AN AI-BASED FRAMEWORK FOR ESTIMATING $PM_{2.5}$ USING SATELLITE AEROSOL OPTICAL DEPTH AND METEOROLOGICAL DATA IN A COASTAL INDUSTRIAL REGION OF THAILAND

Maliwan THABTHONG [1], ***Teerawong LAOSUWAN*** [1,4*], ***Satith SANGPRADID*** [2,4*], ***Yannawut UTTARUK*** [3,4], ***Jenjira PIAMDEE*** [1], ***Piyatida AWICHIN*** [1,4], ***Titipong PHOOPHATHONG*** [1,4] & ***Maharaja SINGHARAJ*** [5]

DOI: 10.21163/GT_2026.212.07

ABSTRACT

Monitoring fine particulate matter ($PM_{2.5}$) in coastal industrial areas remains difficult, mainly because ground-based monitoring stations are sparse and unevenly distributed. In this study, we developed a practical AI-based framework to estimate daily $PM_{2.5}$ concentrations by combining satellite-derived Aerosol Optical Depth (AOD) with commonly available reanalysis-based meteorological data. The analysis focused on Rayong Province, Thailand, an industrial coastal region where atmospheric processes and pollution sources are highly variable. Daily $PM_{2.5}$ measurements from monitoring stations were merged with MODIS/MAIAC AOD and meteorological variables obtained from ERA5 and ERA5-Land. The meteorological inputs included air temperature, relative humidity, wind speed, and boundary layer height (BLH). Five machine learning models were tested, including Multiple Linear Regression (MLR), Support Vector Regression (SVR), Artificial Neural Networks (ANN), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). Feature selection was carried out using Recursive Feature Elimination with Cross-Validation and Mutual Information, and model performance was evaluated using five-fold cross-validation. Among the tested models, Random Forest showed the best overall performance, achieving an R^2 value of 0.70, followed by XGBoost and ANN. The results also indicate that nonlinear models consistently performed better than linear regression, suggesting that the relationship between $PM_{2.5}$ and atmospheric variables is not linear. Wind speed emerged as the most influential predictor, with relative humidity and BLH also playing important roles. In particular, the $PM_{2.5}$–meteorology relationships exhibited clear threshold behavior under low wind speed conditions. Overall, the proposed GeoAI-based framework provides a practical and transferable approach for spatial $PM_{2.5}$ estimation in complex coastal industrial environments, offering an effective solution for air quality assessment in regions with sparse ground monitoring networks.

Keywords: *$PM_{2.5}$ estimation; Aerosol Optical Depth; Machine learning; Spatial air pollution; Coastal industrial region.*

[1]*Department of Physics, Faculty of Science, Mahasarakham University, Maha Sarakham 44150, Thailand, 670102560001@msu.ac.th (MT), teerawong@msu.ac.th (TL), jenjira@msu.ac.th (JP), 680102620002@msu.ac.th (PA), 640102620003@msu.ac.th (TP).*

[2] *Department of Geoinformatics, Faculty of Informatics, Mahasarakham University, Maha Sarakham 44150, Thailand, satith.s@msu.ac.th (SS).*

[3] *Department of Biology, Faculty of Science, Mahasarakham University, Maha Sarakham 44150, Thailand, yannawut.u@msu.ac.th (YU).*

[4] *Greenhouse Gas Research and Operations Center, Mahasarakham University, Maha Sarakham 44150, Thailand.*

[5] *Nampapa Nakhoneluang, Vientiane Capital Water Supply State Enterprise, Kaisone Road, Xaysettha District, Vientiane City, Laos, mesha_coltd@yahoo.com (MS).*

**Corresponding authors: teerawong@msu.ac.th (T.L.), satith.s@msu.ac.th (S.S)*

1. INTRODUCTION

Air pollution is a global environmental problem that poses serious threats to human health, ecosystems, and economic sustainability (Laosuwan et al., 2023; Kovács & Haidu, 2024). Among various air pollutants, fine particulate matter with an aerodynamic diameter of less than 2.5 μm ($PM_{2.5}$) has been identified by the World Health Organization as one of the most harmful pollutants due to its ability to penetrate deep into the alveoli and enter the bloodstream (World Health Organization, 2021). Long-term exposure to $PM_{2.5}$ has been associated with chronic inflammation, respiratory diseases, cardiovascular disorders, and an increased risk of cancer (Sun et al., 2022). At the global scale, $PM_{2.5}$ exposure is recognized as a leading cause of premature mortality, contributing to millions of early deaths each year and imposing substantial burdens on public health systems and economic productivity (Caplin et al., 2019). In Thailand, $PM_{2.5}$ pollution has shown a worsening trend over recent years, particularly in urban centers and industrial regions such as Bangkok and the eastern part of the country (Rotjanakusol et al., 2024). Rayong Province represents a critical hotspot due to its concentration of petrochemical industries, power plants, and major transportation networks, which continuously contribute to elevated $PM_{2.5}$ levels (Kawichai et al., 2022). Seasonal variability is also pronounced, with higher $PM_{2.5}$ concentrations typically observed between November and March. This pattern is strongly influenced by unfavorable meteorological conditions, including low wind speed, frequent temperature inversion, and open biomass burning activities, as reported by the Pollution Control Department.

Despite increasingly strict air pollution control policies, including emission standards for industrial and vehicular sources, vehicle inspection programs, restrictions on open burning, and national action plans, fine particulate pollution remains persistent in many areas (Kovács & Haidu, 2021). This persistence highlights the ongoing challenge of effectively controlling $PM_{2.5}$ pollution (Archer et al., 2024). To improve air quality monitoring, government agencies have expanded ground-based monitoring networks. However, the limited spatial coverage of monitoring stations remains a major constraint, preventing comprehensive representation of air pollution conditions across heterogeneous landscapes (Rosales et al., 2025). Consequently, recent air quality research has increasingly emphasized the use of satellite-based observations, particularly Aerosol Optical Depth (AOD), as a proxy for aerosol loading in the atmospheric column. AOD has been shown to correlate with near-surface $PM_{2.5}$ concentrations under certain meteorological conditions (Sorek-Hamer et al., 2020). Numerous studies have demonstrated that integrating satellite-derived AOD with meteorological variables such as air temperature, relative humidity, wind speed, and boundary layer height (BLH) can substantially improve $PM_{2.5}$ estimation accuracy (Nakapan et al., 2022). When combined with statistical and machine learning models, including Multiple Linear Regression, Geographically Weighted Regression, Random Forest, and Extreme Gradient Boosting, these approaches are capable of capturing nonlinear relationships and atmospheric complexity (Wang et al., 2022; Li et al., 2025; Phoophathong et al., 2025).

In recent years, the integration of artificial intelligence with geographic data has become an effective approach for spatial impact analysis (Aroonsri & Sangpradid, 2021; Intarat et al., 2025; Intarat et al., 2026). Such geo-AI frameworks offer new opportunities to better understand spatial variability, seasonal dynamics, and meteorological controls of $PM_{2.5}$ particularly in complex coastal industrial ;environments where conventional monitoring remains insufficient (Wong et al., 2024). Although numerous studies have applied satellite-derived aerosol optical depth (AOD) and meteorological variables to estimate $PM_{2.5}$ concentrations (Schneider et al., 2020; Zaman et al., 2021), several important gaps remain. First, many existing studies focus on large metropolitan areas or inland regions, while coastal industrial environments receive comparatively less attention. Coastal areas are influenced by complex atmospheric processes, including land–sea interactions, variable boundary layer dynamics, and diverse industrial emission sources. These factors can weaken the AOD– $PM_{2.5}$ relationship and reduce model reliability if not explicitly considered (Tang et al., 2022).

Second, previous research often emphasizes prediction accuracy without sufficient discussion of spatial interpretation. In many cases, machine learning models are treated as black boxes, and their

results are not fully linked to underlying geographical processes or atmospheric mechanisms (Zaman et al., 2021). This limits their value for understanding the spatial variability and seasonal behavior of $PM_{2.5}$, particularly in regions characterized by heterogeneous land use and meteorological conditions. Third, while machine learning techniques such as Random Forest and XGBoost have shown promising performance in $PM_{2.5}$ estimation (Chen & Guestrin, 2016), comparative evaluations across multiple models are still limited in Southeast Asian contexts. Most studies rely on a single modeling approach or short time periods, which restricts the generalizability of the findings. In addition, feature selection strategies and their influence on model performance are often underexplored (Li et al., 2016).

Finally, few studies explicitly frame $PM_{2.5}$ estimation as a geographical artificial intelligence (GeoAI) problem. Although the integration of artificial intelligence with spatial and atmospheric data has demonstrated strong potential for enhancing geographical understanding of air pollution patterns (Liu et al., 2022), this perspective remains underdeveloped in coastal industrial regions of Thailand. To address these gaps, this study aims to develop an AI-based geographical framework for estimating daily $PM_{2.5}$ concentrations in Rayong Province, Thailand. The specific objectives are to: (1) integrate satellite-derived AOD and meteorological variables for spatial $PM_{2.5}$ estimation; (2) compare the performance of multiple machine learning models; and (3) analyze the influence of key atmospheric variables on $PM_{2.5}$ variability in a coastal industrial setting.

2. MATERIALS AND METHODS

2.1. Study Area

Rayong Province (**Fig. 1**) is located in eastern Thailand. Its terrain consists of coastal plains along the Gulf of Thailand, combined with higher inland areas toward the interior of the province.

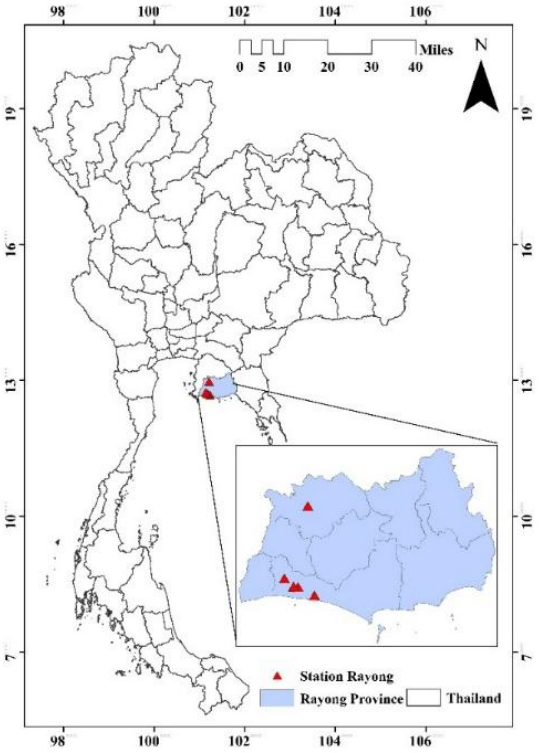

Fig. 1. Location of the study area in Rayong Province, Thailand, showing the five ground-based $PM_{2.5}$ monitoring stations and surrounding geographical features.

