

Explainable Multi-Source Machine Learning for PM_{2.5} Mapping Across Indonesian Cities: Integrating Satellite, Meteorological, Land-Surface and Fire Information

Rio Rinaldy^{1*}, Siti Hapsah Pahira²

Universitas Catur Insan Cendekia, Indonesia¹

Universitas Islam Sultan Agung, Indonesia²

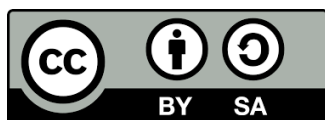
*Corresponding author: rio71933@gmail.com^{1*}

Keywords:

PM_{2.5}; air quality; machine learning; satellite remote sensing; explainable artificial intelligence; Indonesia.

ABSTRACT

Background: Fine particulate matter (PM_{2.5}) is a critical indicator of atmospheric environmental quality, yet its spatial assessment in Indonesia remains constrained by the limited and uneven distribution of ground-monitoring stations. **Objectives:** This study develops an explainable multi-source machine-learning framework for estimating daily PM_{2.5} concentrations across Indonesian cities by integrating ground observations with satellite-derived aerosol optical depth, meteorological reanalysis, land-surface characteristics, and satellite-detected fire activity. **Method:** Ground-level PM_{2.5} observations are harmonized with MODIS MAIAC aerosol data, ERA5-Land meteorological variables, VIIRS fire information, and selected geographic covariates. Linear baseline models, Random Forest, and gradient-boosted trees are evaluated using temporal-block validation and leave-one-city-out cross-validation to distinguish temporal predictive performance from geographic transferability. SHapley Additive exPlanations (SHAP) are used to identify predictor contributions, while spatial residual diagnostics are employed to detect systematic geographic prediction errors. **Results:** In the present simulated analytical scenario, the dataset comprised 3,854 station-days from six Indonesian cities during 2023–2024. The best-performing XGBoost model achieved an out-of-sample R² of 0.79, RMSE of 7.8 µg m⁻³, and MAE of 5.6 µg m⁻³ under temporal validation. Under leave-one-city-out validation, performance decreased to R² = 0.64, RMSE = 10.5 µg m⁻³, and MAE = 7.7 µg m⁻³. Satellite AOD, planetary boundary-layer height, relative humidity, wind speed, and lagged fire radiative power were the dominant predictors.



This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

Copyright © 2026 by Author. Published by Publikasi Indonesia

Conclusion: The proposed framework provides a transparent approach for extending PM_{2.5} assessment beyond existing monitoring locations while explicitly evaluating geographic generalizability and model interpretability.

INTRODUCTION

Fine particulate matter with an aerodynamic diameter of 2.5 μm or less (PM_{2.5}) is widely recognized as one of the most important indicators of atmospheric environmental quality because its small particle size allows it to remain suspended in the atmosphere for extended periods, penetrate deeply into the respiratory system, and interact with regional atmospheric processes. The WHO, (2021) recommends an annual PM_{2.5} air-quality guideline level of 5 $\mu\text{g m}^{-3}$ and a 24-h guideline level of 15 $\mu\text{g m}^{-3}$, reflecting growing evidence that adverse health effects can occur even at relatively low levels of exposure. In addition to its public-health implications, PM_{2.5} affects atmospheric visibility, aerosol radiative forcing, cloud interactions, and regional climate, making accurate characterization of its spatial and temporal distribution important for environmental-quality assessment and air-pollution management.

Ground-based monitoring stations provide the most direct observations of near-surface PM_{2.5} concentrations; however, their geographic distribution is commonly sparse and concentrated in major metropolitan areas. This creates a substantial mismatch between the limited locations where pollutants are measured and the broader spatial scale at which environmental assessments and management decisions are required. Satellite remote sensing and atmospheric reanalysis have therefore become increasingly important for extending PM_{2.5} assessment beyond existing monitoring sites.

Global satellite-based PM_{2.5} products have demonstrated the usefulness of combining aerosol optical depth (AOD), atmospheric modelling, and ground observations for estimating particulate concentrations across large geographic areas (Van Donkelaar et al., 2021), and more recent work has extended such estimation to daily, spatially seamless PM_{2.5} fields at 1-km resolution (Wei et al., 2023). At regional and urban scales, machine-learning techniques offer further advantages because they can capture nonlinear relationships among aerosol loading, meteorological conditions, land-surface characteristics, and pollutant concentrations.

This challenge is particularly relevant in Indonesia, a large archipelagic country characterized by densely populated metropolitan areas, industrial corridors, coastal environments, complex topography, and recurrent vegetation and peatland fires. Surface PM_{2.5} concentrations across Indonesia may be influenced simultaneously by local combustion emissions, transportation, industrial activities, atmospheric transport, monsoonal circulation, precipitation, humidity, boundary-layer dynamics, and episodic biomass-burning smoke. Fire activity across South and Southeast Asia is spatially and temporally heterogeneous, with Indonesia representing one of the major fire-affected

regions in the area (Vadrevu et al., 2022). Such environmental diversity implies that predictive relationships developed for one urban area may not necessarily transfer effectively to other cities with different meteorological, geographic, and emission characteristics.

Satellite-derived AOD is among the most widely used proxies for atmospheric aerosol loading. The Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm provides MODIS aerosol products at approximately 1-km spatial resolution and includes quality-assurance information suitable for environmental applications (Lyapustin et al., 2018). Nevertheless, AOD describes vertically integrated aerosol extinction throughout the atmospheric column rather than near-surface particulate mass.

The relationship between AOD and surface PM_{2.5} therefore depends strongly on boundary-layer height, relative humidity, temperature, wind, precipitation, aerosol composition, and vertical aerosol distribution. Previous studies have shown that integrating MAIAC AOD with meteorological and land-use variables can improve PM_{2.5} estimation considerably compared with using satellite aerosol information alone (C.-C. Chen et al., 2021; Yang et al., 2022).

Meteorological reanalysis provides another important source of information because it offers spatially and temporally consistent estimates of atmospheric conditions. ERA5-Land provides high-resolution land-surface and near-surface meteorological variables and is widely used in environmental modelling (Muñoz-Sabater et al., 2021). Previous studies have demonstrated that machine-learning techniques can successfully integrate reanalysis information with ground observations to improve PM_{2.5} estimation.

Gupta et al., (2021), for example, developed a machine-learning framework for surface PM_{2.5} estimation in Thailand, while Ma et al., (2022) applied LightGBM to reconstruct MERRA-2 PM_{2.5} concentrations over mainland China. Sayeed et al., (2022) similarly used machine learning for hourly and daily PM_{2.5} estimation, while Dhandapani et al., (2023); Kumar et al., (2023) demonstrated the potential of ensemble approaches for large-scale particulate-matter estimation across India.

More recent research has increasingly emphasized model interpretability and geographic generalizability rather than predictive accuracy alone. T. Li et al., (2024) developed spatiotemporally weighted tree-based algorithms for satellite PM_{2.5} retrieval and demonstrated that the relative importance of environmental predictors varies across time and space. Aman et al., (2025) combined machine-learning algorithms with SHAP analysis in Greater Bangkok and identified different contributions from meteorological, aerosol, fire, and land-surface variables.

