AIMS Environmental Science

AIMS Environmental Science, 13(3): 528–549.
DOI: 10.3934/environsci.2026022
Received: 31 March 2026
Revised: 17 June 2026
Accepted: 02 July 2026
Published: 09 July 2026

https://www.aimspress.com/journal/environmental

Research article

Performance analysis of classical time series and machine learning models for PM2.5 and PM10 forecasting: A case study at an air quality monitoring station in Udon Thani, Thailand

Winai Meesang[1], Erawan Baothong[1], Krit Somkantha[2] and Wilaiporn Kultangwattana[2,*]

[1] Department of Environmental Health, Faculty of Science, Udon Thani Rajabhat University, Thailand
[2] Department of Computer Science and Information Technology, Faculty of Science, Udon Thani Rajabhat University, Thailand

* **Correspondence:** Email: wilaiporn.ku@udru.ac.th.

Abstract: Air pollution forecasting is a critical component of air quality management, especially in areas experiencing persistent particulate matter issues. However, in practice, the performance of forecasting models depends on local climatic conditions and human activities in each area, resulting in varying forecasting patterns based on spatial contexts. This research aimed to compare the performance between time-series methodologies and machine learning in forecasting hourly PM2.5 and PM10 concentrations, focusing on analysis at air quality monitoring stations in Udon Thani province, Thailand, a medium-sized urban area influenced by both local emissions and regional meteorological patterns. This study used hourly datasets collected from January 1, 2023, to November 30, 2024, comprising a total of 16,800 records obtained from the pollution control department's monitoring station at Nong Phra Chak Park. Since missing values are frequently encountered in real-world data collection, a linear interpolation method was employed to handle these gaps and enhance the efficiency of the proposed methods. The tested models included three time-series methods—seasonal autoregressive integrated moving average (SARIMA), seasonal naïve, and Holt–Winters exponential smoothing (ETS)—and four machine learning methods: random forest (RF), support vector regression (SVR), artificial neural networks (ANN), and extreme gradient boosting (XGBoost). The study utilized 80% of the data for training and 20% for testing, with performance evaluated through MAE, RMSE, R^2, MASE, and sMAPE indicators. The findings reveal that while time-series models can effectively reflect basic seasonal structures, they possess limited capability in capturing

short-term fluctuations and sudden pollution surges, with MAE ranging from 8.096 to 14.106. In contrast, machine learning models consistently demonstrated lower forecasting errors, particularly the support vector regression methodology, which provided the most accurate and stable performance for both PM2.5 (MAE = 2.4899) and PM10 (MAE = 3.7340). This demonstrates its efficiency in modeling nonlinear relationships and short-term dynamics under the environmental conditions of Udon Thani province. The experimental results suggest that forecasting models aligned with local characteristics can yield reliable predictions, which will benefit environmental agencies and policymakers in developing air quality surveillance systems and implementing mitigation measures that effectively reflect regional specificities.

Keywords: air pollution forecasting; PM2.5 and PM10; time-series models; machine learning and short-term forecasting

1. Introduction

Air pollution, particularly particulate matter smaller than 2.5 μm (PM2.5) and 10 μm (PM10), remains a significant environmental and public health issue in many countries around the world. Particulate matter affects people by penetrating the respiratory system and bloodstream, impacting health in both the short and long term. Numerous studies report that high exposure to PM2.5 and PM10 is associated with increased morbidity and mortality from respiratory and cardiovascular diseases [1]. Therefore, the World Health Organization (WHO) has established particulate matter concentration guidelines to be used as a criterion for assessing public health risks [2]. Monitoring, measuring, and forecasting PM2.5 and PM10 concentrations plays a crucial role in supporting air quality management and environmental policy formulation. Accurate forecasting can help provide early warnings, reduce public health impacts, and enhance the effectiveness of air pollution control measures. Time-series models are one widely used approach as they adequately describe the temporal structure and seasonal patterns of particulate matter data [3]. However, air pollution data is complex and nonlinear, influenced by varying weather factors, pollution sources, and other human activities in different areas, resulting in forecasting model performance that can vary depending on the spatial context. Some research, therefore, focuses on developing models that can better accommodate nonlinearity and data volatility, especially by creating models that are appropriate for the specific context of each area [4]. From these limitations, a comparative study of the performance of predictive models under different spatial contexts is important to support the development of an air quality monitoring system and pollution management measures that are appropriate and responsive to the local environment.

Forecasting air quality, particularly at the hourly scale, plays a critical role in supporting early warning systems, regulating pollution-related activities, and facilitating public health planning. Accurate short-term forecasts enable authorities to anticipate pollution episodes and implement timely mitigation measures to reduce population exposure. However, air pollution time-series data are inherently complex, characterized by strong seasonal components, short-term fluctuations, nonlinear behavior, and interactions with meteorological and geographical factors. These characteristics vary significantly across locations, making air quality forecasting a challenging task. Traditional statistical approaches, especially time-series models, have been widely employed for particulate matter forecasting due to their ability to capture temporal dependence and seasonal patterns. For example, seasonal autoregressive integrated moving average (SARIMA) models have demonstrated satisfactory

performance in modeling PM10 and PM2.5 concentrations in urban environments with relatively stable emission patterns [5]. Nevertheless, such models often struggle to adequately represent abrupt pollution events and nonlinear relationships commonly observed in real-world air quality data. To address these limitations, machine learning techniques have been increasingly applied in air quality forecasting studies. Artificial neural networks and other data-driven approaches have shown improved performance by learning complex nonlinear relationships between historical pollutant concentrations and influencing factors [6,7]. Empirical evidence suggests that these methods can outperform traditional statistical models, particularly in short-term forecasting scenarios where rapid variations in pollutant levels occur. Despite these advances, no single forecasting approach consistently delivers optimal performance across all spatial contexts. Model effectiveness remains highly dependent on local environmental conditions, emission sources, and climatic variability [8]. This highlights the necessity of evaluating multiple modeling techniques and selecting forecasting approaches that are appropriate for specific geographical and environmental settings.

Recent studies have increasingly adopted hybrid modeling approaches for particulate matter forecasting to address the nonlinear and high-dimensional characteristics of air quality data. By integrating multiple techniques, hybrid frameworks can more effectively capture temporal structures and spatial influences, leading to improved PM2.5 and PM10 prediction accuracy. For instance, ARIMA–LSTM hybrids have demonstrated significant error reductions in PM2.5 forecasting in Bangkok [9], while hybrid frameworks combining autoencoders, dilated convolutional networks, and gated recurrent units have enhanced prediction performance across multiple locations in Taiwan [10]. Other studies have reported promising results from integrating nonlinear autoregressive models with ARMA for short-term air pollution forecasting. Overall, hybrid models often outperform single-model approaches, particularly during high pollution episodes [11–13]. Although hybrid models have demonstrated significant potential in atmospheric science, this study aims to evaluate the performance of traditional time-series and fundamental machine learning models to establish a benchmark for the Udon Thani region. Establishing this benchmark is a crucial step in understanding the predictive capability for local pollution patterns before implementing more complex architectures, which may be prone to overfitting in data-limited scenarios. However, their performance remains highly dependent on data characteristics and spatial context, and gains are sometimes marginal or achieved at the expense of increased model complexity. Therefore, this study proposes a forecasting approach tailored to a specific spatial context, aiming to demonstrate improved effectiveness, stability, and practical applicability.

Udon Thani Province, a key area in northeastern Thailand, experiences periodic air pollution, particularly during the dry season, and open burning. Its climate, topography, and economic activities differ from major cities or industrial zones, resulting in unique patterns and behaviors for PM2.5 and PM10 data. Historically, the majority of air quality research in Udon Thani has primarily focused on particulate matter concentration monitoring, source apportionment, and statistical health impact assessments. However, a critical research gap remains regarding high-resolution hourly forecasting, which is essential for proactive air quality management. Therefore, our study presents a localized investigation that develops a region-specific hourly forecasting framework. This research aims to establish a reliable benchmark for local pollution patterns by systematically evaluating traditional time-series and machine learning models, thereby facilitating a shift from mere environmental monitoring to actionable forecasting. This study aims to forecast hourly PM2.5 and PM10 concentrations from air quality monitoring stations in Udon Thani Province, Thailand. It utilizes standard indicators such as MAE, RMSE, R^2, MASE, and sMAPE to evaluate model performance across multiple dimensions. The research findings are expected to help understand the advantages and limitations of each forecasting approach and provide appropriate policy recommendations for selecting

models for air quality management in Udon Thani Province and similar areas.

