Songklanakarin J. Sci. Technol.
47 (2), 141–150, Mar. – Apr. 2025

http://www.sjst.psu.ac.th

Original Article

Predictive models for evaluating daily $PM_{2.5}$ concentrations using climate factors in the upper north of Thailand

Kuntalee Chaisee[1, 2] and Kamonrat Suphawan[1*]

[1] *Department of Statistics, Faculty of Science, Chiang Mai University, Mueang, Chiang Mai, 50200 Thailand*

[2] *Data Science Research Center, Faculty of Science, Chiang Mai University, Mueang, Chiang Mai, 50200 Thailand*

Received: 25 July 2024; Revised: 7 January 2025; Accepted: 28 March 2025

Abstract

This study evaluated the potential of statistical models to predict daily $PM_{2.5}$ concentrations in Upper Thailand using daily climate data on air pressure, temperature, humidity, rainfall, evaporation, wind speed, and direction, as predictors. Four statistical methods were employed: Multiple Linear Regression (MLR), Quantile Regression (QR), Generalized Additive Models (GAMs), and Support Vector Regression (SVR). Humidity emerged as the most influential climate factor on $PM_{2.5}$, especially in cool and hot seasons. SVR outperformed other models in prediction accuracy, while GAMs showed promise in specific provinces. Despite limitations indicated by R^2 values, this research demonstrates the potential of utilizing statistical modeling and accessible climate data for $PM_{2.5}$ prediction in regions lacking air quality monitoring equipment, but with access to real-time or short-term forecasted climate data.

Keywords: $PM_{2.5}$ concentrations, climate factors, correlation, predictive models, upper north Thailand

1. Introduction

Fine particulate matter ($PM_{2.5}$), an invisible yet insidious form of air pollution, poses a serious threat to human health, with exposure directly linked to increased risks of cardiovascular and respiratory diseases (Hayes *et al*., 2020; Liu *et al*., 2017; Pope, Coleman, Pond, & Burnett, 2020; Pun, Kazemiparkouhi, Manjourides & Suh, 2017; Ren *et al*., 2021; Slama *et al*., 2019). While air quality monitoring stations offer crucial real-time data on $PM_{2.5}$ concentrations, their geographic coverage, particularly in regions like northern Thailand, is often limited. Forecasting models present a promising solution for predicting $PM_{2.5}$ levels even in areas without extensive monitoring infrastructure. Current models typically rely on a combination of air quality data (e.g., sulfur dioxide, nitrogen dioxide) and meteorological data (Bensalam, 2024; Gulati *et al*., 2023; Sirithian & Thanatrakolsri, 2022; Zaman, Kanniah, Kaskaoutis & Latif, 2021). However, the availability of detailed air quality data can be a constraint, especially in regions like northern Thailand. In contrast, meteorological data is widely accessible, often in real-time or short-term forecasts, making it a valuable resource for timely $PM_{2.5}$ predictions.

This study focuses on the upper region of northern Thailand, which is particularly vulnerable to severe $PM_{2.5}$ episodes, especially during the hot season. We aim to evaluate the effectiveness of four distinct models for predicting daily $PM_{2.5}$ concentrations in this region using solely meteorological data.

Multiple linear regression (MLR) is broadly used to investigate the relationship between two or more variables. Due to its simplicity and interpretability in capturing the relationship between independent and dependent variables, MLR is often a good starting point for air pollution prediction. It has been successfully used in past studies to examine $PM_{2.5}$ concentrations (Amnuaylojaroen, 2022; Bekesiene, Meidute-Kavaliauskiene, & Vasiliauskiene, 2021; Kliengchuay *et al*., 2021; Lesar, & Filipčić, 2021). Another alternative type of regression analysis is quantile regression (QR), which is useful

*Corresponding author
Email address: kamonrat.s@cmu.ac.th

when the focus is on capturing relationships beyond just the mean values of the variables, such as the median or other quantile, providing a more meticulous understanding of $PM_{2.5}$ variations. We also consider the models for nonlinearity. Generalized additive models (GAMs) offer greater flexibility for modeling non-linear relationships and they have shown promise in previous air pollution research (Zeng, Jaffe, Qiao, Miao, & Tang, 2020). Finally, we apply support vector regression (SVR), which is another powerful tool for non-linear modeling. It has been successfully applied in air quality prediction studies (Chen, Yang, Du, & Huang, 2021; Mogollón-Sotelo *et al*., 2021; Weizhen *et al*., 2014; Zaman, Kanniah, Kaskaoutis & Latif, 2021).

By investigating these models, this study seeks to identify the most accurate and reliable approach for $PM_{2.5}$ prediction in the upper region of northern Thailand. The findings will not only advance our understanding of the relationship between climate and air pollution but also provide critical information for public health interventions and decision-making during periods of high $PM_{2.5}$ levels.

2. Materials and Methods

2.1 Data

This study analyzes the relationship between $PM_{2.5}$ and various climate factors in Thailand's upper north region. Daily average $PM_{2.5}$ data (micrograms per cubic meter, $\mu g/m^3$) were collected from the Pollution Control Department (Ministry of Natural Resources and Environment). Climate data, including air pressure (hPa), temperature (°C), relative humidity (%), rainfall (mm), evaporation (mm), wind speed (m/s), and wind direction (degrees), were obtained from the Northern Meteorological Center (Ministry of Digital Economy and Society).

The study focuses on eight provinces in the upper north: Chiang Mai, Chiang Rai, Lampang, Lamphun, Mae Hong Son, Nan, Phayao, and Phrae. Data from Uttaradit was excluded due to recent station installation, as shown in Figure 1. The timeframe spans from January 1st, 2018, to December 31st, 2022. Daily $PM_{2.5}$ and climate data were merged based on corresponding dates and locations. It's important to note that some stations had missing data in early 2018, resulting in variations in data points per province as shown in Table 1. The analysis considers seasonal variations defined by the Thai Meteorological Department: wet (June-October), cool (November-February), and hot (March-May).