This geographical setting strongly influences local air circulation patterns, particularly the interaction between sea breezes and land breezes, which play an important role in the dispersion and accumulation of air pollutants. The regional climate is governed by the tropical monsoon system and can be divided into three main seasons: the hot season (March–May), the rainy season (June–October), and the cool season (November–February).

During the cool season, Rayong Province frequently experiences stable atmospheric conditions characterized by low wind speeds and nighttime temperature inversion. These conditions limit vertical air mixing and suppress pollutant dispersion, allowing fine particulate matter ($PM_{2.5}$) to accumulate near the surface. Statistical observations show that $PM_{2.5}$ concentrations are typically highest between December and February, which is associated with the southward extension of high-pressure systems from mainland China. Rayong Province is also home to the Map Ta Phut Industrial Estate and several surrounding industrial clusters. Emissions from petrochemical plants, power generation facilities, large-scale transportation networks, and community activities represent major sources of $PM_{2.5}$. When combined with coastal meteorological influences and seasonal atmospheric stability, these emission sources contribute to pronounced spatial and temporal variability in air quality across the province.

2.2. Data Sources and Modeling Framework

This study integrates ground-based air quality data with satellite observations and meteorological data to develop daily $PM_{2.5}$ prediction models for Rayong Province. The overall modeling framework is illustrated in **Fig. 2**. All satellite and meteorological variables were accessed and processed using the Google Earth Engine (GEE) platform to ensure consistency in spatial and temporal resolution. Satellite-based variables included Aerosol Optical Depth (AOD) and Land Surface Temperature (LST) derived from Terra/MODIS products.

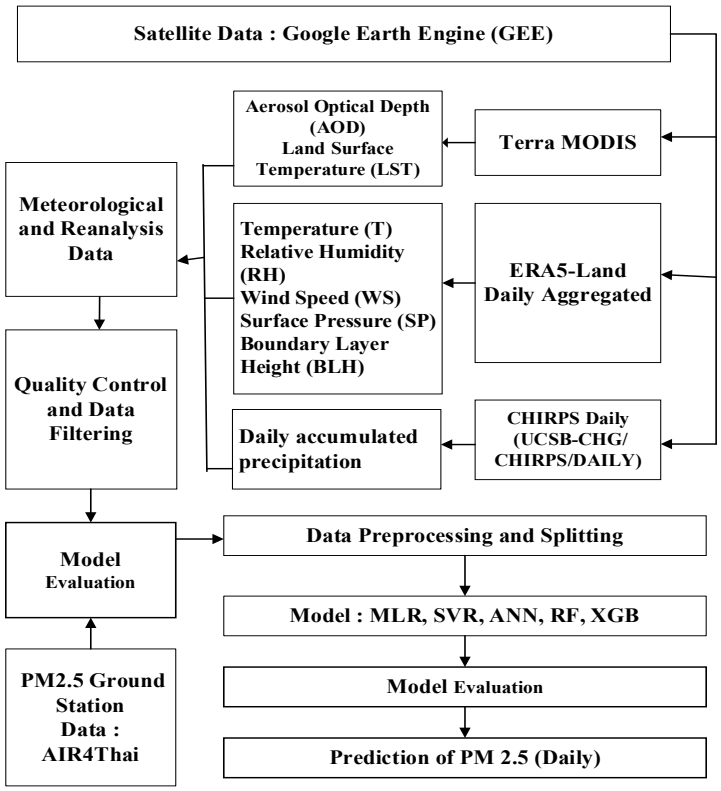

Fig. 2. Overview of the proposed framework for estimating daily $PM_{2.5}$ concentrations, including data acquisition, preprocessing, feature selection, model development, and validation.

Meteorological variables were obtained from the ERA5 and ERA5-Land datasets provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). These variables included air temperature, dew point temperature, relative humidity, wind speed, and boundary layer height (BLH), which are known to influence $PM_{2.5}$ formation and dispersion processes. All datasets were temporally matched and spatially processed to generate daily predictor variables. The processed satellite and meteorological variables were then used as independent inputs for statistical and machine learning models. The modeling approaches included simple linear regression, Multiple Linear Regression (MLR), and selected machine learning algorithms. Model performance was evaluated using standard error metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Based on the evaluation results, the best-performing model was selected to generate spatial distribution maps of daily $PM_{2.5}$ concentrations across Rayong Province.

2.2.1. Ground-Based $PM_{2.5}$ Monitoring Data

Daily PM2.5 concentration data were obtained from the Pollution Control Department (PCD) of Thailand. The measurements were conducted using the Beta Attenuation Method, following the standards of the United States Environmental Protection Agency (USEPA). Hourly and daily air quality data are publicly available through the AIR4Thai monitoring system (https://air4thai.pcd.go.th). In this study, daily $PM_{2.5}$ data were collected for the period from 1 November 2024 to 28 February 2025. The selected study period corresponds to the winter season in Thailand, during which $PM_{2.5}$ concentrations are typically elevated due to weak wind conditions, shallow boundary layer height, and frequent temperature inversions that limit atmospheric dispersion. Focusing on this season allows the models to capture pollution accumulation processes and evaluate performance under the most critical air quality conditions. A total of five ground-based monitoring stations located across Rayong Province were used for data collection (see **Fig. 2**). These stations included: (28t) Pluak Daeng District Public Health Office, (29t) Map Ta Phut Subdistrict Health Promoting Hospital, (30t) Rayong Provincial Agricultural Office, (31t) Rayong Field Crops Research Center, and (74t) Rayong Provincial Government Center. The selected monitoring stations represent different urban and industrial settings within the province. The ground-based $PM_{2.5}$ observations served as reference data for model development and performance evaluation.

2.2.2. Satellite and Meteorological Data

This study integrated satellite-derived variables with spatial meteorological data to support $PM_{2.5}$ estimation in a coastal region characterized by high variability in wind conditions and boundary layer height, such as Rayong Province.

Table 1.

Summary of satellite and meteorological datasets used in this study.

Variable	Data source (Google Earth Engine)	Unit	Spatial resolution	Temporal resolution
Aerosol Optical Depth (AOD)	MCD19A2.061: Terra & Aqua MAIAC Land Aerosol Optical Depth (Band: Optical_Depth_047)	Unitless	1 km	Daily
Land Surface Temperature (LST)	MOD11A1.061 Terra Land Surface Temperature and Emissivity (Band: LST_Day_1km)	Kelvin	1 km	Daily
2 m air temperature	ERA5-Land Daily Aggregated – ECMWF Climate Reanalysis (Band: temperature_2m)	Kelvin	0.1° × 0.1°	Daily
2 m dew point temperature	ERA5-Land Daily Aggregated – ECMWF Climate Reanalysis (Band: dewpoint_temperature_2m)	Kelvin	0.1° × 0.1°	Daily
U-component of wind at 10 m	ERA5-Land Daily Aggregated – ECMWF Climate Reanalysis (Band: u_component_of_wind_10m)	m s^{-1}	0.25° × 0.25°	Daily
V-component of wind at 10 m	ERA5-Land Daily Aggregated – ECMWF Climate Reanalysis (Band: v_component_of_wind_10m)	m s^{-1}	0.25° × 0.25°	Daily
Boundary layer height (BLH)	ERA5 Hourly – ECMWF Climate Reanalysis (Band: boundary_layer_height)	m	0.25° × 0.25°	Hourly

Meteorological variables included 2 m air temperature, 2 m dew point temperature, wind speed calculated from the u- and v-components, and boundary layer height (BLH). All satellite and meteorological datasets were extracted and processed using the Google Earth Engine (GEE) platform. This approach ensured spatial and temporal continuity of the data and helped reduce gaps associated with limited ground-based monitoring coverage. The max_pixels parameter was applied during data extraction to prevent processing errors related to large data requests. A summary of all variables and data sources used in this study is provided in **Table 1**.

2.2.3. Meteorological and Reanalysis Data

Ground-based $PM_{2.5}$ data, AOD, and LST from MODIS/MAIAC were combined with meteorological data from ERA5 and ERA5-Land. Spatial–temporal collocation was applied using the Nearest Neighbour method. All variables were converted to a daily resolution. Only days with complete data from all sources were used. This step helped reduce model uncertainty. Air temperature from ERA5 was originally provided in Kelvin (K). It was converted to degrees Celsius (°C) using Equation (1).

$$T_{(^{o}C)} = T_{(K)} - 273.15 \tag{1}$$

After data preprocessing, derived meteorological variables were calculated for model input. These included relative humidity (RH) and wind speed (WS). RH was calculated using the Magnus–Tetens equation, as shown in Equation (2):

$$RH(\%) = 100 \times \frac{e(T_d)}{e(T)} \tag{2}$$

in this equation, $e(T)$ and $e(T_d)$ represent vapor pressure in hPa

T is the air temperature at 2 m.

T_d is the dew point temperature at 2 m.

Wind speed (WS) was calculated from horizontal wind components using Equation (3):

$$WS = \sqrt{U^2 + V^2} \tag{3}$$

in this equation, WS is wind speed (m/s).

U is the west–east wind component (m/s).

V is the south–north wind component (m/s).

These derived variables are important for describing atmospheric processes. They influence the accumulation and dispersion of $PM_{2.5}$. They were used together with other meteorological variables. This improves spatial and temporal analysis accuracy.

2.3. Data Preprocessing and Quality Control

Before statistical analysis and model development, all satellite and meteorological datasets were preprocessed. All data were reprojected to the UTM Zone 48N coordinate system. A 10 km buffer was created around each $PM_{2.5}$ monitoring station. Daily values of each variable were extracted using median zonal statistics. This approach reduces spatial noise and outliers caused by sensor location uncertainty and terrain effects.

AOD data that did not meet MAIAC quality criteria were removed. Outliers were detected using the interquartile range (IQR) rule, and missing values, including NaN and infinite values, were excluded using Boolean masking. After data cleaning, the multivariate dataset included AOD, temperature, relative humidity (RH), wind speed (WS), and boundary layer height (BLH), with 468 samples remaining. The dataset was split into a training set (80%) and a test set (20%). All independent variables were standardized using the StandardScaler, with scaling parameters learned from the training set only to prevent data leakage.