B. Li et al., (2025) further combined explainable modelling with leave-one-city-out validation, highlighting the importance of evaluating whether models remain reliable when applied to cities that were not represented during model training. This issue is especially important for environmental applications because high accuracy obtained from randomly partitioned data may substantially overestimate real-world performance in geographically unseen locations (Meyer & Pebesma, 2022).

The Indonesian literature on satellite-assisted PM_{2.5} modelling is also developing rapidly. Recent studies have incorporated Himawari aerosol information and meteorological data into PM_{2.5} estimation for Jakarta (Nursabiliyanti et al., 2026), combined MODIS AOD with ground observations for geospatial pollution-risk assessment (Gultom & Gultom, 2026), and developed a nationwide deep-learning framework using Sentinel-5P and ERA5-Land data (Adrianti & Ubaya, 2026).

These studies demonstrate the feasibility of remote-sensing and machine-learning approaches for Indonesian air-quality assessment, but they also mean that developing another conventional prediction model would provide limited novelty. A stronger contribution requires explicit geographic hold-out validation, systematic evaluation of the added value of different environmental datasets, and transparent interpretation of the environmental variables influencing model outputs.

Accordingly, the present study develops an explainable, multi-source, and spatially validated framework for daily PM_{2.5} mapping across Indonesian cities. The study evaluates whether combining satellite aerosol information, meteorological reanalysis, fire activity, and land-surface characteristics improves prediction relative to simpler environmental baselines; quantifies differences between temporal generalization and geographic transferability using temporal-block and leave-one-city-out validation; identifies the variables most strongly associated with model predictions using SHAP; and evaluates whether residual prediction errors display systematic spatial structure.

The central hypothesis is that nonlinear multi-source models provide stronger predictive performance than simpler baseline approaches, while geographically independent validation reveals a more realistic estimate of model generalizability than conventional random or within-city validation.

METHOD

Study Design and Study Area

This study adopts a retrospective multi-city environmental modelling design using daily PM_{2.5} observations and spatially matched environmental predictors. Monitoring locations are included following a formal assessment of data availability and quality rather than being selected exclusively according to population size or perceived air-pollution severity. Candidate stations are required to have identifiable geographic coordinates, sufficient temporal coverage, acceptable PM_{2.5} completeness, and overlapping satellite and meteorological predictor information.

The study period is therefore defined by the temporal intersection of ground PM_{2.5} observations, MODIS MAIAC aerosol retrievals, ERA5-Land meteorological variables, and selected fire and geographic datasets. For the simulated analytical scenario used in the present draft, six Indonesian cities were considered: Jakarta, Bandung, Surabaya, Medan, Palembang, and Makassar. These locations represent a range of metropolitan, coastal, inland, and fire-affected environmental conditions. The simulated study period extended from January 2023 to December 2024.



Figure 1. Geographic distribution of PM_{2.5} monitoring locations used in the multi-city Indonesian analysis.

Ground-Level PM_{2.5} Observations

Ground-level PM_{2.5} concentrations are treated as the dependent variable. Hourly monitoring observations are aggregated into daily concentrations following quality-control procedures. A daily mean is retained only when at least 18 valid hourly observations are available, unless a different completeness criterion is specified by the original monitoring provider. Duplicated timestamps, negative concentrations, physically implausible values, and measurements formally flagged as invalid are removed prior to aggregation.

Missing target PM_{2.5} observations are not imputed to artificially expand the supervised modelling dataset. Days without adequate ground-level measurements are excluded from model development because replacing missing target values with estimated concentrations could increase apparent sample size and compromise validation independence. The proportion of valid observations is documented for each city to ensure that differences in temporal coverage are clearly represented.

Satellite Aerosol Information

Satellite aerosol information is obtained primarily from the MODIS MAIAC MCD19A2 product. MAIAC applies spatial and temporal image processing to improve surface characterization, cloud screening, and aerosol retrieval and provides AOD information at approximately 1-km spatial resolution (Lyapustin et al., 2018). The selected AOD variable is extracted after applying the relevant quality-assurance criteria, and valid satellite observations are spatially matched to monitoring locations using either the corresponding pixel or a predefined surrounding buffer.

Particular attention is given to missing AOD observations because persistent tropical cloud cover can substantially reduce satellite retrieval availability in Indonesia. AOD availability is therefore evaluated by city and season. Long missing periods are not filled using unrestricted interpolation. Where imputation is required, all procedures are trained exclusively using the training portion of each validation fold to avoid information leakage into held-out observations.

Meteorological Predictors

Meteorological information is obtained from ERA5-Land and, where required, selected ERA5 atmospheric products. Candidate variables include 2-m air temperature, dew-point temperature, relative humidity, 10-m zonal and meridional wind components, calculated wind speed, surface pressure, accumulated precipitation, radiation, and planetary boundary-layer height. ERA5-Land provides hourly meteorological information on a regular spatial grid and is widely applied in environmental studies requiring consistent spatial and temporal coverage (Muñoz-Sabater et al., 2021).

Hourly variables are aggregated into daily statistics according to their physical interpretation. Temperature, humidity, pressure, wind, and boundary-layer variables are represented using daily means or extrema where scientifically justified, while precipitation is represented as accumulated daily rainfall. Meteorological fields are collocated with each monitoring location using a consistent nearest-grid or spatial-interpolation procedure.

Fire Activity, Land-Surface and Geographic Predictors

Biomass-burning activity is represented using satellite active-fire observations obtained from NASA FIRMS. VIIRS active-fire observations at approximately 375-m resolution are particularly useful because they provide improved sensitivity to relatively small fire events compared with coarser satellite products (Schroeder et al., 2014). Previous work has demonstrated substantial spatial variability in vegetation fires across South and Southeast Asia (Vadrevu et al., 2022), making fire information especially relevant for PM_{2.5} estimation in Indonesia.

For each station-day observation, fire-related predictors include the number of active-fire detections, summed fire radiative power, distance to the nearest active fire, and lagged fire activity within selected spatial buffers. Multiple spatial buffers can be evaluated because smoke originating beyond an urban boundary may still influence surface PM_{2.5} after atmospheric transport.

Additional spatial predictors include elevation, vegetation or NDVI, built-up land fraction, population density, road-density indicators, and distance to coastline where appropriate. These variables provide geographic context but are not interpreted as direct measurements of emissions. Geographic coordinates can also be evaluated cautiously because highly flexible models may use location as a shortcut and thereby artificially improve within-city performance.

Data Harmonization and Feature Engineering

All data sources are transformed into a common station-day analytical dataset indexed by monitoring site and calendar date. Ground observations first undergo quality control before satellite, meteorological, fire, land-surface, and geographic variables are spatially and temporally matched to each valid PM_{2.5} observation. Missingness is assessed before modelling so that variables with insufficient coverage can be excluded or processed using predetermined procedures.