Figure 2 illustrates the seasonal variations in $PM_{2.5}$ concentrations across the eight provinces throughout the year. It reveals that the hot season (red) exhibits a considerably higher $PM_{2.5}$ compared with the wet season (green). Notably, Chiang Mai, Chiang Rai, and Mae Hong Son experience peak $PM_{2.5}$ of over 300 $\mu g/m^3$ during the hot season, which substantially exceeds standard levels.

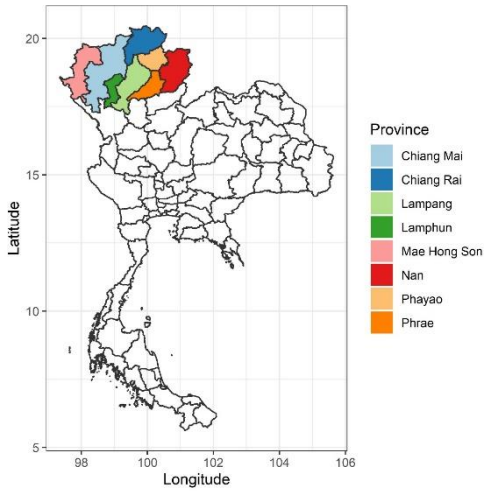

Figure 1. Map of Thailand in which the highlighted regions are the upper northern provinces

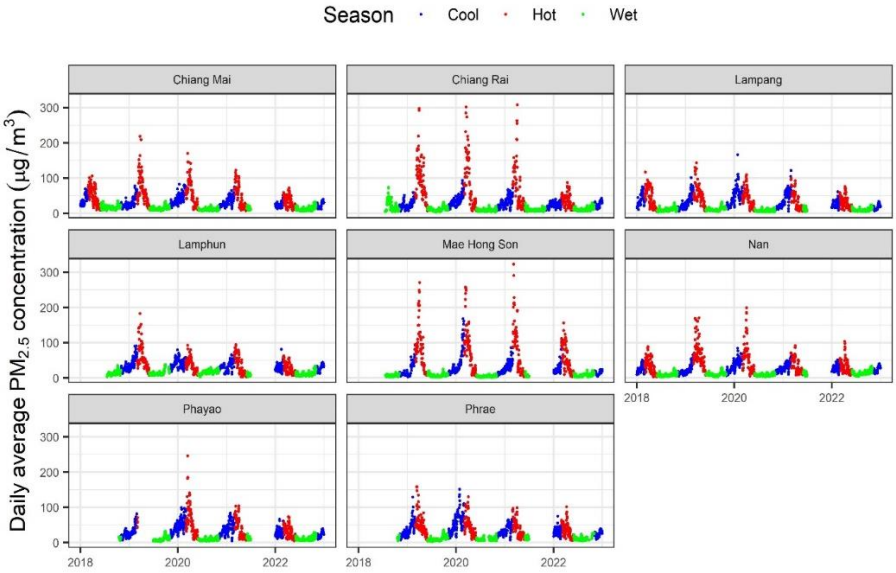

Figure 2. Time series plot of daily average $PM_{2.5}$ concentrations

Table 1 also presents the descriptive statistics for $PM_{2.5}$ concentrations across eight provinces, grouped by season. The mean and standard deviation of $PM_{2.5}$ were highest during the hot season, exceeding those of the wet and cool seasons. Air pollution is particularly severe during the hot or dry season when farmers routinely burn agricultural fields, resulting in trapping smoke and other pollutants close to the ground.

2.2 Methodology

2.2.1 Correlation analysis

Correlation analysis is used to examine a linear relationship between two variables: X and Y, and to assess the relationship. For a given dataset $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$, the sample correlation coefficient, r, is defined as

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

and n is the number of observations. The correlation coefficient defined in Equation (1) ranges from +1 to -1. A value close to +1 indicates a strong positive correlation, whereas a value near -1 indicates a strong negative correlation. A value around 0 indicates a weak or negligible relationship between the two variables.

2.2.2 Multiple linear regression

Multiple linear regression (MLR) attempts to predict a dependent variable, Y, by assuming a linear relationship with independent variables, $X_1, X_2, \dots, X_k$. The model can be expressed as

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_j x_{ij} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, \dots, n, \; j = 1, \dots, k \tag{2}$$

where x_{ij} is the value of i^{th} observation of j^{th} independent variable, y_i is the value of i^{th} observation of the dependent variable, $\beta_0, \beta_1, \dots, \beta_k$ are regression coefficient parameters, ϵ is random error assumed to be $\epsilon_i \sim N(0, \sigma^2)$ and n is the number of observations. The regression parameters can be estimated by the least squares and maximum likelihood methods, yielding the fitted equation.

2.2.3 Quantile regression

We can use quantile regression to estimate the τ quantile of the dependent variable, called quantile regression (QR). The τ value shows the quantile and its value is between 0 and 1. The 0.5 quantile or $\tau = 0.5$ is the median regression meaning that 50% of the data are less than the value of the median. Similarly, the 0.25 and 0.75 quantiles are values such that 25% and 75% of the data are smaller than these values, respectively. The regression model for τ quantile is defined as

$$Q_r(y_i) = \beta_0(\tau) + \beta_1(\tau) x_{i1} + \cdots + \beta_k(\tau) x_{ik}, \tag{3}$$

where $Q_r(.)$ is the estimated quantile point for the τ quantile. The $\beta_0(\tau), \beta_1(\tau), \dots, \beta_k(\tau)$ are regression coefficient parameters for τ quantile regression and are estimated by solving

$$min \sum_{i=1}^{n} \rho_\tau \left(y_i - \beta_0(\tau) - \sum_{j=1}^{k} \beta_j(\tau) x_{ij}\right), \tag{4}$$

where $\rho_\tau(r) = \tau \, max(r, 0) + (1 - \tau) \, max(-r, 0)$ is referred to as a check function (Koenker, & Hallock, 2001).

