EASR

Engineering and Applied Science Research

https://www.tci-thaijo.org/index.php/easr/index

Published by the Faculty of Engineering, Khon Kaen University, Thailand

Ensemble machine learning-based PM2.5 modeling using hotspot counts (0-1000 km) reflecting Chiang Mai, Thailand's extreme pollution

Rati Wongsathan* and Apimook Sabkam

Department of Electrical Engineering, Faculty of Engineering and Technology, North-Chiang Mai University, Chiang Mai 50230, Thailand

Received 22 January 2025
Revised 29 May 2025
Accepted 11 June 2025

Abstract

Persistent local and transboundary smog has critically elevated PM2.5 levels in Northern Thailand over the past decade, resulting in significant health risks. The spatial distribution of hotspot counts, indicative of biomass burning and smoke dispersion, demonstrates a strong correlation with PM2.5 concentration patterns, underscoring the importance of incorporating such data into air quality analyses. This study integrates hotspot data to capture both temporal dynamics and external influences in PM2.5 prediction models. The importance of lagged hotspot counts within 100–1000 km of Chiang Mai—ranked as the world's most polluted city during the study period—and lagged ground-level PM2.5 is assessed using Lasso regularization. The analysis reveals that the cumulative effects of hotspots extend their influence on air quality in Chiang Mai up to approximately 700 km. Advanced tree-based ensemble machine learning methods, including Random Forest (RF), Gradient Boosting (GB), and Extreme Gradient Boosting (XGBoost), are implemented alongside the Long Short-Term Memory (LSTM) deep learning model to evaluate their predictive performance. This approach provides a novel framework for PM2.5 modeling in Northern Thailand. Five key features with specific day lags were identified for modeling. These include PM2.5 at lag 1, short-range hotspots within 100 km at lags 1 to 3, mid-range hotspots at 200 and 400 km at lags 2 to 4, and long-range hotspots beyond 700 km at lag 5. Incorporating hotspot data improved model performance by approximately 20%, as evidenced by error metrics and residual analysis. Among the models tested, GB outperformed XGBoost, RF, and LSTM, achieving the highest R^2 (0.97), lowest RMSE (5.49), MAE (2.08), and MAPE (5.8%), along with near-zero MBE and minimal MdAE (0.48). Statistical validation confirmed the model's reliability with no significant bias.

Keywords: PM2.5 Prediction model, Hotspot analysis, Random Forest, Gradient Boosting, Extreme Gradient Boosting, LSTM

1. Introduction

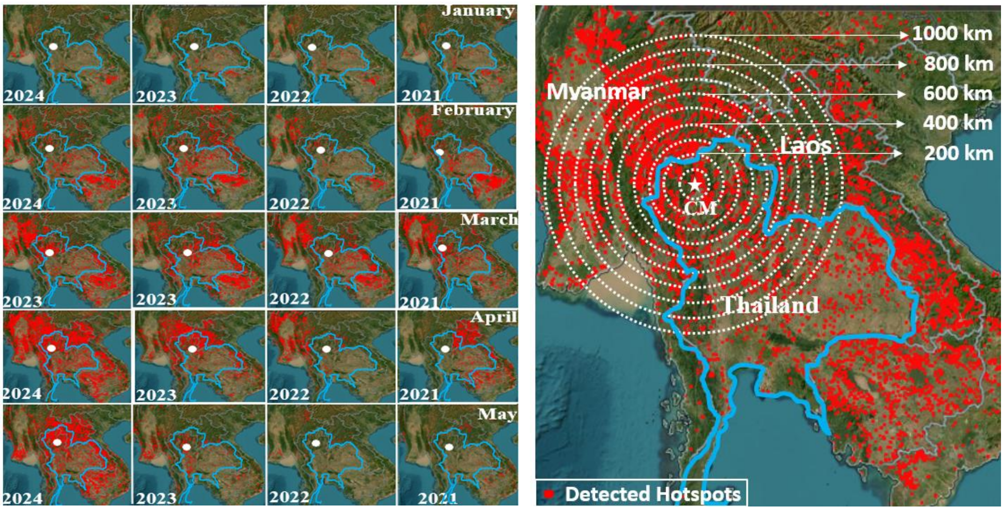
PM2.5 is a major issue in northern Thailand, particularly Chiang Mai. On March 27, 2024, Chiang Mai recorded the world's highest PM2.5 level 151 μg/m^3 far exceeding the WHO guideline of 15 μg/m^3 and Thailand's Pollution Control Department (PCD) standard of 37.5 μg/m^3. This severe pollution is driven by local fires and transboundary haze from Myanmar and Laos, as shown by a backward trajectory model [1]. The Positive Matrix Factorization (PMF) model indicates Chiang Mai's PM2.5 consists of 51% biomass burning from the Thai-Myanmar border and 26.4% local dust and traffic [2]. In addition, the Chiang Mai Basin's geography, surrounded by mountains, hinders pollution dispersal during the dry season, with 33-50% of smog linked to sources in Myanmar and Laos [3, 4]. Consistent open burning patterns and numerous hotspots, peaking in March, are observed annually from January to May [5] (see Figure 1(a), retrieved from NASA's FIRMS for 2021-2024).

The seasonal PM2.5 pattern mirrors hotspots distribution within 1000 km of Chiang Mai (Figure 1(b)), linking burning activities to pollution (Figure 2). Hotspots within 100 km indicate local burning, while those 300–1000 km reflect transboundary pollution. PM2.5 levels from past trend (2015), mid-trend (2020), and current trend (2023) strongly correlate with hotspots, particularly within 300–600 km. Beyond 800 km, correlation weakens due to dispersion, highlighting combined local and long-range impacts on Northern Thailand's air quality [6].

PM2.5 modeling employs two main approaches: (1) physics-based models (e.g., WRF-Chem, HYSPLIT, and CTMs), and (2) data-driven methods, including regression (e.g., MLR, and Lasso), statistical models (e.g., ARIMA and ARIMAX), and machine learning (ML) (e.g., SVM, NN, and LSTM). Predictors like meteorological data, aerosol optical depth (AOD), toxic gases (NO_x, SO_2, CO, O_3), and land-use data support research but are costly, sensitive to variations, and can destabilize forecasts. High-dimensional models often add complexity without improving accuracy. Simpler methods, like ARIMA, can outperform complex models such as LSTM when using only lagged PM2.5 data (e.g., in [7]). However, most studies focus on PM2.5 predictors, with limited research on region-specific sources like biomass burning, vehicle, and industry. In Northern Thailand, where fire is a major PM2.5 source, WRF-Chem with updated emission factors improves accuracy [8]. HYSPLIT prediction benefit from including biomass burning parameters [5]. Combining hotspot counts and meteorological data in a deep neural network (DNN) enhances PM10 prediction [9].

*Corresponding author.
Email address: rati@northcm.ac.th

doi: 10.14456/easr.2025.42



(a) (b)

Figure 1 (a) Monthly hotspot distribution (red dots) from January to May across different years (2021–2024), showing seasonal variations in fire activity. (b) Spatial distribution of hotspots within a 1000 km radius from Chiang Mai (CM), Thailand, highlighting transboundary influences from Myanmar and Laos

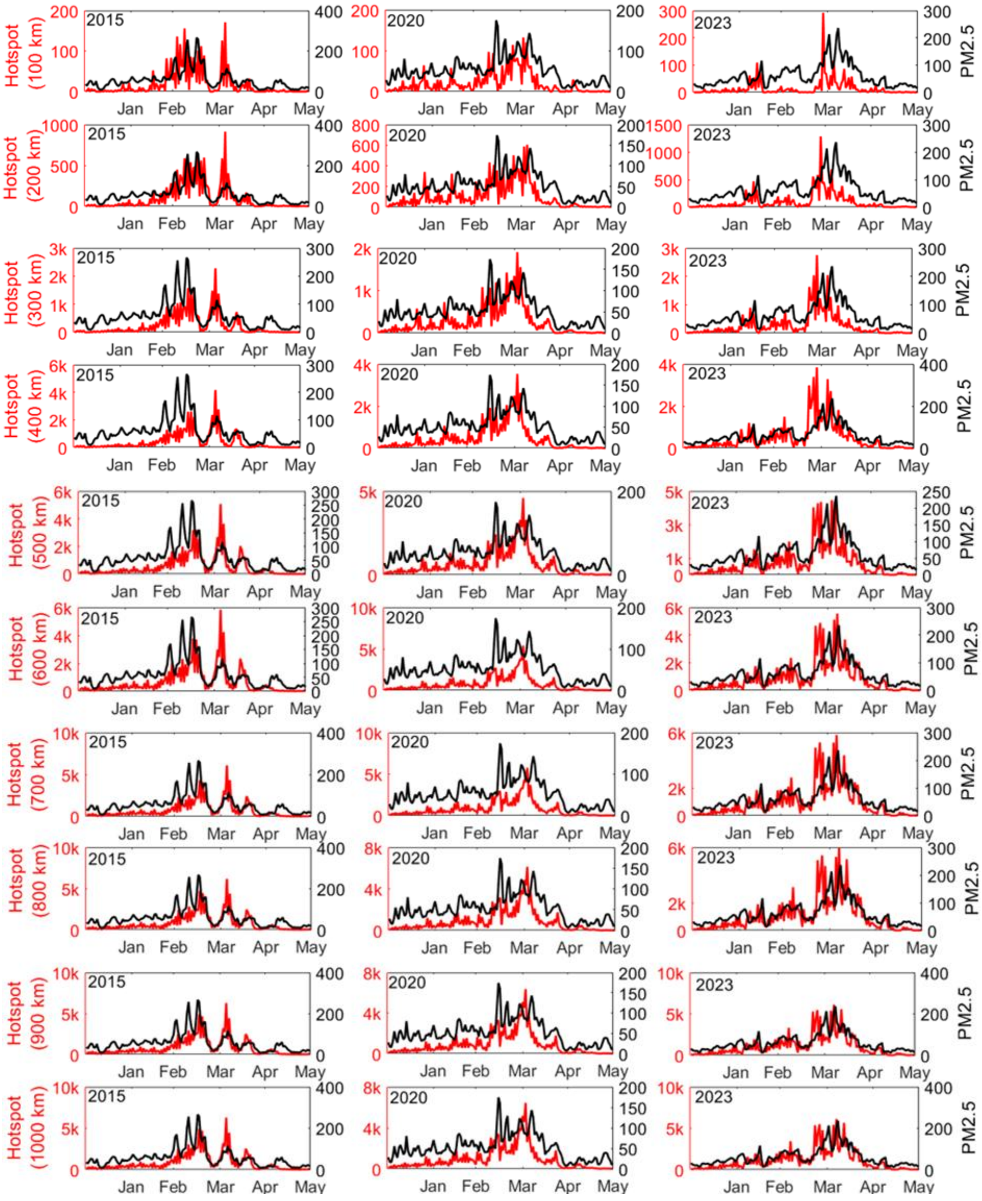


Figure 2 Correlation of PM2.5 levels and hotspot count at increasing distances from Chiang Mai between 0–1000 km in past (2015), mid-(2020), and current (2023) trends

Recent PM2.5 prediction research has utilized ML techniques. Traditional models (e.g., SVM, ANNs, and k-NN) improved with preprocessing methods like normalization and feature selection, highlighting data quality [10]. Deep learning models (CNNs, RNNs, LSTMs, and GRUs) offered higher accuracy but required large datasets and high computational power [11]. Hybrid models, such as CNN-LSTM, boosted accuracy and generalizability [12]. Tree-based models like Random Forest (RF) [13], Generalized Additive Models (GAM) [14], and Extreme Gradient Boosting (XGBoost) [15] capture nonlinear relationships in PM2.5 data, with hybrid models outperforming individual models [14, 16], albeit with increased complexity. However, no studies have applied these methods to PM2.5 forecasting in Northern Thailand or used hotspot as predictors, with RF showing superior performance in meteorological and altitude-based models [17].