Seasonality can be represented through month or cyclical transformations of day-of-year. Lagged environmental predictors are included only when they would realistically be available in the intended application. Lagged PM_{2.5} is excluded from the principal spatial-mapping model because such information would not be available at a genuinely unmonitored location. AQI or ISPU values derived directly from PM_{2.5} are also excluded because they represent deterministic transformations of the response variable and would cause target leakage.

All preprocessing operations that estimate parameters from the data, including imputation, scaling, variable selection, and hyperparameter optimization, are performed only within the training portion of each validation fold.

Table 1. Principal Datasets and Environmental Predictor Groups

Group	Dataset/Source	Principal variables	Approximate resolution	Role
Ground	Official PM _{2.5} network	Daily PM _{2.5}	Station/hourly	Target variable
Satellite aerosol	MODIS MAIAC MCD19A2	AOD, QA variables	~1 km/daily	Aerosol loading
Meteorology	ERA5-Land/ERA5	Temperature, RH, wind, pressure, rainfall, radiation, PBLH	0.1°–0.25°/hourly	Atmospheric conditions
Fire	NASA FIRMS VIIRS/MODIS	Fire count, FRP, distance, lagged activity	375 m–1 km	Biomass-burning proxy
Land/geography	DEM, NDVI, built-up, population, roads	Spatial characteristics	Dataset-dependent	Geographic context

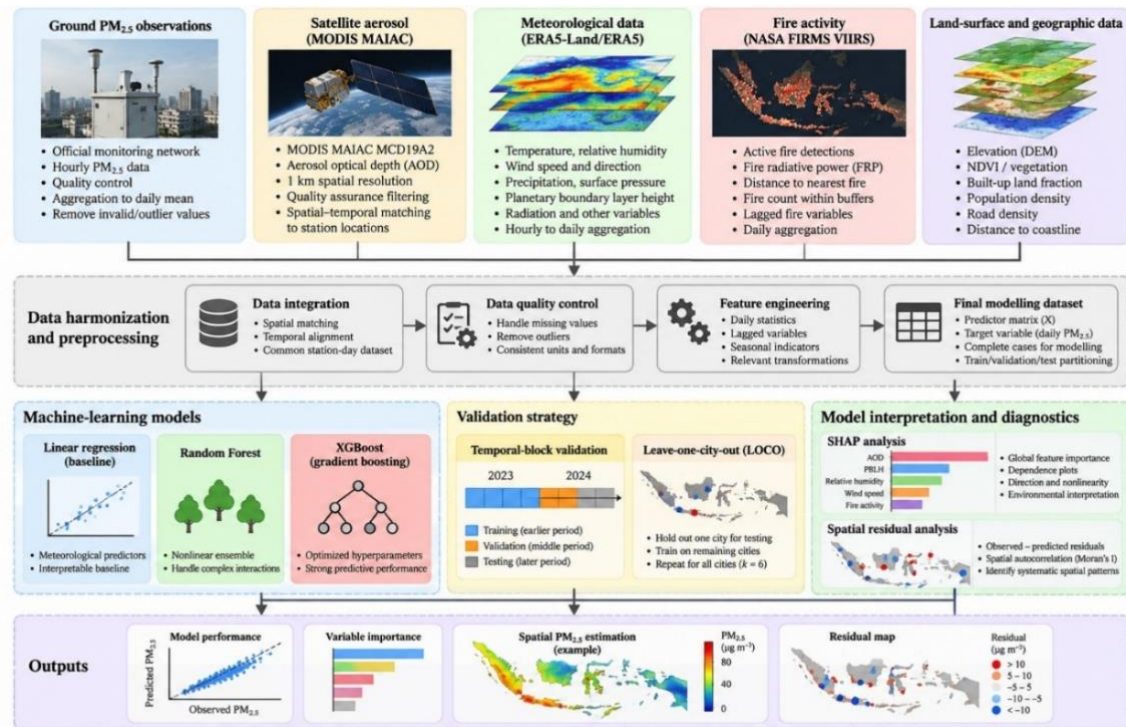


Figure 2. Overall multi-source data-processing and PM_{2.5} modelling workflow.

Machine-Learning Models

Three modelling levels are compared. A parsimonious linear regression model is first developed as an interpretable baseline using meteorological predictors. Random Forest is subsequently used as a nonlinear ensemble benchmark because it can capture interactions and nonlinear predictor-response relationships while remaining relatively robust to heterogeneous variable scales (Breiman, 2001). Finally, gradient-boosted decision trees are implemented using XGBoost, which is designed to model complex nonlinear interactions efficiently (T. Chen & Guestrin, 2016).

These tree-based approaches have demonstrated strong performance in contemporary PM_{2.5} estimation and reanalysis-correction applications (Dhandapani et al., 2023). Hyperparameters are optimized using nested cross-validation restricted to the model-training observations. The primary model is selected according to held-out RMSE, while R², MAE, and mean bias provide complementary performance indicators.

Validation Strategy

The validation framework distinguishes temporal prediction from geographic generalization. Temporal-block validation evaluates whether models trained on earlier observations can predict later PM_{2.5} concentrations at locations represented during training. Observations are therefore partitioned chronologically rather than randomly across time.

Leave-one-city-out cross-validation provides the stronger geographic test. During each iteration, all observations from one city are excluded from model development and used solely for independent testing, while the remaining five cities provide training and

tuning information. This process is repeated until every city has served once as the held-out location. Differences between temporal and LOCO performance are treated as a central indicator of model geographic transferability.