Table 1. Descriptive statistics of PM2.5 concentrations

Province	Start date	Season	Number of observations	Mean	SD	Min	Max
Chiang Mai	Jan 1, 2018	Cool	537	32.05	14.53	12	83
		Hot	457	49.74	34.38	8	219
		Wet	633	13.81	5.27	7	35
Chiang Rai	Jul 18, 2018	Cool	539	27.94	13.79	7	94
		Hot	366	68.94	60.38	7	308
		Wet	691	12.27	7.89	4	74
Lampang	Jan 1, 2018	Cool	540	33.53	21.49	6	167
		Hot	460	40.42	26.23	6	143
		Wet	640	10.10	4.51	4	33
Lamphun	Jul 20, 2018	Cool	468	35.38	14.13	11	91
		Hot	360	39.55	27.60	3	183
		Wet	538	13.80	6.27	3	38
Mae Hong Son	Jul 21, 2018	Cool	478	26.27	23.15	4	168
		Hot	355	69.08	60.68	4	323
		Wet	588	6.97	3.78	2	25
Nan	Jan 1, 2018	Cool	538	27.26	13.44	6	85
		Hot	453	44.22	33.56	4	200
		Wet	636	10.73	5.08	4	34
Phayao	Oct 17, 2018	Cool	469	34.50	18.52	5	99
		Hot	275	42.48	33.83	5	246
		Wet	456	10.12	5.14	3	32
Phrae	Oct 17, 2018	Cool	475	37.65	22.18	8	151
		Hot	339	41.64	26.57	6	158
		Wet	378	11.01	5.38	3	36

2.2.4 Generalized additive model

A generalized additive model (GAM) eases the assumption of normality and linearity between dependent and independent variables required in the MLR. GAM assumes that the mean of the dependent variable depends on independent variables through a non-linear function. It uses smoothing techniques to model the shape of a relationship which is not entitled to take a particular form such as linear or exponential.

The GAM model is defined as

$$y_i = s_1(x_{i1}) + s_2(x_{i2}) + \cdots + s_k(x_{ik}), \quad i = 1, \ldots, n, \tag{5}$$

where $s_j()$ is a smoothing function which corresponds to an associated independent variable x_j for $j = 1, \ldots, k$. The $s_j()$ function or smoothing terms are spline functions of a single independent variable with smoothing parameters (Binder & Tutz, 2008).

2.2.5 Support vector regression

Support vector regression (SVR) is a machine learning technique used for regression tasks. It is useful when dealing with non-linear relationships between independent and dependen variables. For a given dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $x_i = (x_{i1}, \ldots, x_{ik})$ is a vector of independent variables for ith observation. The SVR creates a hyperplane

$$y_i = f(x_i) = w^T x_i + b, \tag{6}$$

where w is a coefficient vector and b is an intercept. The two boundary lines are constructed from the hyperplane with margin $\pm\varepsilon$. The distances between the data points outside the boundary lines and the boundary lines are denoted by ξ and ξ^*. The SVR aims to find a function $f(x)$ that minimizes $w^T w$ while having a maximum deviation of ε from the actual targets for all the data

$$\textit{minimize} \ \ \frac{1}{2} w^T w + C \sum_{i=1}^{n} (\xi_i, \xi_i^*), \tag{7}$$

$$\textit{constraints} \ \ y_i - w^T x_i - b \leq \varepsilon + \xi_i, \tag{8}$$

$$w^T x_i + b - y_i \leq \varepsilon + \xi_i^*, \tag{9}$$

$$\xi_i, \xi_i^* \geq 0 \tag{10}$$

(Awad, & Khanna, 2015).

2.2.6 Packages and programming

This work utilizes R for statistical analysis, employing various model fitting techniques. The function *lm* from base R is used to fit a linear model and estimate the coefficients for the MLR. Unlike traditional regression, we use the *rq* function from the *quantreg* package to perform QR analysis for various quantiles of the dependent variable, providing a more comprehensive picture. We employ the *gam* function to fit a GAM, allowing for the capture of non-linear relationships using a smoothing function. The *svm* function from the *e1071* package implements SVR, a kernel-based method suitable for handling non-linear relationships and potentially high-dimensional data depending on the hyperplane used in the analysis.

3. Results and Discussion

This study employed models; MLR, QR with $\tau = 0.5$ and $\tau = 0.75$ denoted QR0.5 and QR0.75, respectively, GAM and SVR are employed to predict the daily $PM_{2.5}$ concentrations in eight provinces across three seasons in Upper Thailand. Daily climate data on air pressure, temperature, humidity, rainfall, evaporation, wind speed, and wind direction served as independent variables or inputs. A total of 120 models (5 models x 8 provinces x 3 seasons) were generated. Prior to model interpretation, multicollinearity among climate variables was assessed as it can lead to unreliable coefficient estimates and significance levels, especially when the independent variables are highly correlated. The Variance Inflation Factor (VIF) was calculated for each climate variable, with a VIF > 5 indicating a potential issue. In this study, the maximum VIF observed was 2.5, suggesting no substantial multicollinearity among the climate variables included in our models.

The results can be divided into two main subsections. In Subsection 3.1, we show and discuss the relationship between $PM_{2.5}$ and climate data using correlation coefficients. In Subsection 3.2, we assess the performance of models across the different provinces and seasons. To evaluate the predictive performance, the root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) are calculated to assess the model's accuracy. The coefficient of determination (R^2) is used to evaluate the goodness of fit.

3.1 Correlation

Studies have revealed that climate factors can influence $PM_{2.5}$ concentrations, with both positive and negative correlations observed. These effects can exhibit substantial variation depending on the region and season, making it a necessity to incorporate these factors into any analysis.

As an initial step, scatterplots can be used to visualize the relationships between climate factors and $PM_{2.5}$. For instance, Figures 3 and 4 illustrate the relationships between humidity and evaporation, respectively, with humidity generally showing a positive correlation and evaporation showing a negative correlation. In addition, during the hot season, $PM_{2.5}$ concentrations appear more dispersed compared to the cool and wet seasons, likely due to the wider range of $PM_{2.5}$ values observed during this period. However, the relationships between climate variables and $PM_{2.5}$ seem to be non-linear, with no clear trend across all three seasons.

