Atmospheric Environment 375 (2026) 122006

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

A comparative study on CMAQ-based predictors and reanalysis AOD for improving $PM_{2.5}$ estimation in Thailand using machine learning

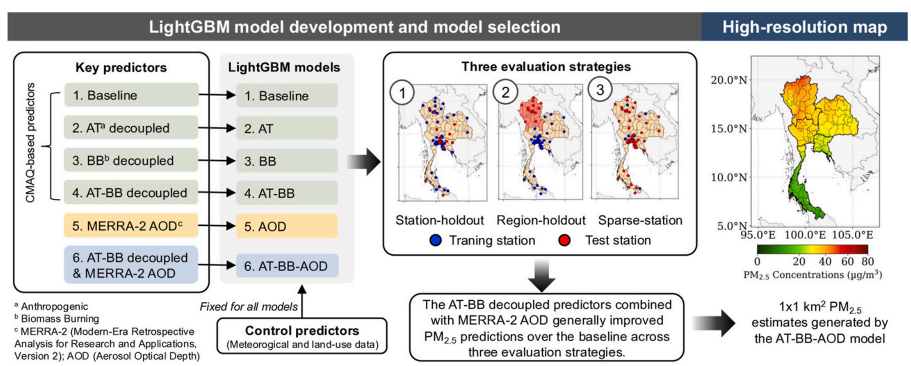
Supitcha Sukprasert [a,*], Pongpisit Thanasutives [b], Shin Araki [a], Katsushige Uranishi [c], Tomohito Matsuo [a], Hikari Shimadera [a]

[a] Graduate School of Engineering, The University of Osaka, 2-1 Yamadaoka, Suita, Osaka, 565-0871, Japan
[b] RIKEN Center for Advanced Intelligence Project (AIP), 1-5 Yamadaoka, Suita, Osaka, 565-0871, Japan
[c] Department of Life Science and Biotechnology, The University of Kitakyushu, 1-1 Hibikino, Wakamatsu, Kitakyushu, 808-0135, Japan

HIGHLIGHTS

- High spatiotemporal $PM_{2.5}$ maps in Thailand are produced by the CMAQ-LightGBM model.
- CMAQ-derived source-decoupled $PM_{2.5}$ contributions are used as LightGBM predictors.
- Three evaluation strategies were used to assess LightGBM spatial generalizability.
- Source-decoupled predictors are most effective when applied within known regions.
- Reanalysis AOD offers a practical alternative when CMAQ simulations are unavailable.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
AOD
Chemical transport model
CMAQ
Machine learning
$PM_{2.5}$
Thailand

ABSTRACT

High-resolution data on fine particulate matter ($PM_{2.5}$) are essential for air quality management and health risk assessments, yet monitoring stations in Thailand remain sparse. This study improves daily $PM_{2.5}$ estimates by identifying effective predictors from the Community Multiscale Air Quality (CMAQ) model and reanalysis Aerosol Optical Depth (AOD) using a Light Gradient Boosting Machine framework. Six predictor sets were tested: (1) one baseline CMAQ predictor with all emission sources, (2–4) three source-decoupled CMAQ predictors isolating individual contributions from anthropogenic (AT), biomass burning (BB), and combined (AT-BB) emissions, (5) one AOD from MERRA-2, and (6) a hybrid AT-BB-AOD predictor. Models were evaluated under station-holdout, region-holdout, and sparse-station strategies. Under the station-holdout evaluation, the AT-BB model outperformed the baseline, and performance further improved when combined with AOD, surpassing all other models. The sparse-station evaluation confirmed that decoupled predictors remained effective even under limited training station coverage. Region-holdout results showed that the effectiveness of decoupled predictors depended on the alignment of feature distributions between training and test regions, and the generalizability of learned feature interactions. Findings suggest that decoupled predictors are most effective