2.4. Statistical Methods

Descriptive statistics were used to summarize the distribution and variability of $PM_{2.5}$ concentrations and meteorological variables during the study period. The analysis included measures of central tendency, data dispersion, and variability. The calculated statistics consisted of the mean, standard deviation, 95% confidence interval, minimum, maximum, and coefficient of variation (CV). These indicators are commonly used in environmental and atmospheric studies to describe air quality characteristics.

The mean value was calculated to represent the average concentration of each variable, as shown in Equation (4).

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4}$$

The standard deviation was used to describe data dispersion around the mean, as defined in Equation (5).

$$s = \sqrt{\frac{1}{n-1}}\sum_{i=1}^{n} (xi - \bar{x})^2 \tag{5}$$

The 95% confidence interval was computed to estimate the uncertainty of the mean value, following Equation (6).

$$CI_{95\%} = \bar{x} \pm t_{(1-\alpha/2,n-1)}\left(\frac{s}{\sqrt{n}}\right) \tag{6}$$

The minimum and maximum values were used to describe the data range, as shown in Equation (7).

$$x_{\min} = \min(x_i) x_{\max} = \max(x_i) \tag{7}$$

The coefficient of variation (CV) was calculated to assess relative variability, as defined in Equation (8).

$$CV(\%) = \frac{s}{\bar{x}} \times 100 \tag{8}$$

2.5. Theoretical Framework of PM2.5 Estimation Using Satellite and Meteorological Data

2.5.1. Atmospheric Processes Governing Near-Surface $PM_{2.5}$ Concentrations

Near-surface $PM_{2.5}$ concentrations result from interactions among emission sources, physical and chemical aerosol processes, and atmospheric dynamics in both horizontal and vertical directions. These processes are especially important in coastal industrial areas. Such regions are influenced by land–sea breeze circulation, wind speed variability, and changes in boundary layer height (BLH). From a theoretical perspective, near-surface $PM_{2.5}$ concentration can be described by the balance between emission rate (E) and atmospheric mixing volume. The mixing volume is mainly controlled by BLH and horizontal transport processes. This conceptual relationship can be expressed as shown in Equation (9):

$$C_{PM\,2.5}\alpha \frac{E}{BLH} \times f(WS) \tag{9}$$

Low BLH and low wind speed (WS) limit pollutant dilution. This leads to significant accumulation of $PM_{2.5}$ near the surface. This mechanism explains why the relationship between PM2.5 and meteorological variables is often nonlinear. It also indicates interaction effects among multiple variables acting simultaneously.

2.5.2. Physical Meaning of Aerosol Optical Depth (AOD) and Its Relation to $PM_{2.5}$

Aerosol Optical Depth (AOD) is a vertically integrated measure of the total aerosol loading in the atmospheric column. In physical terms, AOD represents the cumulative effect of the aerosol extinction coefficient along the vertical direction. This relationship can be expressed as shown in Equation (10):

$$AOD = \int_0^\infty \beta_{ext}(z)dz \tag{10}$$

Because $PM_{2.5}$ is a near-surface mass concentration, the relationship between AOD and $PM_{2.5}$ depends on the vertical structure of the atmosphere. It is also affected by boundary layer height, aerosol vertical distribution, and transport processes. Previous studies have shown that boundary layer height (BLH) plays a key role in linking columnar AOD to near-surface $PM_{2.5}$.This effect is especially strong in tropical and coastal regions.

2.5.3. Hygroscopic Growth and the Role of Relative Humidity (RH)

Relative humidity (RH) strongly influences aerosol optical properties and particle size through hygroscopic growth. Under high RH conditions, aerosol particles absorb water and increase in size. This process enhances light scattering and increases the extinction coefficient, even without a corresponding increase in $PM_{2.5}$ mass near the surface. The effect of hygroscopic growth can be described using a growth factor, as shown in Equation (11):

$$f(RH) = \frac{\beta_{ext}(RH)}{\beta_{ext}(dry)} \tag{11}$$

At high RH, aerosol water uptake increases particle size and scattering efficiency. As a result, AOD can increase without a proportional rise in near-surface $PM_{2.5}$ concentration. This physical mechanism has been widely used to explain $PM_{2.5}$–AOD relationships under humid atmospheric conditions. It is particularly relevant in Southeast Asia and East Asia. These findings support the physical basis of humidity-dependent statistical relationships between $PM_{2.5}$ and AOD.

2.5.4. Non-linear and Threshold Effects of Wind Speed on $PM_{2.5}$ Dispersion

Wind speed plays a key role in atmospheric pollutant dispersion. This concept is a fundamental part of atmospheric dispersion theory. However, many studies have shown that the relationship between $PM_{2.5}$ and wind speed (WS) is not linear. Clear threshold behavior has been widely reported. From a theoretical perspective, the response of $PM_{2.5}$ concentration to wind speed can be described as Equation (12):

$$\frac{\partial C_{pm2.5}}{\partial WS} < 0, \left| \frac{\partial^2 C_{pm2.5}}{\partial WS^2} \right| \neq 0 \tag{12}$$

2.5.5. Interaction Effects Between Wind Speed and Boundary Layer Height

Interaction effects between horizontal transport and vertical mixing are well explained by boundary-layer meteorology theory. Under conditions of low wind speed and low boundary layer height (BLH), atmospheric mixing volume is strongly limited. This leads to pollutant accumulation

near the surface. When either WS or BLH increases, pollutant dispersion becomes more effective. As a result, near-surface $PM_{2.5}$ concentrations decrease significantly. These interaction effects have been widely examined using statistical approaches. Examples include two-way ANOVA and interaction modeling. Such methods are used to quantify the combined effects of meteorological variables on $PM_{2.5}$.

2.6. Machine Learning Models

This study developed and evaluated machine learning models to predict daily $PM_{2.5}$ concentrations in Rayong Province using an integrated dataset of ground-based observations and satellite–meteorological variables. The predictor variables included Aerosol Optical Depth (AOD), land surface temperature (LST), relative humidity (RH), wind speed (WS), and boundary layer height (BLH). The use of multiple models enables performance comparison under highly nonlinear conditions and strong spatial variability commonly observed in urban–industrial coastal environments. Previous studies have reported that nonlinear and ensemble models generally outperform linear models in $PM_{2.5}$ estimation. Based on this evidence, both linear and nonlinear models were evaluated in this study.

2.6.1. Data Preprocessing and Splitting

Data extracted from buffer-based zonal statistics and quality control procedures were converted into a numerical format suitable for machine learning analysis. All non-numeric values, missing values (NaN), and infinite values were removed. Predictor variables and the target variable ($PM_{2.5}$) were combined into a structured tabular dataset. The dataset was randomly divided into a training set (80%) and a test set (20%) using a fixed random seed to ensure reproducibility. For models sensitive to feature scaling, including Support Vector Regression (SVR) and Artificial Neural Networks (ANN), feature standardization was applied using a StandardScaler fitted on the training set only. The same transformation was then applied to the test set to prevent data leakage. These steps ensured that the dataset was fully prepared for model development.

2.6.2. Model Description

Multiple Linear Regression (MLR) was used as a baseline model to represent linear relationships between $PM_{2.5}$ and the explanatory variables, including AOD, LST, RH, WS, and BLH. Model coefficients were estimated using the Ordinary Least Squares (OLS) method. The MLR formulation can be expressed as Equation (13):

$$PM_{2.5} = \beta_0 + \sum_{i=1}^{5} \beta_i f_i + \varepsilon \tag{13}$$

where β_0 is the intercept, β_i represents the regression coefficient of each independent variable, f_i denotes the predictor variables, and ε is the random error term. Model reliability was assessed by examining multicollinearity using the Variance Inflation Factor (VIF) and by analyzing residual distributions to test assumptions of homoscedasticity, independence, and normality. Although MLR offers high interpretability, linear models often show lower predictive accuracy under strongly nonlinear conditions.

Support Vector Regression (SVR) was applied to model nonlinear relationships using the margin maximization principle and the ε-insensitive loss function. All input features were standardized prior to training. Key hyperparameters, including the penalty parameter *C,* kernel coefficient γ, and ε, were optimized using grid search combined with k-fold cross-validation. The radial basis function (RBF) kernel was used to capture complex nonlinear patterns.

Artificial Neural Networks (ANN), implemented as a Multilayer Perceptron (MLP), were used to model complex nonlinear interactions among satellite and meteorological variables. The network employed multiple hidden layers with ReLU activation functions to enhance nonlinearity. The model was trained using the Adam optimizer with Mean Squared Error (MSE) as the loss function. Early

stopping was applied to reduce overfitting, and key hyperparameters were tuned to ensure stable convergence.

Random Forest (RF) is an ensemble tree-based model constructed using bootstrap aggregation and random feature selection. This structure makes the model robust to noise and effective in learning nonlinear relationships. RF does not require feature scaling and can provide feature importance measures, which are useful for interpreting the influence of individual variables on PM2.5 concentrations.

XGBoost is a gradient boosting model widely used in air quality and environmental studies. It incorporates both L1 and L2 regularization to reduce overfitting and is well suited for structured tabular data. XGBoost can effectively capture complex nonlinear patterns and interaction effects, often resulting in high predictive accuracy for $PM_{2.5}$ estimation.

2.6.3. Hyperparameter Tuning and Cross-Validation

Hyperparameter tuning is a critical step in machine learning model development, as model performance depends strongly on appropriate parameter selection. In this study, hyperparameters for each model were optimized using grid search combined with 5-fold cross-validation. This approach reduces bias associated with a single data split and improves model generalization performance. The optimal hyperparameter settings for each machine learning model are summarized in **Table 2**. The selected configurations reflect a balance between model complexity and predictive stability, particularly for nonlinear and ensemble models applied to atmospheric datasets with strong interaction and threshold effects.

Table 2.

Optimal hyperparameters for each machine learning model.

Model	Optimal Hyperparameters
Multiple Linear Regression (MLR)	No hyperparameter tuning performed
Support Vector Regression (SVR)	(C = 10), kernel = RBF
Artificial Neural Network (ANN)	Hidden layers = (100, 100), activation = ReLU
Random Forest (RF)	n_estimators = 400, max_depth = 30, min_samples_leaf = 1
XGBoost Regressor (XGBR)	learning rate = 0.01, max_depth = 5, n_estimators = 500

Note: Hyperparameter names follow standard XGBoost notation.