To explore the relationships among all factors comprehensively, we present the coefficients of correlation and significance of the test in Table 2. The relationship between $PM_{2.5}$ and climate variables differs across seasons. The wet season shows the weakest relationship, while the cool and hot seasons show stronger relationships. Relative humidity is the most strongly correlated climatic factor with $PM_{2.5}$ levels during both the cool and hot seasons, exhibiting correlation coefficients of -0.55 and -0.51, respectively. This suggests an inverse relationship, where higher relative humidity corresponds to lower $PM_{2.5}$ concentrations during these seasons. However, relative humidity shows no noticeable correlation with $PM_{2.5}$ levels during the wet season, suggesting their relationship is minimal under such climatic conditions. Furthermore, evaporation emerges as the second most

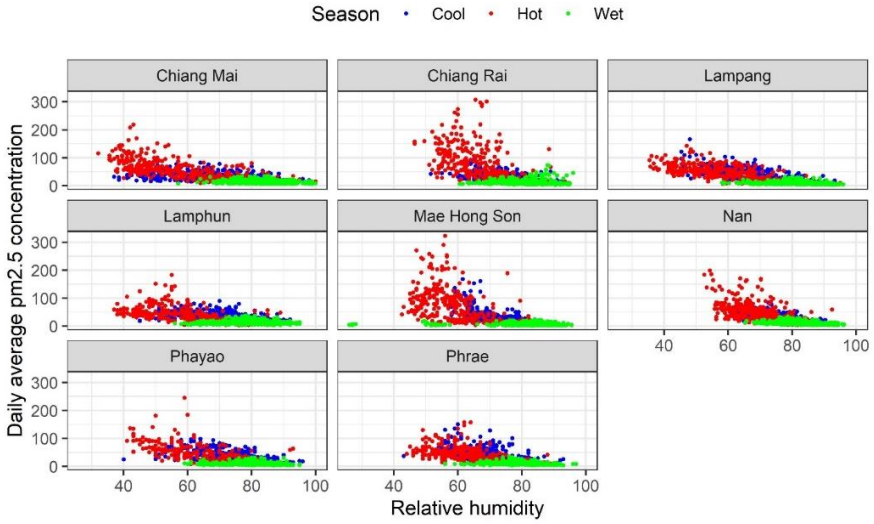

Figure 3. Scatterplots of $PM_{2.5}$ concentrations and relative humidity

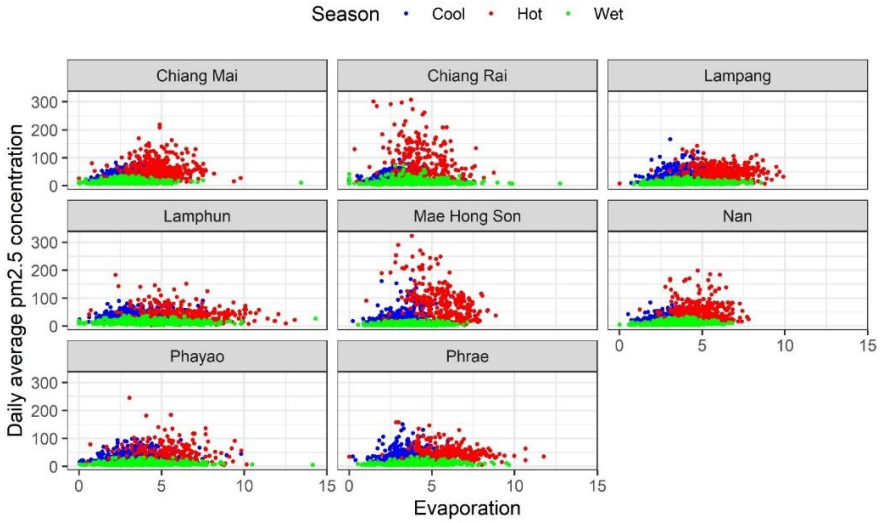

Figure 4. Scatterplots of $PM_{2.5}$ concentrations and evaporation

correlated variable with $PM_{2.5}$ in the cool season, with a correlation value of 0.27, but shows no discernible relationship during the hot and wet seasons. Conversely, rainfall amount is the second most correlated variable with $PM_{2.5}$ in the hot seasons with a correlation coefficient of -0.23, yet exhibits no appreciable association in the cool and wet seasons. Moreover, air pressure and temperature demonstrate notable correlations with $PM_{2.5}$ during the wet season, with values of 0.28 and 0.13, respectively. Finally, wind speed and wind direction exhibit relatively weak relationships with $PM_{2.5}$ across all three seasons.

When considering seasonal variations, we observe that the correlations between $PM_{2.5}$ concentrations and climate variables differ across seasons. During the cool season, relative humidity and evaporation emerge as the most influential climatic variables related to $PM_{2.5}$. Relative humidity exhibits a moderate negative correlation, while evaporation shows a weak positive correlation. These results suggest that low relative humidity and high evaporation rates contribute to elevated $PM_{2.5}$ during the cool season. In the hot season, relative humidity and rainfall amount are the most strongly correlated variables associated with $PM_{2.5}$, both displaying negative correlations. This implies that low relative humidity and low precipitation levels lead to increased $PM_{2.5}$ concentrations during this period.

Conversely, in the wet season, air pressure stands out as the primary climatic variable related to $PM_{2.5}$, exhibiting a moderate positive correlation. Consequently, higher air pressure values correspond to higher $PM_{2.5}$ levels. It is noteworthy that the wet season is generally considered the off-season for $PM_{2.5}$, characterized by the lowest concentrations and the least variation compared to the cool and hot seasons. Hence, $PM_{2.5}$ during the wet season are less influenced by climatic factors, resulting in no or weak relationships between most climatic variables and $PM_{2.5}$ levels. Therefore, it is recommended to model $PM_{2.5}$ separately for each season, as the correlations with climate variables vary across different times of the year.