* Corresponding author.
E-mail addresses: supitcha.suks@ea.see.eng.osaka-u.ac.jp (S. Sukprasert), pongpisit.thanasutives@riken.jp (P. Thanasutives), araki@ea.see.eng.osaka-u.ac.jp (S. Araki), uranishi@kitakyu-u.ac.jp (K. Uranishi), matsuo@see.eng.osaka-u.ac.jp (T. Matsuo), shimadera@see.eng.osaka-u.ac.jp (H. Shimadera).

https://doi.org/10.1016/j.atmosenv.2026.122006
Received 9 December 2025; Received in revised form 12 March 2026; Accepted 4 April 2026
Available online 5 April 2026
1352-2310/

when applied within known regions. For out-of-domain applications, the inclusion of training data from regions with diverse emission profiles and meteorological conditions is recommended to fully exploit the benefits of decoupled predictors. The AOD model achieved comparable accuracy to that of the baseline, serving as a practical alternative when CMAQ simulations are unavailable; however, it showed greater underestimation at high $PM_{2.5}$ concentrations. Therefore, CMAQ-based predictors with source decoupling are preferred for high-accuracy $PM_{2.5}$ estimation.

1. Introduction

Fine particulate matter ($PM_{2.5}$) is a major pollutant of concern due to its ability to carry toxic components, such as heavy metals and polycyclic aromatic hydrocarbons, contributing to adverse health outcomes (Chen et al., 2024a; Wang et al., 2025). Thailand has long struggled with elevated $PM_{2.5}$ concentrations. In 2024, daily averages ranged from 21.9 to 218.6 μg/m^3, with most values exceeding the national standard of 37.5 μg/m^3, equivalent to the World Health Organization's interim target 3 (Pollution Control Department, 2025). Anthropogenic (AT) emissions are a dominant source of $PM_{2.5}$ in Thailand, particularly in densely populated areas (e.g., Bangkok), while biomass burning (BB) is another major contributor (Hassan Bran et al., 2024; Thongsame et al., 2025). BB-derived $PM_{2.5}$ arises both locally and from transboundary transport from Laos, Myanmar, and Cambodia (Chantaraprachoom et al., 2024). BB-derived $PM_{2.5}$ exacerbates severe pollution episodes in Northern Thailand during the peak burning season (March to April), when daily averages can exceed 300 μg/m^3 in Chiang Rai (Bran et al., 2022).

$PM_{2.5}$ particles are small enough to penetrate deep into the lungs, causing respiratory irritation, inflammation, and an increased risk of lung cancer (Chen et al., 2024a). Toxic substances adsorbed onto their surfaces enter the bloodstream, contributing to cardiovascular diseases through mechanisms such as oxidative stress and inflammation (Feng et al., 2023). Health Risk Assessment (HRA) studies in Thailand's industrial and traffic zones have detected heavy metals (e.g., Cr, Ni, and Mn) in $PM_{2.5}$, highlighting potential health risks to the population (Ahmad et al., 2022; Kawichai et al., 2022). In Chiang Mai province, Northern Thailand, elevated $PM_{2.5}$ exposure has been strongly linked to hospital admissions for dermatitis, conjunctivitis, stroke, and lung cancer (Jarernwong et al., 2023). Yet, many HRA studies rely on province- or city-level averages from sparse Air Quality Monitoring (AQM) networks, overlooking local variability. This can lead to either underestimation or overestimation of health risks. Higher-resolution $PM_{2.5}$ data are therefore essential to strengthen HRA studies and guide more effective policy development.

Chemical Transport Models (CTMs) have long been used to address the limitation of sparse monitoring data, as they provide spatially and temporally continuous estimates of air pollutants, including $PM_{2.5}$. However, CTMs depend on external inputs such as boundary conditions, meteorology, and emission inventories, all of which can introduce biases (Chen et al., 2023; Solazzo and Galmarini, 2016; Thongsame et al., 2024; Xu et al., 2021). In addition, biases can arise from physical parameterizations (e.g., chemical mechanisms and deposition) and numerical parameterizations (e.g., grid resolution and time step) within the model itself (Mallet and Sportisse, 2006; Solazzo and Galmarini, 2016). To address these limitations, machine learning (ML) approaches have been increasingly applied over the past decade. In $PM_{2.5}$ estimation, a common strategy is to use CTM-simulated $PM_{2.5}$ as a predictor in ML models, usually supplemented with meteorological and land-use information. Among ML methods, tree-based algorithms such as Random Forest and Gradient Boosting are commonly employed for their efficiency and simplicity (Bi et al., 2022; Tang et al., 2024; Thongthammachart et al., 2021, 2023).