2. Materials and methods

2.1. Characteristics of Udon Thani Province and data sources

Udon Thani Province is located in the northeastern region of Thailand and serves as an important urban and economic center in the upper northeastern part of the country. The province has a tropical climate with three main seasons: the hot season (March–May), the rainy season (June–October), and the cool season (November–February). Seasonal meteorological conditions, including temperature, wind patterns, and rainfall, influence the dispersion and accumulation of airborne particulate matter. Understanding the seasonal characteristics of pollution is highly critical, as the region frequently experiences severe pollution events, particularly from late winter to summer (December to April). These intense pollution spikes are primarily driven by regional emission sources, specifically agricultural burning, transboundary haze, and temperature inversions.

Air quality data used in this study were obtained from the official air quality monitoring station operated by the Pollution Control Department under the Ministry of Natural Resources and Environment. The monitoring station is located at Nong Prajak Public Park in Udon Thani Province, representing an urban environment surrounded by residential areas and transportation activities. The geographical location of the monitoring station used in this study is illustrated in Figure 1.

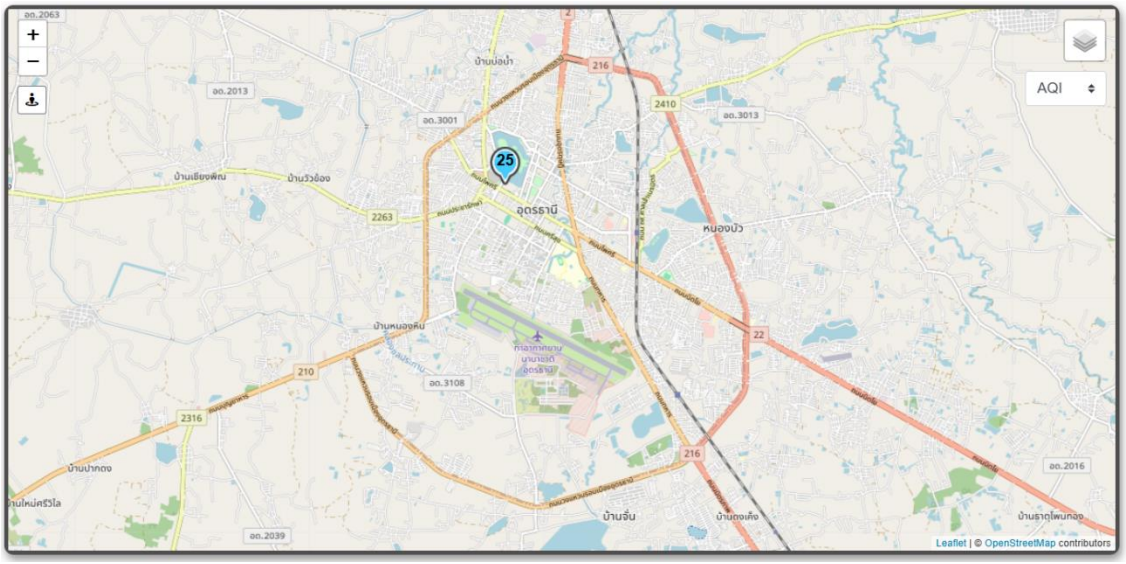

Figure 1. Location of the air quality monitoring station at Nong Prajak Public Park, Udon Thani Province, Thailand.

The dataset used in this research consists of hourly measurements of particulate matter concentrations, including PM2.5 and PM10, collected from January 1, 2023, to November 30, 2024, resulting in a total of 16,800 observations. The air quality data were obtained from the national air quality monitoring database publicly available through the Air4Thai platform provided by the Pollution Control Department [14]. These measurements were recorded using standardized air quality monitoring instruments that comply with national monitoring guidelines. The dataset includes timestamp information together with PM2.5 and PM10 concentration values, which serve as the

primary variables for the forecasting models developed in this study. A summary of the dataset characteristics is presented in Table 1. Prior to model development, the collected data were examined and prepared to ensure consistency and reliability for further analysis. The processed dataset was subsequently used as input for both time-series forecasting models and machine learning approaches in order to evaluate their performance in predicting short-term variations in particulate matter concentrations.

Table 1. Example of the hourly PM2.5 and PM10 dataset collected from the monitoring station.

No.	Date and time	PM10 ($\mu g/m^3$)	PM2.5 ($\mu g/m^3$)
1	01/01/2023 01:00	36	30
2	01/01/2023 02:00	35	26
3	01/01/2023 03:00	33	24
4	01/01/2023 04:00	33	26
5	01/01/2023 05:00	33	27
6	01/01/2023 06:00	32	28
7	01/01/2023 07:00	31	27
8	01/01/2023 08:00	32	28
...	...	...	...
16,797	30/11/2024 20:00	58	27.8
16,798	30/11/2024 21:00	61	31.5
16,799	30/11/2024 22:00	56	30.6
16,800	30/11/2024 23:00	61	37.3

2.2. Data preprocessing

During the collection of air quality data from real monitoring stations, missing or incomplete data may occur at certain time periods due to several factors, such as sensor malfunctions, communication system failures, or temporary interruptions in monitoring equipment. Specifically, in this study, the missing data accounted for 311 hourly observations (1.85%) for PM2.5 and 224 hourly observations (1.33%) for PM10. These gaps occurred as sporadic, short-term intervals, and such missing values may affect the accuracy and reliability of time-series analysis and forecasting results. Therefore, in this study, a data preprocessing step was performed to handle incomplete observations before conducting further analysis. In this study, the linear interpolation method was applied to estimate missing values in the hourly PM2.5 dataset. Linear interpolation is a widely used technique for estimating unknown values located between two known data points in a time-series dataset. The method assumes that the change between two adjacent observations can be approximated by a straight line [15,16]. Suppose that a missing value occurs at time t, which lies between two known observations y_{t1} and y_{t2}, where $t_1 < t < t_2$. The estimated value y_t can be calculated using the following equation:

$$y_t = y_{t1} + \frac{(y_{t2} - y_{t1})}{(t_2 - t_1)}(t - t_1) \quad (1)$$

where y_t represents the estimated value at time t, y_{t1} and y_{t2} are the known observations before and after the missing point, and t_1 and t_2 denote the corresponding time indices of the known observations.

This method estimates missing data based on the trend between neighboring observations, allowing the dataset to maintain continuity and consistency. The completed dataset was then used as input for subsequent time-series forecasting and machine learning model development. Missing data resulting from technical interruptions were handled using linear interpolation. Extreme concentration

values were retained rather than processed as outliers, as these represent genuine environmental events crucial to forecasting severe pollution episodes. Removing these would have compromised the model's reliability during critical periods.

Following the initial data preprocessing, this research applied a strict chronological time-series splitting method to define the experimental structure, thereby preventing data leakage or look-ahead bias. The continuous dataset was divided sequentially, where the first 80% of the chronologically timestamped historical data was designated as the training set to construct both the traditional time-series models and machine learning models. Meanwhile, the remaining 20% of the chronological data was reserved as the testing set to validate the performance of the proposed models.

2.3. Experimental design and model evaluation

This study proposes a forecasting framework for predicting hourly concentrations of particulate matter, specifically PM2.5 and PM10, using both time series and machine learning approaches. The experimental design focuses on the development and comparison of forecasting models based on hourly air quality observations obtained from the monitoring station in Udon Thani Province. The dataset consists of 16,800 hourly observations of PM2.5 and PM10 collected from January 1, 2023, to November 30, 2024, which serve as the primary input variables for the forecasting models. To evaluate the predictive capability of the models, the dataset was divided into two subsets. Specifically, 80% of the data were used for model training and parameter calibration, while the remaining 20% were reserved for testing and validation. In this study, two groups of forecasting approaches were implemented and compared. The first group consists of time-series models, including seasonal autoregressive integrated moving average (SARIMA), seasonal naïve, and Holt–Winters exponential smoothing. The second group consists of machine learning models, namely random forest, support vector regression (SVR), artificial neural networks, and extreme gradient boosting. These models were selected because they represent widely used techniques for environmental time-series forecasting and possess different capabilities in capturing temporal patterns and nonlinear relationships in air pollution data. To evaluate the forecasting performance, five statistical evaluation metrics were employed, including mean absolute error (MAE), root mean square error (RMSE), coefficient of determination (R^2), mean absolute scaled error (MASE), and symmetric mean absolute percentage error (sMAPE) [17,18].