This study estimates ground-level PM2.5 in Chiang Mai (2020–2023) using training data from 2011–2019. It employs tree-based ensemble ML algorithms (RF, GB, XGBoost), an LSTM model, and optimal subset predictors. Hotspots within 100–1000 km serve as key exogenous variables. Feature selection used PACF analysis and Lasso regularization to identify optimal lags for hotspot and PM2.5 data. Models are evaluated with error metrics and residual analysis, and comparisons with and without hotspot data and relevant studies validate the approach. Key contributions include:

1. PM2.5 levels in Northern Thailand show a strong relationship with hotspot activity, especially from regions 300 to 600 kilometers away. These areas contribute significantly to air pollution in Chiang Mai.
2. The spatial pattern of hotspots reveals low activity within 100 to 200 kilometers, a steep increase between 300 and 400 kilometers (mainly in Thailand, followed by Laos and Myanmar), and a peak at 500 to 600 kilometers. Beyond 600 kilometers, contributions gradually decline. This highlights the combined impact of both nearby emissions and long-range pollution transported from neighboring countries.
3. Ensemble tree-based machine learning models—Gradient Boosting (GB), Extreme Gradient Boosting (XGBoost), and Random Forest (RF)—demonstrate stronger predictive performance than deep learning models like Long Short-Term Memory (LSTM) when using hotspot data as the main external input for PM2.5 prediction.

2. Data analysis

2.1 Data and data processing

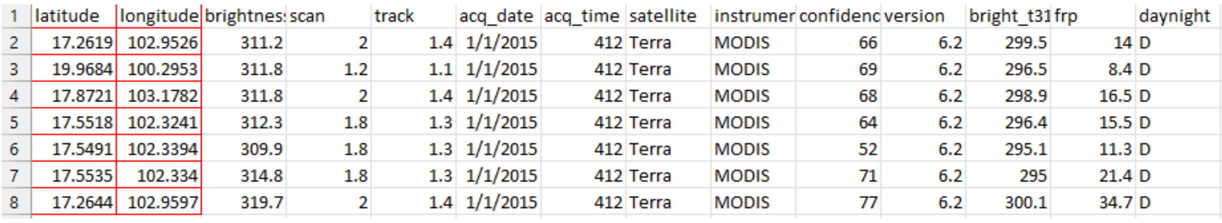
This analysis uses raw data with latitude and longitude coordinates from MODIS/Aqua Terra Thermal Anomalies/Fire locations to identify hotspots (Figure 3). Hotspot data for Thailand, Myanmar, and Laos, provided in latitude and longitude, were converted to distances from Chiang Mai's city center (18.7883° N, 98.9853° E) using the Haversine formula,

$$d = D \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\Delta\lambda}{2}\right)}\right), \tag{1}$$

where ϕ_1 and ϕ_2 are the latitudes of the two points on Earth, λ_1 and λ_2 are their respective longitudes, $\Delta\phi$ and $\Delta\lambda$ are their differences, $\Delta\phi=|\phi_1-\phi_2|$, $\Delta\lambda=|\lambda_1-\lambda_2|$ and D is Earth's diameter.

This equation calculates the great-circle distance between two points on a sphere using latitude and longitude. It accounts for Earth's curvature, making it ideal for geographic distance measurements. The formula incorporates sine and cosine functions to handle angular differences in latitude (ϕ) and longitude (λ). The *arcsine* function helps determine the angular separation, which is then scaled by Earth's diameter (~12742 km) to convert it into a real-world distance.

Missing data were resolved, and hotspots (100–1000 km) were counted using Pandas, a Python data analysis library by applying loops and the unique() function for 2015–2023. Daily PM2.5 data were obtained from ground stations 35T (City Hall) and 36T (City High School), provided by Thailand's PCD. PM2.5 and hotspot counts were normalized to [-1, 1] to ensure comparability, reduce bias, and adapt to temporal changes.



1	latitude	longitude	brightnes:	scan	track	acq_date	acq_time	satellite	instrumer	confidenc	version	bright_t31	frp	daynight
2	17.2619	102.9526	311.2	2	1.4	1/1/2015	412	Terra	MODIS	66	6.2	299.5	14	D
3	19.9684	100.2953	311.8	1.2	1.1	1/1/2015	412	Terra	MODIS	69	6.2	296.5	8.4	D
4	17.8721	103.1782	311.8	2	1.4	1/1/2015	412	Terra	MODIS	68	6.2	298.9	16.5	D
5	17.5518	102.3241	312.3	1.8	1.3	1/1/2015	412	Terra	MODIS	64	6.2	296.4	15.5	D
6	17.5491	102.3394	309.9	1.8	1.3	1/1/2015	412	Terra	MODIS	52	6.2	295.1	11.3	D
7	17.5535	102.334	314.8	1.8	1.3	1/1/2015	412	Terra	MODIS	71	6.2	295	21.4	D
8	17.2644	102.9597	319.7	2	1.4	1/1/2015	412	Terra	MODIS	77	6.2	300.1	34.7	D

Figure 3 Raw data of latitude and longitude for geographic coordinates from MODIS/Aqua Terra Thermal Anomalies/Fire locations, indicating hotspots detected in Thailand during 2015, retrieved from NASA's Fire Information for Resource Management System (FIRMS)

Figure 4 shows local and regional hotspot counts. Activity is low within 100–200 km, rising sharply at 300–400 km, led by Thailand, followed by Laos and Myanmar. Peaks occur at 500–600 km, with Laos nearing Thailand and Myanmar growing steadily. Beyond 700 km, hotspots decline, with Thailand dominant, Laos moderate, and Myanmar lowest. Activity peaks seasonally from February to April, driven by agricultural burning and climate factors. Within 100 km, hotspots are local, with PM2.5 levels below 50/day and moderate peaks in February and March. At 200 km, Thailand dominates, with Myanmar's influence growing and strong PM2.5 correlations in late February to early March. Between 300–400 km, Myanmar's hotspot activity drives PM2.5 peaks during the February–April season, with Laos contributing less. From 500–700 km, Myanmar leads in hotspot counts, with strong PM2.5 correlations, while Thailand contributes moderately and Laos shows late-season activity. At 800–1000 km, Myanmar remains dominant, closely aligning with PM2.5 peaks, while Laos shows minimal increases. The data highlights Myanmar's key role in long-range transboundary smoke as Thailand's influence diminishes.

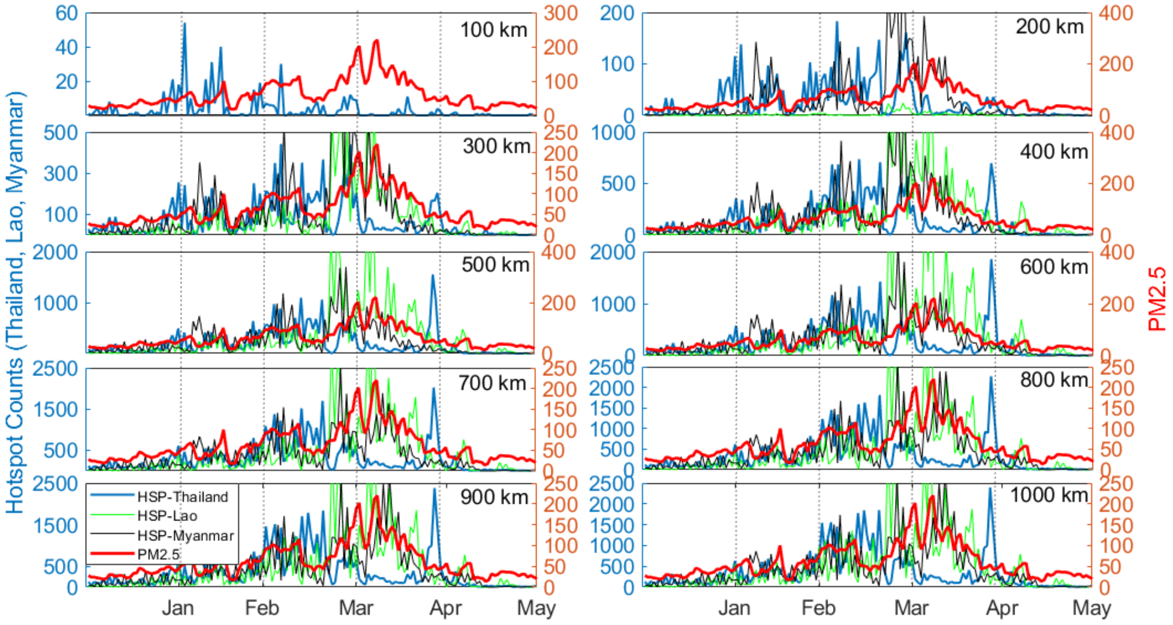


Figure 4 Hotspot counts in Thailand (blue), Laos (green), and Myanmar (black) at different distances (100–1000 km) from Chiang Mai, and their correlation with PM2.5 levels (red) over time (January–May) where the left y-axis represents hotspot counts, while the right y-axis (in red) represents PM2.5 levels

2.2 Hotspot counts analysis

2.2.1 Monthly trends and seasonal patterns

Monthly hotspot trends (2015–2023) around Chiang Mai (Figure 5) show a March peak, aligning with the dry season's agricultural burning and forest fires contributing to PM2.5. January and February have low activity, marking the pre-burning phase, while counts decrease as the burning season ends in April and May. This pattern can be categorized as follows,

- Pre-burning phase (January–February): Low hotspot activity.
- Peak burning season (March–April): Significant increase in hotspots, especially in March.
- Post-burning phase (April–May): Gradual decline in hotspot counts.

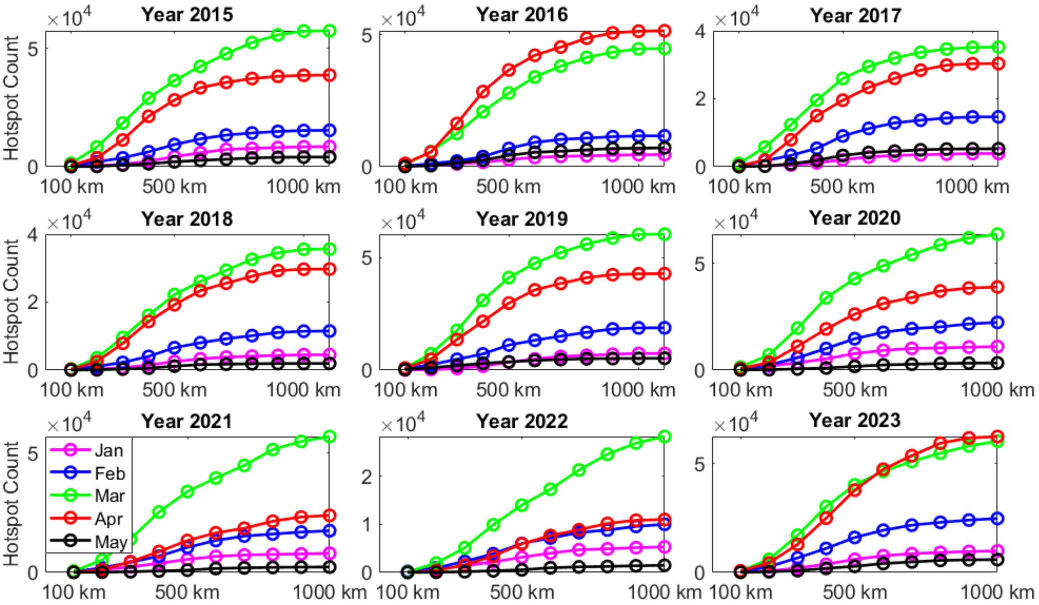


Figure 5 Monthly trends of cumulative hotspot count against distance from Chiang Mai, Thailand (2015–2023)

2.2.2 Spatial distribution of hotspots

Hotspot distribution by distance reveals steep increases within 0–500 km due to shorter transport times, particularly in March and April. Beyond 500 km, regional contributions remain notable, aided by strong winds or atmospheric stability, but diminish with distance. This spatial breakdown enhances resolution for air quality predictions, highlighting the influence of both local emissions and regional pollution from Myanmar and Laos on Chiang Mai's air quality.

