Satellite and Reanalysis-Based Estimation of Surface $PM_{2.5}$ Concentrations Using Bayesian Averaging and Random Forest Models

Tran, T. A.,[1*] Le, T. K.,[1] Nguyen, T. L.[2] and Thai, H. A. C.[3]
[1]Thu Dau Mot University, Ho Chi Minh City, Vietnam, E-mail: antt@tdmu.edu.vn*
[2]National Economics University, Hanoi, Vietnam
[3]Eastern International University, Ho Chi Minh City, Vietnam
**Corresponding Author*

DOI: https://doi.org/10.52939/ijg.v22i10.5238

Abstract
Air pollution characterized by elevated fine particulate matter ($PM_{2.5}$) is a critical environmental and public health challenge in rapidly industrializing regions where ground-based monitoring networks remain sparse. This study evaluates and compares two predictive models - a probabilistic linear ensemble via Bayesian Model Averaging (BMA) and a non-parametric machine learning algorithm using Random Forest (RF) - to estimate surface $PM_{2.5}$ concentrations across Binh Duong province, located in the North of Ho Chi Minh City, Vietnam during the dry seasons of 2024 - 2025. The models integrate multi-source satellite observations, including MODIS Aerosol Optical Depth (AOD), Sentinel-5P TROPOMI tropospheric pollutants (NO_2, SO_2, and CO), and ERA5 meteorological reanalysis variables. Model parameters and predictive robustness were optimized and evaluated using repeated 5-fold cross-validation. The cross-validation results revealed that the Random Forest model outperformed the BMA model, achieving an R^2 of 0.703, MAE of 4.594 ± 0.48 μg/m³, RMSE of 5.608 ± 0.62 μg/m³, and MAPE of 39.621%. In comparison, the BMA model yielded a validation R^2 of 0.626, MAE of 4.839 μg/m³, RMSE of 5.992 μg/m³, and MAPE of 40.852%. The moderate MAPE values (~40%) indicate that while the framework is not suitable for strict regulatory compliance enforcement, it serves as an effective screening and spatial assessment tool. Satellite-derived surface maps successfully reproduced regional spatial gradients, highlighting pronounced $PM_{2.5}$ concentrations in southern industrial corridors (Di An, Thuan An, and Thu Dau Mot) contrasting with lower levels in northern rural districts. These findings demonstrate the practical utility of integrating satellite remote sensing with machine learning for dry-season air quality screening and environmental management in data-limited developing regions.

Keywords: $PM_{2.5}$ Estimation, Aerosol Optical Depth, Sentinel-5P, ERA5 Reanalysis, Random Forest, Bayesian Model Averaging, Binh Duong, Ho Chi Minh City

1. Introduction

The World Health Organization (WHO) identified air pollution and climate change among the top ten threats to global health in 2019 [1]. According to WHO, up to 97% of cities in low- and middle-income countries with populations over 100,000 do not meet international air quality guidelines [2]. In Southeast Asia, rapid economic growth, urbanization, and industrial development have increased air pollution, especially fine particulate matter ($PM_{2.5}$), which consists of particles with a diameter of 2.5 μm or smaller [3]. Owing to their small size, $PM_{2.5}$ particles can travel deep into the lungs and enter the bloodstream. This can cause inflammation, oxidative stress, and problems with blood vessels. Long-term exposure to $PM_{2.5}$ is associated with a higher risk of premature death and hospital admissions, particularly from chronic obstructive pulmonary disease, stroke, and ischemic heart disease [4][5][6] and [7].

Currently, ambient air quality monitoring in Vietnam relies primarily on ground-based observation networks operated by provincial Departments of Natural Resources and Environment (DONRE) and regional environmental monitoring centers [8] and [9]. Although automated reference stations provide high-precision, continuous measurements of criteria pollutants, their spatial coverage is exceedingly sparse due to substantial capital installation, routine recalibration, and maintenance costs. As a result, existing ground

monitoring stations do not cover many suburban, industrial, and rural areas. This creates important gaps in the data needed to assess population exposure to air pollution across the study area.

To bridge these spatial monitoring gaps, satellite-based Earth observation has emerged as an indispensable methodology [10] [11] and [12]. Satellite-derived Aerosol Optical Depth (AOD) measures total column aerosol light extinction, serving as an effective physical proxy for surface particulate matter. However, converting satellite-based AOD data into accurate ground-level $PM_{2.5}$ concentrations is challenging. The relationship between AOD and $PM_{2.5}$ changes across different locations and times because of variations in the height of the planetary boundary layer, the vertical distribution of aerosols, relative humidity, and aerosol composition. In addition, satellite observations are available only at specific overpass times, while ground monitoring stations collect data continuously. This difference in observation times can cause temporal mismatches between satellite and ground measurements. Therefore, meteorological factors and other relevant variables need to be considered in multivariate models to improve the accuracy of $PM_{2.5}$ estimation.

Recent advances in machine learning (ML) and statistical regression have significantly enhanced satellite-based $PM_{2.5}$ retrieval by coupling AOD with multi-source meteorological reanalysis and atmospheric trace gas products [13][14] and [15]. Statistical ensemble techniques such as Bayesian Model Averaging (BMA) have demonstrated notable strength in accounting for model uncertainty, exploring candidate model spaces, and determining the posterior inclusion probabilities (PIP) of competing predictors [16] [17] and [18]. Concurrently, non-parametric algorithms such as Random Forest (RF) are widely favored due to their ability to capture complex non-linear relationships and high-order predictor interactions without imposing rigid distributional assumptions [19][20] and [21]. While RF provides superior predictive flexibility for complex non-linear systems, BMA provides transparent probabilistic interpretation and parsimonious variable selection. A systematic comparison of these two fundamentally different paradigms provides valuable scientific insights into whether non-linear flexibility outweighs linear probabilistic interpretability under data-scarce tropical monsoon conditions [22][23] and [24]. In order to evaluate the model performances, the k-Fold Cross-Validation is a resampling technique used in predictive models by ensuring every data point has a chance to be used for both training and testing [25] [26] and [27]. It provides a much more robust estimate of a model's generalization ability than a traditional single train-test split, which can be heavily biased depending on how the data happens to fall.

Cloud-based computational platforms such as Google Earth Engine (GEE), combined with open-source scientific computing environments, have revolutionized the processing of large scale geospatial archives [28]. The integration of MODIS AOD, atmospheric measurements from the Sentinel-5P TROPOMI instrument, and meteorological data from the ERA5 reanalysis provides a useful multi-source framework for regional environmental assessment. The primary objective of this study is to develop a multi-source satellite and reanalysis-based modeling framework for estimating surface $PM_{2.5}$ concentrations across Binh Duong province, located in North of Ho Chi Minh City, Vietnam during the dry seasons (December - April) of 2024 - 2025, while systematically comparing the predictive performance and physical interpretations of Random Forest and Bayesian Model Averaging. The analysis focuses on the dry season because frequent cloud cover and heavy rainfall during the wet season can reduce the quality of optical satellite observations. These conditions can also lead to missing data and errors in satellite-based measurements. The study aims to: (1) evaluate the importance of satellite AOD, Sentinel-5P trace gases, and ERA5 meteorological variables for $PM_{2.5}$ estimation; (2) evaluate model performance and generalizability using cross-validation; and (3) produce continuous 1 km resolution $PM_{2.5}$ distribution maps for the dry season to support local air-quality management and public health planning.

2. Study Area and Data Used

2.1 Study Area

The study area encompasses Binh Duong Province, located in the North of Ho Chi Minh City, Vietnam between latitudes 10°51'46"N to 11°30'00"N and longitudes 106°20'00"E to 106°58'00"E, covering an administrative area of approximately 2,694.4 km² (Figure 1). Binh Duong is bordered by Binh Phuoc Province to the north, Ho Chi Minh City to the south, Dong Nai Province to the east, and Tay Ninh Province to the west. Following the administrative reorganization implemented by the Vietnamese Government in July 2025, Binh Duong Province was merged into Ho Chi Minh City. However, because the analysis in this study is based on data from the 2024 - 2025 dry season, the study area is defined according to the former administrative boundary of Binh Duong Province.