Table 2 also presents the correlations among climate variables. It appears that some climate variables are not related, while others seem to have a relationship, but it differs in different seasons. For instance, air pressure and temperature are correlated in the cool and hot seasons, but not in the wet season. On the other hand, temperature and humidity are correlated in the hot and wet seasons, but not in the cool season, which is similar to the relationship between temperature and rain amount in the cool season. However, humidity and evaporation exhibit a relatively strong relationship in every season.

3.2 Predictive performance

To evaluate the predictive performance of the five models, we analyze the agreement between predicted and observed $PM_{2.5}$ concentrations using the test dataset. Figures 5-7 display the plots between predicted and observed $PM_{2.5}$ for the three seasons. It is important to highlight that the QR0.75 models specifically forecast the third quantile of $PM_{2.5}$. Consequently, their predicted values tend to be higher compared to those generated by other models.

In Figure 6, during the hot season, there is notable agreement between model predictions and observed values, as evidenced by the scattering of dots along the diagonal for all provinces. This observation aligns with the average R^2 values presented in Table 3, where a higher R^2 value, closer to 1, signifies a superior fit for the model. Essentially, this suggests that climate variables account for a larger portion of the variability in $PM_{2.5}$ concentrations. Moreover, the range of R^2 values across all models spans from 0.32 to 0.69 with SVR demonstrating the highest R^2 values ranging from 0.54 to 0.69, positioning it as the top-performing model. However, the other models also exhibit relatively strong performance. Moreover, the predictions for the cool season are similar to those of the hot season, but with slightly lower degrees of agreement. The R^2 for the SVR model ranges from 0.362 to 0.634, which is the highest among the models evaluated.

In contrast to the hot and cool seasons, the wet season exhibits the least agreement between model predictions and observed values, as shown in Figure 7. Here, the dots scatter along the diagonal primarily for a low range of $PM_{2.5}$. However, the models tend to underestimate the high $PM_{2.5}$ level. This discrepancy suggests a challenge in accurately predicting $PM_{2.5}$ levels during the wet season.

Overall, the prediction results in three seasons reveal that the predicted values for high $PM_{2.5}$ concentration regions are consistently lower than the observed values across all models. This suggests that our models consistently underestimate high $PM_{2.5}$ values.

To evaluate the predictive errors, Table 4 presents the root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) scores for assessing the predictive performance of the models. Smaller values for these metrics indicate better model performance. During the cool season, the SVR generally outperforms other models across most provinces, except for Phayao. In Phayao, the RMSE of GAM is lower than that of SVR. In the hot season, the assessments indicate that SVR performs better than other models in most provinces, except for Mae Hon Son, Nan, and Phayao. In these provinces, the RMSEs of GAM are lower than those of SVR. In the wet season, considering RMSEs, GAM and SVR exhibit similar performance, both surpassing MLR, QR0.5, and QR0.75. However, MAE and MAPE consistently show that SVR outperforms other models across all provinces.

Table 2. Correlation coefficients and significance of the correlation test

Cool	$PM_{2.5}$	AvgPress	AvgTemp	AvgHumid	Rainamnt	Evapor	WindSpeed	WindDirection
$PM_{2.5}$	1.00							
AvgPress	-0.15*	1.00						
AvgTemp	-0.01	-0.43*	1.00					
AvgHumid	-0.55*	0.18*	0.02	1.00				
Rainamnt	-0.09*	-0.06*	0.01	0.23*	1.00			
Evapor	0.27*	-0.23*	0.31*	-0.54*	-0.06*	1.00		
WindSpeed	0.07*	-0.11*	-0.05*	-0.32*	0.14*	0.25*	1.00	
WindDirection	0.11*	-0.15*	0.06*	-0.14*	0.01	0.12*	0.03	1.00
Hot	$PM_{2.5}$	AvgPress	AvgTemp	AvgHumid	Rainamnt	Evapor	WindSpeed	WindDirection
$PM_{2.5}$	1.00							
AvgPress	0.09*	1.00						
AvgTemp	-0.05*	-0.49*	1.00					
AvgHumid	-0.51*	0.13*	-0.43*	1.00				
Rainamnt	-0.23*	0.06*	-0.20*	0.39*	1.00			
Evapor	-0.05*	-0.30*	0.65*	-0.47*	-0.04*	1.00		
WindSpeed	-0.20*	-0.14*	0.05*	0.04*	0.31*	0.23*	1.00	
WindDirection	-0.08*	-0.15*	0.08*	-0.09*	-0.05*	0.08*	0.16*	1.00
Wet	$PM_{2.5}$	AvgPress	AvgTemp	AvgHumid	Rainamnt	Evapor	WindSpeed	WindDirection
$PM_{2.5}$	1.00							
AvgPress	0.28*	1.00						
AvgTemp	-0.13*	-0.23*	1.00					
AvgHumid	-0.10*	0.07*	-0.64*	1.00				
Rainamnt	-0.13*	-0.15*	-0.19*	0.36*	1.00			
Evapor	0.03*	-0.10*	0.55*	-0.54*	-0.10*	1.00		
WindSpeed	0.01	-0.22*	-0.14*	-0.05*	0.12*	0.13*	1.00	
WindDirection	-0.02	-0.26*	0.03*	-0.08*	-0.01	0.07*	0.19*	1.00

*means significance at level 0.05

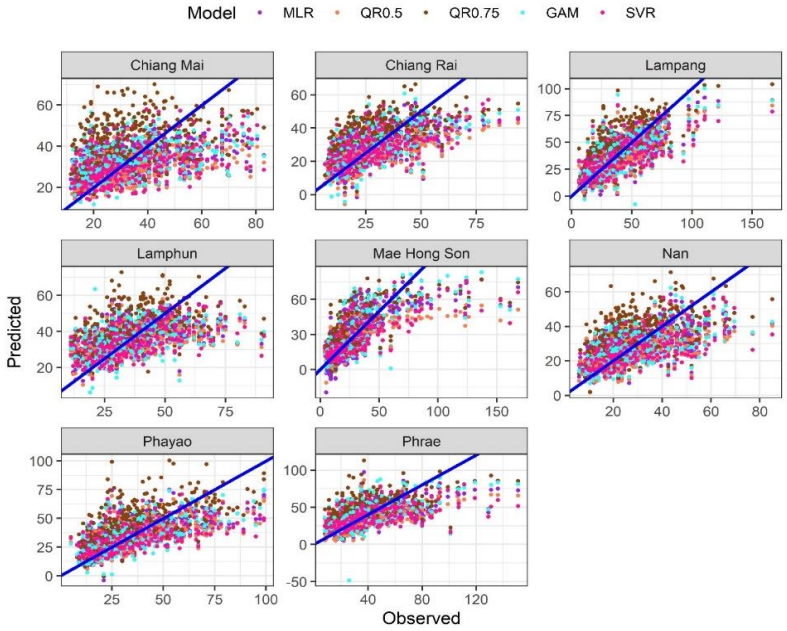

Figure 5. Predicted and observed $PM_{2.5}$ concentrations in the test set for the cool season