2.2.3 Inter-annual variability and driving factors

Hotspot intensity and distribution exhibit significant inter-annual variability, with years like 2015, 2019, and 2023 showing higher counts compared to 2017, 2020, and 2022, correlating with PM2.5 profiles (Figure 6). Climatic conditions, fire management policies, and socio-economic factors likely drive these variations, with 2022 standing out due to strong fire control measures that reduced hotspots despite peaks in other years. Annual PM2.5 concentrations (μg/m³) from 2015–2023 align with hotspot activity patterns (Figure 6), especially in March–April during peak biomass burning. Significant PM2.5 and hotspot peaks occurred in 2015 and 2019, while 2020 and 2021 saw decreased PM2.5 levels despite moderate hotspot activity, likely due to reduced human activity and biomass burning during the COVID-19 pandemic, underscoring the link between biomass burning and air quality deterioration.

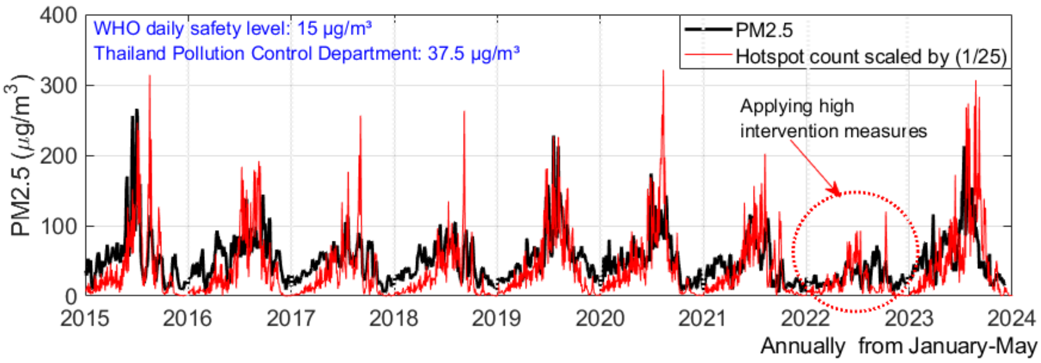


Figure 6 Annual PM2.5 levels recorded from January to May (2015–2024) in Chiang Mai, Thailand, compared with total hotspot counts from Thailand, Myanmar, and Laos (hotspot counts scaled by 1/25)

In 2022, despite moderate hotspot activity, stricter fire control measures reduced PM2.5 levels, highlighting the effectiveness of policy in mitigating air pollution [18]. In contrast, 2023 saw elevated PM2.5 levels despite high hotspot counts, indicating that the intensity of biomass burning had a stronger impact on air quality.

2.2.4 Hotspot distance and air quality impact

Table 1 shows that nearby hotspots (100–300 km) and moderate-distance hotspots (400–700 km) had a stronger influence on air quality in 2015, 2019, and 2023, with evolving dynamics over time. The gradual impact of distant hotspots beyond 700 km is also evident. The cumulative effect highlights the far-reaching impact of hotspot activity on Chiang Mai's air quality, extending up to approximately 700 km.

Table 1 Percentage increase in PM2.5 levels at varying distances and their increments.

Distance	2015		2019		2023	
	%Hotspot increasing	%Increment	%Hotspot increasing	%Increment	%Hotspot increasing	%Increment
100 km	1.63%	-	4.02%	-	2.66%	-
200 km	2.83%	+1.2%	2.78%	-1.24%	1.74%	-0.92%
300 km	2.94%	+0.11%	5.05%	+1.03%	2.57%	+0.83%
400 km	4.41%	+1.47%	7.61%	+2.56%	3.76%	+1.19%
500 km	5.59%	+1.18%	8.28%	+0.67%	4.00%	+0.24%
600 km	6.55%	+0.95%	9.29%	+1.01%	4.66%	+0.66%
700 km	7.63%	+1.08%	10.31%	+1.02%	5.27%	+0.61%
800 km	7.81%	+0.18%	10.63%	+0.32%	5.86%	+0.59%
900 km	8.36%	+0.55%	11.27%	+0.64%	6.24%	+0.38%
1000 km	8.87%	+0.51%	12.20%	+0.93%	6.76%	+0.52%

2.3 Feature selection

2.3.1 Identifying optimal lag features

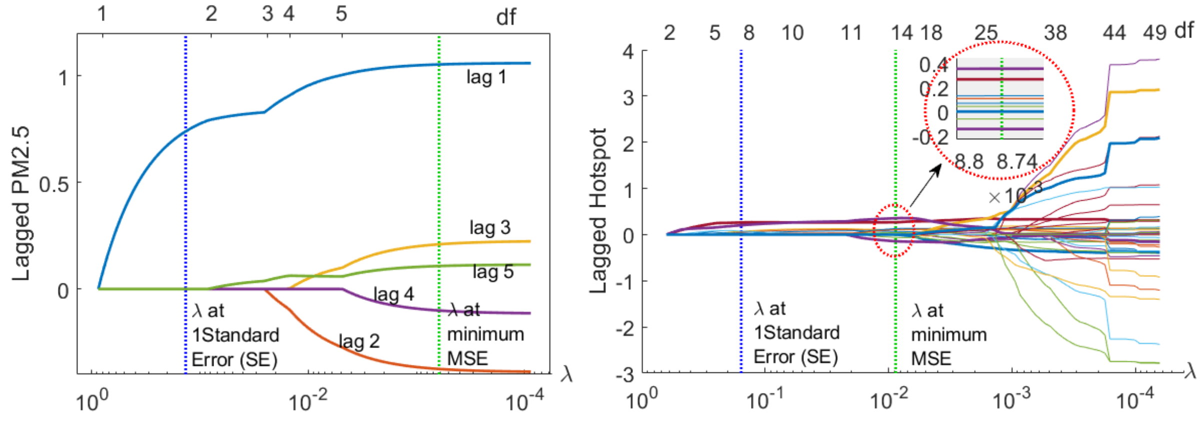
To identify the most relevant predictors for PM2.5 forecasting, both statistical and machine learning methods were employed. First, time-series analysis techniques such as the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) were used to explore how past values of PM2.5 relate to current levels. This helped identify meaningful lag intervals for PM2.5 itself. To evaluate lagged hotspot variables—measuring fire activity at various distances from Chiang Mai—each distance band (e.g., 100 km, 200 km, up to 1000 km) was tested at multiple lag intervals (up to five days), reflecting possible delays in pollutant transport due to atmospheric movement.

For systematic lag selection, the TimeSeriesSplit method from the scikit-learn library in Python was used to perform cross-validation that respects the chronological nature of time-series data. Lasso regression, a regularized linear model, was then applied to evaluate combinations of lagged PM2.5 and hotspot features. By penalizing less relevant variables, Lasso selects only those features that meaningfully contribute to predictions while avoiding overfitting. Model performance was evaluated using Mean Squared Error (MSE) during cross-validation. The final selection of features—those offering the best predictive performance—was determined based on the lowest validation error. Once selected, these features were used to train and test multiple models, and their performance was assessed using standard evaluation metrics such as Root Mean Squared Error (RMSE) and the coefficient of determination (R^2).

This combined approach ensured that the selected lags—both for local PM2.5 trends and for hotspot activity at different distances—reflected real-world pollution transport patterns and improved prediction accuracy.

2.3.2 Feature selection process

Lagged variables for PM2.5 and hotspot data were first explored using ACF and PACF to identify significant autocorrelations over time. These initial findings were then refined using Lasso regression, a method that selects only the most important predictors while penalizing less informative ones. Figure 7 illustrates the Lasso path for both lagged PM2.5 values (Figure 7 (a)) and hotspot counts at various lags and distances (Figure 7 (b)), highlighting two key points on the regularization path: the value of λ at the first standard error (1-SE) and the value at the minimum mean squared error (MSE).



(a) (b)

Figure 7 Lasso path for (a) lagged PM2.5 variables and (b) lagged Hotspot variables.

In Figure 7(a), lag 1 of PM2.5 stands out clearly with a large coefficient across a wide range of λ values. This confirms it as the most influential predictor—indicating that the previous day's PM2.5 level (PM2.5(*t*–1)) is strongly correlated with the current day's reading. Although other lags such as lag 3 and lag 5 appear as λ decreases, their contributions are notably smaller. The 1-SE criterion supports the selection of only lag 1 for simplicity and interpretability, whereas the minimum MSE criterion would also include lag 5, suggesting a more complex but potentially more accurate model. Nonetheless, the dominance of lag 1 aligns with common findings in time-series air quality forecasting.

In Figure 7(b), the selection of lagged hotspot variables is more complex, as it involves a larger number of features spanning different distances and time lags. Initially, most hotspot coefficients remain close to zero, but as λ decreases (moving toward the right of the graph), several variables begin to emerge as relevant predictors—particularly around λ values between 10^{-2} and 10^{-3}. This range is zoomed in and annotated in the figure for clarity. The emergence of non-zero coefficients here indicates the model's identification of hotspot influences from different distances and lag days. Specifically, this shows how Lasso pinpoints key lagged hotspots, such as those occurring within 100 km over the past 1–3 days and more distant fires (e.g., 200–400 km or beyond 700 km) with lagged effects up to 5 days.

Together, these results form the basis for selecting robust and meaningful lagged predictors, balancing interpretability (under 1-SE) and prediction accuracy (at minimum MSE). These selected features are carried forward and analyzed in more detail in next Sub-Section 2.3.3 and Section 4.1, where the effects of short-, mid-, and long-range fire activity on PM2.5 levels in Chiang Mai are systematically examined.

2.3.3 Key lagged variables and their influence

Table 2 summarizes these variables by year (2015–2019 training data) and overall.