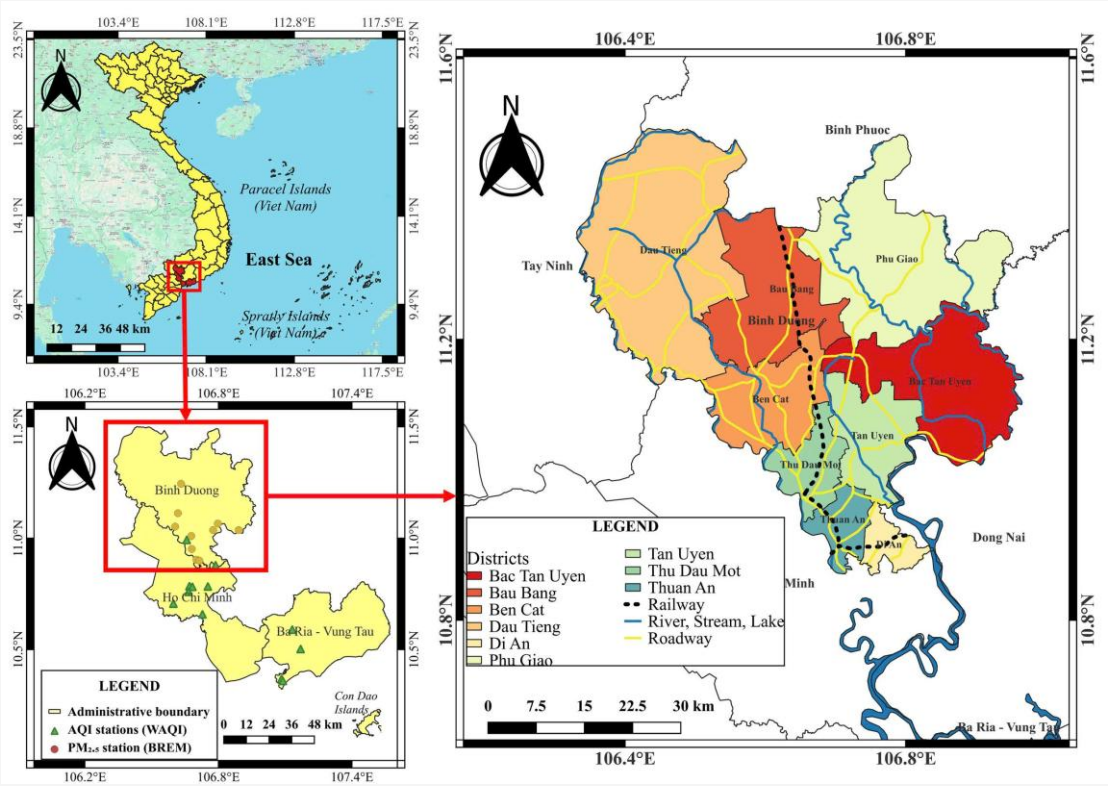


Figure 1: Location of Binh Duong province and the $PM_{2.5}$ monitoring stations

To increase the number of observations and improve model performance, data from the former Binh Duong Province, Ho Chi Minh City, and Ba Ria - Vung Tau Province were used to train and optimize the predictive models. In the second stage, the spatial analysis and mapping of $PM_{2.5}$ concentrations were conducted only within the former Binh Duong Province boundary.

Binh Duong has experienced rapid industrial and economic development and has become an important industrial center in southern Vietnam. The province has 29 industrial parks and 12 industrial clusters distributed across its nine administrative districts [29]. Industrial activities such as manufacturing, chemical processing, electronics production, and furniture manufacturing are mainly concentrated in the southern urban areas, including Di An, Thuan An, and Thu Dau Mot City. These activities can produce significant air emissions from industrial facilities, diesel generators, and heavy-duty vehicles used for transporting goods along major roads. In contrast, the northern areas, including Dau Tieng, Phu Giao, and Bac Tan Uyen districts, have lower population densities and are mainly characterized by rubber plantations, agricultural land, and forested areas. The study area experiences a tropical monsoon climate characterized by two distinct seasons: a rainy season spanning May to November and a dry season extending from December to April [29][30] and [31]. The dry season is characterized by high ambient temperatures (daily maxima frequently exceeding 35°C), low relative humidity (often dropping below 50%), prolonged sunshine hours, and prevailing northeasterly and easterly monsoon winds. These meteorological conditions promote intense photochemical reactions, atmospheric stagnation, and thermal boundary-layer inversions that inhibit pollutant dispersion, resulting in persistent particulate accumulation. Consequently, fine particulate pollution peaks annually during the dry season, creating severe respiratory health risks for urban populations.

2.2. $PM_{2.5}$ Concentrations Data

Ground-level $PM_{2.5}$ monitoring observations were assembled from two complementary monitoring sources: 10 regulatory reference stations operated by the Ho Chi Minh City Center of Natural Resources and Environmental Monitoring (HCEM) (https://moitruonghcm.vn/) [32] and 12 continuous monitoring stations aggregated via the World Air Quality Index (WAQI) project (https://aqicn.org/) [33] across the study domain (Figure 1). The HCEM stations represent regulatory-grade monitoring facilities operating automated and calibrated according to Vietnamese national standards (QCVN

05:2023/BTNMT). The WAQI nodes include both regulatory stations and community-based sensors that undergo EPA-standardized quality assessment protocols [34][35] and [36]. To ensure data validity and eliminate sensor anomalies, a rigorous screening procedure was conducted: (1) temporal synchronization: ground observations were matched within a ±1 hour temporal window of satellite daytime overpass times (Terra at ~10:30 and Aqua/Sentinel-5P at ~13:30 local solar time); (2) anomaly screening: negative readings, persistent zero values, and extreme values (>500 μg/m³) were removed; and (3) completeness filtering: daily records required at least 75% valid hourly data. After quality screening, a total of 569 spatio-temporally matched samples across the 2024 - 2025 dry seasons (December 2023 to April 2024, and December 2024 to April 2025) were compiled for model training and validation.

2.3 Remote Sensing Data

Remote sensing datasets and meteorological reanalysis fields were accessed and preprocessed using Google Earth Engine (GEE), a high-performance cloud computing platform for regional scale geospatial data analysis [37]. GEE enables parallel extraction, spatial aggregation, and masking of massive multi-sensor archives, ensuring computational reproducibility. Aerosol Optical Depth (AOD) was derived from the MODIS MCD19A2 Version 6.1 Level-2 gridded product generated by the Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm at 1 km spatial resolution [38][39] and [40]. The MCD19A2 product utilizes calibrated radiance from both Terra and Aqua platforms, retrieving AOD over land at the blue (0.47 μm) and green (0.55 μm) wavelengths. Multi-Angle Implementation of Atmospheric Correction (MAIAC) AOD was selected over conventional MODIS Collection products because its 1 km spatial resolution is essential for resolving localized aerosol gradients across heterogeneous urban and industrial landscapes. Furthermore, MAIAC's multi-angle time-series approach substantially mitigates surface reflectance contamination over bright surfaces and mixed vegetation, offering superior accuracy in tropical environments.

Tropospheric trace gas column densities were acquired from the TROPOspheric Monitoring Instrument (TROPOMI) on the Copernicus Sentinel-5 Precursor (Sentinel-5P) satellite [41] and [42]. TROPOMI provides daily global coverage across UV, visible, near-infrared, and shortwave infrared spectral bands with a native nadir resolution of 3.5 × 7 km². Three tropospheric trace gas products were investigated via their official GEE image collections: Nitrogen Dioxide (NO_2) using 'COPERNICUS/S5 P/NRTI/L3_NO_2' (variable 'tropospheric_NO_2_colu mn_number_density') [43]; Sulfur Dioxide (SO_2) using 'COPERNICUS/S5P/OFFL/ L3_SO_2' (variable 'SO_2_column_number_density') [44]; and Carbon Monoxide (CO) using 'COPERNICUS/S5P/ NRTI/L3_CO' (variable 'CO_column_number_den sity') [45]. These trace gases which are measured as column number density in mol/m², serve as physical proxies for primary combustion emissions, vehicular traffic, and industrial activity, providing vital precursor information for secondary aerosol formation (Table 1). Meteorological reanalysis data were retrieved from the ECMWF ERA5 hourly dataset on single levels at a native spatial resolution of 0.25° × 0.25° (~31 km) [46].

Table 1: Remote sensing and reanalysis-based data used in the study

Data Type	Level	Temporal Resolution	Spatial Resolution (km)	Format	References
MODIS AOD	2	Daily	1 x 1	Raster	[39]
Sentinel-5P Nitrogen Dioxide (NO_2)	3	Daily	7 × 3.5	Raster	[43]
Sentinel-5P Sulphur Dioxide (SO_2)	3	Daily	7 × 3.5	Raster	[44]
Sentinel-5P Carbon Monoxide (CO)	3	Daily	7 × 3.5	Raster	[45]
ERA5 Meteorological reanalysis data	Single	Hourly	~31 x 31	Raster	[46]

Hourly ERA5 fields between 09:00 and 15:00 local time were aggregated into daytime mean values to match the satellite overpass window. Key meteorological covariates included air temperature (T), dewpoint temperature (T_d), total precipitation, surface pressure, eastward wind component (U), and northward wind component (V) [47] and [48]. Relative humidity (RH) and horizontal wind speed (WS) were derived according to Equations 1 and 2 [49]:

$$RH = \frac{e^{\frac{17.625T_d}{243.04+T_d}}}{e^{\frac{17.625T}{243.04+T}}} \times 100$$

Equation 1

Where T is the air temperature (°C), and T_d is the dew point temperature (°C).

$$WS = \sqrt{U^2 + V^2}$$

Equation 2

where U is the eastward wind component (m/s), and V is the northward wind component (m/s).

Because multi-sensor datasets have disparate native spatial resolutions (MODIS AOD at 1 km, Sentinel-5P at 3.5 × 7 km², and ERA5 meteorology at ~31 km), all predictor grids were resampled to a common 1 km grid via bilinear interpolation matching the MCD19A2 AOD grid [38][39] and [40]. It is essential to clarify that while bilinear interpolation produces a continuous, uniform grid necessary for pixel-level predictive modeling, it does not create fine-scale micro-meteorological phenomena. Resampling ERA5 and Sentinel-5P provides a smooth spatial background for regional meteorological and trace gas fields. The spatial uncertainty resulting from downscaling coarse reanalysis data is recognized as an inherent scale constraint in regional satellite modeling.

3. Methodology

The overall methodology integrates satellite remote sensing, atmospheric reanalysis, and machine learning into a four-stage workflow (Figure 2): (1) data collection and quality screening of ground $PM_{2.5}$, MODIS AOD, Sentinel-5P trace gases, and ERA5 meteorological data; (2) multi-sensor spatio-temporal matching and 1 km grid harmonization; (3) model training, hyperparameters optimization, and repeated 5-fold cross-validation for Random Forest and Bayesian Model Averaging; and (4) continuous pixel-level raster simulation and environmental interpretation across Binh Duong Province.

3.1 Bayesian Model Average Method

The Bayesian Model Averaging (BMA) method utilizes the Bayes Factor (BF) and the Bayesian Information Criterion (BIC) to measure the balance between model complexity and predictive ability in order to select the optimal model. BMA is a coherent statistical framework designed to address model selection uncertainty by evaluating an ensemble of candidate regression models rather than conditioning inference on a single arbitrarily chosen model [18]. In linear $PM_{2.5}$ modeling, let Y denote the dependent variable (surface $PM_{2.5}$ concentration) and x_1, x_2, ..., x_p represent candidate predictor variables, represented as Equation 3:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon$$

Equation 3

Where Y is the dependent variable (surface $PM_{2.5}$ concentration in μg/m³), x_j (j = 1, ..., p) are candidate predictor variables, β_0 is the intercept, β_j are regression coefficients, and ε is the residual error term.