Mean absolute error (MAE) measures the average magnitude of prediction errors without considering their direction:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{2}$$

where y_i represents the observed value, $\hat{y}_i$ denotes the predicted value, and nnn is the number of observations.

Root mean square error (RMSE) provides a measure of the square root of the average squared differences between predicted and observed values:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

Coefficient of determination (R^2) evaluates the proportion of variance in the observed data explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

where $\bar{y}$ is the mean of the observed values.

Mean absolute scaled error (MASE) is a scale-independent evaluation metric that measures the magnitude of forecasting errors while allowing consistent comparison across datasets with different scales.

$$MASE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|y_i - \hat{y}_i|}{\frac{1}{n-1}\sum_{j=2}^{n}|y_j - y_{j-1}|}\right) \tag{5}$$

Symmetric mean absolute percentage error (sMAPE) measures the relative forecasting error in percentage terms while reducing bias caused by large values:

$$sMAPE = \frac{100}{n}\sum_{i=1}^{n}\left(\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2}\right) \tag{6}$$

These evaluation metrics provide a comprehensive assessment of model performance in terms of absolute error magnitude, variance explanation, and relative forecasting accuracy. The results obtained from these metrics are presented and analyzed in Section 3.

2.4. Time-series forecasting models

Time-series forecasting is a data analysis technique that utilizes sequential data points over time to develop models for predicting future values, relying on historical patterns and relationships such as trends, seasonality, and random variations. This methodology is extensively applied in environmental and air quality research, as air pollution measurement data typically exhibits time-series characteristics with temporal change patterns. In this study, time-series forecasting was applied to predict the hourly concentrations of PM2.5 and PM10. The data, having undergone preprocessing and missing value management in the previous stage, were utilized to develop the forecasting models. Subsequently, the dataset was partitioned into a training set and a testing set for model construction and performance evaluation, respectively. Given that the data used in this research are hourly, the seasonal period was defined as s = 24, corresponding to the diurnal fluctuation patterns of the data. The models implemented in this investigation include seasonal autoregressive integrated moving average (SARIMA), seasonal naïve model, and Holt–Winters exponential smoothing, with the specific details of each method described in the following subsections.

2.4.1. Seasonal autoregressive integrated moving average (SARIMA)

The seasonal autoregressive integrated moving average (SARIMA) [19,20] is a statistical model developed for the analysis and forecasting of time-series data that possesses seasonal components. It is an extension of the autoregressive integrated moving average (ARIMA) model, designed to incorporate seasonal structures into the modeling process, thereby enabling a more effective explanation of data patterns that recur over specific intervals. The form of the SARIMA model is shown in the following equation:

$$SARIMA(p,d,q)(P,D,Q)_s \tag{7}$$

where p is the order of the autoregressive (AR), d is the degree of differencing, and q is the order of the moving average (MA). The parameters P, D, and Q correspond to the seasonal autoregressive order, seasonal differencing order, and seasonal moving average order, respectively. The parameter s indicates the seasonal period of the time series.

Using the backshift operator notation, the SARIMA model is shown in the following equation:

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^t_{yt} = \Theta_Q(B^s)\Theta_q(B)\epsilon_t \quad (8)$$

where B is the backshift operator defined as $B_{yt}=y_{t-1}$, y_t is the observed value at time t, ε_t is the random error term at time t, p is the order of the non-seasonal autoregressive (AR) component, d is the order of non-seasonal differencing, q is the order of the non-seasonal moving average (MA) component, P is the order of the seasonal autoregressive component, D is the order of seasonal differencing, Q is the order of the seasonal moving average component, and s denotes the seasonal period of the time series.

The polynomials of the autoregressive and moving average components are defined as

$$\phi_{p(B)} = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \quad (9)$$

$$\theta_{q(B)} = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \quad (10)$$

Similarly, the seasonal components can be written as

$$\Phi_{P(B^s)} = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps} \quad (11)$$

$$\Theta_{Q(B^s)} = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs} \quad (12)$$

In this study, the SARIMA model is applied to hourly air pollution data, where the seasonal period is defined as s = 24, corresponding to the daily cycle observed in the dataset. The parameters of the model are estimated using maximum likelihood estimation, and the fitted model is subsequently used to generate forecasts for the testing dataset.

2.4.2. Seasonal naïve model

The seasonal naïve model [21,22] is a statistical forecasting method that assumes future predicted values are equal to the actual observed values from the previous seasonal cycle. Its primary advantages include low computational complexity, being parameter-free, and serving as an effective baseline model. However, a significant drawback is its inability to account for long-term trends or stochastic events, as it relies solely on historical data based on seasonal periodicity. In this study, the model was implemented to forecast hourly PM2.5 and PM10 concentrations based on the daily cycle using the following equation:

$$\hat{y}_{T+h} = y_{T+h-s(k+1)} \quad (13)$$

where $\hat{y}_{T+h}$ represents the forecasted particulate matter concentration at time h beyond the training horizon, y denotes the actual observed historical value, T is the final time point of the training dataset, s is the seasonal period, and k is an integer representing the number of full seasonal cycles completed prior to time T + h.

Regarding the experimental configuration in this research, the seasonal period (s) was defined as 24, directly corresponding to the hourly measurement frequency and the 24-hour cycle of air pollutants in Udon Thani province. The forecast horizon (h) was configured to span the entire duration of the testing dataset, approximately 3360 hours (20% of the total observations), by cyclically repeating the final 24-hour observed values from the training set to establish a baseline for comparative performance analysis.

2.4.3. Holt–Winters exponential smoothing

The Holt–Winters exponential smoothing method, or triple exponential smoothing [23,24], is an advanced forecasting technique designed to effectively handle time-series data characterized by both trend and seasonal components. Its primary advantage lies in its adaptability to recent changes in data behavior. However, a notable limitation is its sensitivity to outliers, which may lead to over-smoothing and potentially hinder the model's ability to accurately capture peak concentrations during critical pollution episodes. In this study, the additive seasonality [or ETS(A,N,A)] framework was selected to forecast hourly particulate matter concentrations, as represented by the following set of equations:
Level equation:

$$L_t = \alpha(y_t - S_{t-s}) + (1 - \alpha)L_{t-1} \tag{14}$$

Seasonal equation:

$$S_t = \gamma(y_t - L_t) + (1 - \gamma)S_{t-s} \tag{15}$$

Forecast equation:

$$\hat{y}_{t+k} = L_t + S_{t-s+k} \tag{16}$$

where L_t represents the level of pollution concentration at time t, S_t denotes the seasonal component, y_t is the actual observed concentration, $\hat{y}_{t+k}$ signifies the forecasted value at k steps into the future, s is the seasonal period, and α and γ are the smoothing parameters for the level and seasonal components, respectively.

Regarding the parameter configuration in this study, the level smoothing constant (α) and the seasonal smoothing constant (γ) were both set to 0.3. This specific calibration was chosen to maintain an optimal balance between responsiveness to recent data fluctuations and the preservation of historical seasonal patterns. The seasonal period (s) was defined as 24 hours to align with the cycle inherent in the pollution data. Furthermore, the initial values for the level (L_1) and the seasonal components ($S_{1:S}$) were derived from the averages and deviations observed during the first 24-hour window of the training dataset, ensuring the model's predictive accuracy from the onset of the forecasting process.

2.5. Machine learning models

In the past decade, applying machine learning (ML) models has become an important method for data classification and forecasting. This is because these methods can effectively learn complex and nonlinear relationships from large amounts of data, unlike traditional statistical models that usually require set assumptions about data structure beforehand. Machine learning methods can adapt to the nature of the data through a data-driven learning process, leading to better accuracy in forecasting and analyzing complex environmental systems. The ability of these algorithms helps improve learning from datasets to make environmental predictions more accurate and reliable [25–29]. For this reason, this research proposes efficient machine learning models, including random forest regression, support vector Regression, and artificial neural network, to develop and compare their ability to forecast PM2.5 and PM10 concentrations in the Udon Thani province.