To improve the prediction of PM2.5 levels, this study considers the impact of forest fires and agricultural burning activities. Using only past PM2.5 values (called time lags) may not be sufficient and can limit the model's ability to explain real-world causes. Therefore, the number of fire hotspots is also included as an input variable alongside past PM2.5 data. To select the most relevant lag periods of PM2.5 and hotspot activity—especially those occurring at different distances from Chiang Mai—a statistical method called Lasso is used. This method helps identify key predictors while avoiding overfitting. The selection follows two criteria: the 1-SE and the minimum MSE. The selected lags are summarized in Table 2 and categorized as follows:

1. *Selected time lags for PM2.5*: Lasso regression using the 1-SE rule consistently selected PM2.5(*t*–1) across all years (2015–2019), as it shows the strongest autocorrelation with current PM2.5 levels and effectively captures short-term local influences—dominant in urban air quality dynamics. The minimum MSE criterion also selected PM2.5(*t*–5), aiming for lower error by capturing more complex patterns. However, the added value of PM2.5(*t*–5) was marginal. As shown in Figure 7(a), the coefficients of PM2.5(*t*–1) and PM2.5(*t*–5) differ significantly, supporting the decision to retain only PM2.5(*t*–1) to keep the model simple, interpretable, and practically usable. This finding is consistent with previous research, which also identifies PM2.5(*t*–1) as the most effective and widely used lag for capturing daily air pollution trends.
2. *Selected time lags and distances for hotspot counts*: To better understand how fire-related emissions contribute to PM2.5 levels in Chiang Mai, this study examines not only the immediate past values of PM2.5 but also fire hotspot counts at different distances and time lags. Since smoke from biomass burning does not reach Chiang Mai instantly, incorporating these lagged hotspot variables helps the model capture real-world delays in pollutant transport caused by meteorological factors such as wind direction, wind speed, and atmospheric conditions. Using Lasso regression with two criteria—the 1-SE rule and the

minimum MSE (Figure 7(b))— the most informative lagged hotspot features were categorized into three spatial ranges with corresponding temporal lags:

1) Short-range hotspots (within 100 km, lags 1–3 and 200 km, lags 2–4): Fires close to Chiang Mai show a strong and immediate influence on PM2.5 levels, particularly at lag 1. This indicates that emissions from nearby fires can elevate PM2.5 concentrations as early as the next day due to minimal transport time. The selection of lags 1 to 3 under both 1-SE and minimum MSE criteria reinforces the urgency of localized fire response and real-time monitoring in the surrounding region. In addition, hotspot activity within 200 km—selected at lags 2 to 4—indicates that slightly more distant fires may influence Chiang Mai's air quality with a delay of two to four days. This lag is consistent with the time required for smoke and fine particulate matter to travel from regional sources under typical wind and weather conditions. These findings highlight the necessity for regional coordination beyond the immediate vicinity of Chiang Mai, especially during peak burning seasons when fire activity is high in nearby provinces or border regions.
2) Mid-range hotspots (400 km, lags 2–4): Hotspot counts from 200 km and 400 km (lags 2–4) emerged as significant predictors, especially under the minimum MSE criterion. These suggest that smoke from regional sources—such as fires in northern Thailand or western Laos—can reach Chiang Mai within two to four days. These patterns align with known meteorological behaviors, such as regional wind transport, and emphasize the need for coordinated efforts across provinces and neighboring countries.
3) Long-range hotspots (beyond 700 km, lag 5): Hotspots located more than 700 km away—up to 1000 km in regions such as parts of Laos or eastern Myanmar, where the number of hotspots remains relatively steady beyond 700 km—were selected at lag 5, highlighting their delayed but noticeable impact on air quality. This long-range effect supports the evidence of transboundary pollution and the influence of regional biomass burning events several days after ignition. It also underscores the importance of cross-border air quality management and regional early warning systems.

Together, these results show that short-, mid-, and long-range fire activity all contribute to air pollution in Chiang Mai, each with distinct transport timelines. The adaptive selection of these lagged hotspot variables using Lasso demonstrates how the model reflects real-world atmospheric processes and helps reveal when and where fire-related emissions are most likely to affect urban air quality. This insight is vital for developing timely and geographically targeted air quality management strategies—both within Thailand and in collaboration with neighboring countries. Further analysis of the lagged effects of biomass burning hotspots on PM2.5 levels is presented in Section 4.1.

Table 2 Important lagged variables for PM2.5 and Hotspot contributions (2015-2019) based on Lasso analysis at 1-SE and minimum MSE

Year	Important lagged variables			
	1-Standard Error (SE)		Minimum MSE	
	PM2.5	Hotspot at distance	PM2.5	Hotspot at distance
2015	1	-100 km (lag 2 and 3) -200 km (lag 1 and 2)	1-5	-100 km (lag 2-5), -200 km (lag 1-4) -400 km (lag 3-5), -1000 km (lag 1)
2016	1	-100 km (lag 2-3), -200 km (lag 1-4), -300 km (lag 4) -600 km (lag 3) -700 km (lag 5)	1, 4, 5	-100 km (lag 2-5), -200 km (lag 1-5) -300 km (lag 1-4), -400 km (lag 1) -500 km (lag 1-2, 4), -600 km (lag 3-5) -800 km (lag 3-4), -900 km (lag 2) -1000 km (lag 1)
2017	1	-100 km (lag 1) -200 km (lag 2 and 5) -300 km (lag 1) -600 km (lag 1)	1-5	-100 km (lag 1-2,4), -200 km (lag 2,5) -300 km (lag 2), -400 km (lag 3) -500 km (lag 3), -600 km (lag 2) -700 km (lag 1)
2018	1	-200 km (lag 1-3) -900 km (lag 1)	1	-100 km (lag 1-2, 4), -200 km (lag 1-4) -300 km (lag 4-5), -400 km (lag 2) -500 km (lag 5), -600 km (lag 3-4) -1000 km (lag 1, 3, 5)
2019	1	-200 km (lag 2) -400 km (lag 1-2) -700 km (lag 1)	1, 5	-100 km (lag 5), -200 km (lag 2-4) -400 km (lag 1-2), -700 km (lag 1) -1000 km lag 5
Over all	1	-100 km (lag 1-3) -200 km (lag 1-4)	1-5	-100 km (lag 1-3, 5), -200 km (lag 2-5) -300 km (lag 5), -400 km (lag 2-4) -700 km (lag 1), -1000 km (lag 1,5)

2.3.4 Meteorological and policy impacts on feature variability

These year-to-year variations highlight the complex interplay between meteorological conditions (e.g., wind patterns and rainfall), fire activity, and emission sources, which complicates PM2.5 modeling. Year-specific data captures localized dynamics, shaped by unique conditions each year, while combined data reveals robust, generalized predictors but may overlook annual anomalies, emphasizing the need to balance specificity and generalization.

2.3.5 Implications for PM2.5 modeling

Under the 1-SE criterion, PM2.5 consistently shows lag 1 significance, with hotspots at 100 km (lags 1–3), 200 and 400 km (lags 2–4), and beyond 700 km (lag 5) being key predictors. The Minimum MSE approach identifies a broader range of influential hotspots across 100 km to 1000 km, with short-range distances (100–200 km) dominating and mid- to long-range distances (300–1000 km) showing variability depending on meteorological and burning conditions. This underscores the need to integrate both short- and long-range hotspot influences in PM2.5 modeling to capture local and regional pollution effects. Lagged variables vary significantly across years, limiting the applicability of fixed models trained on 2015–2019 data for predicting 2020–2023 levels. Temporal factors like climate, burning practices, and policy interventions (e.g., reduced burning in 2020 and 2022) drive these variations. Thus, while combined data identifies generalized patterns, accounting for yearly dynamics enhances accuracy for exceptional years.

For model training, focus on PM2.5 lag 1 and hotspots at 100 km (lags 1–3), 200 km (lags 2–4), 400 km (lags 2–4), and beyond 700 km (lag 5), and during testing on 2020–2023 data, monitor anomalies to balance generalization with specificity for reliable predictions.

3. Ensemble ML and LSTM models for PM2.5 prediction

Accurate prediction of PM2.5 concentrations is essential for effective air quality management and public health protection. To achieve this, advanced ML models, including RF, GB, XGBoost, and LSTM, can be used, utilizing lagged PM2.5 data and hotspot counts (H) as key predictors, i.e., $X = PM2.5(t–n, \ldots, t), H(t–m, \ldots, t)$, where m and n represent the maximum lags for PM2.5 and hotspot counts.

3.1 RF model

RF is a tree-based ensemble learning method that independently combines multiple decision trees to reduce variance and improve accuracy. The prediction for PM2.5 is represented by the relationship

$$PM2.5(t+1) = \frac{1}{M}\sum_{i=1}^{M} Tree_i(PM2.5(t-n, \ldots, t)) \tag{2}$$

where $Tree_i(X) = \sum_{j=1}^{Leaf} w_j \cdot 1\,(X \in R_j)$ is a recursive partitioning model that predict PM2.5(t+1) by splitting the input feature $X = PM2.5(t–n,\ldots,t) \in R_j$ into smaller subsets based on threshold values, R_j represents the region assigned to leaf node j, $1(X\in R_j)$ is an indicator function that equals 1 if X belongs to region R_j, otherwise 0, $Leaf$ is the number of terminal nodes (leaf nodes) in the tree, and w_j is the training parameter in R_j for the leaf node j.

Trees are independent and uncorrelated. RF employs several hyperparameters such as the number of trees (M), maximum tree depth (max_{depth}), minimum samples per split (min_{split}), minimum samples per leaf (min_{leaf}), and maximum features per split ($max_{feature}$) to control model complexity and prevent overfitting.

3.2 GB model

GB improves predictive accuracy by combining weak learners, typically decision trees, through iterative training and gradient descent. The prediction for PM2.5 is obtained by summing the contributions of all weak learners,

$$PM2.5(t+1) = \sum_{m=1}^{M} \alpha_m Tree_m(X) \tag{3}$$

GB assigns weights (α_m) to each tree, training them sequentially to minimize errors and focus on important features. Hyperparameter tuning, including M, learning rate (η), max_{depth}, min_{split}, min_{leaf}, and controls bias and variance (*Subsample*), optimizing performance and preventing overfitting.

3.3 XGBoost model

XGBoost is an optimized version of GB that incorporates advanced L2 regularization techniques to reduce overfitting and enhance computational efficiency. When applied to PM2.5 prediction, XGBoost prediction can be represented as

$$PM2.5(t+1) = \sum_{i=1}^{M} Tree_i(X) + L_2 \tag{4}$$

where L2 regularization terms is expressed as

$$L_2 = \sum_{t=1}^{N}(PM2.5_{Actual,t} - PM2.5_{Predict,t})^2 + \sum_{m=1}^{M}\Omega(Tree_m), \tag{5}$$

$\Omega(\mathbf{Tree}_m)$ represents the regularization term used to balance predictive accuracy with model simplicity related to the number of leaf nodes and the magnitude of leaf weights as follows

$$\Omega(Tree_m) = \gamma T_m + \frac{1}{2}\lambda||\omega||^2, \tag{6}$$

N is the total number of data points, T_m is the number of leaves in $\mathbf{Tree}_m$, ω is weights of the leaf nodes, γ is controlling penalty for adding a leaf, and λ is regularization parameter controlling leaf weights' magnitude. The common hyperparameters of XGBoost are similar to those of GB.

3.4 LSTM model

LSTM is specialized of recurrent neural network (RNN) designed to capture long-term dependencies in sequential data. For PM2.5 prediction, LSTM uses X to learn patterns in air quality. The key components include,

- Forget gate controls which information to retain or to discard from the memory activated by Sigmoid activation function (σ),

$$f_t = \sigma(W_f \cdot [h_{t-1}, X_t] + b_f). \tag{7}$$

The σ squashes values between 0 and 1, where values close to 0 mean forgetting, and values close to 1 mean retention.

- Input gate determines what new information to store in the memory,

$$i_t = \sigma(W_i \cdot [h_{t-1}, X_t] + b_i). \tag{8}$$

The sigmoid decides the importance of new input. Values close to 1 mean new information is added; values close to 0 mean it is ignored.

- Memory update unit updates the cell state using the forget and input gates,

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t, \tag{9}$$

where C_{t-1} is previous cell state, $C_t = \tanh(W_C \cdot [h_{t-1}, X_t] + b_C)$is a Candidate cell state, proposed new information processed via the hyperbolic tangent activation, the f_t scales C_{t-1}, controlling how much past information remains, and i_t scales $\tilde{C}_t$, determining how much new information is added.