All candidate models were initially assumed to have the same prior probability. The posterior probability of each model was then calculated using the BIC, where a lower BIC indicates a better balance between model fit and model complexity [18]. The importance of each predictor was evaluated using the Posterior Inclusion Probability (PIP). PIP indicates how frequently a predictor appears in the models that have strong support from the observed data. A high PIP means that the predictor is likely to be important for explaining $PM_{2.5}$ variation, while a low PIP indicates weaker evidence for its contribution. In this study, 11 predictor variables were considered, resulting in 2^{11}=2,048 possible combinations of predictors. These models were evaluated using the BMA package in R to focus on models with strong statistical support. Although BMA can combine predictions from multiple models using their posterior probabilities, this study selected the model with the lowest BIC and the highest posterior probability for the final spatial prediction. Bayesian Model Averaging (BMA) plays an important role in air quality estimation by addressing model uncertainty and improving prediction reliability. Instead of relying on a single model, BMA combines multiple candidate models and assigns weights to each based on their posterior probabilities. This allows the final prediction to reflect the contribution of different models, leading to more robust and stable estimates of air pollutants such as $PM_{2.5}$ [16] and [17].

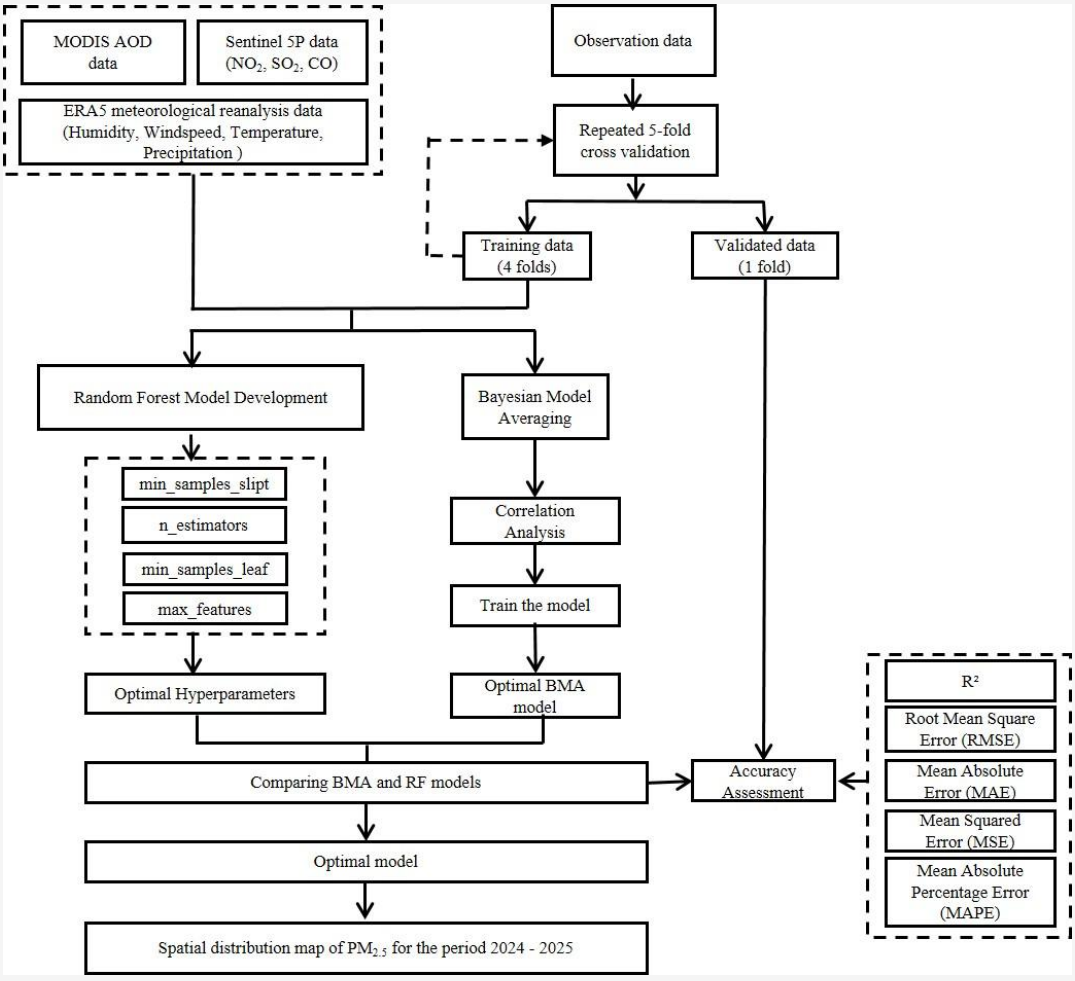


Figure 2: Flowchart of data processing on PM2.5 estimation

3.2 The Random Forest (RF) Method

Random Forest (RF) is a non-parametric ensemble machine learning method that combines multiple decision trees using bootstrap sampling (bagging) and random selection of predictor variables [19]. Feature sampling is referred to as column sampling, while data point sampling is known as row sampling. Each tree is trained using both row and column subsets. This approach enhances the robustness of the model, allowing it to perform well when predicting unseen data [19]. The Random Forest prediction can be mathematically represented as Equation 4 [19]:

$$RF = \frac{1}{K}\sum_{k=1}^{K} h_k(x) \qquad \text{Equation 4}$$

Where:

RF is the final prediction of the Random Forest model

K is the total number of decision trees in the forest

$h_k(x)$ is the prediction of the k-th decision tree for the input x

x is the input vector, consisting of independent variables (features)

The Random Forest model was constructed by combining multiple decision trees, each trained on a bootstrap sample randomly drawn with replacement from the training dataset (Bagging technique). The main parameters tuned in this study included the number of trees in the forest (n_estimators), the minimum number of samples required at each leaf node (min_samples_leaf), the minimum number of samples required to split an internal node (min_samples_split), and the number of predictor variables randomly selected at each node split (max_features) [21]:

n_estimators: this parameter represents the number of decision trees in the random forest. Increasing the number of trees generally enhances the stability and accuracy of the model by reducing variance; however, it also leads to longer training and prediction times due to higher computational demands [21].

min_samples_leaf: this parameter specifies the minimum number of samples required to be present in a leaf node, which represents the terminal node of a decision tree. It controls the minimum size of each leaf and affects the granularity of the model. A very small value may cause the model to fit random fluctuations in the data, resulting in overfitting, while a larger value ensures that each leaf contains more samples, leading to smoother predictions and reduced noise, albeit at the cost of potentially losing fine-grained details [21].

min_samples_split: this parameter defines the minimum number of samples required to split an internal node. It determines how deeply the decision trees can grow during training. A smaller value allows the trees to split more frequently and become deeper, enabling the model to capture more detailed patterns but increasing the risk of overfitting. Conversely, a larger value restricts the number of splits, leading to simpler trees and potentially better generalization performance [21].

max_features: this parameter in the Random Forest model represents the number of predictor variables randomly selected at each node split. This parameter controls the degree of randomness among trees and was optimized through cross-validation to improve model generalization [21].

In this study, the Random Forest (RF) model was implemented in Python 3.10 using several supporting packages, including scikit-learn v1.2, pandas, and geopandas. The model validation and accuracy assessment were also performed using Python-based procedures. Finally, the estimated $PM_{2.5}$ concentrations were exported in raster format and visualized as spatial distribution maps using the QGIS environment.

3.3 K-fold Cross Validation and Accuracy Assessment

To evaluate model performance and stability on independent data, a 5-fold cross-validation procedure was used. The 569 spatio-temporally matched samples were divided into five nearly equal groups, with 114 samples in each of the first four folds and 113 samples in the fifth fold. In each iteration, four folds were used to train and calibrate the model, while the remaining fold was used for validating. This process was repeated five times so that each fold was used once as the validation set. It is important to acknowledge that standard random K-fold cross-validation on spatial-temporal data can introduce potential data leakage due to spatial autocorrelation among monitoring stations located in close geographical proximity [25] and [26]. While Spatial Block Cross-Validation or Leave-Location-Out Cross-Validation (LLOCV) provides a more rigorous test of pure spatial extrapolation to completely unmonitored regions [50], the relatively sparse distribution of reference stations across the study area precluded reliable spatial blocking. Consequently, the cross-validation metrics reported in this study represent regional spatial-temporal interpolation performance under operational monitoring network conditions.