2.5.1. Random forest regression (RF)

The random forest model is an effective ensemble learning algorithm [30,31]. This method relies

on the principle of bootstrap aggregating (bagging) to create a set of decision trees (h) that are diverse and independent of each other. The advantage of this approach is its ability to reduce the variance of the overall model without significantly increasing the bias. This results in high robustness against noise in the data and a better ability to reduce the risk of overfitting compared to a single decision tree. However, the disadvantage of the random forest algorithm is its structural complexity, which makes the interpretability of physical relationships more difficult than traditional linear models. During the training process, the model uses bootstrap sampling to create subsets of data L_k from the main dataset L by sampling with replacement, according to the equation:

$$L_k = \ Bootstrap\big(L, N_{samples}\big) \tag{17}$$

Each constituent tree is trained on its respective L_k subset. To ensure de-correlation between trees, a random sub-spacing technique is applied at every node split. The final prediction ($\hat{y}$) is derived by aggregating the individual outputs from all N trees through an averaging process:

$$\hat{y}(x) = \frac{1}{N}\sum_{k=1}^{N} h_k(x, \Theta_k) \tag{18}$$

where h_k represents the output of the k tree, and Θ_k denotes the independent identically distributed random vector used for each bootstrap iteration.

A pivotal attribute that renders the random forest (RF) framework exceptionally suitable for forecasting PM2.5 and PM10 concentrations of pollutants characterized by significant atmospheric volatility is its inherent mechanism for ensemble variance reduction. This is achieved by aggregating the predictions of multiple decorrelated decision trees. Mathematically, if we denote the variance of an individual tree as σ^2 and the average pairwise correlation between any two trees as ρ, the total variance of the RF model can be evaluated using the following expression:

$$Var(\hat{y}) = \ \rho\sigma^2 + \frac{1-\rho}{N}\sigma^2 \tag{19}$$

where $\mathrm{Var}(\hat{y})$ signifies the total ensemble variance, ρ represents the mean inter-tree correlation, σ^2 is the variance of a single estimator, and N denotes the total number of trees within the forest.

In this paper, to ensure maximum predictive stability and precision for the hourly dataset obtained from the Udon Thani monitoring station, the ensemble size N was strategically configured at 200 trees. According to the theoretical framework above, as N increases, the second term of the equation $\frac{1-\rho}{N}\sigma^2$ asymptotically diminishes toward zero, effectively constraining the system's total variance primarily by the correlation coefficient ρ. Furthermore, the implementation of feature sub-spacing—specifically by randomly selecting a subset of input predictors (Lags [1, 2, 3, 24]) during node construction—plays a critical role in minimizing ρ. This decorrelation strategy empowers the RF model in this study to robustly distinguish genuine pollutant signals from stochastic noise induced by localized activities or erratic meteorological shifts. Consequently, this leads to a substantial reduction in generalization error and enhances the model's overall robustness when processing highly uncertain environmental time-series data.

2.5.2. Support vector regression (SVR)

The support vector regression (SVR) [32,33] model is a highly effective supervised learning algorithm for handling nonlinear regression problems, based on the fundamental principle of finding a hyperplane in a high-dimensional space to create a prediction boundary within a specified error range ($\in$).

A key advantage of SVR is its ability to control model complexity through the penalty parameter (*C*) and its robustness against outliers; however, its performance is sensitive to the choice of kernel function and requires data scaling for accurate distance calculations between data points. Regarding the kernel function and mapping, since PM2.5 and PM10 data have complex nonlinear relationships with historical factors (Lags [1,2,3,24]), this research selected the radial basis function (RBF) or Gaussian kernel to transform the data into a high-dimensional feature space, allowing the SVR to create a regression curve that fits the behavior of the pollutants according to the following mathematical form:

$$K(x_i, x_j) = \ exp\left(-\gamma \left|\left|x_i - \ x_j\right|\right|^2\right) \tag{20}$$

where $K(x_i, x_j)$ denotes the kernel function quantifying the similarity between input vectors, $\left|\left|x_i - \ x_j\right|\right|^2$ represents the squared Euclidean distance between two data points, and γ is the kernel parameter (gamma) that regulates the influence radius of the support vectors.

To ensure maximum efficiency in the RBF kernel function, this research applied data standardization before the training process. Since the input variables for each dimension (Lags) may have different value ranges, standardization transforms the data to have a mean of 0 and a standard deviation of 1. This process prevents variables with higher numerical values (such as dust levels during critical periods) from dominating other variables and also helps accelerate the convergence of the optimization process during model training. The primary objective of the SVR model is to identify an optimal hyperplane that minimizes the regularized risk function:

$$min\frac{1}{2}\left|\left|w\right|\right|^2 + \ \ C \ \ \sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{21}$$

where $\left|\left|w\right|\right|^2$ controls the smoothness or complexity of the model, *C* is the limit of the box (penalty parameter) that manages the balance between model simplicity and acceptable error, and ξ_i, ξ_i^* are the excess variables representing the deviations of the training sample that are outside the zone insensitive to changes in $\in$ of pollution concentration. The final predicted *f(x)* is calculated using the following extension:

$$f(x) = \ \sum_{i=1}^{n}(\alpha_i - \ \alpha_i^*)K(x_i, x) + \ \ b \tag{22}$$

where α_i and α_i^* are the Lagrange multipliers associated with the support vectors, and *b* is the bias term determined during the optimization phase.

In this paper, the SVR model was structured to detect air quality changes in Udon Thani province, specifically utilizing the radial basis function (RBF) kernel due to its suitability for modeling complex and nonlinear environmental data relationships. Additionally, data standardization was established as a crucial step to balance the scales of historical lag data at [1,2,3,24] hours, preventing variables with higher numerical values from dominating the calculations. This process, applied to 16,800 hourly observations, effectively accelerates the convergence to optimal parameters during the model construction process.

2.5.3. Artificial neural network (ANN)

The artificial neural network (ANN) [34,35] in this paper is designed based on the feedforward backpropagation architecture to serve as a nonlinear function approximator for forecasting future

particulate matter concentrations. Its primary advantage is its high flexibility in capturing patterns within data characterized by hourly volatility, performing better than many other statistical models. However, a disadvantage of the ANN is its sensitivity to the initial parameter values, which may cause the convergence to get trapped in a local minimum if an efficient training algorithm is not utilized.

The computational process begins by feeding historical pollutant indices at lags of *t-1*, *t-2*, *t-3*, and *t-24* hours as input features ($y_{\{t-i\}}$). Within the hidden layer, each neuron computes a weighted sum of these inputs, incorporates a bias term, and applies a nonlinear activation function to facilitate feature extraction, as defined by the following expression:

$$h_j = \sigma\left(\sum_{i\in\{1,2,3,24\}} w_{\{ji\}} y_{\{t-i\}} + b_j\right) \tag{23}$$

where h_j represents the output from the neuron in the hidden layer, $y_{\{t-i\}}$ denotes the particulate matter concentrations at various time lags, $w_{\{ji\}}$ are the weights linking the historical data to the neuron, and b_j is the neuron's bias. Subsequently, all outputs from the hidden layer are passed to the output layer to calculate the predicted particulate matter concentration for the current time ($\hat{y}_t$) according to the following summation equation:

$$\hat{y}_t = \mathrm{f}_{\{\mathrm{out}\}}\left(\sum_{j=1}^{10} w_j h_j + \mathrm{B}_{\{\mathrm{out}\}}\right) \tag{24}$$

where $\hat{y}_t$ is the forecasted PM2.5 or PM10 concentration, w_j represents the weights connecting the hidden layer to the output node, and $\mathrm{B}_{\{\mathrm{out}\}}$ is the output layer bias.

In this paper, to suit the air pollution values in Udon Thani province, the model was configured with a single hidden layer consisting of 10 neurons and utilized the Levenberg–Marquardt algorithm [36] for training. This algorithm is highly efficient in updating weights (*w*) and biases (*b*) through the calculation of the Jacobian matrix to rapidly minimize the mean squared error (MSE). As a result, the model can effectively learn the patterns of increasing or decreasing particulate matter influenced by the cycle.