- Output gate determines the final output,

$$o_t = \sigma(W_o \cdot [h_{t-1}, X_t] + b_o), \tag{10}$$

where h_{t-1} is previous hidden state, $X_t = PM2.5(t-n,\ldots,t)$ is current input, $\odot$ represents element-wise multiplication, and W's and b's are learnable weights and biases. After determining o_t, the final hidden state is computed as $h_t = o_t \odot \tanh(C_t)$, where C_t is the update cell state. The $\tanh(C_t)$ scales the cell state between -1 and 1, while o_t determines how much of $\tanh(C_t)$ contributes to h_t.

Key LSTM hyperparameters for accurate PM2.5 prediction include the number of layers (L), units per layer (U), and dropout rate (d) to control complexity and overfitting. Training parameters like learning rate (η), batch size (B), and optimizer (O) optimize weight updates, while L2 penalty (λ) and gradient clipping (C) prevent overfitting and gradient issues.

To evaluate model performance, metrics such as root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), mean bias error (MBE), median absolute error (MdAE), and R^2 should be used, along with t-statistic and p-value of residual analysis. Sensitivity analysis is also crucial for assessing the model's responsiveness to specific lags and distances, allowing for refinement of inputs as needed. This comprehensive approach ensures a balance between capturing general trends and adapting to year-specific variations, enhancing the model's predictive accuracy and practical applicability.

3.5 Hyperparameters tuning

Tuning hyperparameters helps balance bias and variance, enhancing model performance and preventing overfitting. Grid search explores predefined parameter ranges for different models. The optimized hyperparameters of the PM2.5 prediction models are shown in Table 3.

Table 3 Summarize the optimized hyperparameters using grid search.

Hyperparameters	Range	Optimum value			
		RF	GB	XGB	LSTM
The number of trees or estimators (M)	[100, 1000]	100	100	100	-
Maximum tree depth (max_{depth})	[3, 10] and 'None'	'None'	6	6	-
Minimum samples per leaf (min_{leaf})	[1, 5]	1	2	2	-
Minimum samples per split (min_{split})	[2, 10]	2	5	5	-
Maximum features per split ($max_{feature}$)	[Auto, SQRT, Log2]	Auto	Auto	Auto	-
Boostrap	[True, False]	True	False	True	-
Learning rate (η)	[0.01, 0.3]	-	0.1	0.1	-
Controls variance ($Subsample$)	[0.5, 1]	-	0.8	0.8	-
Weights of the leaf nodes (ω)	[1, 10]	-	5	5	-
Controlling penalty (γ)	[0, 10]	-	2	2	-
Regularization parameter (λ)	[0, 10]	-	0.5	0.5	-
The number of layers (L)	[1, 3]	-	-	-	2
Units per layer or neurons (U)	[32, 512]	-	-	-	50
Dropout rate (d)	[0.1, 0.5]	-	-	-	0.3
Learning rate (η)	[0.0001, 0.01]	-	-	-	0.001
Batch size (B)	[16, 128]	-	-	-	32
Optimize weight updates (O)	-	-	-	-	1
L2 penalty (λ)	[10^{-5}, 10^{-2}]	-	10^{-3}	10^{-3}	-
Gradient clipping (C)	[0.1, 5.0]	-	1	1	1

4. Results and discussion

4.1 Analysis of lagged effects of biomass burning hotspots on PM2.5 levels

This study clearly demonstrates that forest and agricultural fires, even those occurring several days prior, play a crucial role in shaping Chiang Mai's air quality. By analyzing the time-lagged influence of biomass burning hotspots on PM2.5 concentrations, it is evident that emissions do not affect the air immediately but continue to impact pollution levels for several days, depending on their distance from the city.

1) *Short-range hotspots*: Hotspots within 100 kilometers of Chiang Mai often influence PM2.5 levels the very next day, which is why lag 1 (one-day delay, $H(t–1)$) for local hotspots was consistently selected in the model. This reflects how local burning—such as open field fires—has an almost immediate effect on the air residents breathe. In 2015, some nearby fires continued to influence air quality two to three days after the event, supporting the inclusion of lags 2 and 3 in the hotspot variables for that year. These findings reinforce the importance of incorporating time-lagged hotspot variables into PM2.5 prediction models, as they help capture the real-world delay between fire activity and pollution impact. They also highlight that reducing local burning can result in rapid and tangible improvements in air quality—an insight especially useful for city and provincial authorities aiming for timely interventions during haze episodes. Policymakers should implement short-term interventions, such as immediate burning bans, in alignment with the observed short-distance hotspot effects on PM2.5 levels in Chiang Mai. Subdistrict and District Administrative Organizations Community, leaders, and local residents can play a vital role by monitoring fire activity and reporting incidents to authorities, enabling enforcement actions such as fines or legal prosecution. This collaborative approach can help reduce the occurrence of local hotspots and mitigate their immediate impact on air quality.
2) *Mid-range hotspots*: At the regional level, fires occurring 200 to 400 kilometers away also have a strong influence on Chiang Mai's air quality, but with more delayed effects. This is reflected in the model's selection of hotspot lags up to 5 days for these distances. For instance, in 2016, hotspots located around 200 km away were found to impact PM2.5 levels from lag 1 through lag 5, $H(t–1)–H(t–5)$—showing how pollutants from regional fires can persist and accumulate in the atmosphere over several days. These lag patterns were essential for capturing the extended influence of regional biomass burning in the prediction model. The findings emphasize that PM2.5 pollution is not confined to local sources alone and highlight the importance of inter-provincial coordination. Fires in one province can degrade air quality in another days later, reinforcing the need for joint mitigation strategies across neighboring regions. This situation often arises from unauthorized forest burning by local residents. In the short term, strict enforcement of existing laws is essential. Inter-provincial coordination should be led by provincial-level authorities, such as Provincial Administrative Organizations and governors, to formulate and implement collaborative policies aimed at addressing the issue at a broader scale. In the long term, sustainable solutions should involve raising public awareness about the broader health and environmental impacts, alongside providing alternative livelihood options that are both appropriate and economically viable for local communities.
3) *Long-range hotspots*: Even more striking, fires in distant areas such as northern Laos or eastern Myanmar—located over 700 kilometers away—can still degrade air quality in Chiang Mai, despite only a slight increase in hotspot counts beyond 700 kilometers. This is reflected in the model's selection of long-range hotspot variables with lags of up to 5 days, particularly in years like 2018 when hotspots at 1000 km influenced PM2.5 levels at lags 1, 3, and 5. These delayed but measurable impacts confirm that Chiang Mai's air quality is affected by transboundary pollution during peak burning seasons. Including these lagged, long-distance variables in the model is essential to capture the full scope of PM2.5 sources. This information can be used by public health agencies to issue timely warnings and recommend protective measures during periods when a high number of hotspots are detected in neighboring countries, thereby enabling more effective public health responses. These findings highlight the need for international cooperation and shared air quality management strategies among ASEAN countries, as pollution carried by wind does not stop at national borders.

Importantly, the pattern and distance of influence vary each year. In years with intense biomass burning, such as 2016 and 2018, Chiang Mai experienced pollution from both nearby and distant fires. In contrast, years with fewer fires, like 2017, saw more localized impacts. These findings highlight that effective PM2.5 management requires adaptive strategies tailored to each year's burning intensity. Finally, the study shows that the effects of fires can persist for up to five days. Close-range fires have the strongest and fastest effects, while long-range fires take longer to reach the city but still significantly raise PM2.5 levels. These insights are important for designing early-warning systems and pollution forecasts—helping authorities take preventive actions in advance.

4.2 Performance of the proposed models for PM2.5 prediction with hotspot variable

To improve the accuracy of PM2.5 prediction, hotspot count variables were incorporated into the models alongside lagged PM2.5 values. Specifically, we included PM2.5 at lag 1 and hotspot counts at three distinct spatial-temporal ranges: short-range hotspots (100 km with lags 1–3), mid-range hotspots (200–400 km with lags 2–4), and long-range hotspots (beyond 700 km with lag 5). These selections reflect the time it takes for fire-related emissions to travel and affect air quality in Chiang Mai, influenced by wind and atmospheric conditions.

4.2.1 Visual performance of the models

Figures 8 through 11 present the results for four predictive models: Long Short-Term Memory (LSTM), Random Forest (RF), Gradient Boosting (GB), and XGBoost. Each figure includes four panels:

(Panel a) Scatter plots of predicted versus observed PM2.5 values
(Panel b) Residuals plotted against predicted values
(Panel c) Histograms of residuals
(Panel d) Residuals over time

The scatter plots show that all models demonstrate high accuracy, as the points align closely with the ideal prediction line (the red dashed diagonal). This indicates strong agreement between observed and predicted PM2.5 concentrations. The residual plots (Panel b's) show a random distribution around zero, suggesting that the models do not systematically overestimate or underestimate the values. Histogram plots of residuals (Panel c's) reveal that errors are symmetrically distributed. Notably, XGBoost has the narrowest residual spread, followed by GB and RF, suggesting progressively smaller prediction errors across these models. Temporal plots of residuals (Panel d's) also indicate consistent performance, with no significant time-related biases.

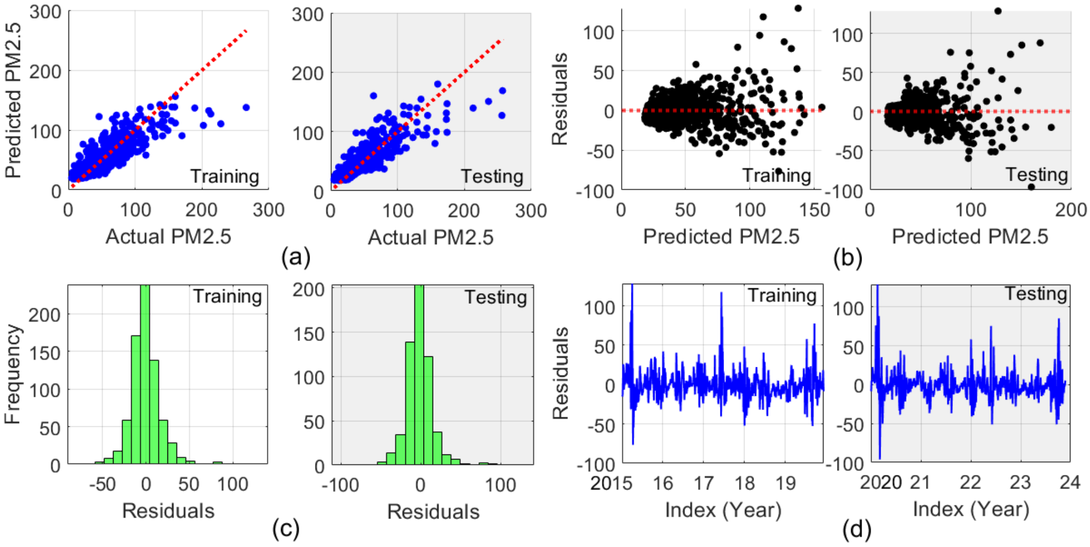


Figure 8 Performance evaluation of LSTM model for PM2.5 prediction, (a) Scatter plot of predicted against actual PM2.5 values, (b) Residuals against predicted PM2.5, (c) Histogram of residuals, and (d) Residuals plotted against index (Year)