In the model evaluation stage, the predictive performance of the proposed BMA and RF models was assessed using the Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2). These performance metrics were calculated according to Equations 5 to 9:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|$$

Equation 5

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}\right)^2$$

Equation 6

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}\right)^2}$$

Equation 7

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Equation 8

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

Equation 9

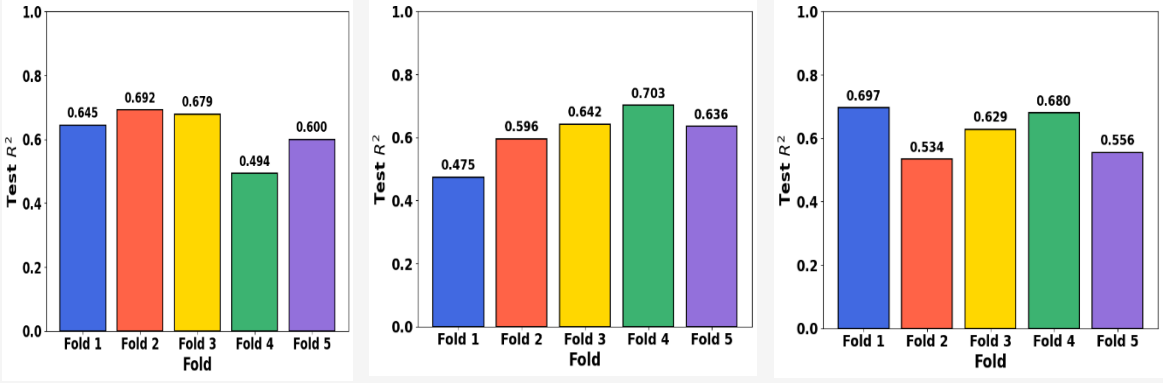
Where n is the total number of validation observations; y_i is the ground truth observed $PM_{2.5}$ concentration ($\mu g/m^3$); $\hat{y}_i$ is the predicted $PM_{2.5}$ concentration ($\mu g/m^3$); and $\bar{y}$ is the arithmetic mean

of observed concentrations (µg/m³). Units for the evaluation metrics are: MAE and RMSE are expressed in µg/m³, MSE is expressed in (µg/m³)², MAPE is expressed in percentage (%), and R^2 is dimensionless.

4. Results and Discussions

4.1 Random Forest Model Performance for $PM_{2.5}$ Estimation

The $PM_{2.5}$ observational dataset was partitioned into five folds of nearly equal size. This balanced distribution of samples across the folds minimizes the potential influence of sample size imbalance on model evaluation, thereby ensuring the objectivity and reliability of the 5-fold cross-validation procedure. During the validation process, each fold was sequentially designated as the test set, while the remaining four folds were used for model training. This approach provides a comprehensive assessment of the model's predictive performance on independent datasets, reduces the dependence of evaluation results on a single train - test partition, and enhances the robustness of the estimated model performance (Figure 3). The evaluation results from the three independent repetitions showed that the Test R^2 values ranged from 0.475 to 0.703. Specifically, the Test R^2 values ranged from 0.494 to 0.692 in the first repetition (Figure 3(a)), from 0.475 to 0.703 in the second repetition (Figure 3(b)), and from 0.534 to 0.697 in the third repetition (Figure 3c), indicating relatively consistent predictive performance across repeated evaluations. Analysing from all of validation folds, the mean test $R^2 = 0.612 \pm 0.084$, validation RMSE = 5.608 ± 0.62 µg/m³, and MAE = 4.594 ± 0.48 µg/m³. Although variations in Test R^2 were observed among the folds, the magnitude of these differences was relatively small, and no substantial decline in model performance was detected in any repetition. The third repetition exhibited the narrowest range of Test R^2 values, suggesting greater stability in predictive performance across the five folds. Overall, all Test R^2 values exceeded 0.47, with the majority of folds achieving values above 0.60, demonstrating satisfactory predictive capability and consistent model performance across different validation subsets. These results indicate that the proposed model possesses good generalization ability and is only minimally affected by variations in data partitioning during the cross-validation process. In the Random Forest (RF) model, optimizing hyperparameters plays a crucial role in determining model performance [21]. Table 2 presents the results of hyperparameter tuning conducted to determine the optimal configuration for the Random Forest model. The number of trees (n_estimators) was tested within a wide range (1–500), with the optimal value identified as 200.



(a) (b) (c)

Figure 3: Test R^2 values for each fold across three independent repetitions of 5-fold cross-validation: (a) first repetition, (b) second repetition and (c) third repetition

Table 2: Hyperparameters tunning in Random Forest model for PM2.5 estimation

Hyperparameters	Adjustment range	Optimal parameters
n_estimators	1 – 500	200
min_samples_leaf	1 – 10	4
min_samples_split	2 – 10	2
max_features	1 – 11	11

This relatively large number of trees suggests that increasing ensemble size improves model stability and predictive performance, while reducing variance.The parameter min_samples_leaf was optimized at 4, indicating that each terminal node contains at least four samples. This helps prevent overfitting by avoiding overly specific splits and ensures better generalization. Similarly, min_samples_split was selected as 2, which allows the model to fully explore potential splits while still maintaining flexibility in capturing data patterns. For max_features, the optimal value was 11, corresponding to the total number of input variables. This result implies that using all available features at each split yields the best performance for this dataset, highlighting the importance of incorporating the full set of predictors in modeling $PM_{2.5}$. It is obvious that the selected hyperparameters reflect a balance between model complexity and generalization ability, enabling the Random Forest model to effectively capture nonlinear relationships while minimizing overfitting (Table 2). Figure 4 illustrates the relative importance of the predictor variables in the Random Forest model. AOD emerged as the most influential variable, contributing approximately 62% of the total feature importance, indicating that satellite-derived aerosol information is the primary driver of $PM_{2.5}$ estimation. This finding is consistent with the well-established relationship between aerosol optical depth and near-surface particulate matter concentrations. CO was the second most important predictor (approximately 6–7%), whereas meteorological variables, including the U and V wind components, wind speed, dew point temperature, and relative humidity, exhibited moderate contributions ranging from approximately 3% to 5%. These variables reflect the effects of atmospheric transport, dilution, and moisture conditions on $PM_{2.5}$ distribution. By comparison, NO_2, precipitation, air temperature, and SO_2 showed relatively minor contributions (<3%), suggesting that they provide only limited additional information beyond that captured by AOD and the major meteorological variables. While variable importance can exhibit sensitivity to correlated features, these results demonstrate that the model is physically anchored in satellite aerosol optical depth while utilizing trace gases and meteorology for secondary atmospheric modulation.

Figure 5 displays the scatter plots between observed and predicted $PM_{2.5}$ concentrations for the training and validation datasets. On the training set (Figure 5(a)), the RF model achieved R^2 = 0.831, RMSE = 4.202 µg/m³, MAE = 3.015 µg/m³, and MAPE = 21.230%. On the independent validation set (Figure 5(b)), performance showed an expected degradation to R^2 = 0.703, RMSE = 5.608 µg/m³, MAE = 4.594 µg/m³, and MAPE = 39.621%. The decline in R^2 reflects out-of-sample error and the non-parametric nature of tree ensembles, which naturally fit training residuals more closely. The relatively high validation MAPE of 39.6% also indicates that the model has greater relative errors for individual predictions. Therefore, although the RF model provides reasonable overall predictive performance and is useful for identifying spatial patterns and regional variations in $PM_{2.5}$, its predictions should be interpreted with caution when precise concentration estimates are required. The model is more suitable for regional-scale screening and spatial characterization than for direct regulatory or compliance applications.

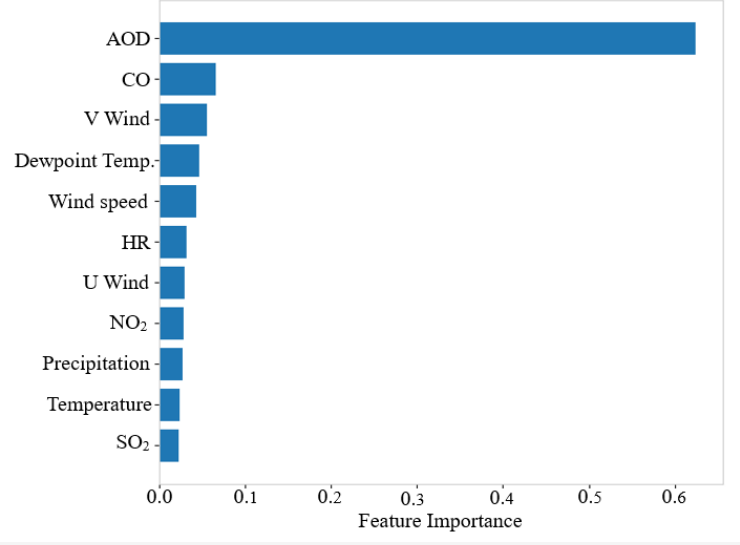
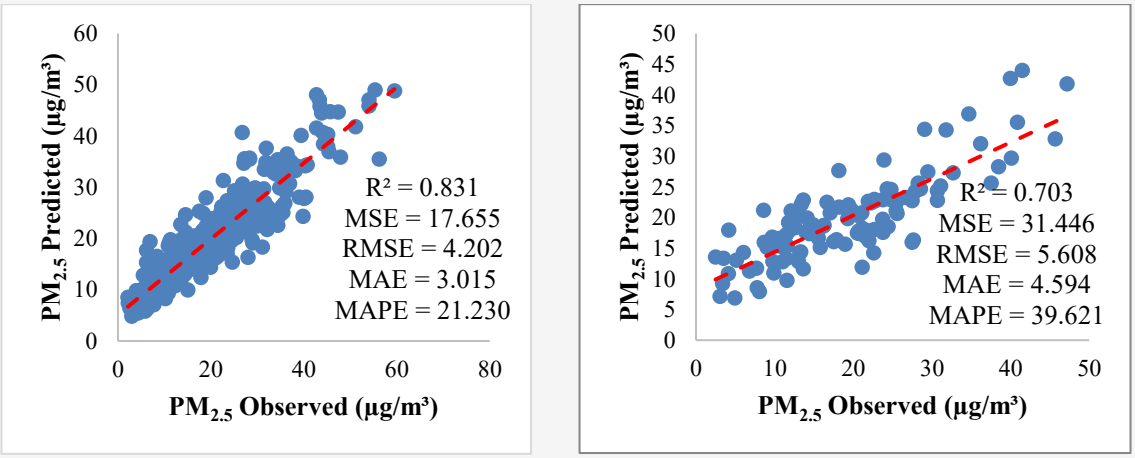