2.5.4. Extreme gradient boosting (XGBoost)

The XGBoost model is a machine learning algorithm based on ensemble learning [37,38], which extends the structure of decision trees under the gradient boosting framework. This framework emphasizes the sequential collaboration of weak learners to construct a highly accurate strong learner. Its key strengths lie in predictive efficiency, the capability to handle nonlinear relationships within data, and the reduction of overfitting issues through regularization techniques. The underlying mechanism of XGBoost relies on creating multiple decision trees sequentially (sequential learning), where each individual tree is built to correct the residual errors generated by the preceding trees, thereby enabling the model to continuously improve its accuracy. The XGBoost model can be formulated as follows:

$$\mathrm{Obj}(\theta) = \sum_{\mathrm{i}=1}^{\mathrm{n}} L(y_i, \hat{y}_i) + \sum_{\mathrm{k}=1}^{\mathrm{K}} \Omega(f_k) \tag{25}$$

where $L(y_i, \hat{y}_i)$ represents the loss function used to measure the discrepancies between the actual observed outcomes (y_i) and the predicted values ($\hat{y}_i$), while $\Omega(f_k)$ denotes the regularization term,

which functions to control the complexity of the decision tree structure to prevent overfitting, which is shown in the following equation:

$$\Omega(f) = \gamma \mathrm{T} + \frac{1}{2}\lambda \sum_{\mathrm{j}=1}^{\mathrm{T}} w_j^2 \tag{26}$$

where T is the number of leaf nodes in each tree, w_j represents the score weight of leaf node j, and γ and λ denote the penalty parameters utilized to control the model complexity and prevent overfitting.

In the forecasting process, the XGBoost model receives the lagged features as an input dataset and sequentially adds decision trees one by one in an additive manner. Each new tree is constructed to learn from and minimize the residual errors from the preceding trees, utilizing the first- and second-order derivatives of the loss function. In this research, the learner type was specified as gbtree, utilizing regression with squared loss (reg:squarederror) as the objective function. The learning rate was set at 0.1, the maximum tree depth was limited to 6 levels, and the total number of boosting rounds for tree construction was set at 100 iterations to effectively control the model structure and prevent overfitting.

3. Results

This section presents the analysis results and performance evaluation of various models for forecasting PM2.5 and PM10 concentrations in Udon Thani province. The presentation structure is divided into four main parts: Starting with the analysis of basic statistical characteristics of air pollution data to initially understand data behavior and distribution, followed by the forecasting results from traditional time-series models; then, the predictions obtained from machine learning models, and finally, a comparative analysis of the forecasting performance among all models for air quality in Udon Thani province.

3.1. Descriptive statistical analysis of air pollutants

In this paper, the basic statistical characteristics of the air pollution data were studied to evaluate the behavior and volatility of the dataset in Udon Thani province. This preliminary statistical analysis helps identify the scale of the pollution problem in the area and serves as a crucial step in verifying assumptions regarding data distribution. Table 2 presents the statistical values derived from the hourly concentrations of PM2.5 and PM10 recorded by air quality monitoring stations in Udon Thani throughout the study period, comprising a total of 16,800 continuous data points. The analysis results reveal significant fluctuations in pollution levels.

Data show that the mean concentrations of PM2.5 and PM10 are 36.83 μg/m^3 and 23.15 μg/m^3, respectively, with maximum values reaching 198.00 μg/m^3 and 154.00 μg/m^3, both of which exceed air safety standards. Furthermore, the high standard deviation relative to the mean reflects a high level of variability in the dataset, which is linked to local pollution events. Distribution analysis confirms that both pollutants exhibit a right-skewed distribution, with skewness coefficients of 1.47 for PM2.5 and 1.29 for PM10. Meanwhile, the kurtosis values (4.92 and 5.99) indicate a leptokurtic distribution, suggesting a higher probability of extreme values compared to a normal distribution. Additionally, the Kolmogorov–Smirnov test yielded a p-value significantly below 0.001, confirming that the air quality data is not normally distributed. Consequently, this research proposes time-series and machine learning methods to compare with environmental pollution data in the area, which exhibits nonlinear relationships and non-normal distributions.

Table 2. Summary statistics for hourly PM2.5 and PM10 data.

Statistical parameters	PM2.5 (μg/m^3)	PM10 (μg/m^3)
Sample size (n)	16,800	16,800
Minimum	0	0
Maximum	198	154
Mean	36.83	23.15
Std. deviation (SD)	25.22	18.39
Skewness	1.29	1.47
Kurtosis	4.92	5.99
Normality (p-value)	<0.001*	<0.001*

*Note: Significant at the 0.05 level.

3.2. Forecasting performance of time-series models

This section tests the forecasting performance of three classical time-series models: seasonal autoregressive integrated moving average, seasonal naïve, and Holt–Winters exponential smoothing. A dataset was split into two parts: 80% for model training and 20% for testing, to reflect the models' forecasting capabilities on previously unseen data. Figures 2 and 3 illustrate the capability of each method in tracking the trends and seasonal patterns of the data. In particular, these figures show a comparison between the actual values and the forecasting results from all three models. Furthermore, to ensure clarity in visualizing the graph lines due to the large volume of data, a sample of 5% of the total test dataset is displayed.

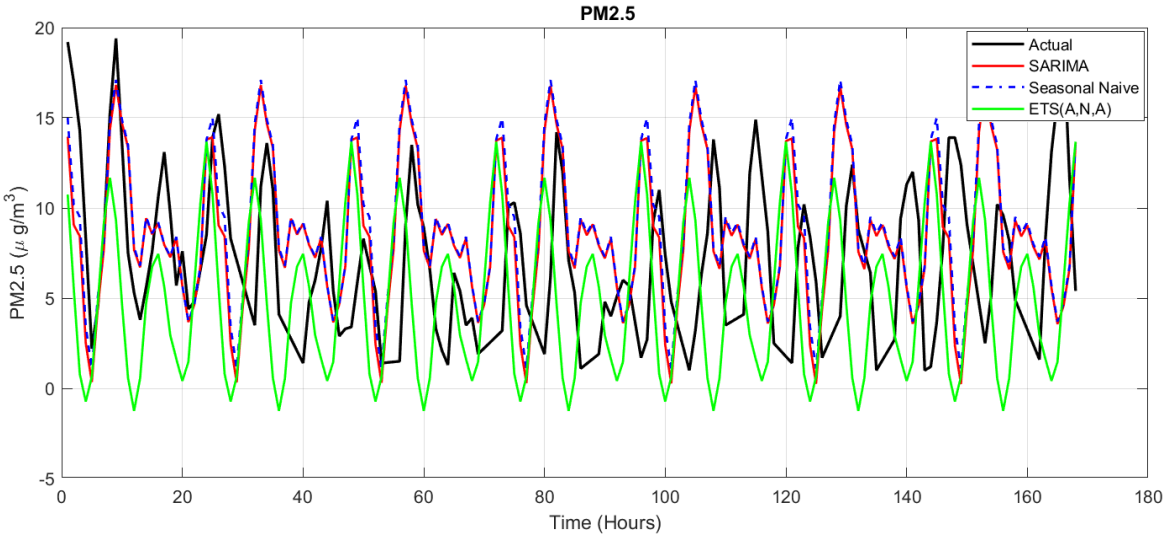

Figure 2. Sample of actual values and predicted values for PM2.5 (time-series methods).