Table 2 presents the optimal hyperparameter configurations obtained through grid search with 5-fold cross-validation. The ANN achieved optimal performance with two hidden layers and ReLU activation, enabling effective learning of nonlinear relationships. SVR performed best with C = 10 using the RBF kernel, which supports flexible nonlinear function approximation. The ensemble tree-based models, RF and XGBoost, showed improved stability when a large number of trees were used. In particular, the low learning rate of XGBoost (0.01) helped reduce overfitting and allowed gradual model optimization.

2.7. Model Evaluation Metrics

Model performance in predicting daily $PM_{2.5}$ concentrations was evaluated using several quantitative metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). These metrics describe different aspects of prediction error between observed and predicted values. MAE measures the average magnitude of prediction errors and is less sensitive to large outliers. MSE and RMSE assign greater weight to large errors, making them more sensitive to extreme deviations. RMSE is particularly useful because it retains the same physical unit as $PM_{2.5}$. The R^2 metric represents the proportion of variance in observed $PM_{2.5}$ concentrations explained by the model. The definitions of these metrics are given as Equation (14)-(17).

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y - y_i| \tag{14}$$

Mean Squared Error; (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y - y_i)^2 \tag{15}$$

Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y - y_i)^2} \tag{16}$$

Coefficient of Determination (R^2)

$$R^2 = \frac{(\hat{y}_i - \bar{y}_i)^2}{(y_i - \hat{y}_i)^2} = 1 - \frac{(y_i - \hat{y}_i)^2}{(\hat{y}_i - \bar{y}_i)^2} \tag{17}$$

where y_i represents the observed $PM_{2.5}$ concentration, $\hat{y}_i$ is the model-predicted value, $\bar{y}_i$ is the mean of the observed values, and n is the total number of observations.

3. RESULTS AND DISCUSSION

3.1. Descriptive Statistics and Seasonal Variability of $PM_{2.5}$

3.1.1. Temporal and Seasonal Variations

Fig. 3 shows daily $PM_{2.5}$ concentrations measured at individual monitoring stations in the study area. The data cover period from 1 November 2024 to 28 February 2025. Clear temporal variability is observed across all stations, indicating short-term fluctuations and differences among locations. These variations reflect the combined influence of local emission sources and meteorological conditions, consistent with previous findings (Bran et al., 2024). Several pollution episodes are evident during the study period. $PM_{2.5}$ concentrations frequently exceeded 60–80 μg/m^3, with peak values above 80 μg/m^3 at some stations. These high-concentration events mainly occurred during periods of low wind speed and stable atmospheric conditions, which limited pollutant dispersion. Seasonal differences are also apparent, with higher $PM_{2.5}$ levels generally observed during the dry season compared to the wet season.

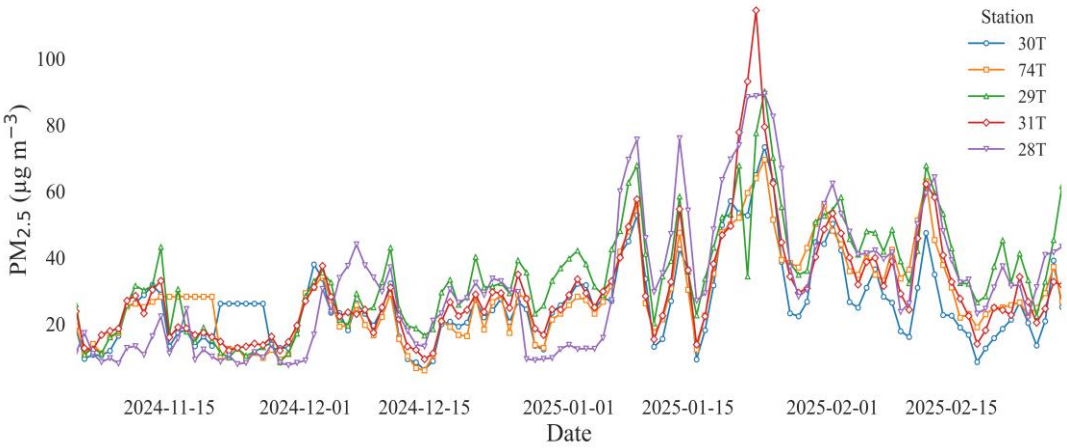

Fig. 3. Time series of observed daily $PM_{2.5}$ concentrations at the five monitoring stations during the study period.

Spatial differences among monitoring stations are clearly visible in **Fig. 4**. Some stations consistently recorded higher $PM_{2.5}$ concentrations than others, suggesting the influence of nearby emission sources and local land-use characteristics. In contrast, stations located farther from industrial areas showed relatively lower concentration levels. Overall, the temporal patterns indicate that $PM_{2.5}$ concentrations in the study area are strongly influenced by short-term meteorological variability as well as seasonal atmospheric conditions. These findings highlight the importance of incorporating meteorological and spatial factors into $PM_{2.5}$ estimation models, particularly in coastal industrial regions.

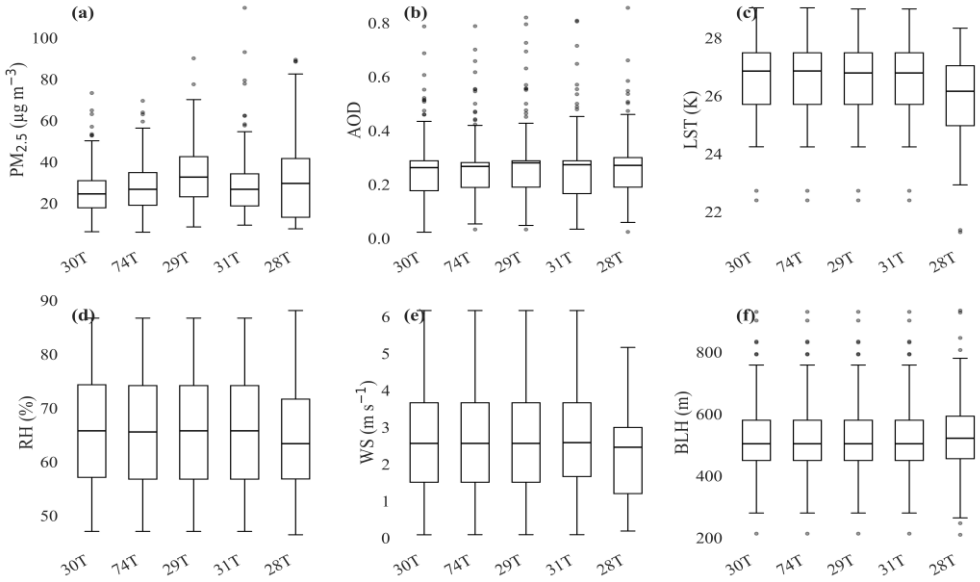

Fig. 4. Box plots showing the distribution of daily PM2.5 concentrations across the five monitoring stations during the study period.

3.1.2. Descriptive Statistics and Distribution Characteristics

Descriptive statistics and variability of $PM_{2.5}$ concentrations were evaluated together with satellite-derived and meteorological variables to characterize baseline pollution conditions during the winter season. **Table 3** summarizes the mean, standard deviation, 95% confidence interval, minimum, maximum, and coefficient of variation (CV) for all variables considered in this study. Overall, Table 3 highlights that $PM_{2.5}$ and AOD exhibit substantially higher variability than temperature, while relatively low mean wind speed and boundary layer height indicate meteorological conditions favorable for pollutant accumulation.

As shown in **Table 3**, the mean $PM_{2.5}$ concentration during the winter period was 30.05 μg/m^3 (95% CI: 28.73–31.37), which is substantially higher than the World Health Organization annual guideline, reflecting persistent air pollution in the study area. The high coefficient of variation (CV = 54.2%) further indicates strong day-to-day fluctuations and frequent episodic pollution events. Such variability is commonly associated with unfavorable meteorological conditions, including weak wind speed and shallow boundary layers, and is consistent with regional patterns reported in Thailand and Southeast Asia (Hassan Bran et al., 2024).

The mean AOD value was 0.27 with similarly high variability (CV = 50%), indicating strong fluctuations in aerosol loading within the atmospheric column. This variability is comparable to that observed for near-surface $PM_{2.5}$ and suggests a heterogeneous vertical aerosol distribution influenced by industrial emissions and regional transport processes. However, the large variability in AOD also highlights limitations in using satellite data alone as a direct proxy for surface $PM_{2.5}$, particularly under complex vertical atmospheric structures.

Table 3.

Descriptive statistics of $PM_{2.5}$ and explanatory variables during the study period.

Variables	Mean ± SD	95% CI	Minimum	Maximum	CV (%)
PM2.5 (µg/m³)	30.05 ± 16.29	[28.73, 31.37]	6	114.5	54.2
AOD	0.27 ± 0.14	[0.26, 0.28]	0.02	0.86	50
Temperature (°C)	26.48 ± 1.29	[26.38, 26.59]	21.3	29.04	4.9
WS (m/s)	2.53 ± 1.31	[2.42, 2.64]	0.08	6.16	51.7
RH (%)	65.63 ± 10.52	[64.77, 66.48]	46.43	88.12	16
BLH (m)	521.29 ± 122.89	[511.31,531.27]	209.44	934.21	23.6

In contrast, air temperature exhibited a narrow distribution (CV = 4.9%), reflecting the relatively stable winter climate in eastern Thailand controlled primarily by regional-scale air masses. This stability suggests that temperature plays a secondary role in explaining $PM_{2.5}$ variability compared with dynamic atmospheric factors. Wind speed (WS) and boundary layer height (BLH) showed relatively low mean values (2.53 m s^{-1} and 521 m, respectively) but high variability, indicating limited horizontal and vertical dispersion capacity. Under these stagnant conditions, both transport and mixing are suppressed, leading to enhanced accumulation of near-surface $PM_{2.5}$ concentrations.

In contrast, the background station (28T) displays a lower median and a narrower interquartile range, indicating weaker direct anthropogenic influence. However, the presence of outliers at this station suggests that $PM_{2.5}$ levels are occasionally dominated by regional-scale transport rather than local emissions. AOD shows similar median values across all stations, indicating that columnar aerosol loading is largely controlled by regional atmospheric processes rather than local land-use characteristics. Nevertheless, the presence of extreme AOD values at some stations reflects episodic high aerosol loading, potentially associated with transboundary haze events or aerosol accumulation under stagnant atmospheric conditions.