Figure 4: Variable Importance Analysis in RF model

Table 3: BMA Model training results

Variable	PIP*	Expected Value	Standard Deviation	Model 1	Model 2	Model 3	Model 4	Model 5
Intercept	100	17.4	5.67	18.211	19.519	16.624	16.787	17.343
AOD	100	5.86×10^{-2}	2.99×10^{-3}	0.058	0.059	0.059	0.059	0.060
Temperature	6.7	2.47×10^{-2}	0.162	-	-	-	-	-
Precipitation	6.2	-1.17×10^{-3}	5.77×10^{-3}	-	-	-0.019	-	-
Humidity Relative	84.8	-0.182	8.87×10^{-2}	-0.219	-	-0.190	-0.220	-
Wind Speed	86.2	-0.632	0.344	-0.730	-0.586	-0.784	-	-
Dewpoint Temperature	18.2	-0.135	0.17	-	-0.827	-	-	-0.844
U Wind	9.7	1.53×10^{-2}	0.143	-	-	-	0.426	-
V Wind	88.7	-1.06	0.4185	-1.200	-	-1.220	-1.265	-
NO_2	2.9	-1.72×10^{2}	2.25×10^{-3}	-	-	-	-	-
SO_2	2.8	14.5	2.21×10^{2}	-	-	-	-	-
CO	100	185	44.4	184.120	181.194	177.180	189.837	196.894
nVar	-	-	-	5	4	6	5	3
R^2	-	-	-	0.566	0.556	0.567	0.562	0.549
BIC	-	-	-	-348.846	-344.876	-344.521	-344.509	-343.637
Posterior probability	-	-	-	0.537	0.074	0.062	0.061	0.04



a) Training data (4 folds) b) Validated data (1 fold)

Figure 5: Scatter plot of RF model (a) training data, and (b) validated data

4.2Assessing The Correlation between $PM_{2.5}$ Concentration and Variables from Remote Sensing Data Using BMA Model

Table 3 presents the estimation results of the independent variables in the Bayesian Model Averaging (BMA) regression model. The PIP (Posterior Inclusion Probability) value reflects the probability of a variable being included in the model, indicating the importance of the variable in explaining the dependent variable, which includes AOD, NO_2, SO_2, CO, Temperature, Precipitation, Wind Speed, and Humidity. The correlation matrix in Figure 6 shows that $PM_{2.5}$ has a strong positive correlation with AOD (r = 0.70), a moderate correlation with CO (r = 0.34), negative correlations with Windspeed (r = -0.25) and Dewpoint (r = -0.23). The correlation with SO_2 is almost negligible (r = -0.01) (Figure 6). Meteorological factors such as temperature, humidity, and rainfall have a weak influence on $PM_{2.5}$ (r = -0.17, -0.11, and +0.01). These results suggest that atmospheric pollutant variables exhibited higher predictive importance than meteorological variables in the developed models (Figure 6). The Bayesian Model Averaging analysis evaluated all 2,048 possible subset regressions across the 11 candidate predictors, identifying five top candidate models (Table 3). AOD and CO demonstrated Posterior Inclusion Probabilities (PIP) of 100%, providing unequivocal posterior evidence for their inclusion in the linear model space. V wind component (PIP = 88.7%), wind speed (PIP = 86.2%), and relative humidity (PIP = 84.8%) also

exhibited substantial posterior support. It should be emphasized that PIP measures posterior evidence for predictor inclusion under the specified prior structure rather than establishing direct causal importance. The negative coefficients of wind speed and the V wind component suggest that stronger winds are associated with lower $PM_{2.5}$ concentrations, likely because wind enhances the dispersion and transport of air pollutants during the dry season. Among candidate models, Model 1 achieved the lowest BIC (-348.846) and the highest posterior probability (0.537). While BMA provides a full ensemble average, Model 1 was adopted for explicit spatial simulation. The resulting empirical regression equation was established as in Equation 10:

$$PM2.5 = 18.211 + 0.058AOD - 0.219RH - 0.730WS - 1.2Vwind + 184.12CO \quad \text{Equation 10}$$

The relationship between the observed $PM_{2.5}$ concentrations and the values estimated by the Bayesian Model Averaging (BMA) model is presented in Figure 7. During the training phase (Figure 7(a)), the model achieved a coefficient of determination of $R^2 = 0.566$, indicating that approximately 56.6% of the variability in $PM_{2.5}$ concentrations was explained by the selected predictor variables. The corresponding error metrics (MAE = 5.122 µg/m³, MSE = 45.634, RMSE = 6.755 µg/m³, and MAPE = 36.389%) suggest that the BMA model provides a moderate fit to the training dataset. During the validation phase (Figure 7(b)), the model performance improved slightly, with R^2 increasing to 0.626, indicating that approximately 62.6% of the variation in the observed $PM_{2.5}$ concentrations was captured by the model. The prediction errors also decreased, as reflected by MAE = 4.839 µg/m³, MSE = 35.901, and RMSE = 5.992 µg/m³, although the MAPE increased slightly to 40.852%, suggesting relatively larger percentage errors for some observations with low $PM_{2.5}$ concentrations. Overall, the scatter plots show that most predicted values are distributed close to the 1:1 reference line, indicating a reasonable agreement between the estimated and observed $PM_{2.5}$ concentrations. However, deviations remain for several high-concentration samples, implying that the model tends to under or overestimate extreme $PM_{2.5}$ events. These results demonstrate that the BMA model has satisfactory predictive capability and acceptable generalization performance, making it suitable for estimating $PM_{2.5}$ concentrations in the study area, although further improvements are needed to reduce prediction errors under extreme pollution conditions.

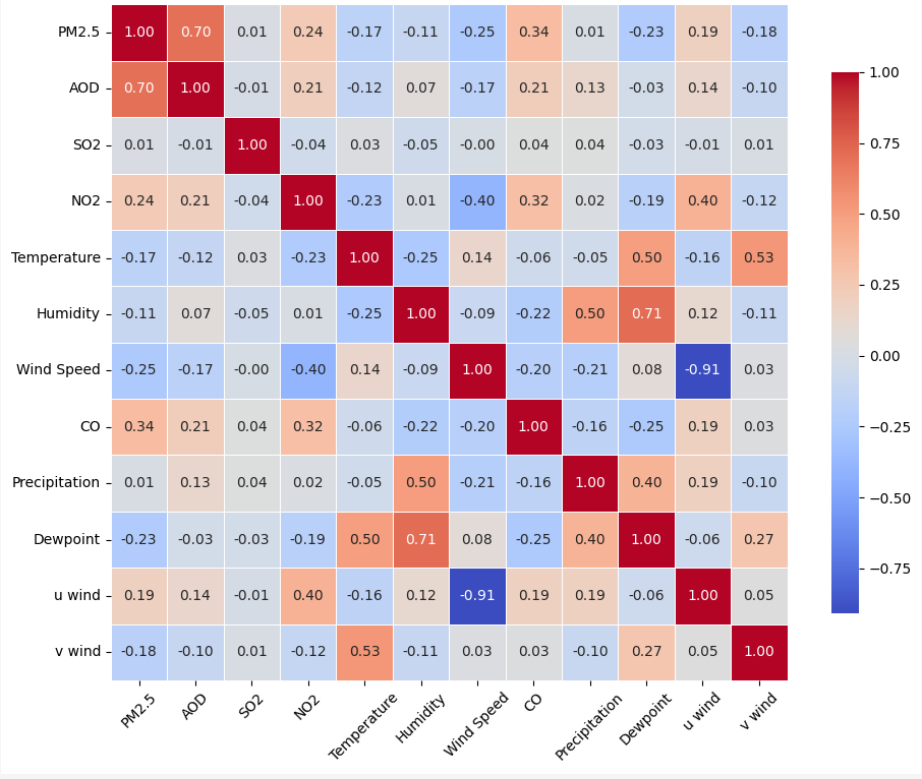
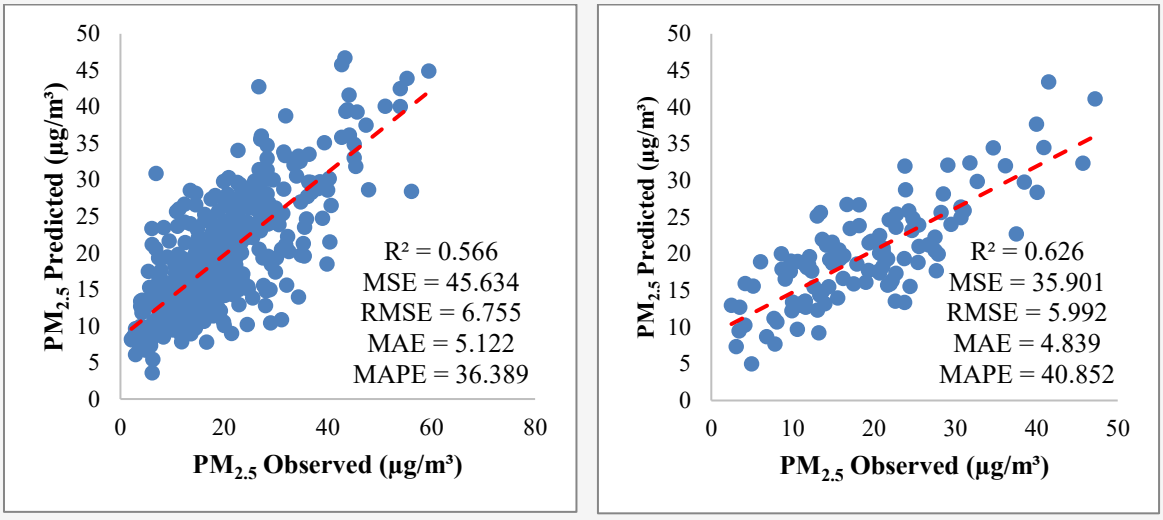


Figure 6: Correlation matrix of $PM_{2.5}$ and independent variables in the BMA model



a) Training data (4 folds) b) Validating data (1 fold)