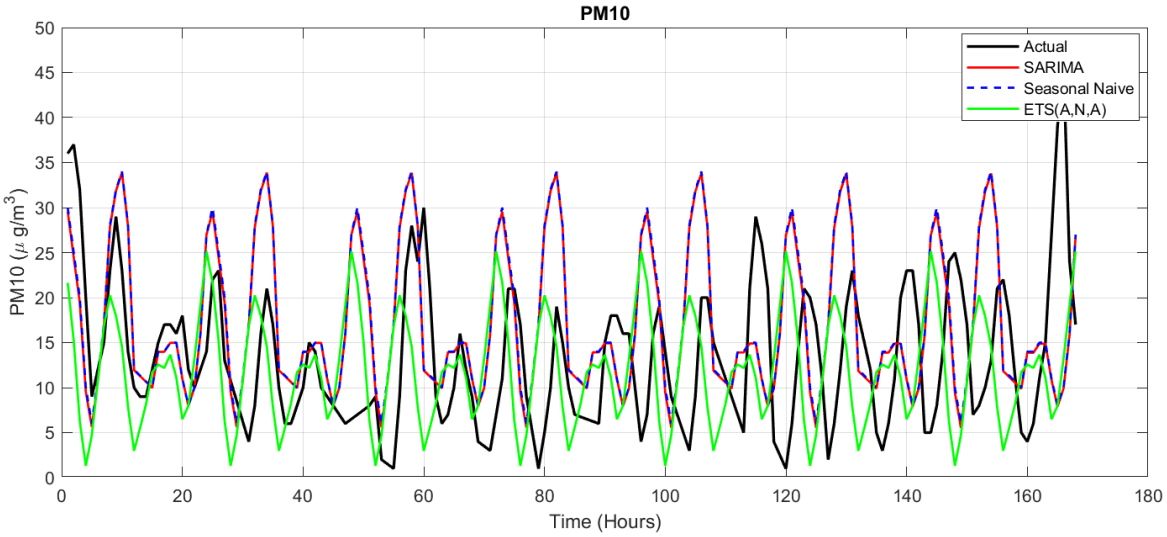

Figure 3. Sample of actual values and predicted values for PM10 (time-series methods).

Next, in addition to the graphical presentation, a comprehensive quantitative evaluation was conducted on the entire test dataset to verify the accuracy of the time-series models. The forecasting performance of the proposed methods was measured using MAE, RMSE, R^2, MASE, and sMAPE, as these metrics provide diverse perspectives on model efficiency. The performance results for PM2.5 and PM10 forecasting are summarized in Tables 3 and 4.

Table 3. Performance comparison of PM2.5 forecasting models (time-series methods).

PM2.5	MAE	RMSE	R^2	MASE	sMAPE
SARIMA	8.8332	12.526	-0.52285	3.3868	79.233
SeasonalNaive	8.096	11.403	-0.26211	3.1041	67.927
ETS_ANA	9.9874	13.524	-0.77513	3.8293	107.76

Table 4. Performance comparison of PM10 forecasting models (time-series methods).

PM10	MAE	RMSE	R^2	MASE	sMAPE
SARIMA	12.983	17.215	-0.60748	3.4679	62.294
SeasonalNaive	12.365	16.38	-0.45529	3.3029	57.421
ETS_ANA	14.106	18.589	-0.87439	3.7679	74.942

3.3. Forecasting performance of machine learning models

This section evaluates the forecasting performance of four efficient machine learning models: random forest, support vector regression, artificial neural network, and extreme gradient boosting. Similarly, the dataset was split into two parts: 80% for model training and 20% for testing, to reflect the models' forecasting capabilities on previously unseen data. Figures 4 and 5 illustrate the capability of each method in tracking the trends and seasonal patterns of the data. In particular, these figures show a comparison between the actual values and the forecasting results from all three models. Furthermore, to ensure clarity in visualizing the graph lines due to the large volume of data, a sample of 5% of the total test dataset is displayed.

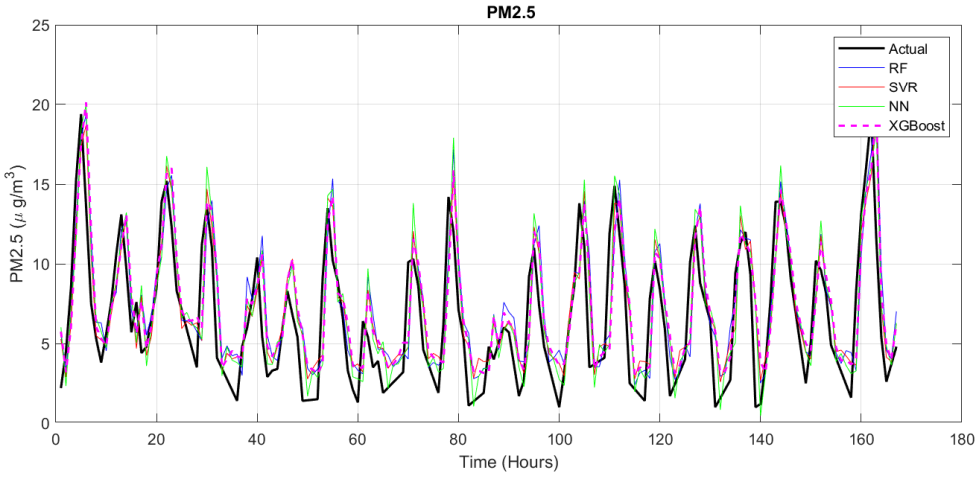

Figure 4. Sample of actual values and predicted values for PM2.5 (machine learning methods).

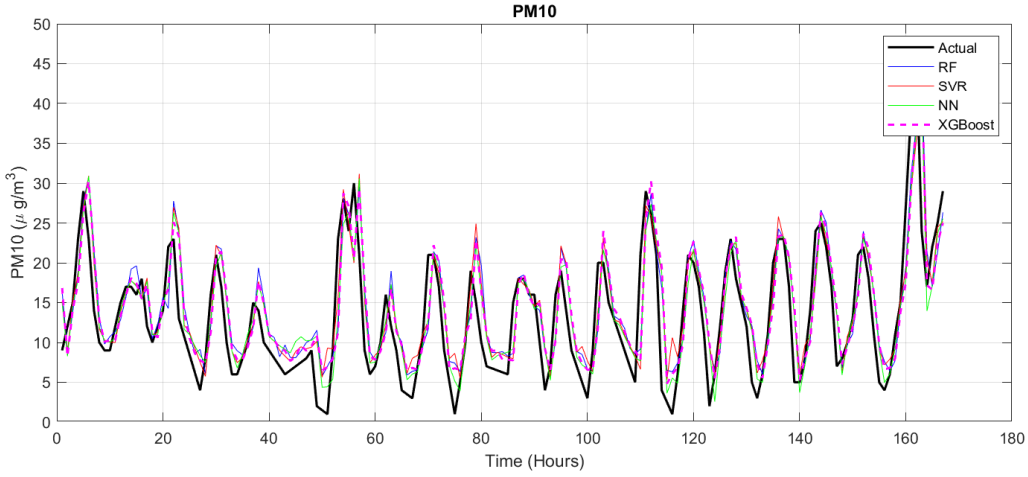

Figure 5. Sample of actual values and predicted values for PM10 (machine learning methods).

Next, to demonstrate the performance of the proposed machine learning models, forecasting efficiency was measured through a set of five standard statistical indicators, consistent with the previous tests. The performance results of the proposed methods are presented in Tables 5 and 6.

Table 5. Performance comparison of PM2.5 forecasting models (machine learning methods).

PM2.5	MAE	RMSE	R^2	MASE	sMAPE
RF	2.5967	3.5768	0.88155	0.99538	26.068
SVR	2.4899	3.4369	0.89064	0.95444	25.319
ANN	2.5154	3.4887	0.88732	0.96423	25.506
XGBoost	2.6321	3.6518	0.87654	1.0090	26.268

Table 6. Performance comparison of PM10 forecasting models (machine learning methods).

PM10	MAE	RMSE	R^2	MASE	sMAPE
RF	3.8977	5.3831	0.85085	1.0401	19.434
SVR	3.7340	5.1571	0.86312	0.99643	18.734
ANN	3.7511	5.1838	0.86169	1.0010	18.713
XGBoost	3.8832	5.4418	0.84758	1.0363	19.321

3.4. Comparative performance analysis between models

This section provides a comparative analysis to identify performance differences between traditional time-series models and machine learning models, leading to the selection of the most suitable algorithm for optimal air quality forecasting in Udon Thani province. From the forecasting results illustrated in Figures 2–5, a clear difference in tracking capability is observed between the two methodological groups. In Figures 2 and 3, which utilize time-series models, although they can capture the main trends and seasonal cycles, the predicted lines often exhibit a lag and fail to accurately respond to severe fluctuations during peak values. In contrast, Figures 4 and 5 demonstrate that machine learning models, particularly artificial neural networks, random forest, and extreme gradient boosting, are highly efficient in adapting to changes in actual data. The forecasting curves of the machine learning group are significantly more closely aligned with the actual observed values. To provide an overview of all test data, Figure 6 shows a time-series comparison for PM2.5 and PM10, displaying the results of the SVR and ETS models from the best-performing machine learning and time-series methods, evaluated on the entire test dataset of 3360 samples. The experimental results demonstrate that the machine learning method effectively captures pollutant fluctuations across both low and high concentration levels. In contrast, the time-series method performs satisfactorily only under low-concentration conditions, while failing to accurately track sharp spikes during extreme pollution events.