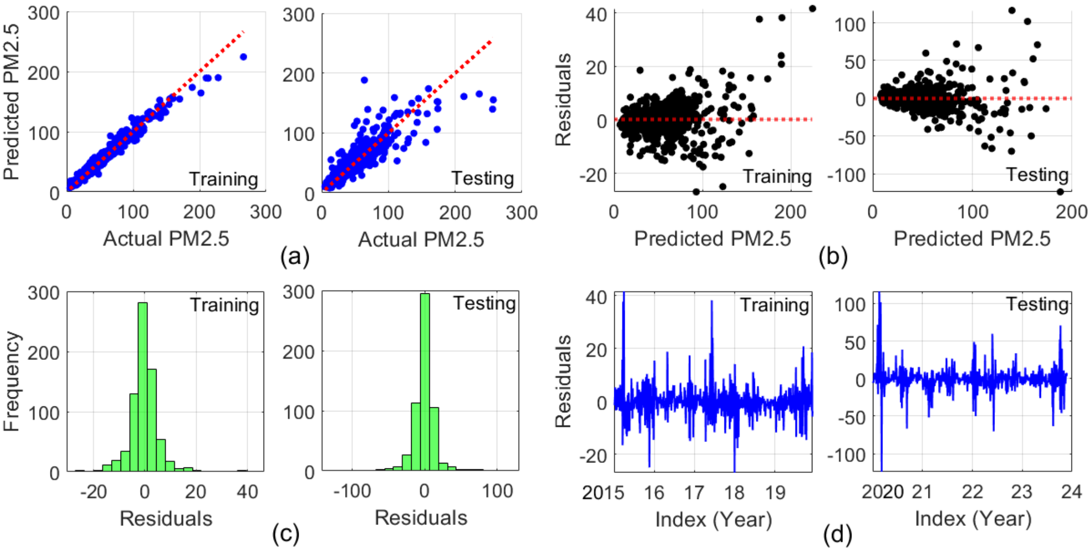


Figure 9 Performance evaluation of RF model for PM2.5 prediction, (a) Scatter plot of predicted against actual PM2.5 values, (b) Residuals against predicted PM2.5, (c) Histogram of residuals, and (d) Residuals plotted against index (Year)

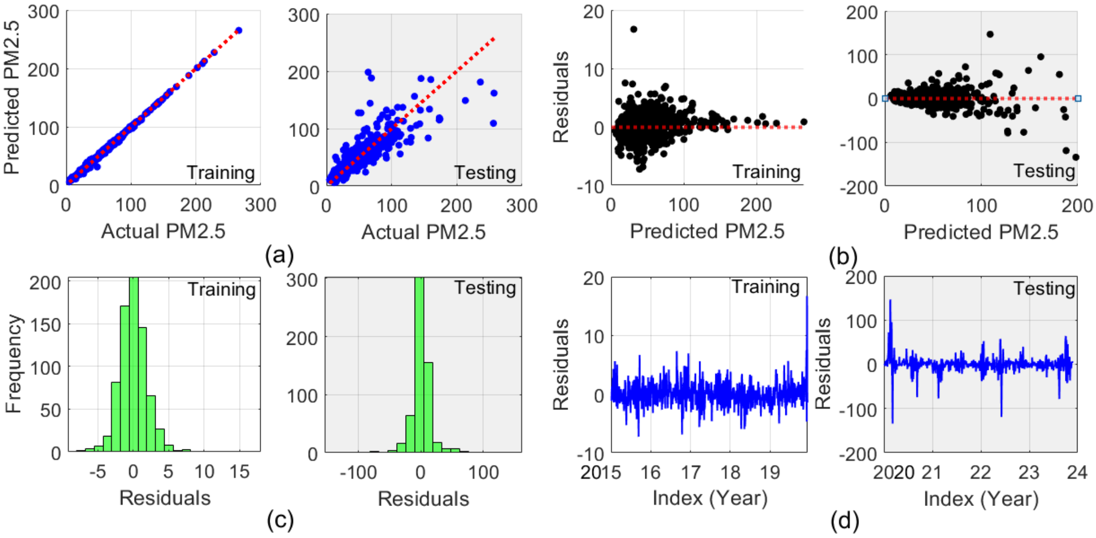


Figure 10 Performance evaluation of GB model for PM2.5 prediction, (a) Scatter plot of predicted against actual PM2.5 values, (b) Residuals against predicted PM2.5, (c) Histogram of residuals, and (d) Residuals plotted against index (Year)

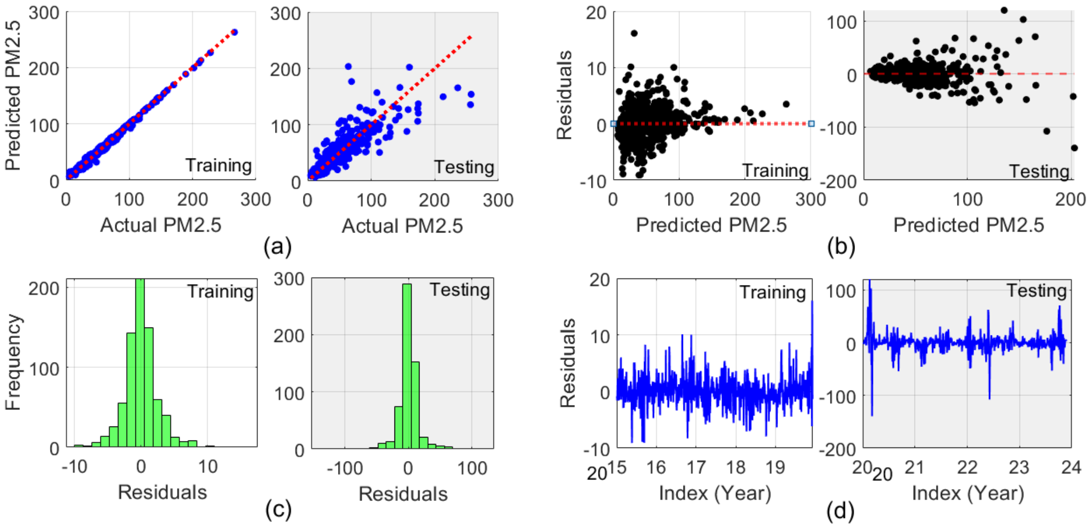


Figure 11 Performance evaluation of XGBoost model for PM2.5 prediction, (a) Scatter plot of predicted against actual PM2.5 values, (b) Residuals against predicted PM2.5, (c) Histogram of residuals, and (d) Residuals plotted against index (Year)

4.2.2 Interpretation of lagged hotspot effects

The inclusion of hotspot variables allows each model to consider the delayed impact of fire activity on PM2.5. This delay is a realistic representation of how smoke and particulates are transported through the atmosphere. The chosen lags serve as interpretable proxies for these transport processes. For instance, fires within 100 km can affect PM2.5 levels as soon as the next day, while fires over 700 km away may take up to five days to have an impact. The integration of these lagged variables gives the models an advantage in recognizing the temporal footprint of emissions across space and time.

4.2.3 Model interpretability and feature influence

Interpretability plays a vital role in model transparency. In the Random Forest (RF) model, we can find out which hotspot variables are most useful for predicting PM2.5 levels. This is done by checking how much each variable helps the model make better decisions. For example, if removing a variable makes the model less accurate, that means it was important. Another way is to see how much a variable helps the model split the data into clearer groups. When a variable makes a big improvement, it gets a higher importance score. This helps us understand which hotspot distances and time lags matter most in predicting air quality.

In the GB and XGBoost models, additional tools like SHAP values (SHapley Additive Explanations) and partial dependence plots are used to understand the contribution of each predictor. SHAP is a Python-based tool developed to explain how machine learning models make decisions. Based on game theory, it assigns each input variable (e.g., a lagged hotspot) a value indicating how much it contributes to the model's prediction, making complex models more transparent and easier to interpret across various fields. For example, SHAP values can tell us whether a fire that happened three days ago and 200 kilometers away is likely to increase PM2.5 levels today, and by how much. These tools don't just show which factors are important; they also explain whether each one pushes the prediction higher or lower, and how strongly. This helps make the model's decisions more transparent and supports clearer, evidence-based understanding of air pollution sources.

The tree-based nature of RF, GB, and XGBoost also allows the models to uncover non-linear interactions and threshold effects that linear models might miss. This capability helps identify specific lags that consistently influence PM2.5 levels, supporting hypotheses about how air pollution events evolve over time due to meteorological transport patterns. Thus, beyond their accuracy, the interpretability of RF, GB, and XGBoost models—when constrained to a subset of well-chosen lagged predictors—provides scientifically meaningful insights into the temporal coupling between fire activity and air quality, supporting both empirical understanding and policy-relevant conclusions.

4.2.4 Comparative model performance

Among all tested models, GB and XGBoost outperform the others in terms of precision and stability. GB shows a tighter residual distribution and reduced error variability compared to RF. XGBoost slightly surpasses GB by producing predictions most closely aligned to observed values and exhibiting the smallest spread in residuals.

Figure 12 compares model performance across the training period (2015–2019) and testing period (2020–2023). LSTM tracks general trends well but struggles with extreme values. RF performs well during training but shows signs of overfitting during testing. GB offers the best balance—tracking real data closely and achieving low error rates. XGBoost also performs well, but with slightly higher errors than GB. Overall, GB stands out as the most robust and consistent model. Statistical analysis confirms these differences. For example, GB significantly outperforms LSTM in terms of residual accuracy (t-stat = -6.80, p-value = 0.023 for MAE; t-stat = -6.90, p-value = 0.018 for MdAE at α = 0.05). It also performs comparably to RF, with no significant differences in error metrics (p > 0.80). In contrast, LSTM and RF show larger residuals, especially in the testing phase.

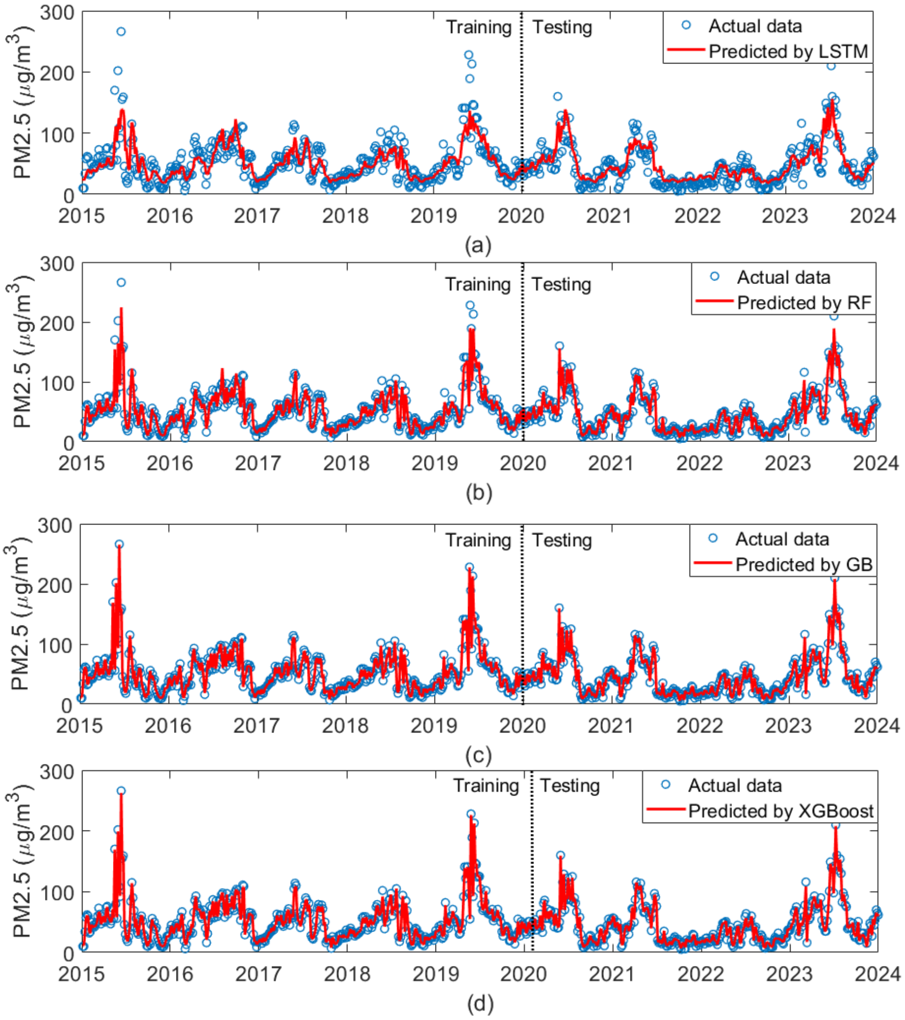


Figure 12 Comparison of PM2.5 predictions using hotspot variables: (a) LSTM, (b) RF, (c) GB, and (d) XGBoost models.