Figure 7: Scatter plot of BMA model on (a) training data, and (b) validated data

Table 4: Statistical results on the accuracy assessment of BMA and RF models

Model	Training set					Validation set				
	MSE (µg/m³)²	RMSE (µg/m³)	MAE (µg/m³)	MAPE (%)	R^2	MSE (µg/m³)²	RMSE (µg/m³)	MAE (µg/m³)	MAPE (%)	R^2
BMA	45.634	6.755	5.122	36.389	0.566	35.901	5.992	4.839	40.852	0.626
RF	17.655	4.202	3.015	21.230	0.831	31.446	5.608	4.594	39.621	0.703

4.3 Evaluate the Accuracy of the Two Models

Table 4 presents the performance of the regression models in predicting $PM_{2.5}$ concentrations. Each model's effectiveness was assessed using MAE, MSE, RMSE, MAPE, and R^2 across training and validation datasets. The results presented in the Table 4 show that both BMA (Bayesian Model Averaging) and RF (Random Forest) models demonstrate quite good estimation capabilities; however, there is a significant difference in accuracy between the two models. The performance comparison shows that the Random Forest (RF) model outperformed the BMA model on the independent validation dataset (Table 4). For the validation data, RF achieved an R^2 of 0.703, RMSE of 5.608 µg/m³, MAE of 4.594 µg/m³, and MAPE of 39.621%, compared with R^2 of 0.626, RMSE of 5.992 µg/m³, MAE of 4.839 µg/m³, and MAPE of 40.852% for BMA. Thus, RF provided a higher explained variance and slightly lower prediction errors than BMA. The training results also show a stronger fit for RF, with an R^2 of 0.831, compared with 0.566 for BMA. However, the difference between the two models was smaller for the validation data, suggesting that RF provided better predictive performance while maintaining reasonable generalization to independent observations. The increase in RMSE from the training to validation dataset was also moderate for RF (from 4.202 to 5.608 µg/m³), indicating some loss of predictive accuracy when applied to unseen data.

To further compare the predictive performance of the two models, a bootstrap test was conducted to assess the difference in R^2 between the RF and BMA models using the validation dataset (Figure 8). The bootstrap analysis demonstrates a clear difference in predictive performance between the RF and BMA models. As shown in Figure 8(a), the bootstrap distribution of ΔR^2 (RF − BMA) is centered on a positive value, with the observed difference of approximately 0.074. The distribution is predominantly located above zero, indicating that RF generally achieves a higher R^2 than BMA across the bootstrap samples. Moreover, the 95% bootstrap confidence interval for ΔR^2 does not appear to include zero, supporting the statistical significance of the difference. This finding is consistent with the bootstrap test result ($p = 0.005$) in Figure 8(b), indicating that the higher validation performance of RF is unlikely to be explained by random sampling variation alone.

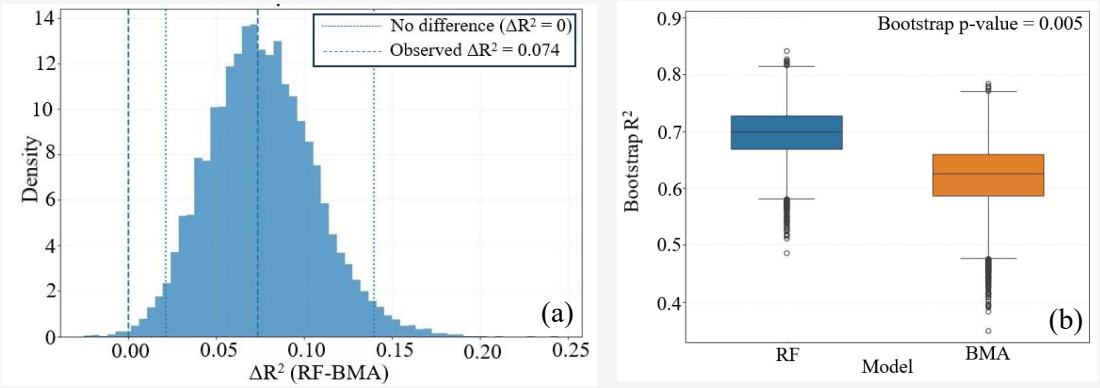


Figure 8: Bootstrap Test on R^2 Values between RF and BMA model.
(a) Bootstrap distribution of ΔR^2; (b) Bootstrap comparison of R^2.

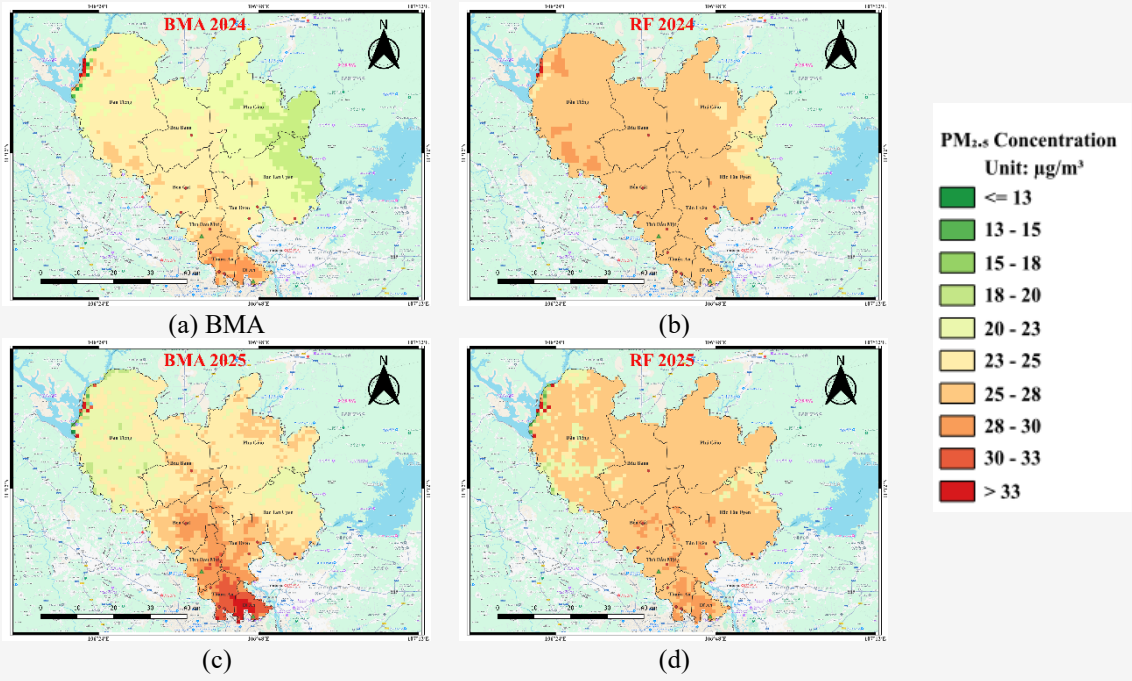


Figure 9: $PM_{2.5}$ concentration maps in Binh Duong area in dry-season of 2024 – 2025:
(a) 2024 dry season BMA, (b) 2024 dry season RF, (c) 2025 dry season BMA, and (d) 2025 dry season RF

The boxplots further show that the bootstrap (R^2) values for RF are generally higher than those for BMA, although there is some overlap between the two distributions. Overall, the results indicate that RF provides improved predictive performance compared with BMA for the validation dataset. However, the relatively high validation MAPE values for both models (39.621% for RF and 40.852% for BMA) indicate that considerable relative errors remain, particularly for observations with low $PM_{2.5}$ concentrations. Therefore, the RF model can be considered the better-performing model for regional $PM_{2.5}$ prediction, while the BMA model remains useful because of its greater interpretability and its ability to identify important predictors and quantify model-selection uncertainty.

4.4 $PM_{2.5}$ Spatial Distribution Map

In order to characterize the spatial distribution on air quality, we mainly focused on Binh Duong province since this area is center of industrial and urbanization development in the Southeast region on Vietnam. After verifying the two regression models, the $PM_{2.5}$ concentration maps were generated for the study area using QGIS software. The resulted maps are displayed in Figure 9. The simulation results of $PM_{2.5}$ concentrations from the two models BMA and RF show significant differences in spatial reflection and

reliability. Comparing the two algorithms, Random Forest generates smoother, more physically plausible spatial gradients, whereas BMA produces sharper localized anomalies due to its high sensitivity to local trace gas inputs. The results shown in Figure 9 indicate that the variation in average $PM_{2.5}$ concentration among different areas in the province exhibits significant differences both spatially and temporally. Highly elevated $PM_{2.5}$ concentrations (>20 μg/m³, with peak localized estimates reaching 25–28 μg/m³) are concentrated across southern Binh Duong, specifically within Di An City, Thuan An City, and Thu Dau Mot City. These southern urban centers are characterized by dense industrial clusters (including Vietnam - Singapore Industrial Parks VSIP I & II, Song Than Industrial Park, and Phu Tan Industrial Zone) and intense heavy-duty diesel freight traffic along National Highway 13 and the My Phuoc - Tan Van expressway. In contrast, northern districts, including Dau Tieng, Phu Giao, and Bac Tan Uyen, exhibit substantially lower $PM_{2.5}$ concentrations (<15 μg/m³), consistent with expansive rubber plantations, agricultural land, higher canopy cover, and low industrial emission density.