Air temperature exhibits narrow distributions and similar medians across all stations, consistent with the stable winter climate indicated in **Table 3**. Spatial temperature differences therefore play a limited role in explaining inter-station differences in $PM_{2.5}$. Relative humidity (RH) shows a wide distribution, particularly at coastal and mixed stations. High RH can enhance secondary aerosol formation and hygroscopic growth, potentially amplifying $PM_{2.5}$ concentrations during specific periods. Wind speed is generally low across stations, with occasional high-value outliers. This pattern indicates that atmospheric dispersion is limited for much of the winter season, especially in industrial and low-lying areas. boundary layer height (BLH) also exhibits a low median and wide variability. Days with boundary layer height (BLH) below approximately 400 m represent conditions of severely restricted vertical mixing and are frequently associated with elevated $PM_{2.5}$ concentrations across multiple stations.

3.1.3. Implications for Seasonal Variability and Subsequent Analysis

The analysis in Section 3.1 indicates that $PM_{2.5}$ concentrations in the study area are characterized by high variability, non-normal distributions, and strong spatial heterogeneity. Although temporal trends across all monitoring stations are generally consistent under the influence of regional-scale factors, the absolute concentration levels are clearly modulated by local land-use characteristics and station types. These features suggest that wintertime $PM_{2.5}$ variability cannot be fully explained by descriptive statistics alone. Instead, $PM_{2.5}$ variability is strongly associated with nonlinear relationships and interaction effects among $PM_{2.5}$, AOD, and meteorological variables. Therefore, Section 3.2 focuses on pairwise and conditional relationships among these variables to reveal underlying mechanisms that may be obscured by traditional linear analysis.

3.2. Pairwise Relationships and Non-linearity

Pairwise analysis between $PM_{2.5}$, AOD, and meteorological variables is an important step for understanding data structure and the underlying mechanisms of air pollution, particularly in coastal regions influenced by multidimensional atmospheric processes. In this study, pairwise scatter plots

were combined with Pearson correlation coefficients to evaluate both the direction and strength of linear relationships. At the same time, this approach allows detection of nonlinear patterns, data clustering, and heteroscedasticity that cannot be captured by summary statistics alone.

To better interpret these nonlinear patterns, interaction effects should also be considered. Conceptually, interaction effects indicate that the influence of one variable depends on the level of another rather than acting independently. For example, wind speed becomes more effective at dispersing pollutants when the boundary layer is shallow, leading to nonlinear responses in $PM_{2.5}$ concentrations.

3.2.1. Correlation Analysis Between PM2.5 and Environmental Variables

The pairwise analysis integrates correlation coefficients with scatter plot patterns to assess linear relationships and to identify signals of non-linearity that may not be evident from correlation values alone. **Fig. 5** presents pairwise scatter plots and Pearson correlation coefficients between $PM_{2.5}$, AOD, and meteorological variables across the five monitoring stations in Rayong Province during 01 November 2024–28 February 2025.

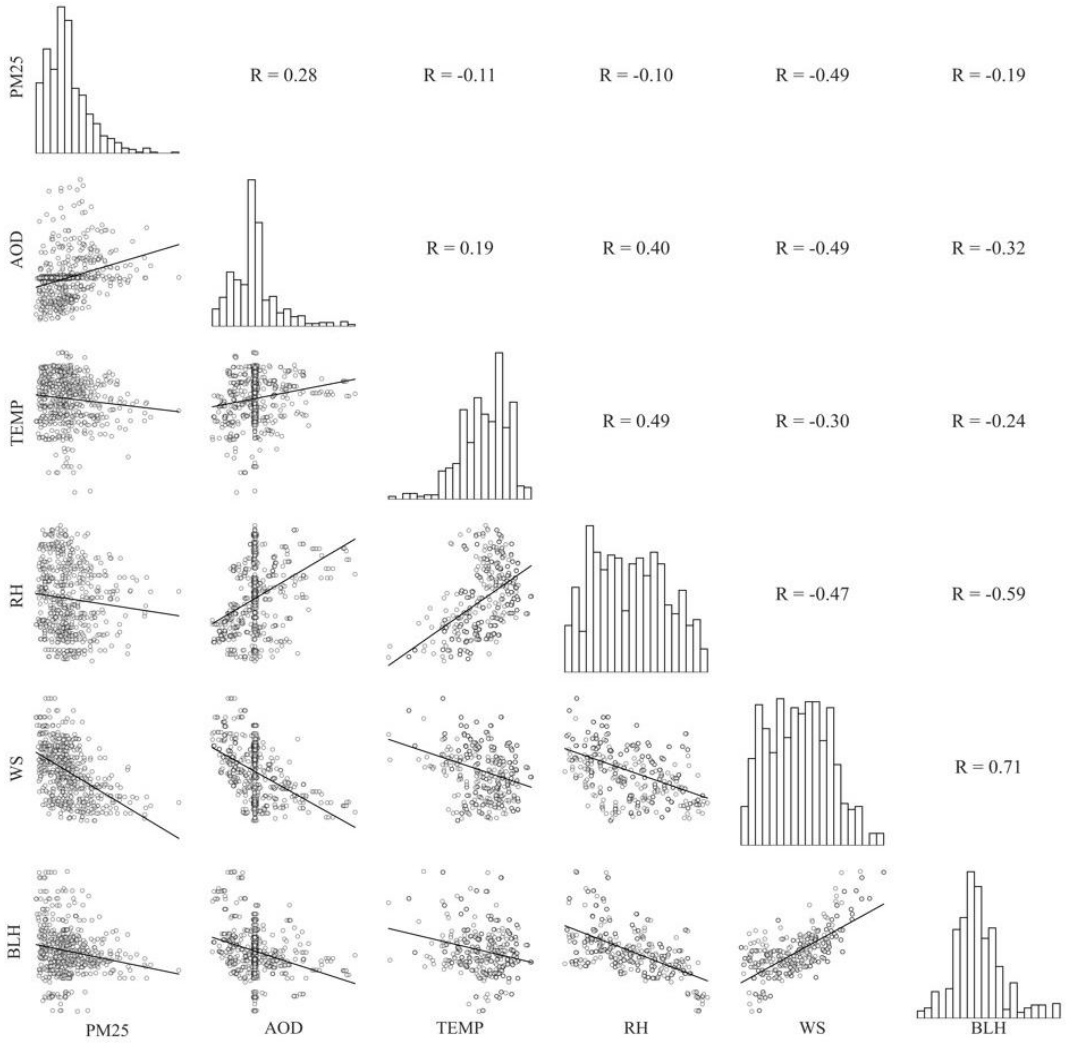

Fig. 5. Pairwise scatter plots illustrating the relationships between $PM_{2.5}$ and key meteorological variables, highlighting nonlinear patterns, threshold behavior, and conditional effects.

The results show that wind speed (WS) has the strongest negative correlation with $PM_{2.5}$ ($r = -0.49$). $PM_{2.5}$ concentrations decrease markedly as WS increases, highlighting the dominant role of horizontal transport and pollutant dilution in coastal environments. However, the scatter pattern is not strictly linear. High $PM_{2.5}$ values are strongly clustered under low WS conditions, indicating a clear threshold-type behavior.

Boundary layer height (BLH) shows a weaker negative correlation with $PM_{2.5}$ ($r = -0.19$) and a wide data spread. This suggests that vertical mixing plays a secondary role compared to horizontal transport, consistent with the stronger influence of WS observed in the analysis. AOD exhibits a moderate positive correlation with $PM_{2.5}$ ($r = 0.28$). However, the scatter plots show substantial variability, with many cases of high $PM_{2.5}$ occurring at moderate AOD levels. This pattern highlights the limitations of using AOD alone as a direct proxy for near-surface $PM_{2.5}$, especially in humid tropical and coastal regions where vertical aerosol stratification and moisture effects are pronounced.

Relative humidity (RH) and air temperature (TEMP) show very weak linear correlations with $PM_{2.5}$ ($r \approx -0.10$ to -0.11), with no clear linear trends in the scatter plots. These variables appear to act as modulating factors rather than direct drivers of PM2.5 concentrations. The non-normal distribution of $PM_{2.5}$ and the presence of numerous outliers further indicate strong nonlinearity in the relationships among $PM_{2.5}$, AOD, and meteorological variables. Spatial differences among monitoring stations further emphasize the influence of local emission sources and site-specific atmospheric processes. These spatial contrasts are consistent with previous studies conducted in Thailand and Southeast Asia.

3.2.2. Non-linear Relationships and Interaction Effects

To confirm and explain the nonlinear and conditional patterns observed in the scatter plots in **Fig. 5**, and to link these patterns with the physical framework described in Section 2.5, this study further examined $PM_{2.5}$ variability across different ranges of key meteorological variables. The analysis focused on threshold behavior and on conditional and interaction effects that cannot be captured by linear correlation coefficients alone.

(1) Threshold Effects: Wind Speed–$PM_{2.5}$ Relationship

Based on the conceptual framework described in Section 2.5.4, which emphasizes the nonlinear influence of horizontal transport on pollutant dispersion, a clear threshold-type relationship was identified between $PM_{2.5}$ concentration and wind speed (WS). Overall, **Table 4** demonstrates that $PM_{2.5}$ levels decrease markedly as wind speed increases, highlighting the dominant role of atmospheric ventilation in reducing near-surface pollutant accumulation.

Table 4.

$PM_{2.5}$ concentration stratified by wind speed.

Wind speed category (m s^{-1})	Mean $PM_{2.5}$ (μg m^{-3})	Difference from baseline (μg m^{-3})
$WS < 1.5$	34.2	Reference
$1.5 \leq WS < 2.5$	29.8	−4.4
$2.5 \leq WS < 3.5$	25.1	−9.1
$WS \geq 3.5$	22.6	−11.6

A distinct threshold was observed at approximately 2–2.5 m s^{-1} . $PM_{2.5}$ concentrations declined rapidly as wind speed (WS) increased from calm to moderate conditions, whereas further increases beyond about 3 m s^{-1} resulted in only marginal additional reductions. When wind speed (WS) < 1.5 m s^{-1}, pollutant accumulation was most pronounced, representing the baseline stagnant condition. As wind speed (WS) increased to 1.5–2.5 m s^{-1} and higher ranges, mean $PM_{2.5}$ concentrations decreased substantially; however, the rate of reduction progressively diminished at stronger wind speeds.

This nonlinear response indicates diminishing marginal dispersion effects under high ventilation conditions and confirms that the $PM_{2.5}$ – wind speed (WS) relationship cannot be adequately

explained by a simple linear assumption. Instead, the results suggest that wind-driven transport exerts the strongest influence under low-wind regimes, where small increases in wind speed (WS) can produce large improvements in air quality.