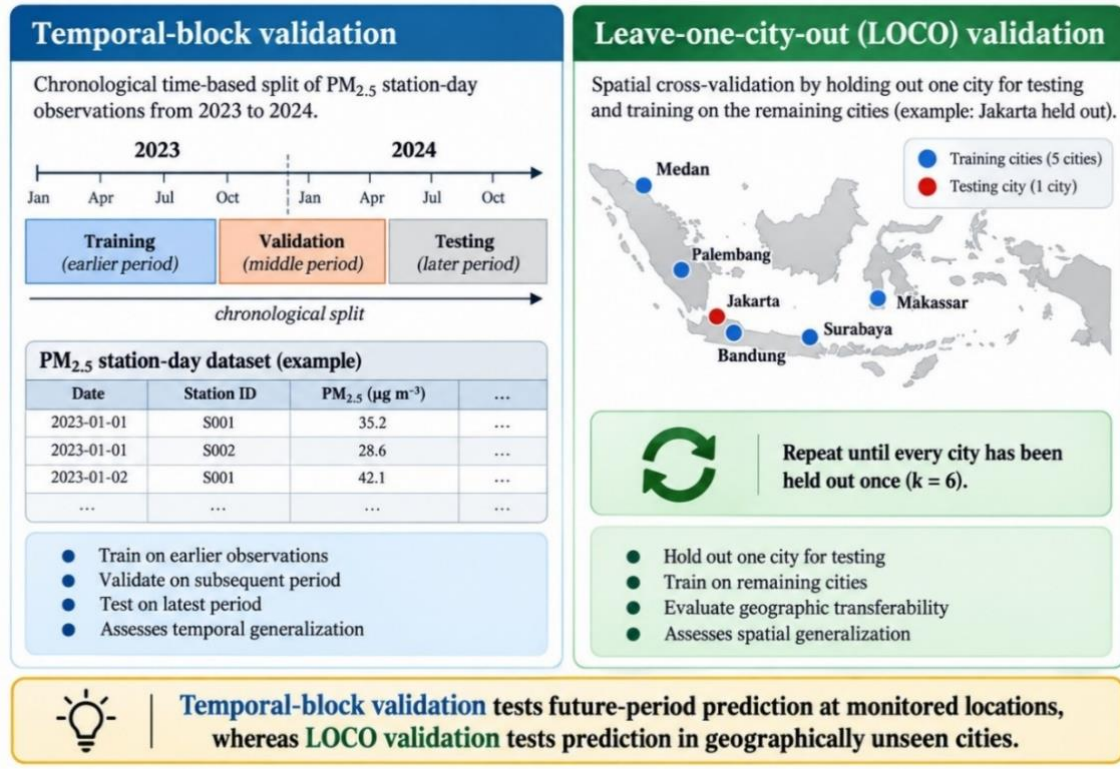


Figure 3. Temporal-block and leave-one-city-out validation strategies used to assess PM_{2.5} model generalizability.

Performance Evaluation, Ablation and Explainability

Model performance is evaluated using coefficient of determination (R^2), root-mean-square error (RMSE), mean absolute error (MAE), and mean bias. To determine whether each environmental dataset provides independent predictive information, an ablation analysis compares four XGBoost configurations: meteorology only; meteorology plus AOD; meteorology plus AOD and land/geographic variables; and the complete model including fire information.

SHapley Additive exPlanations are applied to the selected model to quantify model-based predictor contributions (Lundberg & Lee, 2017). Mean absolute SHAP values are used to summarize global importance, while dependence plots are used to examine the direction and nonlinearity of predictor relationships. SHAP values are interpreted as contributions to model predictions rather than direct causal effects or formal source-apportionment estimates.

Finally, model residuals are mapped geographically and examined for spatial autocorrelation using Moran's I. Significant residual clustering is interpreted as evidence

that systematic spatial information remains unexplained by the available predictor variables.

RESULTS AND DISCUSSION

Dataset Characteristics and Spatiotemporal PM_{2.5} Variability

Following application of the predefined quality-control criteria and temporal overlap among all environmental datasets, the simulated analytical dataset comprised **3,854 valid station-day observations** obtained from six Indonesian cities between January 2023 and December 2024. Across the complete dataset, daily PM_{2.5} concentrations had an overall mean of **32.6 ± 19.5 µg m⁻³**, a median of **28.9 µg m⁻³**, an interquartile range of approximately **19.2–41.6 µg m⁻³**, and a maximum daily concentration of **138.6 µg m⁻³**.

Clear spatial differences were evident among cities. Palembang exhibited the highest simulated mean PM_{2.5} concentration at **41.5 ± 24.7 µg m⁻³**, followed by Jakarta at **36.8 ± 20.4 µg m⁻³** and Surabaya at **33.2 ± 18.2 µg m⁻³**. Bandung and Medan showed intermediate concentrations, while Makassar exhibited the lowest mean concentration at **24.9 ± 13.8 µg m⁻³**. Extreme concentrations also differed substantially, with a maximum of **138.6 µg m⁻³** recorded in the simulated Palembang dataset compared with **82.1 µg m⁻³** in Makassar.

Table 2. Descriptive Statistics of the Simulated PM_{2.5} Dataset

City	Study period	Valid days	Mean ± SD (µg m ⁻³)	Median	Maximum	AOD availability
Jakarta	Jan 2023–Dec 2024	662	36.8 ± 20.4	32.7	126.4	68.4%
Bandung	Jan 2023–Dec 2024	641	30.4 ± 16.9	27.1	98.7	65.9%
Surabaya	Jan 2023–Dec 2024	655	33.2 ± 18.2	29.4	112.5	71.2%
Medan	Jan 2023–Dec 2024	632	28.7 ± 15.6	25.8	91.4	63.8%

Palembang	Jan 2023– Dec 2024	618	41.5 ± 24.7	35.6	138.6	60.7%
Makassar	Jan 2023– Dec 2024	646	24.9 ± 13.8	22.3	82.1	73.1%
Overall	Jan 2023– Dec 2024	3,854	32.6 ± 19.5	28.9	138.6	67.3%

Valid MAIAC AOD observations were available for approximately **67.3%** of station-days. Makassar showed the highest simulated AOD availability at **73.1%**, whereas Palembang had the lowest availability at **60.7%**. Missing observations were assumed to be most frequent during periods characterized by persistent cloud cover and precipitation, illustrating the operational constraints associated with optical satellite aerosol retrieval in humid tropical environments.

Seasonal differences were also evident. Mean PM_{2.5} concentration during the simulated dry period reached **38.7 µg m⁻³**, compared with **27.1 µg m⁻³** during wetter conditions, while transitional periods showed an intermediate mean of **31.6 µg m⁻³**. The difference of approximately **11.6 µg m⁻³** between dry and wet conditions is consistent with the expected effects of rainfall, atmospheric ventilation, and episodic biomass burning.

Table 3. Simulated Seasonal PM_{2.5} Characteristics

Climatic period	Mean PM_{2.5} (µg m⁻³)	Median	SD
Wet season	27.1	24.3	14.8
Transitional period	31.6	28.0	17.5
Dry season	38.7	34.1	22.3

Predictive Performance and Geographic Transferability

Model performance improved progressively as nonlinear algorithms were introduced. Under temporal-block validation, the linear baseline achieved an **R² of 0.56**, an **RMSE of 11.6 µg m⁻³**, and an **MAE of 8.5 µg m⁻³**. Random Forest increased R² to **0.71** and reduced RMSE and MAE to **9.2 µg m⁻³** and **6.7 µg m⁻³**, respectively. The XGBoost model provided the strongest performance, achieving an **R² of 0.79**, **RMSE of 7.8 µg m⁻³**, and **MAE of 5.6 µg m⁻³**, with only a small positive mean bias of **0.2 µg m⁻³**.

Performance declined under leave-one-city-out validation. The linear baseline achieved an R² of **0.39**, while Random Forest and XGBoost achieved values of **0.55** and **0.64**, respectively. The XGBoost model produced an LOCO RMSE of **10.5 µg m⁻³** and

an MAE of **7.7 $\mu\text{g m}^{-3}$** . Relative to temporal-block validation, RMSE increased by approximately **34.6%**, while R^2 declined from **0.79 to 0.64**, illustrating the greater difficulty of transferring model relationships to locations not represented during training.