From the performance analysis of the proposed methods, as shown in Tables 3–6, which represent a quantitative evaluation of the entire test dataset to demonstrate the accuracy of the forecasting models, measurements were taken through MAE, RMSE, R^2, MASE, and sMAPE indicators to obtain multidimensional perspectives on model efficiency. From the evaluation results, it was found that the models in the time-series group exhibited relatively high error values, with MAE ranging from 8.096 to 9.987 for PM2.5 and from 12.365 to 14.106 for PM10. The low R^2 values observed in the time-series models stem from their reliance solely on historical data to predict hourly particulate concentrations, which exhibit high volatility. These models lack essential meteorological variables that serve as primary drivers for pollutant dispersion, limiting their ability to capture sudden, sharp fluctuations. Conversely, the machine learning models demonstrated improved performance, effectively highlighting their capability to extract complex, nonlinear patterns from historical data. Although the numerical differences in performance metrics between machine learning models (e.g., RF, SVR, ANN, XGBoost) are relatively small, indicating comparable capabilities, these models consistently demonstrate a clear hierarchy of performance across all evaluations. Based on this consistent trend, the support vector regression model was the most outstanding by providing the lowest MAE of only 2.4899 for PM2.5 and 3.7340 for PM10, representing an error reduction of approximately 69%–75% compared to the time-series group. Additionally, support vector regression achieved the highest positive R^2 and reduced the sMAPE to just 25.319% (PM2.5) and 18.734% (PM10), indicating higher stability and accuracy in handling volatile data. The study results, therefore, conclude that the support vector regression model is the most suitable and accurate tool for forecasting PM2.5 and PM10

concentrations. This model demonstrates high potential for effective application in early air quality warning systems, specifically at the air quality monitoring station in Udon Thani province, Thailand.

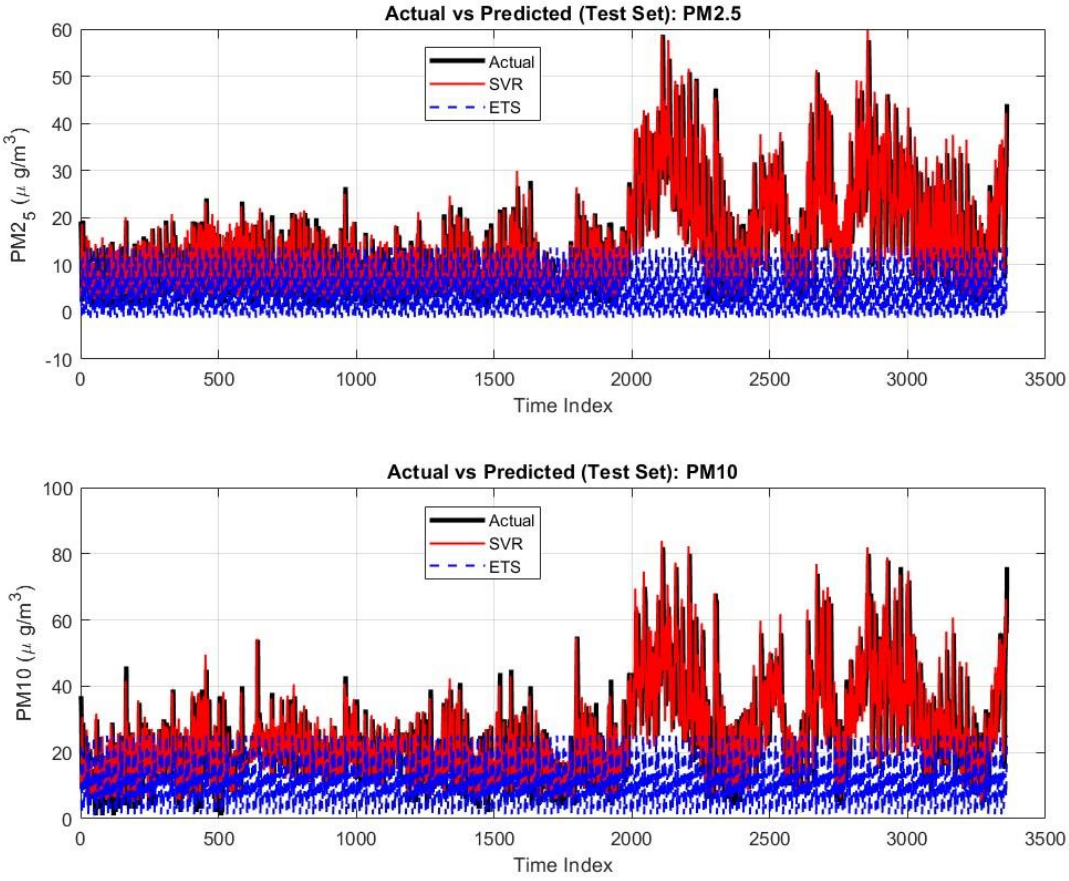

Figure 6. Comparison of actual and predicted values for PM2.5 and PM10 on all test datasets (SVR and ETS methods).

4. Conclusions

The research results show that this study achieved its objective of developing an efficient particulate matter forecasting system suitable for different areas**.** This was achieved by in-depth comparisons between traditional time-series methods and machine learning technologies to predict highly volatile air pollution in air quality monitoring stations in Udon Thani province, Thailand. Given that each area possesses distinct geographical characteristics, certain methods that perform well in one location may not be appropriate for another. From the performance measurement of the proposed methods, when considering the efficiency ranking based on the mean absolute error, it was found that for PM2.5 forecasting, the first ranked is support vector regression (2.4899), followed by artificial neural network (2.5154) and random forest (2.567). In contrast, the time-series group is ranked as seasonal naïve (8.096), seasonal autoregressive integrated moving average (8.8332), and Holt–Winters exponential smoothing (9.9874). For PM10 forecasting, the first rank remains support vector regression (3.7340), followed by artificial neural network (3.7511) and XGBoost (3.8832), while the time-series group is ranked as seasonal naïve (12.365), seasonal autoregressive integrated moving

average (12.983), and Holt–Winters exponential smoothing (14.106).

The knowledge gained from this study reveals that the support vector regression model is the most suitable method for forecasting PM2.5 and PM10 concentrations, providing more efficient results than other approaches. The support vector regression model in this research demonstrates its capability to learn nonlinear data patterns and accurately and stably respond to changes in hourly particulate matter levels under local environmental conditions. This serves as a vital foundation for enhancing the reliability of early air quality warning systems. Beyond the performance ranking of the models, the forecasting capability holds significant practical implications for air quality management, demonstrating that selecting a model aligned with local characteristics enables policymakers and environmental agencies to implement proactive interventions during critical seasons. However, we acknowledge the spatial limitations of this study, as it relies on data from a single monitoring station, which may not fully capture the pollution variability across different areas. Furthermore, while this study primarily focuses on the scope of a single-variable time-series framework, we recognize the decisive role of meteorological factors in pollutant dispersion. Regarding the temporal scope, while nearly two years of hourly data is consistent and sufficient for developing an accurate machine learning model in accordance with the objectives of this research, it represents a relatively short time series in a broader climatological context, which may limit the observation of long-term inter-annual variations. Nevertheless, the researchers plan to integrate key meteorological variables into a multivariate system, expand this framework to a multi-site regional scale covering the upper northeastern region of Thailand, and employ additional validation techniques to further enhance long-term forecasting accuracy and model stability.

Use of AI tools declaration

During the preparation of this work, the authors used Google Gemini and QuillBot in order to improve language phrasing, readability, and grammatical accuracy. After using these tools, the authors closely reviewed and edited the content as appropriate.

Acknowledgments

This research was supported by a research grant from Udon Thani Rajabhat University Research Fund, and Fundamental Fund: Grant 2567FF41SC08-EN5. The authors would like to express their sincere gratitude to the university for its generous financial support, which was instrumental in the successful completion of this study. Additionally, we extend our appreciation to the Air Quality Monitoring Station in Udon Thani, Thailand, for providing the essential data used in this performance analysis.