(2) Conditional Relationships: Influence of Relative Humidity on the $PM_{2.5}$ –AOD Relationship to examine the hygroscopic growth mechanism described in Section 2.5.3, the relationship between $PM_{2.5}$ and AOD was evaluated under different relative humidity (RH) conditions. Overall, Table 6 shows that the strength of the $PM_{2.5}$ AOD association varies systematically with humidity, indicating that moisture plays an important conditional role in linking columnar aerosol loading to near-surface particle concentrations.

Correlation analysis reveals that the $PM_{2.5}$ AOD relationship is not constant across atmospheric states. The correlation coefficient increased under humid conditions, and these differences were statistically confirmed using Fishers Z-test ($p = 0.029$). As summarized in **Table 5**, the strongest association was observed when RH was below 60% ($r = 0.58$), while a statistically significant relationship remained when RH exceeded 70% ($r = 0.31$).

This pattern suggests that aerosol hygroscopic growth enhances the optical response of particles, improving the representativeness of AOD for estimating near-surface $PM_{2.5}$ under moist conditions. Consequently, the conditional influence of humidity helps explain why the overall linear correlation between $PM_{2.5}$ and AOD appears moderate when all observations are analyzed together without stratification.

Table 5.

$PM_{2.5}$ –AOD Pearson correlations by relative humidity (95% CI).

RH Range (%)	PM2.5–AOD Correlation (r)	95% CI	p-value
< 60	0.58	[0.33, 0.75]	< 0.001
60–70	0.31	[−0.04, 0.60]	0.08
> 70	0.31	[0.01, 0.56]	0.04

(3) Interaction Effects between Wind Speed and Boundary Layer Height

Following boundary-layer meteorology theory described in Section 2.5.5, which emphasizes the combined effects of horizontal transport and vertical mixing, this study applied two-way ANOVA to examine interaction effects between wind speed (WS) and boundary layer height (BLH) on $PM_{2.5}$ concentrations. The analysis revealed a statistically significant interaction between wind speed (WS) and boundary layer height (BLH) ($F(1,236) = 8.7$, $p = 0.004$, partial $\eta^2 = 0.035$). $PM_{2.5}$ concentrations increased markedly when both horizontal transport and vertical mixing were simultaneously restricted. In contrast, $PM_{2.5}$ levels decreased significantly when at least one of the two processes favored pollutant dispersion. This interaction highlights the synergistic effects of atmospheric processes and reinforces the inherently nonlinear nature of $PM_{2.5}$ dynamics in coastal industrial environments.

Although the full dataset contains 468 daily observations, the effective sample size used in the two-way ANOVA is smaller. Prior to the analysis, wind speed (WS) and boundary layer height (BLH) were discretized into categorical groups to represent distinct atmospheric regimes. This binning procedure aggregates observations within each category and reduces the degrees of freedom in the statistical model. As a result, the reported degrees of freedom reflect the grouped structure of the data rather than the original number of observations. Such categorization is appropriate for interaction analysis, as it facilitates interpretation of joint effects between wind speed (WS) and boundary layer height (BLH) under physically meaningful atmospheric conditions.

3.3. Feature Selection Results

The results in Section 3.2 show that the relationships between $PM_{2.5}$ and meteorological variables are clearly nonlinear. Threshold behavior and interaction effects are evident for several variable pairs. These patterns agree with the atmospheric framework described in Section 2.5 and indicate that transport, vertical mixing, and aerosol physical–chemical processes cannot be explained by linear

assumptions alone. For this reason, feature selection in this study does not rely on a single statistical method. A hybrid approach is applied. Recursive Feature Elimination with Cross-Validation (RFECV) is used to evaluate how the number of input variables affects model performance, while Mutual Information (MI) is used to quantify nonlinear dependencies and rank the relative importance of predictors. In practical terms, feature importance reflects how strongly each variable contributes to reducing prediction error, meaning that variables with higher importance values exert greater influence on $PM_{2.5}$ estimates. All analyses are based on the preprocessed dataset described in Section 3.2, including AOD, temperature, relative humidity (RH), wind speed (WS), and boundary layer height (BLH).

3.3.1. Recursive Feature Elimination with Cross-Validation (RFECV)

RFECV was applied to evaluate whether incorporating variables related to atmospheric transport, vertical mixing, and aerosol processes improves the ability of the models to explain $PM_{2.5}$ variability. The analysis employed 5-fold cross-validation, and model performance was assessed using the mean cross-validated coefficient of determination (CV R^2). Overall, Table 7 demonstrates that predictive performance consistently improves as additional meteorological variables are introduced, particularly for nonlinear and ensemble models.

As summarized in **Table 6**, the CV R^2 values of ANN, Random Forest, and XGBoost increase markedly with the number of input features, with the largest gains observed after including wind speed (WS) and boundary layer height (BLH). The highest performance is achieved when all five variables are used (k = 5), indicating that these models effectively capture complex nonlinear relationships and interaction effects among meteorological factors. This result is consistent with the wind speed (WS) and boundary layer height (BLH) interaction effects identified in Section 3.2.2. In contrast, the linear Multiple Linear Regression (MLR) model shows only marginal improvement as additional predictors are added. Even after incorporating temperature and relative humidity, performance gains remain limited, and the inclusion of WS and BLH results in only minor increases. These findings highlight the limitations of linear assumptions in representing the nonlinear and threshold behavior of $PM_{2.5}$ in coastal industrial environments and further justify the use of nonlinear machine learning approaches. Intuitively, this procedure identifies the smallest set of variables that still provides strong predictive performance, helping to avoid unnecessary complexity while retaining the most informative environmental drivers.

Table 6.

Cross-validated R^2 from RFECV for different feature sets and regression models.

k	CV R^2 (MLR)	CV R^2 (SVR)	CV R^2 (ANN)	CV R^2 (RF)	CV R^2 (XGBR)
1	0.2215	–0.04	0.12	0.10	0.08
2	0.3567	0.02	0.22	0.23	0.20
3	0.3521	0.11	0.28	0.31	0.28
4	0.3594	0.33	0.49	0.49	0.46
5	0.3873	0.38	0.51	0.54	0.51

3.3.2. Mutual Information (MI)

To further examine variable relevance under non-linear relationships, Mutual Information is used. MI measures how much information each variable provides about $PM_{2.5}$ variability. The MI ranking is RH (f3) > Temperature (f2) > WS (f4) > BLH (f5) > AOD (f1). This order is consistent with the physical mechanisms described in Section 2.5. RH and temperature show high MI values because they directly influence hygroscopic growth and secondary aerosol formation. WS and BLH mainly control atmospheric transport and vertical mixing. Their effects on $PM_{2.5}$ are more conditional than direct. Although AOD has the lowest MI value when considered alone, it still plays a complementary role when combined with meteorological variables. This behavior reflects the fact that AOD represents column-integrated aerosol loading rather than near-surface concentration.

3.3.3. Permutation Importance

Permutation importance is applied to confirm the robustness of the RFECV and MI results. This method is model-agnostic and evaluates performance changes after random shuffling of each variable. The results are summarized in **Table 7**.

Table 7.
Permutation and built-in feature importance for Random Forest and XGBoost models.

Variable	Meaning
f1	Aerosol Optical Depth (AOD)
f2	Temperature
f3	Relative Humidity (RH)
f4	Wind Speed (WS)
f5	Boundary Layer Height (BLH)

Note: Predictor variables are indexed as follows: f1 = AOD, f2 = temperature, f3 = relative humidity (RH), f4 = wind speed (WS), and f5 = boundary layer height (BLH). Feature importance values are shown for both permutation-based and built-in methods for Random Forest and XGBoost models.

Table 7 compares permutation-based and built-in feature importance for the Random Forest and XGBoost models. Relative humidity (f3) and temperature (f2) show consistently high importance across both methods, with low standard deviations, indicating stable and robust contributions to $PM_{2.5}$ estimation. In contrast, wind speed (f4) and boundary layer height (f5) exhibit larger differences between permutation and built-in importance, suggesting that their influence depends more strongly on model structure and interaction effects rather than direct linear contributions. Aerosol optical depth (f1) shows lower but relatively stable importance, supporting its role as a complementary predictor rather than a dominant driver.

3.4. Model Performance Evaluation

3.4.1. Comparison of Model Performance for $PM_{2.5}$ Estimation

To evaluate how well the models capture the nonlinear relationships, threshold behavior, and interaction effects described in Section 2.5 and confirmed in Sections 3.2–3.3, five regression approaches were compared, including Multiple Linear Regression (MLR), Support Vector Regression (SVR), Artificial Neural Networks (ANN), Random Forest (RF), and Extreme Gradient Boosting (XGB). Model performance was assessed using MAE, RMSE, and R^2 on the test dataset (20%). All models were trained after hyperparameter tuning with 5-fold cross-validation to ensure good generalization and reduce the risk of overfitting. The results are summarized in **Table 8.**

Table 8.
Performance comparison of regression models for $PM_{2.5}$ estimation.

Model	MAE	RMSE	R^2	ΔRMSE (%)	ΔR^2
MLR	8.46	11.40	0.50	–	–
SVR	7.90	10.83	0.55	4.96	+0.05
ANN	7.79	9.91	0.62	13.08	+0.12
RF	6.78	8.89	0.70	22.03	+0.20
XGB	7.02	9.44	0.66	17.13	+0.16

Overall, **Table 8** clearly demonstrates that nonlinear and ensemble models consistently outperform the linear MLR baseline, with Random Forest achieving the best overall performance across all evaluation metrics. As shown in **Table 8**, RF yields the lowest MAE and RMSE and the highest R^2 value (0.70), indicating that it explains approximately 70% of the variance in $PM_{2.5}$ concentrations on the test dataset. Compared with MLR, RF reduces RMSE by more than 22% and increases R^2 by nearly 0.20. These improvements are consistent with the nonlinear relationships and interaction effects between $PM_{2.5}$, wind speed, and boundary layer height discussed in Section 3.2, which cannot be fully captured by linear models.

Although the R^2 value of 0.70 indicates strong explanatory power, it should be interpreted within the context of the study area. Rayong Province is characterized by complex emission sources and coastal meteorological influences, which increase the difficulty of $PM_{2.5}$ prediction. Previous studies in urban and industrial environments with similar characteristics typically report R^2 values between 0.60 and 0.80 when combining satellite AOD and meteorological variables. Therefore, the performance achieved here is consistent with related research. When considered together with RMSE and cross-validation stability, Random Forest provides the best balance between accuracy and robustness, making it well suited for subsequent spatial $PM_{2.5}$ mapping applications.