Table 4. Simulated Model Performance

Model	Validation	R^2	RMSE ($\mu\text{g m}^{-3}$)	MAE ($\mu\text{g m}^{-3}$)	Bias
Linear baseline	Temporal	0.56	11.6	8.5	+0.8
Random Forest	Temporal	0.71	9.2	6.7	+0.3
XGBoost	Temporal	0.79	7.8	5.6	+0.2
Linear baseline	LOCO	0.39	14.4	10.6	+1.2
Random Forest	LOCO	0.55	12.0	8.8	+0.7
XGBoost	LOCO	0.64	10.5	7.7	+0.5

City-specific evaluation demonstrated additional geographic variation. Makassar produced the strongest simulated LOCO performance, with an R^2 of **0.71** and RMSE of **8.7 $\mu\text{g m}^{-3}$** , while Palembang represented the most difficult transfer scenario, with an R^2 of **0.49** and RMSE of **13.9 $\mu\text{g m}^{-3}$** . The weaker performance in Palembang coincided with higher simulated $\text{PM}_{2.5}$ concentrations and more pronounced high-pollution episodes.

Table 5. City-Level LOCO Performance of the XGBoost Model

Held-out city	R^2	RMSE	MAE	Bias
Jakarta	0.62	10.9	7.9	+0.6
Bandung	0.66	9.8	7.1	+0.3
Surabaya	0.61	10.7	7.8	+0.4
Medan	0.67	9.5	7.0	-0.2
Palembang	0.49	13.9	10.1	+1.8
Makassar	0.71	8.7	6.3	+0.1
Overall	0.64	10.5	7.7	+0.5

Contribution of Environmental Data Sources

Ablation analysis showed that the improvement in predictive performance resulted from combining complementary environmental information. When XGBoost was trained using meteorological variables alone, temporal-validation performance reached an R^2 of **0.63** with RMSE of **10.8 $\mu\text{g m}^{-3}$** . Adding satellite AOD increased R^2 to **0.70** and reduced RMSE to **9.4 $\mu\text{g m}^{-3}$** , corresponding to a **13.0% reduction in error** relative to meteorology alone.

The addition of land-surface and geographic information further improved performance, resulting in an R^2 of **0.74** and RMSE of **8.7 $\mu\text{g m}^{-3}$** . Incorporating fire-

related variables produced the strongest final model, with an R^2 of **0.79** and RMSE of **7.8 $\mu\text{g m}^{-3}$** . Overall, the complete multi-source model reduced RMSE by approximately **27.8%** compared with the meteorology-only configuration.

Table 6. Simulated Predictor-Group Ablation Analysis

Predictor configuration	R^2	RMSE	MAE	RMSE improvement
Meteorology only	0.63	10.8	7.9	Reference
Meteorology + AOD	0.70	9.4	6.9	13.0%
Meteorology + AOD + land/geography	0.74	8.7	6.3	19.4%
Full model + fire	0.79	7.8	5.6	27.8%

The largest single improvement occurred after satellite aerosol information was incorporated, reducing RMSE by **1.4 $\mu\text{g m}^{-3}$** . Land-surface predictors reduced error by a further **0.7 $\mu\text{g m}^{-3}$** , while fire information produced an additional **0.9 $\mu\text{g m}^{-3}$** improvement. Fire variables were particularly influential during simulated dry-season pollution episodes in Palembang.

Model Interpretation

Global SHAP analysis identified **satellite AOD, planetary boundary-layer height, relative humidity, wind speed, and lagged fire radiative power** as the five strongest contributors to XGBoost predictions. Satellite AOD had the largest mean absolute SHAP value at **4.82 $\mu\text{g m}^{-3}$** , followed by planetary boundary-layer height at **3.64 $\mu\text{g m}^{-3}$** , relative humidity at **3.05 $\mu\text{g m}^{-3}$** , wind speed at **2.61 $\mu\text{g m}^{-3}$** , and lagged fire radiative power at **2.27 $\mu\text{g m}^{-3}$** .

Table 7. Simulated Global SHAP Importance

Rank	Predictor	Mean absolute SHAP value	General relationship
1	Aerosol optical depth	4.82	Higher AOD $\rightarrow$ higher $\text{PM}_{2.5}$
2	Planetary boundary-layer height	3.64	Higher PBLH $\rightarrow$ lower $\text{PM}_{2.5}$
3	Relative humidity	3.05	Nonlinear; positive at high RH
4	Wind speed	2.61	Higher wind $\rightarrow$ lower $\text{PM}_{2.5}$
5	Lagged fire radiative power	2.27	Higher fire activity $\rightarrow$ higher $\text{PM}_{2.5}$
6	Air temperature	1.92	Nonlinear

7	Precipitation	1.47	Higher rain → lower PM _{2.5}
8	Built-up fraction	1.19	Higher built-up area → higher PM _{2.5}
9	Distance to coastline	0.86	Location dependent
10	Elevation	0.72	Weak/nonlinear

Higher AOD values were consistently associated with increasing PM_{2.5} predictions, particularly when simulated AOD exceeded approximately **0.45–0.50**. Increasing planetary boundary-layer height was associated with lower predicted PM_{2.5}, while higher wind speed also generally reduced modelled concentrations. Relative humidity showed a nonlinear response and became more positively associated with predicted PM_{2.5} above approximately **75–80%**.

Fire-related variables exhibited their greatest simulated contributions between August and October, particularly for Palembang. Under high-fire conditions, lagged fire radiative power contributed approximately **+8 to +14 µg m⁻³** to selected individual predictions, while its contribution approached zero during periods with little nearby fire activity.