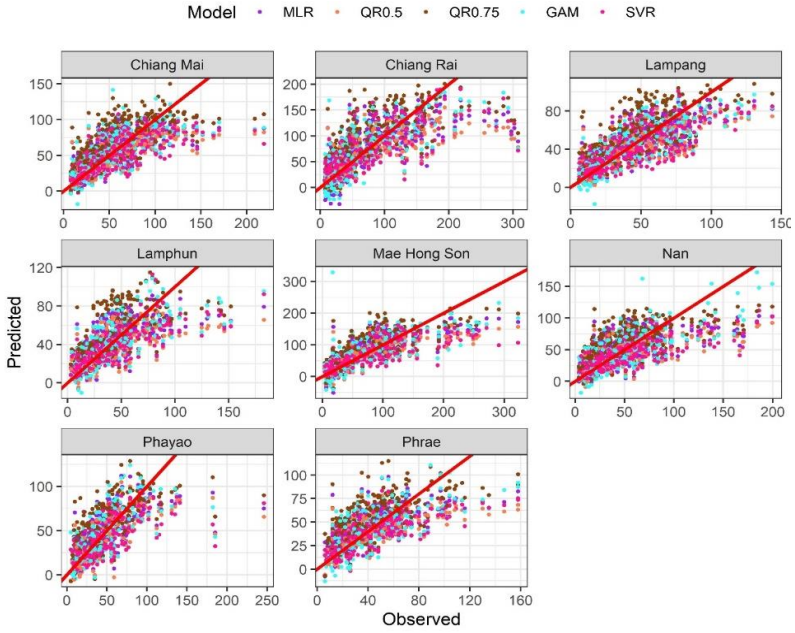

Figure 6. Predicted and observed $PM_{2.5}$ concentrations in the test set for the hot season

In summary, SVR models consistently produce the smallest errors (RMSE, MAE, and MAPE) across all three seasons, while GAM may occasionally outperform SVR. Moreover, the ranges of RMSE, MAE, and MAPE in the hot season are higher than in the cool and wet seasons, in all provinces.

4. Conclusions

In this study, we used climate data, including air pressure, temperature, relative humidity, evaporation, rainfall, wind speed, and wind direction, as predictors to forecast $PM_{2.5}$ concentrations for eight provinces in Upper Northern Thailand from 2018 to 2022. The analysis accounted for seasonal variations (cool, hot, and wet seasons) and employed multiple linear regression (MLR), quantile regression, generalized additive models (GAM), and support vector regression (SVR) to model and predict $PM_{2.5}$ levels. The evaluation metrics RMSE, MAE, MAPE, and R^2 were used to assess model performance.

Our correlation analysis revealed that the relationship between $PM_{2.5}$ and climatic variables varied across seasons,

Table 3. Average R^2 in 5-fold cross validation

Season	Province	MLR	QR0.5	QR0.75	GAM	SVR
Cool						
	Chiang Mai	0.199	0.156	-0.158	0.224	0.362
	Chiang Rai	0.360	0.329	0.141	0.418	0.485
	Lampang	0.589	0.574	0.437	0.626	0.634
	Lamphun	0.255	0.244	0.039	0.282	0.385
	Mae Hong Son	0.526	0.461	0.455	0.574	0.592
	Nan	0.370	0.337	0.108	0.406	0.457
	Phayao	0.430	0.413	0.157	0.474	0.452
	Phrae	0.371	0.326	0.186	0.422	0.454
Hot						
	Chiang Mai	0.580	0.566	0.448	0.587	0.601
	Chiang Rai	0.538	0.506	0.447	0.597	0.628
	Lampang	0.631	0.619	0.496	0.677	0.689
	Lamphun	0.484	0.461	0.319	0.564	0.592
	Mae Hong Son	0.653	0.630	0.571	0.669	0.664
	Nan	0.498	0.473	0.389	0.606	0.544
	Phayao	0.546	0.533	0.408	0.584	0.558
	Phrae	0.498	0.462	0.337	0.559	0.560
Wet						
	Chiang Mai	0.194	0.167	-0.021	0.280	0.336
	Chiang Rai	0.026	-0.084	-0.042	0.098	0.066
	Lampang	0.243	0.200	0.117	0.400	0.380
	Lamphun	0.273	0.240	-0.079	0.383	0.383
	Mae Hong Son	0.121	0.053	-0.111	0.169	0.233
	Nan	0.347	0.307	0.101	0.361	0.418
	Phayao	0.333	0.273	0.101	0.469	0.467
	Phrae	0.339	0.295	0.163	0.435	0.409