From an applied perspective, atmospheric processes rarely follow simple proportional relationships; small changes in meteorological conditions can produce disproportionately large changes in pollution levels. Therefore, nonlinear models are better suited to capture these threshold and accumulation effects.

3.4.2. Analysis of Model Predictions and Errors

To evaluate the quality and reliability of $PM_{2.5}$ predictions, model outputs are compared with ground-based observations using time series and statistical analyses. This evaluation aims to assess overall model accuracy and to identify error patterns under different concentration levels and spatial contexts. Comparisons between observed and Random Forest (RF)–predicted $PM_{2.5}$ are shown in Fig. 6. **Fig. 6** presents time series comparisons of observed and RF-predicted daily $PM_{2.5}$ at the five monitoring stations. The RF model consistently captures temporal trends across all stations at both daily and weekly scales. Multi-day pollution episodes are also well reproduced, indicating that RF effectively learns non-linear relationships between $PM_{2.5}$, AOD, and meteorological variables. Quantitatively, predicted values show the closest agreement with the 1:1 line among all models. The regression slope is 0.89 with an intercept of 3.2 $\mu g/m^3$, indicating low overall bias and strong systematic accuracy. These results are consistent with the lowest RMSE and highest R^2 values obtained by RF on the test dataset (**Table 8**), confirming its overall predictive performance.

However, during high pollution events, RF systematically underestimates $PM_{2.5}$ concentrations. This behavior is most pronounced from January to early February, when atmospheric ventilation is limited in both horizontal and vertical directions. The underestimation appears as reduced variability in predicted values relative to observations and is consistent with the regression slope being lower than one. This limitation reflects the model's reduced ability to capture the full intensity of extreme pollution events, despite accurately representing general trends. Ensemble models combine many decision trees and average their predictions, which improves stability but can also smooth extreme values, making very high pollution episodes harder to reproduce precisely.

This underestimation may be attributed to several factors. First, satellite-derived AOD represents column-integrated aerosol loading and may have limited sensitivity to near-surface pollutant accumulation under shallow boundary layers, leading to weaker signals during severe episodes. Second, rapid and localized emission spikes from industrial or traffic sources may not be fully captured by daily averaged predictors. Finally, machine learning models, particularly ensemble approaches, tend to smooth extreme values during training, which can reduce their ability to reproduce rare high-concentration events. These limitations collectively contribute to the systematic underprediction observed during peak pollution conditions. Station-level analysis reveals clear spatial differences in prediction performance.

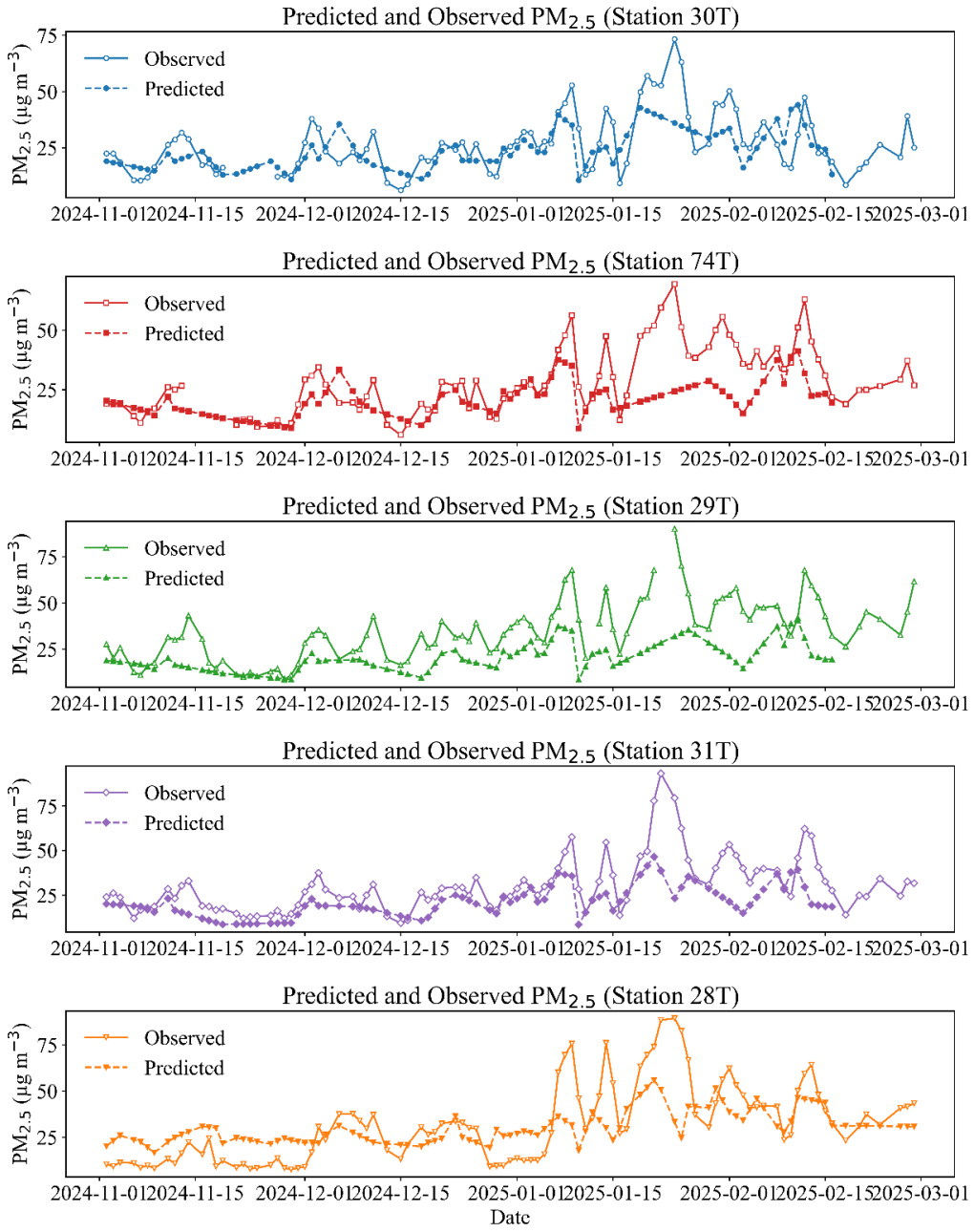

Fig. 6. Comparison between observed and Random Forest–predicted daily $PM_{2.5}$ concentrations at the five monitoring stations, illustrating model performance across different temporal conditions.

The background station (28T) shows the highest agreement between predicted and observed values, with accurate representation of baseline levels and temporal variability and relatively small errors during high concentrations. In contrast, the mixed station (31T) and the coastal station (29T) exhibit the largest discrepancies, particularly during rapid $PM_{2.5}$ increases. These patterns reflect the influence of diverse emission sources, localized transport, and rapidly changing atmospheric processes that are difficult to represent using regional-scale predictors. At the industrial stations (30T and 74T), RF also captures overall $PM_{2.5}$ trends well but continues to underestimate concentrations

during severe pollution episodes, especially when $PM_{2.5}$ increases sharply over short periods. These results suggest that while AOD and meteorological variables effectively represent general accumulation conditions, they remain limited in capturing short-term emission spikes and sub-daily dynamics.

In summary, station-based comparisons indicate that the Random Forest model performs well in representing temporal trends and baseline $PM_{2.5}$ levels across space and time. Prediction errors show clear conditional structures related to pollution intensity and source complexity, with extreme events remaining the main source of uncertainty. These findings support the need for further residual analysis and focused evaluation of model performance during severe pollution events, which are addressed in the following section.

3.5. Spatial Distribution of Satellite-Derived $PM_{2.5}$

The spatial distribution maps of satellite-derived $PM_{2.5}$ from November to February reveal clear spatial and temporal variations in air pollution levels across the study area (**Fig. 7**). $PM_{2.5}$ concentrations show a gradual increase from early winter to mid-winter, indicating progressive accumulation of particulate matter during the study period.

Fig. 7 illustrates that in November, $PM_{2.5}$ concentrations range from approximately 17.92 to 32.35 µg/m^3 Most areas experience moderate air quality to levels that may begin to affect health. Lower concentrations are observed in parts of the southern and southwestern areas, corresponding to air quality conditions with minimal health impacts. In December, $PM_{2.5}$ concentrations increase noticeably, with values ranging from 15.48 to 39.65 µg/m^3. Large portions of the province fall within categories associated with moderate to high health impacts.

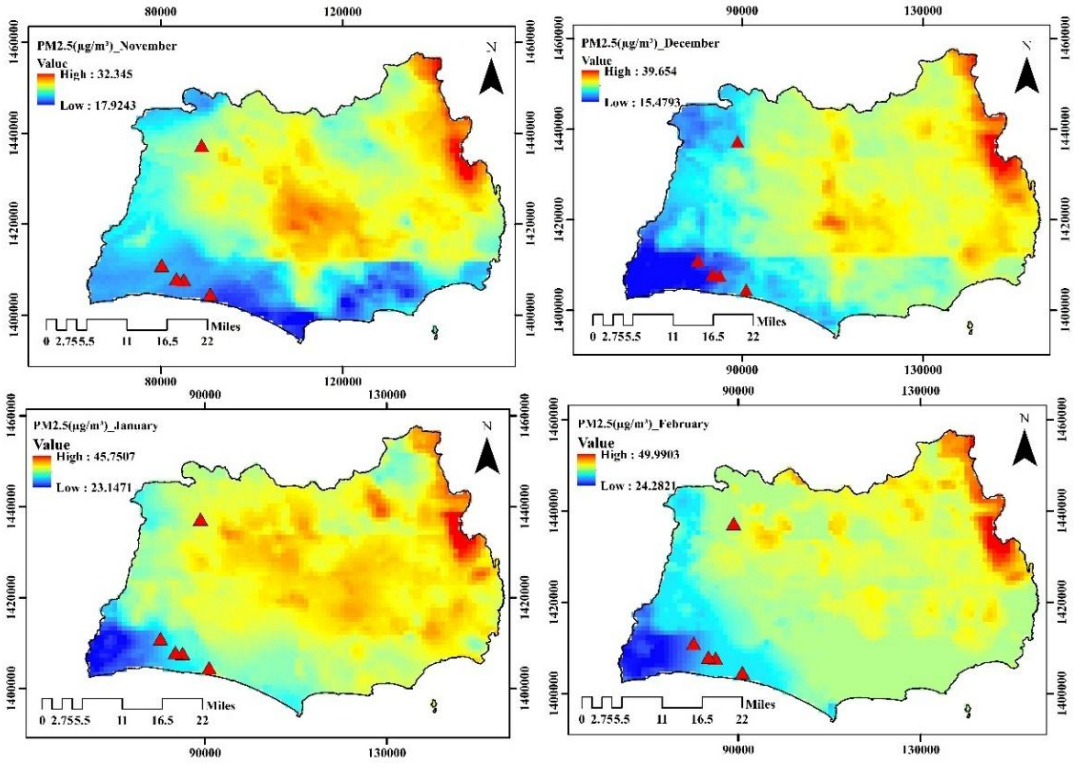

Fig. 7. Spatial distribution of satellite-derived $PM_{2.5}$ concentrations (µg/m^3) in the study area from November 2024 to February 2025, classified using fixed health-based thresholds.