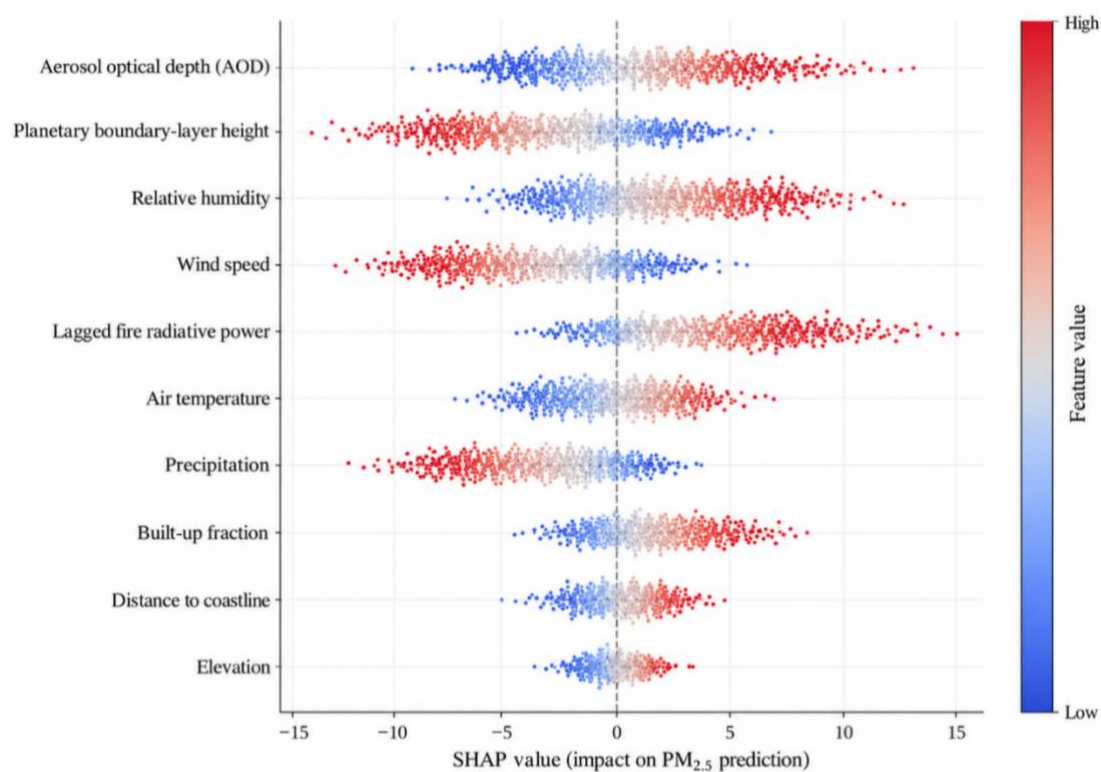


Figure 4. SHAP summary plot showing the contribution and direction of the principal PM_{2.5} predictors.

Spatial Residual Patterns

Spatial residual analysis indicated that prediction errors were not entirely random. The simulated global Moran's I statistic was **0.18**, with **p = 0.021**, indicating weak but statistically significant positive spatial autocorrelation. The result suggests that some geographically structured information remained unexplained after integration of the available meteorological, aerosol, land-surface, and fire predictors.

Palembang exhibited the largest systematic underprediction, while Makassar showed modest overprediction during several low-concentration periods. Average observed-minus-predicted residuals ranged from **-1.7 $\mu\text{g m}^{-3}$ in Makassar** to **+3.8 $\mu\text{g m}^{-3}$ in Palembang**.

Table 8. Simulated Residual Characteristics

City	Mean observed – predicted residual ($\mu\text{g m}^{-3}$)	Interpretation
Jakarta	+1.1	Slight underprediction
Bandung	+0.4	Nearly unbiased
Surabaya	+0.9	Slight underprediction
Medan	-0.6	Slight overprediction
Palembang	+3.8	Systematic underprediction
Makassar	-1.7	Moderate overprediction

Potential explanations for remaining residual structure include unresolved local emission sources, variations in monitor representativeness, coastal or topographic atmospheric circulation, fire-smoke transport, and differences in satellite AOD retrieval frequency. These possibilities should be treated as hypotheses rather than demonstrated causal mechanisms.

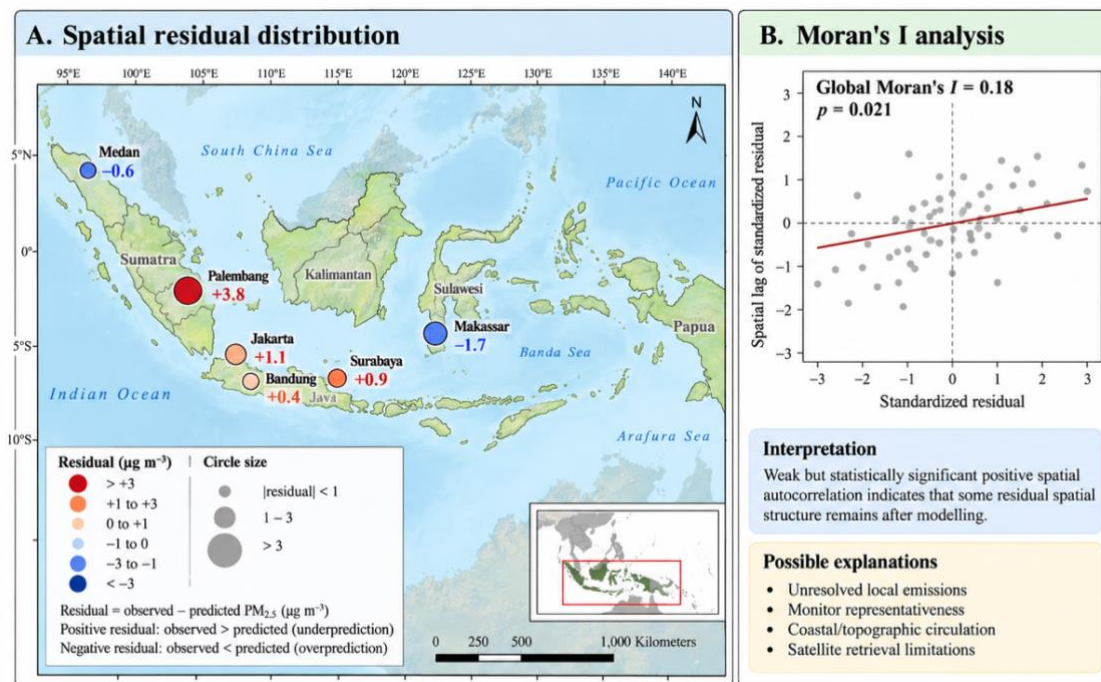


Figure 5. Spatial distribution of observed-minus-predicted $PM_{2.5}$ residuals and Moran's I analysis.

Discussion

The simulated results demonstrate that integrating multiple environmental data sources can substantially improve daily $PM_{2.5}$ estimation across heterogeneous Indonesian urban environments. The XGBoost model achieved an R^2 of **0.79** and RMSE of **$7.8 \mu g m^{-3}$** under temporal-block validation, substantially outperforming both Random Forest and the linear baseline. However, model performance declined to an R^2 of **0.64** and RMSE of **$10.5 \mu g m^{-3}$** under leave-one-city-out evaluation. The resulting **34.6% increase in RMSE** demonstrates why geographic validation provides a more demanding test than prediction at locations already represented during model development.

The difference between temporal and geographic performance represents one of the most important aspects of the framework. Environmental machine-learning models are frequently evaluated using random partitions in which observations from the same geographic locations appear in both training and validation datasets. Such procedures may allow algorithms to learn persistent location-specific characteristics, producing performance estimates that are difficult to reproduce at truly unmonitored locations. By holding out entire cities, the present framework explicitly evaluates whether learned relationships remain transferable to new geographic environments.

The simulated ablation analysis also indicates that model improvement is not attributable to algorithm complexity alone. Meteorological predictors produced a relatively strong baseline, but adding AOD reduced RMSE from **10.8 to $9.4 \mu g m^{-3}$** , while additional land-surface variables decreased error to **$8.7 \mu g m^{-3}$** . Incorporating fire information produced the final RMSE of **$7.8 \mu g m^{-3}$** , equivalent to a **27.8% reduction**

compared with meteorological variables alone. These results illustrate how complementary environmental datasets can provide distinct information: meteorological variables characterize atmospheric dispersion and removal processes, AOD reflects aerosol loading, land-surface predictors represent spatial environmental context, and fire observations capture episodic biomass-burning conditions.