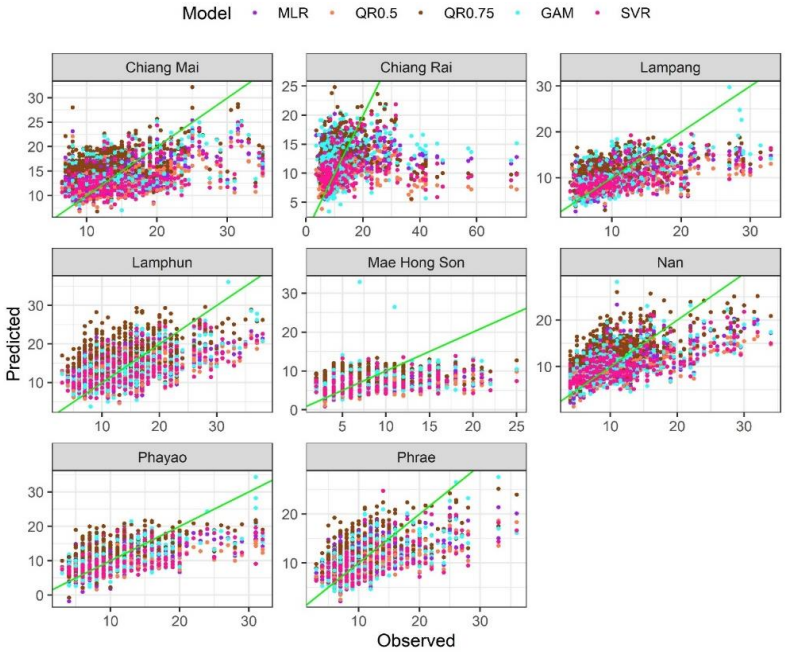

Figure 7. Predicted and observed $PM_{2.5}$ concentrations in the test set for the wet season

aligning with previous research demonstrating seasonal patterns in $PM_{2.5}$ levels and meteorological conditions (Saiohai *et al*., 2023). Relative humidity was the most strongly correlated factor, exhibiting a negative relationship with $PM_{2.5}$, particularly during the cool and hot seasons. This finding aligns with previous studies that reported a strong inverse relationship between relative humidity and $PM_{2.5}$ concentrations (Amnuaylojaroen, 2022; Sirithian & Thanatrakolsri, 2022). Furthermore, we observed a positive correlation between $PM_{2.5}$ and temperature throughout the study period.

Table 4. Average RMSE, MAE, and MAPE in 5-fold cross validation

Province	Model	Cool			Hot			Wet		
		RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
Chiang Mai										
	MLR	12.942	9.927	34.138	22.089	15.626	40.209	4.698	3.476	26.411
	QR0.5	13.298	9.508	29.344	22.480	15.305	36.272	4.782	3.382	24.074
	QR0.75	15.493	12.406	48.279	25.138	18.973	54.818	5.261	4.276	36.152
	GAM	12.715	9.785	34.049	21.837	15.338	38.975	4.424	3.292	24.882
	SVR	11.515	8.155	26.075	21.493	13.948	30.955	4.250	2.922	20.512
Chiang Rai										
	MLR	10.978	8.408	35.399	40.759	28.565	67.587	7.714	4.752	40.957
	QR0.5	11.245	8.232	32.232	42.202	27.246	54.069	8.145	3.995	26.421
	QR0.75	12.731	10.326	49.604	44.503	33.365	87.653	7.987	4.862	42.512
	GAM	10.468	8.024	33.787	38.091	26.505	65.083	7.412	4.653	40.199
	SVR	9.836	7.128	28.698	36.535	22.713	44.868	7.566	3.631	24.327
Lampang										
	MLR	13.682	10.241	36.974	15.837	11.727	39.561	3.892	2.762	28.676
	QR0.5	13.949	10.068	33.660	16.111	11.441	35.388	4.008	2.668	25.486
	QR0.75	15.987	12.581	51.451	18.320	14.162	53.785	4.187	3.276	37.561
	GAM	13.029	9.389	33.823	14.820	11.147	37.773	3.458	2.496	25.640
	SVR	12.867	8.588	28.211	14.471	10.145	31.557	3.523	2.280	21.258
Lamphun										
	MLR	12.099	8.878	28.799	19.600	13.836	49.891	5.293	4.206	36.689
	QR0.5	12.192	8.735	27.237	20.114	13.231	42.092	5.420	4.173	34.721
	QR0.75	13.672	10.709	38.150	22.345	16.850	67.044	6.404	5.201	52.607
	GAM	11.848	8.795	28.750	17.971	12.808	45.561	4.862	3.854	33.569
	SVR	10.961	7.596	24.331	17.431	11.026	36.576	4.882	3.665	30.219
Mae Hong Son										
	MLR	15.847	10.435	50.459	35.324	25.575	71.474	3.534	2.739	46.975
	QR0.5	16.963	9.522	38.746	36.606	24.691	60.352	3.669	2.651	39.584
	QR0.75	16.992	12.012	62.875	39.143	30.046	91.782	3.967	3.320	66.724
	GAM	15.011	9.561	43.167	34.144	24.131	71.600	3.404	2.569	43.708
	SVR	14.657	7.999	33.090	34.798	21.226	46.937	3.295	2.290	34.457
Nan										
	MLR	10.620	8.298	35.311	23.714	16.450	50.136	4.048	3.080	31.252
	QR0.5	10.897	8.154	31.999	24.342	15.907	43.558	4.184	3.003	28.194
	QR0.75	12.594	10.241	49.123	26.014	19.971	70.075	4.696	3.847	44.572
	GAM	10.302	7.980	33.723	20.816	15.358	50.673	4.007	3.021	30.475
Phayao	SVR	9.828	6.828	26.976	22.569	14.268	39.103	3.817	2.727	25.317
	MLR	13.851	10.284	35.010	22.105	15.155	50.239	4.161	3.128	33.884
	QR0.5	14.056	10.149	32.424	22.479	14.762	44.214	4.350	3.011	29.397
	QR0.75	16.791	13.096	51.786	24.792	18.335	72.380	4.826	3.940	48.872
	GAM	13.266	9.945	34.278	21.136	14.489	46.778	3.709	2.791	30.166
	SVR	13.501	9.456	30.211	21.827	13.153	37.035	3.721	2.426	23.075
Phrae										
	MLR	17.278	12.666	39.179	18.710	13.818	44.651	4.350	3.315	33.258
	QR0.5	17.945	12.163	33.513	19.378	13.326	38.152	4.500	3.204	29.675
	QR0.75	19.511	15.565	55.429	21.455	16.957	63.527	4.886	3.967	45.258
	GAM	16.423	11.938	36.909	17.527	13.058	41.881	4.013	3.055	30.730
	SVR	16.135	10.283	27.865	17.486	11.586	31.656	4.099	2.791	25.360

The performance of all models, based on R^2, in the wet season was poor compared with the cool and hot seasons. Although all models performed poorly in the wet season, it may not be a priority to overcome this issue because in Thailand, $PM_{2.5}$ concentrations in are generally low and mostly at safe levels during the wet season, while it is more severe in the cool and hot seasons. SVR consistently outperformed other models in terms of predictive accuracy, and GAM also showed promising results in some provinces, while MLR and QR were inferior. This indicates that the non-linear approaches offered by SVR and GAM are more suitable than linear models.