Comparing the two models, BMA tends to estimate higher $PM_{2.5}$ concentrations than RF in most areas, especially in 2025, demonstrating this model's high sensitivity to local fluctuations. However, according to the overall error assessment results, the RF model is considered to reflect reality more accurately and shows greater stability in simulating $PM_{2.5}$ distribution among areas. Both models capture the slight increasing trend of $PM_{2.5}$ concentrations over time and the clear spatial disparity between urban-industrial areas and rural-ecological regions, that align with provincial industrialization and urbanization gradients. Despite these insights, several methodological limitations should be highlighted. Satellite data (AOD, Sentinel-5P TROPOMI) often have missing values due to clouds, and weather conditions, which can affect the model if not handled appropriately. Additionally, collecting satellite pixel data with the location of ground stations still contains uncertainties, especially when the pixel size is large, and the monitoring data in the study area is still sparse [22]. Some recent studies have proposed newer methods, such as spatiotemporal estimating models with gradient constraints from satellite data, indicating a strong direction for future development [23] and [24]. Therefore, future research should consider: First, applying data assimilation methods that combine data from multiple satellites and ground data; second, applying hybrid models between physics and machine learning to leverage atmospheric knowledge while exploiting big data; third, increasing the density and spatial coverage of monitoring stations would further enhance the robustness and reliability of model evaluation. Future work should also incorporate localized sensor calibration curves, spatial block cross-validation, and physics-guided machine learning architectures to further improve estimation robustness.

5. Conclusions

This study established a satellite and reanalysis-based modeling framework for estimating surface $PM_{2.5}$ concentrations across Binh Duong province, Northern region of Ho Chi Minh City, Vietnam during the dry seasons (December - April) of 2024 and 2025. By integrating 1 km MODIS AOD, Sentinel-5P TROPOMI tropospheric trace gases (NO_2, SO_2, and CO), and ERA5 meteorological variables harmonized to a 1 km grid, the research systematically compared the predictive capability and behavioral characteristics of Random Forest (RF) against Bayesian Model Averaging. The comparative evaluation confirmed that the non-parametric Random Forest model significantly outperformed the linear Bayesian approach, achieving higher explained variance, lower prediction errors, and superior stability across repeated cross-validation folds. AOD served as the predominant physical predictor, while tropospheric CO and boundary-layer wind components provided crucial constraints on emission sources and atmospheric ventilation. Bayesian Model Averaging proved highly effective for identifying posterior inclusion probabilities and selecting an optimal subset of predictors, while Random Forest proved superior for high-dimensional spatial prediction.

Several limitations should be explicitly recognized from this study. First, the framework is strictly confined to dry-season conditions because monsoon cloud cover precludes valid optical satellite AOD retrievals during the wet season. Second, combining regulatory HCEM monitors with uncalibrated community WAQI sensors introduces target noise that could not be fully eliminated through quality screening. Third, standard cross-validation does not account for spatial autocorrelation, meaning reported metrics represent an optimistic bound for unmonitored spatial extrapolation. Fourth, validation MAPE values around 40% restrict the model to regional environmental screening and spatial hazard mapping rather than strict regulatory compliance enforcement.

The generated high-resolution maps translate directly into actionable policy recommendations for the Binh Duong provincial government and environmental management authorities. Given the

pronounced concentration of $PM_{2.5}$ across Di An, Thuan An, and Thu Dau Mot, provincial planning should: (1) mandate dense green vegetative buffer zones between heavy industrial estates and surrounding residential areas; (2) implement targeted roadside emissions inspection and traffic flow optimizations along heavy diesel freight corridors (National Highway 13 and My Phuoc - Tan Van expressway); and (3) strategically expand automated reference monitoring stations into industrial-suburban transition zones to support localized calibration of community low-cost sensor networks.

Future research should focus on four key areas: (1) developing optical–microwave data fusion and deep-learning-based gap-filling methods to reconstruct aerosol fields under cloudy conditions, thereby enabling continuous year-round $PM_{2.5}$ estimation; (2) implementing rigorous Spatial Block Cross-Validation and Leave-Location-Out Cross-Validation to more robustly assess spatial generalizability, particularly across unmonitored regions; (3) establishing collocated calibration and transfer functions for community-based low-cost sensor networks; and (4) integrating localized chemical transport models with satellite-based machine-learning approaches to improve the physical interpretability of the resulting $PM_{2.5}$ estimates.

Acknowledgments

This research is funded by Thu Dau Mot University, Vietnam under grant number DT.22.3-096.

References

[1] World Health Organization. (2019). *Ten Threats to Global Health in 2019.* [Online]. Available: *https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019/* [Accessed: Aug. 18, 2025].

[2] World Health Organization. (2018). *Ambient (Outdoor) Air Quality Database: Summary Results, Update 2018 (Version 6).* Geneva: World Health Organization. [Online]. Available: *https://cdn.who.int/media/docs/default-source/air-quality-database/aqd_2018/.* [Accessed: Aug. 18, 2025].

[3] Le, L. T., Quang, K. B. V., Vo, T. V., Nguyen, T. M. T., Dao, T. V. H. and Bui X. T., (2024). Environmental and Health Impacts of Air Pollution: A Mini-Review. *Vietnam Journal of Science, Technology and Engineering*, Vol. 66 (1), 120-128. https://doi.org/10.31276/VJSTE.66(1).120-128.

[4] Orru, H., Ebi, K. L. and Forsberg, B., (2017). The Interplay of Climate Change and Air Pollution on Health. *Current Environmental Health Reports*, Vol. 4(4); 504-513. https://doi.org/10.1007/s40572-017-0168-6.

[5] Gautam, D. and Bolia, N. B., (2020). Air Pollution: Impact and Interventions. *Air Quality, Atmosphere and Health*, Vol. 13(2); 209-223. https://doi.org/10.1007/s11869-019-00784-8.

[6] Gold, D. R. and Samet, J. M., (2013). Air Pollution, Climate, and Heart Disease. *Circulation*, Vol. 128(21); 411-414. https://doi.org/10.1161/circulationaha.113.003988.

[7] Lin, R. S., Sung, F. C., Huang, S. L., Gou, Y. L., Ko, Y. C. and Shaw, C. K., (2001). Role of Urbanization and Air Pollution in Adolescent Asthma: A Mass Screening in Taiwan. *Journal of the Formosan Medical Association*, Vol. 100(10); 649-655.

[8] Vu, N. P. and Pham, M. H., (2023). Study on the Application of Remote Sensing Technology and Multiple Linear Regression Algorithm in Establishing Fine Particulate Matter ($PM_{2.5}$) Emission Maps. *Journal of Geodesy and Cartography Science (In Vietnamese)*, Vol. 57(9); 39-45. https://doi.org/10.54491/jgac.2023.57.696.

[9] Le, T. N., Nguyen, C. M. D., Nguyen, H. P. and Bui, T. L., (2023). Assessment of Community Health Impacts Due to Short-Term Exposure to $PM_{2.5}$ Pollution: A Case Study in Binh Duong Province. *Vietnam Journal of Hydrometeorology (In Vietnamese)*, Vol. 746; 70-87. http://doi.org/10.36335/VNJHM.2023(746).70-87.

[10] McCarron, A., Semple, S., Braban, C.F., Swanson, V., Gillespie, C. and Price., H. D., (2022). Public Engagement with Air Quality Data: Using Health Behaviour Change Theory to Support Exposure-Minimising Behaviours. *Journal of Exposure Science and Environmental Epidemiology*, Vol. 33(3); 321-331. http://doi.org/10.1038/s41370-022-00449-2.

[11] Yang, H., Wang, W. and Li, G., (2023). Prediction Method of $PM_{2.5}$ Concentration Based on Decomposition and Integration. *Measurement*, Vol. 216. https://doi.org/10.1016/j.measurement.2023.112954.

[12] Hardini, M., Richard, S. A., Asfi, M., Riza Chakim, M. H. and Ayu Sanjaya, Y. P., (2023). Predicting Air Quality Index Using Ensemble Machine Learning. *ADI Journal on Recent Innovation*, Vol. 5(1); 78-86. https://doi.org/10.34306/ajri.v5i1Sp.981.

[13] Zuo, X., Guo, H., Shi, S. and Zhang, X., (2020). Comparison of Six Machine Learning Methods for Estimating $PM_{2.5}$ Concentration Using the Himawari-8 Aerosol Optical Depth. *Journal of the Indian Society of Remote Sensing*, Vol. 48(9); 1277-1287. https://doi.org/10.1007/s12524-020-01154-z.

[14] Yan, X., Zang, Z., Jiang, Y., Shi, W., Guo, Y., Li, D., Zhao, C. and Husi, L., (2021). A Spatial-Temporal Interpretable Deep Learning Model for Improving Interpretability and Predictive Accuracy of Satellite-Based $PM_{2.5}$. *Environmental Pollution*, Vol. 273(10). https://doi.org/10.1016/j.envpol.2021.116459.

[15] Wei, S., Shores, K. and Xu, Y., (2025). A Comparison of Machine Learning-Based Approaches in Estimating Surface $PM_{2.5}$ Concentrations Focusing on Artificial Neural Networks and High Pollution Events. *Atmosphere*, Vol. 16(1). https://doi.org/10.3390/atmos16010048.

[16] Ramli, N., Hamid, H. A., Yahaya, A. S., Ul-Saufie, A. Z., Noor, N. M., Seman, N. A. A., Kamarudzaman, A. N. and Deak, G., (2023). Performance of Bayesian Model Averaging (BMA) for Short-Term Prediction of PM_{10} Concentration in Peninsular Malaysia. *Atmosphere*, Vol. 14(2). https://doi.org/10.3390/atmos14020311.

[17] Qi, H., Ma, S., Chen, J., Sun, J., Wang, L., Wang, N., Wang, W., Zhi, X. and Yang, H., (2022). Multi-Model Evaluation and Bayesian Model Averaging in Quantitative Air Quality Forecasting in Central China. *Aerosol and Air Quality Research*, Vol. 22(5); 210247. https://doi.org/10.4209/aaqr.210247.

[18] Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T., (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, Vol. 14(4); 382-417. https://doi.org/10.1214/ss/1009212814.

[19] Breiman, L., (2001). Random Forests. *Machine Learning*, Vol. 45(1); 5-32. https://doi.org/10.1023/A:1010933404324.