Table 4 Comparison performance of the prediction models with and without Hotspot variable

Model		**Error metrics**						**Residual analysis**	
		R^2	**RMSE**	**MAE**	**MAPE**	**MBE**	**MdAE**	**t-stat**	**p-value (>0.05)**
LSTM	Train	0.62	21.24	13.73	41.35%	5.34	8.98	5.52	0.00
(No Hotspot)	Test	0.54	32.92	23.03	38.19%	11.79	17.06	3.66	0.00
(Hotspot)	Train	0.72	18.07	12.54	35.13%	-0.24	9.29	-0.36	0.72
	Test	0.71	18.07	12.28	34.17%	-1.38	9.10	-1.85	0.06
RF	Train	0.94	8.13	4.76	12.95%	0.21	2.66	0.59	0.55
(No Hotspot)	Test	0.78	13.6	8.99	23.55%	1.13	6.01	0.88	0.37
(Hotspot)	Train	0.97	5.77	3.70	8.91%	-0.20	2.36	-0.95	0.34
	Test	0.76	16.41	9.73	22.9%	-0.48	6.09	-0.72	0.47
GB	Train	0.96	6.54	2.93	8.89%	0.18	1.23	0.67	0.51
(No Hotspot)	Test	0.76	14.19	9.60	25.37%	0.91	6.45	0.68	0.49
(Hotspot)	Train	0.97	5.49	2.08	5.78%	-0.00	0.48	-0.02	0.98
	Test	0.82	12.19	8.48	21.49%	-0.03	5.62	-0.02	0.98
XGBoost	Train	0.96	6.00	3.25	10.4%	0.25	1.65	1.01	0.31
(No Hotspot)	Test	0.82	12.38	9.05	23.26%	1.23	5.71	1.06	0.29
(Hotspot)	Train	0.97	5.90	2.58	7.55%	-0.24	0.86	-0.96	0.34
	Test	0.80	12.91	9.17	22.16%	-1.19	5.83	-0.98	0.33

4.2.5 Model behavior with and without hotspot data

Table 4 compares model performance with and without the inclusion of hotspot variables. The inclusion significantly improves prediction across most models.

- For LSTM, testing R^2 increases by 31.5%, with RMSE and MAE both reduced by around 15%.
- The RF model benefits mainly in training, with 29% and 22% reductions in RMSE and MAE, respectively, though testing RMSE increases slightly.
- GB sees strong gains, with testing R^2 improving by 7.9%, RMSE dropping by 14%, and MAPE decreasing by 16%.
- XGBoost shows minor improvements in training but slightly lower testing R^2.

To confirm that observed performance improvements are meaningful between using and not using the hotspot variable in the prediction models, the statistical significance tests with paired t-tests and calculate confidence intervals are applied to compare the performance metrics across the two scenarios. The results are *t*-stat=3.143 and *p*-value = 0.0235. This result indicates that the p-value is less than 0.05, meaning there is a statistically significant difference between the RMSE values for the models with and without the hotspot variable. The 95% confidence interval for the difference in RMSE is between 0.2341 and 2.1534. This range suggests that the performance improvement is meaningful, and the hotspot variable helps reduce RMSE in a statistically significant way.

Additionally, only LSTM without hotspot variables shows statistically significant bias (*p*-value < 0.05), reinforcing the need for hotspot-informed predictors in time-series air quality models.

4.2.6 Why LSTM Underperformed

Moreover, as shown in Table 4, while LSTM models are commonly effective for modeling temporal dependencies in time-series data, their relatively lower performance in this study—particularly compared to ensemble tree-based models such as Random Forest (RF), Gradient Boosting (GB), and XGBoost—can be attributed to several interrelated factors.

- First, LSTM models generally require large volumes of data to effectively capture long-range dependencies; however, the available dataset, while spanning multiple years, may have been insufficient to fully leverage LSTM's temporal learning capacity.
- Second, the structured nature of the data, which included heterogeneous meteorological and environmental variables, favored the inductive biases of tree-based models that excel in learning complex nonlinear feature interactions without relying on sequential patterns.
- Third, LSTM models are sensitive to hyperparameter tuning, data normalization, and sequence configuration, which may have introduced training instability or hindered generalization. This is evidenced by the higher test error (RMSE = 32.92) and mean bias error (MBE = 11.79) observed in the baseline LSTM model without the Hotspot variable, alongside a statistically significant residual bias ($p < 0.05$). Although the inclusion of Hotspot data substantially improved LSTM performance (R^2 increasing to 0.71, RMSE reducing to 18.07, and residuals becoming statistically non-significant), ensemble models maintained consistently superior performance across both training and testing phases, highlighting their robustness in handling tabular environmental datasets with limited temporal resolution.

4.3 Comparison to related studies

The performance of the proposed ensemble models for PM2.5 prediction in Chiang Mai, Thailand, demonstrates notable accuracy and innovation by leveraging hotspot count variables at varying distances (100–1000 km). This approach is particularly tailored to address Chiang Mai's unique air quality challenges, characterized by biomass burning and transboundary haze, which are less prominent in other geographical regions or studies.

1) Comparison with RF in Thailand [13]

A prior study in Thailand used a RF model with comprehensive inputs—satellite reanalysis (MERRA-2), aerosol data, weather variables, and ground PM2.5 data—to predict daily air pollution. It performed very well overall, especially under varying weather conditions. However, this method depends heavily on large-scale datasets, which may not capture the local and seasonal impact of biomass burning—an important pollution driver in Northern Thailand. Our approach, using ensemble models (GB and XGBoost), focuses directly on fire activity by including hotspot count data from within 100–1000 km. This region-specific information improves predictions during the burning season (March–April), when PM2.5 levels often peak. Key comparison of results are as follows,

Study [13], RF model:

- R^2: 0.98
- RMSE: 3.14 µg/m³
- Input: MERRA-2 reanalysis, meteorological variables, aerosols, ground PM2.5
- Strength: Good for general daily prediction using rich input data
- Limitation: Less effective at capturing fire-related pollution sources

Our study, GB model:

- R^2 (testing): 0.82
- RMSE (testing): 2.93 µg/m³
- Input: Hotspot count (100–1000 km), lagged PM2.5
- Strength: Better for biomass burning season; tracks local and cross-border pollution
- Advantage: High accuracy with simpler, fire-focused inputs

Our model also demonstrated more stable accuracy between training and testing phases, lower residual errors with reduced prediction bias, and stronger relevance to Chiang Mai's seasonal and regional pollution patterns, particularly during high-burning periods.

2) Comparison with Multiple ML Models in the USA [14]

A U.S. study applied multiple machine learning models—ANN, RF, and GB—using diverse inputs such as AOD, meteorological data, land-use characteristics, elevation, and chemical transport model outputs. These models performed strongly due to the availability of rich environmental datasets and comprehensive air quality monitoring systems. In comparison, our study in Chiang Mai focused on a more targeted input: hotspot count data, which directly reflects fire activity, a major source of PM2.5 in Northern Thailand. Despite using fewer and simpler inputs, our ensemble models still achieved competitive results, demonstrating their strength in region-specific and data-limited contexts. Key results comparison:

USA, GB model

- $R^2 = 0.86$
- RMSE = 2.79 µg/m³
- Broad inputs: AOD, weather, land-use, CTM data

Chiang Mai, GB model

- R^2 (training) = 0.97 and R^2 (testing) = 0.82
- RMSE (testing) = 12.19 μg/m³
- Input: Hotspot counts (100–1000 km, lagged days)

While the RMSE is higher in Chiang Mai due to greater PM2.5 variability and fewer input features, the high R^2 shows our model effectively captures pollution patterns, particularly those driven by seasonal biomass burning. This makes our approach well-suited for real-time forecasting in regions with limited data but strong emission seasonality.

3) Comparison with Regional Modeling in China [16]

A study in China modeled PM2.5 concentrations across seven regions using satellite-based aerosol data and spatial clustering. The approach addressed the complexity of a large, heterogeneous area by applying advanced techniques such as data imputation and regional grouping. While effective in handling spatial variability, the model required significant preprocessing and computational resources.

In contrast, our Chiang Mai study used a streamlined method focused on hotspot count variables, which directly indicate biomass burning activity—a key source of PM2.5 in Northern Thailand. Without the need for spatial clustering or complex adjustments, our ensemble models still delivered stronger predictive performance. Key results comparison:

China, regional model

- R^2 = 0.79
- RMSE = 21 μg/m³
- Inputs: Satellite AOD, spatial clustering, imputation

Chiang Mai, GB model

- R^2 (testing) = 0.82
- RMSE (testing) = 12.19 μg/m³
- Input: Hotspot counts (by distance, lagged days)

This shows that hotspot-based modeling can offer higher accuracy with simpler inputs, especially in regions dominated by a single source like biomass burning. It also underscores the practicality of our approach for local-scale PM2.5 prediction in data-limited settings.

4) Comparison with hybrid deep learning CNN-LSTM in Qingdao, China [12]

A study in Qingdao, China, applied a hybrid deep learning model combining CNN and LSTM to predict PM2.5 concentrations. This architecture effectively captured both spatial and temporal features, achieving high predictive performance but requiring more data, computation, and model complexity. In contrast, our Chiang Mai study adopted a simpler and more targeted strategy using hotspot count variables—focused on fire activity across distances and time lags. Despite fewer inputs and lower complexity, our models still delivered competitive results, particularly under biomass-burning conditions. Key results comparison:

Qingdao, CNN-LSTM model

- R^2 = 0.91
- RMSE = 8.216 μg/m³
- Requires complex architecture, high computational cost

Chiang Mai models

- R^2 (range) = 0.71–0.82
- RMSE (range) = 12.19–18.07 μg/m³
- Simple input: Hotspot count variables (by distance and lag)

While the hybrid model outperforms in absolute accuracy, the Chiang Mai models are more practical and interpretable. Their strength lies in computational efficiency and relevance to biomass-burning regions, making them well-suited for real-time monitoring and localized air quality management.

5) Comparison with Municipal PM2.5 Modeling in Shanghai, China [15]

A study in Shanghai used high-resolution air quality data from 16 municipal districts and applied spatial interpolation techniques—Kriging and Inverse Distance Weighting (IDW)—to model PM2.5. These methods performed well due to dense sensor coverage and complex spatial modeling, ideal for large urban environments. In contrast, our Chiang Mai model uses a simpler, fire-focused approach. By relying on hotspot count data (fire detections at various distances), it targets a primary source of pollution—biomass burning—without requiring advanced spatial datasets or urban sensor networks. Key results comparison:

Shanghai model (Kriging/IDW):

- R^2 (training) = 0.90
- RMSE (training) = 3.26 μg/m³
- R^2 (testing) = 0.82
- RMSE (testing) = 3.00 μg/m³
- Requires dense monitoring, urban-scale data, and high computation

Chiang Mai model (GB):

- R^2 (testing) = 0.82
- RMSE (testing) = 12.19 μg/m³
- Uses hotspot data only—regionally targeted and lower complexity

While the Shanghai model excels in a data-rich, urban context, the Chiang Mai approach is more adaptable in regions where pollution is driven by seasonal biomass burning and where monitoring infrastructure is limited. Its simplicity, cost-effectiveness, and relevance to local pollution sources make it especially useful for regional air quality policy and early warning systems.