Acknowledgements

This research is supported by the Department of Statistics, Faculty of Science, Chiang Mai University.

References

Amnuaylojaroen, T. (2022). Prediction of PM2.5 in an urban area of northern Thailand using multivariate linear regression model. *Advances in Meteorology*, *2022*(1), 3190484.

Awad, M., & Khanna, R. (2015). *Efficient learning machines*: *Theories, concepts, and applications for engineers and system designers*. Berkeley, CA: Apress. Retrieved from https://doi.org/10.1007/978-1-4302-5990-9

Bekesiene, S., Meidute-Kavaliauskiene, I., & Vasiliauskiene, V. (2021). Accurate prediction of concentration changes in ozone as an air pollutant by multiple linear regression and artificial neural networks. *Mathematics*, *9*(4), 356.

Bensalam, I., Saelim, R., Samoh, A., Kongbok, N., Toheng, I., & Musikasuwan, S. (2024). An investigation on the influence of meteorological factors on PM2.5 concentration: Towards predictive models for Songkhla, Thailand. *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems* (*ICETSIS*).

Binder, H., & Tutz, G. (2008). A comparison of methods for the fitting of generalized additive models. *Statistics and Computing*, *18*, 87-99.

Chen, N., Yang, M., Du, W., & Huang, M. (2021). PM2.5 estimation and spatial-temporal pattern analysis based on the modified support vector regression model and the 1 km resolution MAIAC AOD in Hubei, China. *ISPRS International Journal of Geo-Information*, *10*(1), 31.

Gulati, S., Bansal, A., Pal, A., Mittal, N., Sharma, A., & Gared, F. (2023). Estimating PM2.5 utilizing multiple linear regression and ANN techniques. *Scientific Reports*, *13*(1), 22578.

Hayes, R. B., Lim, C., Zhang, Y., Cromar, K., Shao, Y., Reynolds, H. R., . . . Thurston, G. D. (2020). PM2.5 air pollution and cause-specific cardiovascular disease mortality. *International Journal of Epidemiology*, *49*(1), 25-35.

Kliengchuay, W., Srimanus, R., Srimanus, W., Niampradit, S., Preecha, N., Mingkhwan, R., . . . Tantrakarnapa, K. (2021). Particulate matter (PM10) prediction based on multiple linear regression: A case study in Chiang Rai Province, Thailand. *BMC Public Health*, *21*, 1-9.

Koenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, *15*(4), 143-156.

Lesar, T. T., & Filipčić, A. (2021). The hourly simulation of PM2.5 particle concentrations using the multiple linear regression (MLR) model for sea breeze in Split, Croatia. *Water, Air, and Soil Pollution*, *232*(7), 261.

Liu, Q., Xu, C., Ji, G., Liu, H., Shao, W., Zhang, C., . . . Zhao, P. (2017). Effect of exposure to ambient PM2.5 pollution on the risk of respiratory tract diseases: A meta-analysis of cohort studies. *Journal of Biomedical Research*, *31*(2), 130.

Mogollón-Sotelo, C., Casallas, A., Vidal, S., Celis, N., Ferro, C., & Belalcazar, L. (2021). A support vector machine model to forecast ground-level PM2.5 in a highly populated city with a complex terrain. *Air Quality, Atmosphere and Health*, *14*, 399-409.

Pope III, C. A., Coleman, N., Pond, Z. A., & Burnett, R. T. (2020). Fine particulate air pollution and human mortality: 25+ years of cohort studies. *Environmental Research*, *183*, 108924.

Pun, V. C., Kazemiparkouhi, F., Manjourides, J., & Suh, H. H. (2017). Long-term PM2.5 exposure and respiratory, cancer, and cardiovascular mortality in older US adults. *American Journal of Epidemiology*, *186*(8), 961-969.

Ren, Z., Liu, X., Liu, T., Chen, D., Jiao, K., Wang, X., . . . Ma, L. (2021). Effect of ambient fine particulates (PM2.5) on hospital admissions for respiratory and cardiovascular diseases in Wuhan, China. *Respiratory Research*, *22*(1), 128.

Saiohai, J., Bualert, S., Thongyen, T., Duangmal, K., Choomanee, P., & Szymanski, W. W. (2023). Statistical PM2.5 prediction in an urban area using vertical meteorological factors. *Atmosphere*, *14*(3), 589.

Sirithian, D., & Thanatrakolsri, P. (2022). Relationships between meteorological and particulate matter concentrations (PM2.5 and PM10) during the haze period in urban and rural areas, northern Thailand. *Air, Soil and Water Research*, *15*.

Slama, A., Śliwczyński, A., Woźnica, J., Zdrolik, M., Wiśnicki, B., Kubajek, J., . . . Franek, E. (2019). Impact of air pollution on hospital admissions with a focus on respiratory diseases: A time-series multi-city analysis. *Environmental Science and Pollution Research*, *26*, 16998-17009.

Weizhen, H., Zhengqiang, L., Yuhuan, Z., Hua, X., Ying, Z., Kaitao, L., . . . Yan, M. (2014). Using support vector regression to predict PM10 and PM2.5. *2014 IOP Conference Series*: *Earth and Environmental Science*.

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.

Zeng, Y., Jaffe, D. A., Qiao, X., Miao, Y., & Tang, Y. (2020). Prediction of potentially high PM2.5 concentrations in Chengdu, China. *Aerosol and Air Quality Research*, *20*(5), 956-965.