Persistently elevated concentrations are observed in the eastern and northeastern areas, reflecting pollutant accumulation under meteorological conditions that limit atmospheric dispersion. $PM_{2.5}$ levels continue to rise in January, reaching a maximum of approximately 45.75 μg/m^3. Most of the study area is classified within high health impact levels, while some locations approach more severe risk categories. Pollution during this month covers a broader spatial extent than in previous months, indicating more widespread and intense air quality degradation during mid-winter.

In February, $PM_{2.5}$ concentrations reach their highest levels of the study period, ranging from about 24.28 to 49.99 μg/m^3. Some areas, particularly in the eastern part of the study area, approach or enter health-hazardous levels. These conditions indicate increased health risks for all population groups and highlight the need for intensified air quality monitoring and management. Overall, the satellite-based spatial analysis shows a consistent temporal increase in $PM_{2.5}$ concentrations and recurring pollution hotspots. The use of fixed health-based classification thresholds enables clear identification of health-vulnerable areas. These results demonstrate the strong potential of satellite-derived data to support spatial air quality assessment and health-oriented air pollution management at the local scale.

4. CONCLUSION

This study developed and evaluated daily $PM_{2.5}$ estimation models for Rayong Province by integrating ground-based monitoring data, MODIS/MAIAC aerosol optical depth (AOD), and meteorological variables from ERA5 and ERA5-Land. The results confirm that this integrated approach can effectively explain $PM_{2.5}$ variability in a coastal industrial area with complex atmospheric dynamics, particularly during the winter season when near-surface pollutant accumulation is favored. Model performance analysis shows that non-linear models, especially Random Forest and XGBoost, clearly outperform linear regression.

This finding indicates that relationships between $PM_{2.5}$ and satellite–meteorological variables are inherently non-linear and involve strong interaction effects. Wind speed plays the dominant role by controlling horizontal transport and dilution, while boundary layer height and relative humidity act as modifying factors. Clear threshold behavior is observed under calm wind conditions and shallow boundary layers. The results demonstrate that combining satellite data with machine learning not only improves spatial coverage beyond ground monitoring networks but also enables detailed mapping of $PM_{2.5}$ distributions. This approach is particularly useful for air quality assessment in coastal industrial regions with limited monitoring stations.

However, model uncertainty increases during severe pollution episodes and at very high $PM_{2.5}$ levels, highlighting the need to incorporate additional process-based information, such as spatial–temporal emission data and upper-atmospheric dynamics. Future work should extend the analysis to full-year periods, evaluate spatial robustness, and test model transferability to other coastal industrial regions. These steps would enhance the reliability and practical application of this framework for regional air quality management and environmental policy, especially in areas vulnerable to $PM_{2.5}$ pollution.

Beyond the study area, the proposed GeoAI framework is readily transferable readily transferable to other coastal or industrial regions with limited monitoring infrastructure, providing a practical tool for improving spatial air quality assessment and supporting evidence-based environmental management.

ACKNOWLEDGMENTS

This research project was financially supported by Mahasarakham University.

REFERENCES

Archer, D., Bhatpuria, D., Nikam, J., & Taneepanichskul, N. (2024). Particulate matter pollution in central Bangkok: Assessing outdoor workers' perceptions and exposure. *Cities & Health*, 1–19. https://doi.org/10.1080/23748834.2024.2390274

Aroonsri, I., & Sangpradid, S. (2021). Artificial neural networks for the classification of shrimp farm from satellite imagery. *Geographia Technica, 16*(2), 149–159. https://doi.org/10.21163/GT_2021.162.12

Bran, S. H., Macatangay, R., Chotamonsak, C., Chantara, S., & Surapipith, V. (2024). Understanding the seasonal dynamics of surface PM2.5 mass distribution and source contributions over Thailand. *Atmospheric Environment, 331*, 120613. https://doi.org/10.1016/j.atmosenv.2024.120613

Caplin, A., Ghandehari, M., Lim, C., Glimcher, P., & Thurston, G. (2019). Advancing environmental exposure assessment science to benefit society. *Nature Communications, 10*(1), 1236. https://doi.org/10.1038/s41467-019-09155-4

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Kovács, K. D., & Haidu, I. (2021). Effect of anti-COVID-19 measures on atmospheric pollutants correlated with the economies of medium-sized cities in 10 urban areas of Grand Est Region, France. *Sustainable Cities and Society*, 74, 103173. https://doi.org/10.1016/j.scs.2021.103173

Kawichai, S., Bootdee, S., Sillapapiromsuk, S., & Janta, R. (2022). Epidemiological study on health risk assessment of exposure to PM2.5-bound toxic metals in the industrial metropolitan of Rayong, Thailand. *Sustainability, 14*(22), 15368. https://doi.org/10.3390/su142215368

Kovács, K. D., & Haidu, I. (2024). Modeling NO2 air pollution variation during and after COVID-19 regulation using principal component analysis of satellite imagery. *Environmental Pollution*, 342, 122973. https://doi.org/10.1016/j.envpol.2023.122973

Laosuwan, T., Uttaruk, Y., & Rotjanakusol, T. (2023). Atmospheric environment monitoring in Thailand via satellite remote sensing: A case study of carbon dioxide. *Polish Journal of Environmental Studies*, 32(4), 3645–3651. https://doi.org/10.15244/pjoes/166170

Li, B., Chen, X., Zhang, W., Li, T., Xing, M., Yang, J., & Han, Z. (2025). Estimation of high-temporal-resolution PM2.5 concentration from 2019 to 2023 using an interpretable deep learning model. *Atmosphere, 16*(12), 1385. https://doi.org/10.3390/atmos16121385

Li, X., Peng, L., Hu, Y., Shao, J., & Chi, T. (2016). Deep learning architecture for air quality predictions. *Environmental Science and Pollution Research, 23*, 22408–22417. https://doi.org/10.1007/s11356-016-7812-9

Liu, Y., Wang, S., Li, T., & Hu, Y. (2022). Geographical artificial intelligence: A paradigm for spatial analysis. *Transactions in GIS, 26*(2), 638–656. https://doi.org/10.1111/tgis.12835

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

Nakapan, S., & Hongthong, A. (2022). Applying surface reflectance to investigate the spatial and temporal distribution of PM2.5 in Northern Thailand. *ScienceAsia, 48*(1), 75–81. https://doi.org/10.2306/scienceasia1513-1874.2022.001

Ponsawansong, P., Prapamontol, T., Rerkasem, K., Chantara, S., Tantrakarnapa, K., Kawichai, S., Li, G., Fang, C., Pan, X., & Zhang, Y. (2023). Sources of PM2.5 oxidative potential during haze and non-haze seasons in Chiang Mai, Thailand. *Aerosol and Air Quality Research, 23*, 230030. https://doi.org/10.4209/aaqr.230030

Rosales, C. M., Bratburd, J. R., Diez, S., Duncan, S., Malings, C., & Pant, P. (2025). Open air quality data platforms for environmental health research and action. *Current Environmental Health Reports, 12*(1), Article 27. https://doi.org/10.1007/s40572-025-00487-6

Rotjanakusol, T., Puckdeevongs, A., & Laosuwan, T. (2024). Relationship assessment between PM10 from the air quality monitoring ground station and aerosol optical thickness. *Geographia Technica, 19*(1), 79–88. https://doi.org/10.21163/GT_2024.191.06

Schneider, R., Vicedo-Cabrera, A. M., Sera, F., Masselot, P., Stafoggia, M., de Hoogh, K., Kloog, I., Reis, S., Vieno, M., & Gasparrini, A. (2020). A satellite-based spatio-temporal machine learning model to reconstruct daily PM2.5 concentrations across Great Britain. *Remote Sensing, 12*(22), 3803. https://doi.org/10.3390/rs12223803

Sorek-Hamer, M., Chatfield, R., & Liu, Y. (2020). Review: Strategies for using satellite-based products in modeling PM2.5 and short-term pollution episodes. *Environment International, 144*, 106057. https://doi.org/10.1016/j.envint.2020.106057

Sun, J., Yan, Z., Zhang, M., & Chen, Y. (2022). Predisposed obesity and long-term metabolic diseases from maternal exposure to fine particulate matter (PM2.5): A review of its effect and potential mechanisms. *Life Sciences, 301*, 121054. https://doi.org/10.1016/j.lfs.2022.121054

Tang, Y., Yang, X., Yang, J., Cai, Z., Han, S., Shi, J., Jiang, M., & Qiu, Y. (2022). Investigation of coastal atmospheric boundary layer and particle by unmanned aerial vehicle under different land-sea temperature. *Aerosol and Air Quality Research, 22*, 220206. https://doi.org/10.4209/aaqr.220206

Wang, J., He, L., Lu, X., Zhou, L., Tang, H., Yan, Y., & Ma, W. (2022). A full-coverage estimation of PM2.5 concentrations using a hybrid XGBoost-WD model and WRF-simulated meteorological fields in the Yangtze River Delta Urban Agglomeration, China. *Environmental Research, 203*, 111799. https://doi.org/10.1016/j.envres.2021.111799

Wong, P.-Y., Su, H.-J., Lung, S.-C. C., Liu, W.-Y., Tseng, H.-T., Adamkiewicz, G., & Wu, C.-D. (2024). Explainable geospatial-artificial intelligence models for the estimation of PM2.5 concentration variation during commuting rush hours in Taiwan. *Environmental Pollution, 349*, 123974. https://doi.org/10.1016/j.envpol.2024.123974

World Health Organization. (2021, September 22). *WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. https://www.who.int/publications/i/item/9789240034228

Zaman, N. A. F. K., Kanniah, K. D., Kaskaoutis, D. G., & Latif, M. T. (2021). Evaluation of machine learning models for estimating PM2.5 concentrations across Malaysia. *Applied Sciences, 11*(16), 7326. https://doi.org/10.3390/app11167326