Previous studies have relied on CTM-simulated $PM_{2.5}$ as a standalone predictor, overlooking contributions from individual emission sources (Bi et al., 2022; Tang et al., 2024; Thongthammachart et al., 2021, 2023), which prevents those models from reaching their best possible performance. To address it, this study introduces, for the first time, a source-decoupling approach that isolates emission contributions using the CTM. By isolating the contributions of different emission sources to $PM_{2.5}$, this decoupling technique provides a more expressive representation of $PM_{2.5}$ concentrations to improve the prediction accuracy of an ML model.

Satellite-derived aerosol optical depth (AOD) offers a practical alternative for estimating ground-level $PM_{2.5}$ concentration when CTM simulations are unavailable, as it is readily accessible and generally correlated with $PM_{2.5}$. AOD has served as a key predictor in statistical and ML models (Chen et al., 2024b; Chen et al., 2021; Liu et al., 2009; Ma et al., 2019), but its utility is often constrained by data gaps from cloud cover and high surface reflectance (Chen et al., 2021; Liu et al., 2009). In regions with poor AOD coverage, its predictive contribution becomes negligible (Chen et al., 2021; Karimian et al., 2023). Missing AOD values are typically filled using geostatistical or ML techniques. Geostatistical approaches rely on the spatiotemporal relationships within the AOD data itself, which often results in lower coverage and limited accuracy (Chen et al., 2020). In contrast, certain ML models can achieve higher coverage and better interpolation quality, yet they require incorporating external data, such as meteorological or cloud fraction information (Chen et al., 2020). Reanalysis AOD products, despite their relatively coarse spatial resolution, offer complete coverage. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) provides a reanalysis AOD product that has shown good agreement with ground-based observations (Gueymard and Yang, 2020; Ohno et al., 2022; Shi et al., 2019), and can therefore be used to estimate surface-level $PM_{2.5}$ concentrations. Nevertheless, the evidence on the feasibility of using reanalysis AOD as a substitute for CTM simulations to estimate surface-level $PM_{2.5}$ in Thailand remains limited.

This study employs source-decoupling to separate AT and BB emission contributions using the Community Multiscale Air Quality (CMAQ) model and incorporates them as key bias-correcting predictors in a Light Gradient Boosting Machine (LightGBM) framework for daily $PM_{2.5}$ estimation in Thailand. Six LightGBM models were developed and compared using varying key predictor sets: (1) Baseline, (2) BB-decoupled, (3) AT-decoupled, (4) AT-BB-decoupled, (5) AOD, and (6) AT-BB-AOD. Model performance was rigorously assessed under three evaluation strategies: a station-holdout evaluation examining full-period, seasonal, high-$PM_{2.5}$, and regional performance; a region-holdout evaluation testing generalizability to completely unseen regions; and a sparse station evaluation investigating robustness under progressively reduced training station coverage. This study therefore represents the first effort to develop a robustly validated model for high-resolution daily $PM_{2.5}$ estimates across Thailand using high-fidelity monitoring data from government agencies and multiple evaluation strategies to assess both predictive accuracy and spatial generalizability under various conditions reflective of real-world use cases.