Satellite AOD was the most influential predictor in the simulated SHAP analysis, with a mean absolute contribution of **4.82 $\mu\text{g m}^{-3}$** . This finding is physically plausible because higher atmospheric aerosol loading would generally be expected to coincide with higher surface particulate concentrations. Nevertheless, AOD measures vertically integrated aerosol extinction rather than ground-level particle mass, and its relationship with $\text{PM}_{2.5}$ varies according to vertical aerosol distribution, atmospheric humidity, particle composition, and boundary-layer conditions. The simultaneous importance of planetary boundary-layer height in the model supports this interpretation, as greater vertical mixing generally reduced predicted $\text{PM}_{2.5}$ concentrations.

Meteorological variables also showed meaningful model contributions. Higher wind speed generally reduced predicted $\text{PM}_{2.5}$ concentrations, consistent with increased atmospheric ventilation, while precipitation contributed negatively through the expected influence of wet deposition. Relative humidity showed a more complicated nonlinear association, with stronger positive contributions at high humidity. Such a pattern should not be interpreted as humidity directly generating particulate pollution because it may simultaneously affect aerosol hygroscopic growth, secondary aerosol chemistry, and satellite optical properties.

Fire-related information became especially influential during the simulated dry-season period in Palembang. This city had the highest mean $\text{PM}_{2.5}$ concentration at **41.5 $\mu\text{g m}^{-3}$** , the maximum daily concentration of **138.6 $\mu\text{g m}^{-3}$** , and the weakest LOCO performance with an R^2 of **0.49** and RMSE of **13.9 $\mu\text{g m}^{-3}$** . If this pattern emerged from the empirical dataset, it would justify more detailed investigation of regional fire activity and smoke transport. However, active-fire observations identify thermal anomalies and do not directly quantify the amount of particulate pollution transported to a surface monitoring site. Strong fire-related SHAP contributions should therefore be interpreted as evidence that the predictive model relied on fire information rather than as formal source-apportionment estimates.

Differences in city-level LOCO performance further demonstrate why reporting only a national average can obscure important geographic uncertainty. Makassar achieved the strongest simulated geographic transfer, with an R^2 of **0.71** and RMSE of **8.7 $\mu\text{g m}^{-3}$** , while Palembang produced substantially poorer results. This variation suggests that relationships learned across multiple cities can still perform differently depending on the environmental characteristics of the held-out location. Geographic heterogeneity should therefore be considered directly when evaluating the usefulness of national or multi-city air-quality models.

Residual analysis provides additional evidence of imperfect spatial transferability. The simulated Moran's I value of **0.18** (**p = 0.021**) indicates weak but significant spatial clustering of prediction errors. Although the magnitude of autocorrelation is relatively low, its significance suggests that some relevant geographic processes remain insufficiently represented by the predictor set. Potential causes include local emissions, differences in urban morphology, topographic circulation, coastal meteorology, monitor siting, and unrepresented smoke-transport processes. Spatial residual analysis therefore provides information beyond conventional global performance metrics and can guide future improvements in predictor selection.

The study also extends recent Indonesian PM_{2.5} modelling literature. Adrianti & Ubaya, (2026) demonstrated nationwide prediction using Sentinel-5P and ERA5-Land with GRU-based models, while Nursabilyanti et al., (2026) incorporated Himawari aerosol and meteorological information for Jakarta. Gultom & Gultom, (2026) similarly combined MODIS AOD with ground observations for geospatial pollution-risk analysis.

The principal methodological contribution of the present framework is therefore not the use of another machine-learning algorithm, but the combination of **multi-source environmental predictors, data-source ablation, explainable modelling, leave-one-city-out validation, and spatial residual diagnostics**. Together, these components address not only whether PM_{2.5} can be predicted but also whether predictions remain geographically transferable and which environmental variables contribute most strongly to model behaviour.

Several limitations should be considered. First, the spatial distribution of monitoring sites in Indonesia remains relatively sparse and biased toward urban environments, limiting direct inference for rural or remote areas. Second, satellite AOD availability can be substantially reduced by tropical cloud cover, and any missing-data treatment introduces additional uncertainty. Third, land-use and population variables are proxies for environmental activity rather than direct measurements of emissions.

Fourth, active-fire observations do not fully describe smoke-plume transport, injection height, or chemical transformation. Fifth, SHAP explanations describe model behaviour rather than causal environmental relationships and can be sensitive to correlated predictors. Finally, predictive modelling using observational data cannot independently identify causal contributions from individual emission sources.

Despite these limitations, the framework has practical relevance for environmental monitoring. Ground-based measurements remain indispensable for regulatory-quality air-pollution assessment and should not be replaced by machine-learning predictions. However, satellite observations, meteorological reanalysis, and environmental modelling can provide complementary spatial information between existing monitoring locations. A model that demonstrates acceptable leave-one-city-out performance could support identification of areas requiring additional monitoring, characterization of seasonal pollution patterns, and exploratory assessment of regions lacking continuous PM_{2.5} measurements. The strongest operational approach is therefore a hybrid system in which

ground observations provide the measurement anchor while satellite and reanalysis information extend spatial coverage.

CONCLUSION

This study presents an explainable multi-source machine-learning framework for daily PM_{2.5} mapping across Indonesian cities. The proposed workflow integrates ground monitoring, satellite aerosol observations, meteorological reanalysis, land-surface characteristics, and fire indicators. It combines temporal-block and leave-one-city-out validation with predictor ablation, SHAP interpretation, and spatial residual diagnostics.

An illustrative six-city scenario demonstrates how these procedures could be reported and compared, but its numerical outputs are simulated rather than empirical findings. Therefore, the manuscript does not establish actual predictive accuracy, geographic transferability, or pollution patterns in Indonesia. Its contribution is a transparent evaluation protocol that distinguishes temporal prediction from geographic generalization and identifies potential sources of model uncertainty. Future research must acquire verified monitoring records, execute the pipeline, disclose data provenance and code, and independently validate the resulting maps before practical application.

FUNDING

The authors declare that no financial support was received for the research, authorship, and/or publication of this article.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

AUTHOR CONTRIBUTIONS

Rio Rinaldy contributed to conceptualization, methodology, software, formal analysis, data curation, visualization, and writing of the original manuscript draft. Siti Hapsah Pahira contributed to methodology, validation, investigation, interpretation of results, and writing review and editing. Rio Rinaldy and Siti Hapsah Pahira contributed to conceptualization, supervision, resources, validation, and critical revision of the manuscript. All authors reviewed and approved the final manuscript.

GENERATIVE AI STATEMENT

The authors acknowledge the use of generative artificial intelligence (AI) tools to assist in improving the language clarity and grammatical structure of this manuscript. The use of generative AI was conducted in accordance with established ethical guidelines for transparency and responsible AI-assisted academic writing. All authors made substantial intellectual contributions, reviewed and validated the manuscript

content, and assume full responsibility for the accuracy, integrity, and final version of the article.