[20] Prajwala, T. R., (2015). A Comparative Study on Decision Tree and Random Forest Using R Tool. *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 4(1); 196-199. https://doi.org/10.17148/IJARCCE.2015.4142.

[21] Jain, S. K. and Gupta, A. K., (2022). Application of Random Forest Regression with Hyper-Parameters Tuning to Estimate Reference Evapotranspiration. *International Journal of Advanced Computer Science and Applications*, Vol. 13(5); 742-750. https://doi.org/10.14569/IJACSA.2022.0130585.

[22] Makhdoomi, A., Sarkhosh, M. and Ziaei, S., (2025). $PM_{2.5}$ Concentration Prediction Using Machine-Learning Algorithms: An Approach to Virtual Monitoring Stations. *Scientific Reports*, Vol. 15. https://doi.org/10.1038/s41598-025-92019-3.

[23] Handschuh, J., Erbertseder, T. and Baier, F., (2023). Systematic Evaluation of Four Satellite AOD Datasets for Estimating $PM_{2.5}$ Using a Random Forest Approach. *Remote Sensing*, Vol. 15(8). https://doi.org/10.3390/rs15082064

[24] Adong, P., Bainomugisha, E. and Dev, S., (2025). Evaluating Machine Learning Methods for $PM_{2.5}$ Estimation Using Satellite AOD, Low-Cost and Reference-Grade Monitors in Kampala. *International Journal of Environmental Science and Technology*, Vol. 22, 15747-15756. https://doi.org/10.1007/s13762-025-06674-0.

[25] Paluang, P., Thavorntam, W., and Phairuang, W. (2024). Application of Multilayer Perceptron Artificial Neural Network (MLP-ANN) Algorithm for PM2.5 Mass Concentration Estimation during Open Biomass Burning Episodes in Thailand. *International Journal of Geoinformatics*, Vol. 20(7), 28–42. https://doi.org/10.52939/ijg.v20i7.3401..

[26] Teodorescu, V. and Brașoveanu, L. O., (2025). Assessing the Validity of k-Fold Cross-Validation for Model Selection: Evidence from Bankruptcy Prediction Using Random Forest and XGBoost. *Computation*, Vol. 13(5). https://doi.org/10.3390/computation13050127.

[27] Hai, P., Minh, T., Luận, V., and Trung, T. (2023). Machine Learning Models for Fine Particulate Matter (PM2.5) Prediction: A Case Study in Bac Ninh Province, Vietnam. *International Journal of Geoinformatics*, *19*(12), 30–50. https://doi.org/10.52939/ijg.v19i12.2975.

[28] Tran, T. A., Nguyen, T. B. N., Pham, T. M. T. and Le, T. K., (2025). Spatio-Temporal Analysis of NO_2 Concentration in Binh Duong Province, Vietnam Using Sentinel-5P Imageries and Google Earth Engine. *The GIS-IDEAS Journal*, Vol. 1(1); 9-16. https://doi.org/10.82667/gisideas.1.1_9.

[29] Binh Duong Statistics Office. (2024). Binh Duong Province Statistical Yearbook 2024. *Statistical Publishing House* (2024), 1-631.

[30] Hess, M., Koepke, P. and Schult, I., (1998). Optical Properties of Aerosols and Clouds: The Software Package OPAC. *Bulletin of the American Meteorological Society*, Vol. 79(5); 831-844. https://doi.org/10.1175/1520-0477(1998)079<0831:OPOAAC>2.0.CO;2.

[31] Ma, Y., Xin, J., Zhang, W., Liu, Z., Ma, Y., Kong, L., Wang, Y., Deng, Y., Lin, S. and He, Z., (2019). Long-Term Variations of the $PM_{2.5}$ Concentration Identified by MODIS in the Tropical Rain Forest, Southeast Asia. *Atmospheric Research*, Vol. 219; 140-152. https://doi.org/10.1016/j.atmosres.2018.12.024

[32] Ho Chi Minh City Center of Natural Resources and Environmental Monitoring. (2025). *Report on Air Quality Monitoring*. [Online]. Available: https://moitruonghcm.vn/ [Accessed: Aug. 18, 2025].

[33] World Air Quality Index Project. (2025). AQICN: Real-Time Air Pollution Information Worldwide. [Online]. Available: https://aqicn.org/ [Accessed: Aug. 18, 2025].

[34] Barkjohn, K. K., Holder, A. L., Frederick, S. G. and Clements, A. L., (2022). Correction and Accuracy of PurpleAir $PM_{2.5}$ Measurements for Extreme Wildfire Smoke. *Sensors*, Vol. 22(24). https://doi.org/10.3390/s22249669.

[35] Raheja, G., Sabi, K. E., Delaplace, M., Westervelt, D. M. and McFarlane, C., (2023). Low-Cost Sensor Performance Intercomparison, Correction Factor Development, and 2+ Years of Ambient $PM_{2.5}$ Monitoring in Accra, Ghana. *Environmental Science and Technology*, Vol. 57(29); 10708-10720. https://doi.org/10.1021/acs.est.2c09264

[36] Farooqui, Z., Biswas, J. and Saha, J., (2023). Long-Term Assessment of PurpleAir Low-Cost Sensor for $PM_{2.5}$ in California, USA. *Pollutants*, Vol. 3(4); 477-493. https://doi.org/10.3390/pollutants3040033.

[37] Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. and Moore, R., (2017). Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone. *Remote Sensing of Environment*, Vol. 202; 18-27. https://doi.org/10.1016/j.rse.2017.06.031.

[38] Shukri, S., Asmat, A., Rahiman, M., Japeiri, A., and Sahwee, Z. (2024). Real-time Vertical Air Quality Monitoring System Development for Urban Scale. *International Journal of Geoinformatics*, Vol. 20(8). https://doi.org/10.52939/ijg.v20i8.3463.

[39] Phan, H. D. P., Le, D. K., Nguyen, T. M. T. and Phan, H. V., (2022). Estimating $PM_{2.5}$ Mass Concentration from MODIS AOD Products in Ho Chi Minh City, Vietnam. *Lecture Notes in Civil Engineering*, Vol. 268; 579-588. https://doi.org/10.1007/978-981-19-3303-5_51.

[40] Vo, Q. B. and Tran, T. V., (2021). An Empirical Relationship between $PM_{2.5}$ and Aerosol Optical Depth from MODIS Satellite Images for Spatial Simulation over Ho Chi Minh City. *Vietnam Journal of Science, Technology and Engineering*, Vol. 63(4); 72-78. https://doi.org/10.31276/VJSTE.63(4).72-78.

[41] Veefkind, J. P., Aben, I., McMullan, K., Forster, H., de Vries, J., Otter, G., Claas, J., Eskes, H. J., de Haan, J. F., Kleipool, Q., van Weele, M., Hasekamp, O., Hoogeveen, R., Landgraf, J., Snel, R., Tol, P., Ingmann, P., Voors, R., Kruizinga, B., Vink, R., Visser, H. and Levelt, P. F., (2012). TROPOMI on the ESA Sentinel-5 Precursor: A GMES Mission for Global Observations of the Atmospheric Composition for Climate, Air Quality and Ozone Layer Applications. *Remote Sensing of Environment*, Vol. 120; 70-83. https://doi.org/10.1016/j.rse.2011.09.027.

[42] Savenets, M., (2021). Air Pollution in Ukraine: A View from the Sentinel-5P Satellite. *Időjárás*, Vol. 125(2); 271-290. https://doi.org/10.28974/idojaras.2021.2.6.

[43] Earth Engine Data Catalog. (2025). *Sentinel-5P NRTI NO2: Near Real-Time Nitrogen Dioxide.* Google Earth Engine. [Online]. Available: https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_NRTI_L3_NO2 [Accessed: Jun. 3, 2025].

[44] Earth Engine Data Catalog. (2025*). Sentinel-5P OFFL SO2: Offline Sulfur Dioxide.* Google Earth Engine. [Online]. Available: https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_OFFL_L3_SO2 [Accessed: Jun. 3, 2025].

[45] Earth Engine Data Catalog. (2025). *Sentinel-5P NRTI CO: Near Real-Time Carbon Monoxide.* Google Earth Engine. [Online]. Available: https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_NRTI_L3_CO [Accessed: Jun. 3, 2025].

[46] Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D. and Thépaut, J. N., (2018). ERA5 Hourly Data on Single Levels from 1940 to Present. *Copernicus Climate Change Service (C3S) Climate Data Store*. [Online]. https://doi.org/10.24381/cds.adbb2d47. [Accessed: July 3, 2025].

[47] Nguyen, P. H., Dao, N. K. and Nguyen, L. S. P., (2024). Development of Machine Learning and Deep Learning Prediction Models for $PM_{2.5}$ in Ho Chi Minh City, Vietnam. *Atmosphere*, Vol. 15(10). https://doi.org/10.3390/atmos15101163.

[48] Pham, C. T., (2023). Evaluation of Variation and Influence of Selected Meteorological Factors on $PM_{2.5}$ Concentrations in a Suburban Site of Hanoi. *TNU Journal of Science and Technology*, Vol. 228(10); 262-270. https://doi.org/10.34238/tnu-jst.7913.

[49] Son, R., Stratoulias, D., Kim, H. C. and Yoon, J. H., (2023). Estimation of Surface $PM_{2.5}$ Concentrations from Atmospheric Gas Species Retrieved from TROPOMI Using Deep Learning: Impacts of Fire on Air Pollution over Thailand. *Atmospheric Pollution Research*, Vol. 14(10). https://doi.org/10.1016/j.apr.2023.101875.

[50] Eneko, L., Giulia, G., Jaione E. E., Jose, M. A. and Andreas, S., (2026). The Importance of Choosing a Proper Validation Strategy in Predictive Models. Part 2: Recipes For (Avoiding) Overfitting-A Tutorial. *Analytica Chimica Acta*, Volume 1384. https://doi.org/10.1016/j.aca.2025.344838.