Equatorial Atmospheric Science and Environment and Humanosphere Science, 2024, pp. 23–33.
[20] S. Shafi and N. Scafetta, "Optimal machine learning techniques for meteorological modeling of $${\textrm{PM}}_{2.5}$$concentration in five major polluted cities of South-East Asia," Nat. Hazards, vol. 121, no. 6, pp. 6981–7025, 2025, doi: 10.1007/s11069-024-07077-z.
[21] W. Wang et al., "Chemical compositions and possible transportation of PM2.5 during two haze periods in a coastal city of the North China Plain," Geol. J., vol. 58(12), pp. 4417–4427, 2023, doi: 10.1002/gj.4850.
[22] I. Odériz, Í. J. Losada, R. Silva, and N. Mori, "Global Assessment of Interannual Variability in Coastal Urban Areas and Ecosystems," Environ. Res. Lett., vol. 19(11), p. 114040, 2024, doi: 10.1088/1748-9326/ad7b5b.
[23] M. Weger and B. Heinold, "Air Pollution Trapping in the Dresden Basin From Gray-Zone Scale Urban Modeling," Atmos. Chem. Phys., 2023, doi: 10.5194/acp-23-13769-2023.
[24] Z. Cui, F. Ren, Q. Wei, and Z. Xi, "What Drives the Spatio-Temporal Distribution and Spillover of Air Quality in China's Three Urban Agglomerations? Evidence From a Two-Stage Approach," Front. Ecol. Evol., 2022, doi: 10.3389/fevo.2022.977598.
[25] P. Sekuła, Z. Ustrnul, A. Bokwa, B. Bochenek, and M. Zimnoch, "Random Forests Assessment of the Role of Atmospheric Circulation in PM10 in an Urban Area With Complex Topography," Sustainability, 2022, doi: 10.3390/su14063388.
[26] S. Gudziunaite, Z. Shabani, L. Weitensfelder, and H. Moshammer, "Time Series Analysis in Environmental Epidemiology: Challenges and Considerations," Int. J. Occup. Med. Environ. Health, 2023, doi: 10.13075/ijomeh.1896.02237.
[27] F. Finazzi, J. Rodeschini, and L. Tedesco, "Discussion on Assessing Predictability of Environmental Time Series With Statistical and Machine Learning Models," Environmetrics, 2025, doi: 10.1002/env.2900.[28] V. M. Hua, T. L. Nguyen, M.-S. Dao, H. D. Nguyen, and B. T. Nguyen, "The Impact of Data Imputation on Air QualityPrediction Problem," PLoS One, 2024, doi: 10.1371/journal.pone.0306303.
[29] A. Abdi, S. Seyedabrishami, and S. O'Hern, "A Two-Stage Sequential Framework for Traffic Accident Post-Impact Prediction Utilizing Real-Time Traffic, Weather, and Accident Data," J. Adv. Transp., 2023, doi: 10.1155/2023/8737185.
[30] T. M. Abegaz, M. B. Ahmed, F. Sherbeny, V. Diaby, H. Chi, and A. A. Ali, "Application of Machine Learning Algorithms to Predict Uncontrolled Diabetes Using the All of Us Research Program Data," Healthcare, 2023, doi: 10.3390/healthcare11081138.
[31] S. Basu, A. Munafo, A.-F. Ben-Amor, S. Roy, P. Girard, and N. Terranova, "Predicting Disease Activity in Patients With Multiple Sclerosis: An Explainable Machine-learning Approach in the Mavenclad Trials," CPT Pharmacometrics \& Syst. Pharmacol., 2022, doi: 10.1002/psp4.12796.
[32] Z. Jinbo, Y. Li, and M. Haitao, "Handling Missing Data of Using the XGBoost-based Multiple Imputation by Chained Equations Regression Method," Front. Artif. Intell., 2025, doi: 10.3389/frai.2025.1553220.
[33] Z. Wang, S. Li, Y. Zhang, X. Wang, S. Liu, and D. Liu, "Built Environment Renewal Strategies Aimed at Improving Metro Station Vitality via the Interpretable Machine Learning Method: A Case Study of Beijing," Sustainability, 2024, doi: 10.3390/su16031178.
[34] K. Kishore, G. Braitberg, N. E. Holmes, and R. Bellomo, "Early Prediction of Hospital Admission of Emergency Department Patients," Emerg. Med. Australas., 2023, doi: 10.1111/1742-6723.14169.
[35] C. Peng, Z. Xie, and X. Jin, "Using Ensemble Learning for Remote Sensing Inversion of Water Quality Parameters in Poyang Lake," Sustainability, 2024, doi: 10.3390/su16083355.
[36] Z. Sun, G. Wang, P. Li, H. Wang, Z. Min, and X. Liang, "An Improved Random Forest Based on the Classification Accuracy and Correlation Measurement of Decision Trees," Expert Syst. Appl., 2024, doi: 10.1016/j.eswa.2023.121549.
[37] P. J. Lakshmi and N. B. M, "Feature Importance Feedback With Deep Q Process in Ensemble-Based Metaheuristic Feature Selection Algorithms," Sci. Rep., 2024, doi: 10.1038/s41598-024-53141-w.
[38] E. EFEOĞLU, "Comparative Performance Analysis of Ensemble Learning Algorithms for Rock Classification," Kastamonu Univ. J. Eng. Sci., 2023, doi: 10.55385/kastamonujes.1355695.

[39] J. Juwariyem, S. Sriyanto, S. Lestari, and C. Chairani, “Prediction of Stunting in Toddlers Using Bagging and Random Forest Algorithms,” Sinkron, 2024, doi: 10.33395/sinkron.v8i2.13448.

[40] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, “Boosting Methods for Multi-Class Imbalanced Data Classification: An Experimental Review,” J. Big Data, 2020, doi: 10.1186/s40537-020-00349-y.

[41] C. Gangavva, J. A. Mangai, and M. Bansal, “An Investigation of Ensemble Learning Algorithms for Fault Diagnosis of Roller Bearing,” 2022, doi: 10.3233/apc220016.

[42] T. Sreenivasulu and G. M. Rayalu, “Enhanced PM2.5 prediction in Delhi using a novel optimized STL-CNN-BILSTM-AM hybrid model,” Asian J. Atmos. Environ., vol. 18, no. 1, p. 25, 2024, doi: 10.1007/s44273-024-00048-7.

[43] Z. He, Q. Guo, Z. Wang, and X. Li, “A Hybrid Wavelet-Based Deep Learning Model for Accurate Prediction of Daily Surface PM2.5 Concentrations in Guangzhou City,” Toxics, vol. 13, no. 4, 2025, doi: 10.3390/toxics13040254.

[44] C. Minutti-Martinez, M. Arellano-Vázquez, and M. Zamora-Machado, “A Hybrid Model for the Prediction of Air Pollutants Concentration, Based on Statistical and Machine Learning Techniques,” in *Advances in Soft Computing*, 2021, pp. 252–264.

[45] P. Kraisitnitikul *et al.*, “Contrastingcompositions of PM2.5 in Northern Thailand during La Niña (2017) and El Niño (2019) years,” *J. Environ. Sci.*, vol. 135, pp. 585–599, 2024, doi: https://doi.org/10.1016/j.jes.2022.09.026.