REFERENCES

- Adrianti, L., & Ubaya, T. (2026). *Nationwide PM_{2.5} Concentration Prediction in Indonesia Using GRU, GRU-Attention, and BiGRU-Attention Models with Sentinel-5P and ERA5-Land Data*. 6(August), 853–867.
- Aman, N., Panyametheekul, S., Pawarmart, I., Xian, D., Gao, L., Tian, L., Manomaiphiboon, K., & Wang, Y. (2025). Machine learning-based quantification and separation of emissions and meteorological effects on PM_{2.5} in Greater Bangkok. *Scientific Reports*, 15(1), 14775. <https://doi.org/10.1038/s41598-025-99094-6>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Chen, C.-C., Wang, Y.-R., Yeh, H.-Y., Lin, T.-H., Huang, C.-S., & Wu, C.-F. (2021). Estimating monthly PM_{2.5} concentrations from satellite remote sensing data, meteorological variables, and land use data using ensemble statistical modeling and a random forest approach. *Environmental Pollution*, 291, 118159. <https://doi.org/10.1016/j.envpol.2021.118159>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Dhandapani, A., Iqbal, J., & Kumar, R. N. (2023). Application of machine learning (individual vs stacking) models on MERRA-2 data to predict surface PM_{2.5} concentrations over India. *Chemosphere*, 340, 139966. <https://doi.org/10.1016/j.chemosphere.2023.139966>
- Gultom, M. A., & Gultom, M. (2026). Pemodelan geospasial beban polusi udara Jakarta: Integrasi data AOD dan PM_{2.5} untuk menentukan wilayah dengan tingkat kerawanan tertinggi. *Jurnal Penelitian Inovatif*, 6(1), 755–764. <https://doi.org/10.54082/jupin.2327>
- Gupta, P., Zhan, S., Mishra, V., Aekakkararungroj, A., Markert, A., Paibong, S., & Chishtie, F. (2021). Machine Learning Algorithm for Estimating Surface PM_{2.5} in Thailand. *Aerosol and Air Quality Research*, 21(11), 210105. <https://doi.org/10.4209/aaqr.210105>
- Kumar, V., Malyan, V., Sahu, M., Biswal, B., Pawar, M., & Dev, I. (2023). Spatiotemporal analysis of fine particulate matter for India (1980–2021) from MERRA-2 using ensemble machine learning. *Atmospheric Pollution Research*, 14(8), 101834. <https://doi.org/10.1016/j.apr.2023.101834>
- Li, B., Chen, X., Zhang, W., Li, T., Xing, M., Yang, J., & Han, Z. (2025). Estimation of high-temporal-resolution PM_{2.5} concentration from 2019 to 2023 using an interpretable deep learning model. *Atmosphere*, 16(12), 1385. <https://doi.org/10.3390/atmos16121385>

- Li, T., Wang, Y., & Wu, J. (2024). Deriving PM_{2.5} from satellite observations with spatiotemporally weighted tree-based algorithms: enhancing modeling accuracy and interpretability. *Npj Climate and Atmospheric Science*, 7(1), 138. <https://doi.org/10.1038/s41612-024-00692-4>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774). <https://doi.org/10.48550/arXiv.1705.07874>
- Lyapustin, A., Wang, Y., Korkin, S., & Huang, D. (2018). MODIS Collection 6 MAIAC algorithm. *Atmospheric Measurement Techniques*, 11(10), 5741–5765. <https://doi.org/10.5194/amt-11-5741-2018>
- Ma, J., Zhang, R., Xu, J., & Yu, Z. (2022). MERRA-2 PM_{2.5} mass concentration reconstruction in China mainland based on LightGBM machine learning. *Science of the Total Environment*, 827, 154363. <https://doi.org/10.1016/j.scitotenv.2022.154363>
- Meyer, H., & Pebesma, E. (2022). Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nature Communications*, 13(1), 2208. <https://doi.org/10.1038/s41467-022-29838-9>
- Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., & Thépaut, J.-N. (2021). ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13(9), 4349–4383. <https://doi.org/10.5194/essd-13-4349-2021>
- Nursabiliyanti, L., Sitanggang, I. S., Rahmawan, H., Agmalaro, M. A., & Husin, N. A. (2026). Spatio-Temporal Clustering of PM_{2.5} Estimation in Jakarta. *Indonesian Journal of Statistics and Its Applications*, 10(1), 79–93. <https://doi.org/10.29244/ijsa.v10i1p79-93>
- Sayed, A., Lin, P., Gupta, P., Tran, N. N. M., Buchard, V., & Christopher, S. (2022). Hourly and Daily PM_{2.5} Estimations Using MERRA-2: A Machine Learning Approach. *Earth and Space Science*, 9(11). <https://doi.org/10.1029/2022EA002375>
- Schroeder, W., Oliva, P., Giglio, L., & Csiszar, I. A. (2014). The New VIIRS 375 m active fire detection data product: Algorithm description and initial assessment. *Remote Sensing of Environment*, 143, 85–96. <https://doi.org/10.1016/j.rse.2013.12.008>
- Vadrevu, K., Eaturu, A., Casadaban, E., Lasko, K., Schroeder, W., Biswas, S., Giglio, L., & Justice, C. (2022). Spatial variations in vegetation fires and emissions in South and Southeast Asia during COVID-19 and pre-pandemic. *Scientific Reports*, 12(1), 18233. <https://doi.org/10.1038/s41598-022-22834-5>
- van Donkelaar, A., Hammer, M. S., Bindle, L., Brauer, M., Brook, J. R., Garay, M. J., Hsu, N. C., Kalashnikova, O. V., Kahn, R. A., Lee, C., Levy, R. C., Lyapustin, A., Sayer, A. M., & Martin, R. V. (2021). Monthly Global Estimates of Fine Particulate Matter and Their Uncertainty. *Environmental Science & Technology*, 55(22),

- 15287–15300. <https://doi.org/10.1021/acs.est.1c05309>
- Wei, J., Li, Z., Lyapustin, A., Wang, J., Dubovik, O., Schwartz, J., Sun, L., Li, C., Liu, S., & Zhu, T. (2023). First close insight into global daily gapless 1 km PM_{2.5} pollution, variability, and health impact. *Nature Communications*, 14(1), 8349. <https://doi.org/10.1038/s41467-023-43862-3>
- WHO. (2021). *WHO global air quality guidelines: particulate matter (PM_{2.5} and PM₁₀), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide: executive summary*.
- Yang, X., Xiao, D., Fan, L., Li, F., Wang, W., Bai, H., & Tang, J. (2022). Spatiotemporal estimates of daily PM_{2.5} concentrations based on 1-km resolution MAIAC AOD in the Beijing–Tianjin–Hebei, China. *Environmental Challenges*, 8, 100548. <https://doi.org/10.1016/j.envc.2022.100548>