Conflict of interest

No potential conflict of interest was reported by the authors.

References

1. Leão MLP, Zhang L, da Silva Júnior FMR (2023) Effect of particulate matter (PM2.5 and PM10) on health indicators: Climate change scenarios in a Brazilian metropolis. *Environ Geochem Hlth* 45: 2229–2240. https://doi.org/10.1007/s10653-022-01331-8

2. World Health Organization, (2021) *WHO global air quality guidelines*: P*articulate matter* (*PM2.5 and PM10*), *ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*, Geneva: World Health Organization. Available from: https://www.who.int/publications/i/item/9789240034228.
3. Rochana K, Wongprachan R (2025) Statistical model for air quality forecasting: A case study of dust particles no larger than 2.5 microns (PM2.5) in Chiang Mai Province. *NKRAFA J Sci Technol* 21: 242–258.
4. Li X, Zhang Y, Wang J, et al. (2023) Long-term forecasting of PM2.5 and PM10 concentrations and analysis of influencing factors. *Sustainability* 16: 19. https://doi.org/10.3390/su16010019
5. Chelani AB, Devotta S (2006) Air quality forecasting using a hybrid autoregressive and nonlinear model. *Atmos Environ* 40: 1774–1780. https://doi.org/10.1016/j.atmosenv.2005.11.019
6. Gardner MW, Dorling SR (1998) Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmos Environ* 32: 2627–2636. https://doi.org/10.1016/S1352-2310(97)00447-0
7. Grivas G, Chaloulakou A (2006) Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece. *Atmo Environ* 40: 1216–1229. https://doi.org/10.1016/j.atmosenv.2005.10.036
8. Zhang Y, Bocquet M, Mallet V, et al. (2012) Real-time air quality forecasting, part I: History, techniques, and current status. *Atmos Environ* 60: 632–655. https://doi.org/10.1016/j.atmosenv.2012.06.031
9. Polsena T, Jitkongchuen D, Thusaranon P (2023) Integrated rearrange processing of hybrid model with weighted values for PM2.5 forecasting. *ICITEE* 213–216. https://doi.org/10.1109/icitee59582.2023.10317678
10. Chiang PW, Horng SJ (2021) Hybrid time-series framework for daily-based PM2.5 forecasting. *IEEE Access* 9: 104162–104176. https://doi.org/10.1109/ACCESS.2021.3099111
11. Chen D, Xu T, Li Y, et al. (2015) A hybrid approach to forecast air quality during high-PM concentration pollution period. *Aerosol Air Qual Res* 15: 1325–1337. https://doi.org/10.4209/AAQR.2014.10.0253
12. Zhang X, Rui X, Xia X, et al. (2015) *A hybrid model for short-term air pollutant concentration forecasting*, In: 2015 IEEE International Conference on Service Operations And Logistics, And Informatics (SOLI), IEEE, 171–175. https://doi.org/10.1109/SOLI.2015.7367614
13. Thanavanich T, Yaibuates M, Duangtang P, et al. (2022) Hybrid algorithms based on historical accuracy for forecasting particulate matter concentrations. *IAES Int J Artif Intell* 11: 1297–1305. https://doi.org/10.11591/ijai.v11.i4.pp1297-1305
14. Pollution Control Department, *Thailand air quality monitoring data*, Ministry of Natural Resources and Environment, Thailand, 2024. Available from: http://air4thai.pcd.go.th/.
15. Smith RL (1986) Time series analysis in acid rain modeling: Evaluation of filling missing values by linear interpolation. *Atmos Environ* 20: 1941–1943. https://doi.org/10.1016/0004-6981(86)90335-5
16. McElroy TS, Politis DN (2022) Optimal linear interpolation of multiple missing values. *Stat Infer Stoch Pro* 25: 471–483. https://doi.org/10.1007/s11203-022-09269-5
17. St-Aubin P, Agard B (2022) Precision and reliability of forecasts performance metrics. *Forecasting* 4: 882–903. https://doi.org/10.3390/forecast4040048
18. Quiroz-Flores JC, et al. (2023) Forecasting analysis using machine learning models and statistical evaluation metrics. *Int J Eng Trend Technol* 71: 39–45. https://doi.org/10.14445/22315381/IJETT-V71I2P205

19. Liu D (2024) The prediction and analysis of global climate change based on SARIMA. *Appl Comput Eng* 40: 268–273. https://doi.org/10.54254/2755-2721/40/20230665
20. Mahanta N, Talukdar R (2024) Forecasting of electricity consumption by seasonal autoregressive integrated moving average model in Assam, India. *Int J Energy Econ Policy* 14: 651–658. https://doi.org/10.32479/ijeep.16506
21. Jiang X, Xu L, Cui Y (2018) Seasonal model and its application in short-term forecasting. *Int Conf Appl Math* 194–196. https://doi.org/10.2991/AMMSA-18.2018.39
22. Li X, Petropoulos F, Kang Y (2023) Improving forecasting by subsampling seasonal time series. *Int J Prod Res* 61: 976–992. https://doi.org/10.1080/00207543.2021.2022800
23. Pan R (2010) *Holt–winters exponential smoothing*, Wiley Encyclopedia of Operations Research and Management Science. https://doi.org/10.1002/9780470400531.EORMS0385
24. Lima S, Gonçalves AM, Costa M (2019) Time series forecasting using Holt-Winters exponential smoothing: An application to economic data. *AIP Conf Proc* 2186: 090003. https://doi.org/10.1063/1.5137999
25. Alotaibi E, Nassif N (2024) Artificial intelligence in environmental monitoring: In-depth analysis. *DIAI* 4: 87. https://doi.org/10.1007/s44163-024-00198-1
26. Milutinović M (2024) *Machine learning in environmental monitoring*, Facta Universitatis, Series: Working and Living Environmental Protection, 21: 155–164. https://doi.org/10.22190/fuwlep241029014m
27. Hsieh WW (2025) *Machine learning in environmental and climate science*: *Overview and introduction*, Oxford Research Encyclopedia of Climate Science. https://doi.org/10.1093/acrefore/9780190228620.013.952
28. Shalu (2023) Environmental monitoring with machine learning. *EPRA Int J Multidiscip Res* 9: 208–212. https://doi.org/10.36713/epra13330
29. Peng N (2023) Application of machine learning techniques in environmental governance: A review. *Adv Eng Technol Res* 7: 528. https://doi.org/10.56028/aetr.7.1.528.2023
30. Liu D, Fan Z, Fu Q, et al. (2019) Random forest regression evaluation model of regional flood disaster resilience based on the whale optimization algorithm. *J Clean Prod* 250: 119468. https://doi.org/10.1016/j.jclepro.2019.119468
31. Rani KU, Venkataramana K (2025) Improved random forest regression for prediction. *Int J Sci Technol Eng* 13: 2084–2089. https://doi.org/10.22214/ijraset.2025.67722
32. Hernández N, Biscay RJ, Talavera I (2007) *Support vector regression methods for functional data*, In: Progress in Pattern Recognition, Image Analysis and Applications, New York: Springer, 564–573. https://doi.org/10.1007/978-3-540-76725-1_59
33. Abaszade M, Effati S (2018) Stochastic support vector regression with probabilistic constraints. *Appl Intell* 48: 243–256. https://doi.org/10.1007/S10489-017-0964-6
34. Sonu, Bhokal RP (2017) Study of artificial neural network. *Int J Math Trend Technol* 47: 253–259. https://doi.org/10.14445/22315373/IJMTT-V47P535
35. Rose A (2024) How do artificial neural networks work. *J Adv Sci Technol* 20: 172–177. https://doi.org/10.29070/ttrkmm98
36. Bergou EH, Diouane Y, Kungurtsev V (2020) Convergence and complexity analysis of a Levenberg-Marquardt algorithm for inverse problems. *J Optimiz Theory App* 185: 927–944. https://doi.org/10.1007/s10957-020-01666-1
37. Sheridan RP, Wang WM, Liaw A, et al. (2016) Extreme gradient boosting as a method for quantitative structure–Activity relationships. *J Chem Inf Model* 56: 2353–2360. https://doi.org/10.1021/acs.jcim.6b00591

38. 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, New York: Association for Computing Machinery, 785–794. https://doi.org/10.1145/2939672.2939785

AIMS Press