4.4 Comparison of physics-based and machine learning approaches for PM2.5 prediction

Traditional physics-based models like WRF-Chem and HYSPLIT (used in studies such as [5] and [8]) are commonly used to forecast air pollution. These models simulate how pollutants move through the atmosphere, how they react chemically, and how weather patterns affect their spread. For example, WRF-Chem combines weather data with chemical processes to estimate pollutant levels in both space and time. HYSPLIT, on the other hand, tracks how pollutants travel from their source to other regions, which is especially useful for understanding long-distance pollution from neighboring countries.

Although these models offer detailed and scientifically grounded results, they require a large amount of input data (such as emissions inventories and weather simulations) and need powerful computers to run. This can be a major barrier for real-time forecasting in regions that lack extensive data.

In contrast, the machine learning approach used in this study offers a simpler, data-driven alternative. Instead of relying on complex physical equations, the model learns patterns directly from observed data—such as hotspot counts and past PM2.5 levels. This allows it to predict pollution levels without needing detailed information about weather or chemical reactions. It also runs faster and can work well in areas with limited data.

A key strength of this approach is its ability to account for both local and cross-border pollution. By including hotspot data at different distances and time lags, the model captures how faraway fires still affect local air quality. In fact, results show that adding these distance-based hotspot predictors improves prediction accuracy by about 20%.

While physics-based models are still essential for understanding the full science behind air pollution, machine learning offers a fast, flexible, and cost-effective tool for real-time forecasting—especially in places affected by seasonal burning. In the future, combining both approaches could offer even better accuracy by blending scientific understanding with data-driven prediction.

4.5 Data preprocessing techniques and their impact on PM2.5 prediction

The primary missing data issue occurred in hotspot measurements due to sensor malfunctions, while PM2.5 data remained largely available, particularly during smoke haze events. To ensure data consistency, the rows with missing hotspot values are removed before merging datasets. However, this approach may introduce bias by reducing the number of available training samples, potentially affecting model generalization. Future studies could explore imputation techniques, such as multiple imputation or deep-learning-based approaches, to reconstruct missing hotspot data and minimize data loss.

PM2.5 measurements can be affected by short-term fluctuations due to sensor noise and environmental factors. To mitigate this, in this study, the Kalman filter is used for noise reduction via ‘pykalman’ in Python, which enhances the stability of the time series data. However, the effectiveness of Kalman filtering depends on the accuracy of the process and measurement noise covariances, which could introduce subtle biases if not optimally tuned. Future research could explore adaptive filtering techniques or hybrid approaches combining statistical and deep learning-based denoising methods to further refine PM2.5 predictions.

PM2.5 sensors may exhibit systematic biases due to environmental conditions, calibration errors, or variations in sensor sensitivity. These biases can influence model predictions, particularly in extreme pollution events. While our study relied on standardized sensor data, future work could incorporate sensor fusion techniques, bias correction models, or machine learning approaches to dynamically adjust for sensor errors and improve data reliability.

4.6 Impact of international policies on transboundary pollution

International policies are vital in addressing transboundary pollution, particularly in regions like Southeast Asia, where air pollution regularly crosses national borders. In response, Thailand, Laos, and Myanmar jointly launched such as the CLEAR Sky Strategy (2024–2030) on 29 October 2024, aiming to mitigate regional haze pollution caused by agricultural burning, forest fires, and industrial sources. This strategy outlines coordinated actions such as fire control, hotspot reduction, air quality forecasting, sustainable land management, law enforcement, and inter-agency collaboration. By integrating these elements, the plan enhances regional capacity to address rising PM2.5 levels and improve air quality.

For predictive modeling, the integration of regionally coordinated policy data—such as hotspot counts, fire risk maps, open burning statistics, areas under sustainable agricultural management, and real-time air quality monitoring and forecasting systems—can significantly enhance model accuracy. These data sources reflect measurable outcomes of international policy interventions, such as reductions in fire incidents following enforcement actions or declining PM2.5 trends corresponding to coordinated policy periods. Incorporating both spatial and temporal datasets into predictive models not only improves forecasting performance but also enables systematic evaluation of the linkage between policy measures and environmental or public health outcomes. This is particularly valuable for policy monitoring, formulating effective regional mitigation strategies, and informing long-term public health planning.

4.7 Future study

By addressing these areas, future studies can further refine ensemble models, ensuring greater accuracy and reliability in PM2.5 prediction for improved air quality management.

1. *Adding more input data*: Including more types of information—such as aerosol optical depth (AOD), weather data (like wind and temperature), and land-use patterns (e.g., forests, urban areas)—can help the model capture key PM2.5 factors.
2. *Combining advanced data sources*: Combining satellite and ground-based data through data fusion helps fill gaps and improve the accuracy of pollution estimates, offering a more complete view of air quality across large areas.
3. *Using advanced model structures*: Some models, like CNNs (Convolutional Neural Networks) and LSTMs (Long Short-Term Memory networks), are good at spotting patterns in space and time. Combining them into one model (a hybrid model) could lead to better predictions by capturing both where and when pollution events happen.
4. *Adapting to different regions*: Adding tools like spatial clustering—grouping similar areas together based on geography or pollution behavior—can improve prediction accuracy across diverse environments, such as mountains, valleys, or city centers.
5. *Making models easier to understand and faster to use*: It’s important that these models are not just accurate but also explainable and efficient. Tools from explainable artificial intelligence (XAI), such as in [19], can show which inputs have the biggest effect on predictions. This makes the models more transparent and useful for decision-makers.
6. *Testing the models in new places*: Lastly, trying these models in other regions or countries will help researchers see if the models work well in different environments. This step is important for making sure the approach can be used more widely.

5. Conclusions

This study presented a comprehensive analysis of PM2.5 pollution in Chiang Mai, Thailand, employing advanced ensemble machine learning and deep learning models with hotspot data as a critical predictor. By incorporating hotspot counts from within a

100–1000 km radius, the models effectively captured localized and transboundary pollution dynamics, significantly enhancing prediction accuracy. Gradient Boosting emerged as the most effective model, achieving an R^2 of 0.97 and the lowest error metrics. Key findings revealed that PM2.5 levels near Chiang Mai transition from being dominated by local hotspots within 200 km to transboundary contributions from Myanmar beyond 300 km, emphasizing the need for both localized mitigation and regional cooperation. Hotspots in Thailand significantly influenced PM2.5 levels within 200 km, while Myanmar became the primary contributor beyond 300 km, particularly during the February–April haze season. The analysis further highlighted the cumulative effect of hotspot activity, extending its impact on air quality up to approximately 700 km. Incorporating distance-specific hotspot data, including PM2.5 lag 1 and hotspots at 100 km (lags 1–3), 200–400 km (lags 2–4), and beyond 700 km (lag 5), proved essential for accurate modeling and effective air quality management. These findings underscore the importance of targeted seasonal interventions, regional collaboration, and the integration of spatially resolved fire data in addressing complex air pollution dynamics.

6. References

[1] Jainontee K, Pongkiatkul P, Wang YL, Weng RJF, Lu YT, Wang TS, et al. Strategy design of PM2.5 controlling for Northern Thailand. Aerosol Air Qual Res. 2023;23(6):220432.

[2] Chansuebsri S, Kolar P, Kraisitnitikul P, Kantarawilawan N, Yabueng N, Wiriya W, et al. Chemical composition and origins of PM2.5 in Chiang Mai (Thailand) by integrated source apportionment and potential source areas. Atmos Environ. 2024;327:120517.

[3] Amnuaylojaroen T, Kaewkanchanawong P, Panpeng P. Distribution and meteorological control of PM2.5 and its effect on visibility in Northern Thailand. Atmosphere. 2023;14(3):538.

[4] Sirithian D, Thanatrakolsri P. Relationships between meteorological and particulate matter concentrations (PM2.5 and PM10) during the haze period in urban and rural areas. Northern Thailand. Air Soil Water Res. 2022;15:1-15.

[5] Sirimongkonlertkun N. Assessment of long-range transport contribution on haze episode in Northern Thailand, Laos, and Myanmar. IOP Conf Ser: Earth Environ Sci. 2018;151:012017.

[6] Hongthong A, Nanthapong K, Kanabkaew T. Biomass burning emission inventory of multi-year PM10 and PM2.5 with high temporal and spatial resolution for Northern Thailand. ScienceAsia. 2022;48:302-9.

[7] Wongrin W, Chaisee K, Suphawan K. Comparison of statistical and deep learning methods for forecasting PM2.5 concentration in Northern Thailand. Pol J Environ Stud. 2023;32(2):1419-31.

[8] Thongsame W, Henze DK, Kumar R, Barth M, Pfister G. Evaluation of WRF-Chem PM2.5 simulations in Thailand with different anthropogenic and biomass-burning emissions. Atmos Environ: X. 2024;23:100282.

[9] Chairungreuang C, Wongsathan R. Forecasting PM10 using a deep neural network. Songklanakarin J Sci Technol. 2021;43(3): 687-95.

[10] Mohammadi F, Teiri H, Hajizadeh Y, Abdolahnejad A, Ebrahimi A. Prediction of atmospheric PM2.5 levels using machine learning techniques in Isfahan, Iran. Sci Rep. 2024;14(1):2109.

[11] Kristiani E, Lin H, Lin JR, Chuang YH, Huang CY, Yang CT. Short-term prediction of PM2.5 using LSTM deep learning methods. Sustainability. 2022;14(4):2068.

[12] Bai X, Zhang N, Cao X, Chen W. Prediction of PM2.5 concentration based on a CNN-LSTM neural network algorithm. PeerJ. 2024;12:e17811.

[13] Gupta P, Zhan S, Mishra V, Aekakkararungroj A, Markert A, Paibong S, et al. Machine learning algorithm for estimating surface PM2.5 in Thailand. Aerosol Air Qual Res. 2021;21:210105.

[14] Di Q, Amini H, Shi L, Kloog I, Silvern R, Kelly J, et al. An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution. Environ Int. 2019;130:104909.

[15] Wang Z, Wu X, Wu Y. A spatiotemporal XGBoost model for PM2.5 concentration prediction and its application in Shanghai. Heliyon. 2023;9(12):e22569.

[16] Xiao Q, Chang HH, Geng G, Liu Y. An ensemble machine-learning model to predict historical PM2.5 concentrations in China from satellite data. Environ Sci Technol. 2018;52(22):13260-9.

[17] Buya S, Usanavasin S, Gokon H, Karnjana J. An estimation of daily PM2.5 concentration in Thailand using satellite data at 1-kilometer resolution. Sustainability. 2023;15(13):10024.

[18] United Nations, Asian and Pacific Centre for Transfer of Technology, United Nations Economic and Social Commission for Asia and the Pacific. Technological interventions and gaps in air pollution control in Bangkok. Bangkok: United Nations; 2024.

[19] Dwivedi R, Dave D, Naik H, Singhal S, Omer R, Patel P, et al. Explainable AI (XAI): Core ideas, techniques and solutions. ACM Computing Surveys. 2023;55(9):1-33.