2. Methodology

2.1. Study area and ground-based air quality measurements

The study area covers Thailand, divided into five regions (Central, Eastern, Northern, Northeastern, and Southern) based on the Thai

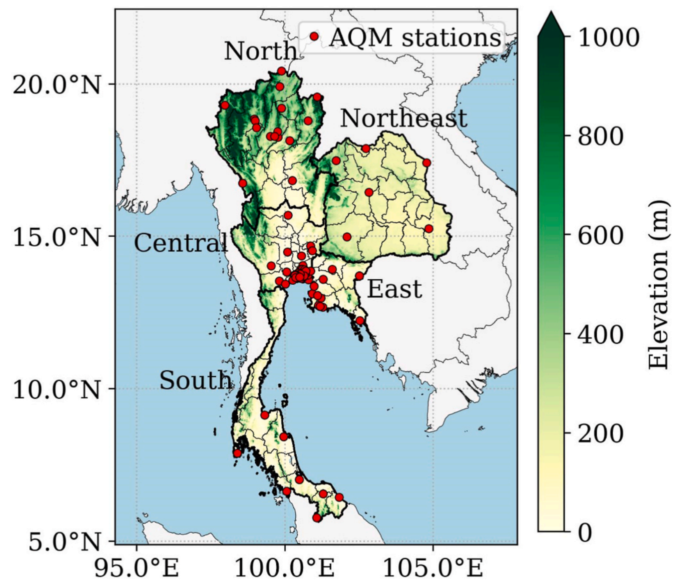


Fig. 1. The study area.

Table 1
Summary of predictor variables.

Predictor variables	Unit	Original resolution[a]
Key predictors		
Reanalysis Aerosol Optical Depth (AOD)	Unitless	0.5° × 0.625°
Baseline (CMAQ_PM25)	μg/m^3	15 km
BB-decoupled (CMAQ_PM25_BB and CMAQ_PM25_No_BB)	μg/m^3	15 km
AT-decoupled (CMAQ_PM25_AT and CMAQ_PM25_No_AT)	μg/m^3	15 km
AT-BB decoupled (CMAQ_PM25_BB, CMAQ_PM25_AT, and CMAQ_PM25_Othr)	μg/m^3	15 km
AT-BB-AOD (AT-BB decoupled and AOD)	μg/m^3, unitless	15 km, 0.5° × 0.625°
Control predictors		
Major road density	km/km^2	-
Minor road density	km/km^2	-
Built-up area ratio	Unitless	10 m
Forest cover ratio	Unitless	30 m
Rice-harvested area	ha	5 arcmin
Population density	people/km^2	1 km
Elevation	m	90 m
WRF_PBL (WRF-simulated planetary boundary layer)	m	15 km
WRF_WSPD (WRF-simulated wind speed)	m/s	15 km
WRF_RH (WRF-simulated relative humidity)	%	15 km
WRF_Temp (WRF-simulated ground-level temperature)	K	15 km

[a] All predictors were resampled to a uniform 1-km spatial resolution before LightGBM development.

Meteorological Department classification (Fig. 1). The Central region is a lowland sloping southward to the Gulf of Thailand, bordered by western mountain ranges. The Eastern region combines plains and mountains, with coastlines along the Gulf to the west and south. The Northern region is mountainous with north–south high ranges forming the headwaters of major rivers. The Northeastern region is a high plain sloping southeast, where rainfall is lower in the west due to a rain shadow from surrounding ranges. The Southern region forms a narrow peninsula between the Gulf of Thailand and the Andaman Sea, characterized by coastal plains and high rainfall throughout the year.

Ground-based measurements of $PM_{2.5}$, wind speed, relative humidity, and temperature were obtained from the Pollution Control

Table 2
Summary of CMAQ simulation cases and emission decoupling calculations.

	Description	Label
CMAQ running case		
C1	All emissions included	CMAQ_PM25
C2	AT emissions excluded in D2	CMAQ_PM25_No_AT
C3	BB emissions excluded in D2	CMAQ_PM25_No_BB
Decoupling individual emission contributions		
C1 – C2	AT contributions	CMAQ_PM25_AT
C1 – C3	BB contributions	CMAQ_PM25_BB
C3 – (C1 – C2)	Other contributions	CMAQ_PM25_Othr

Department and Bangkok Metropolitan Administration for 2019 to 2021. Hourly observations were aggregated into daily averages, calculated only for days with more than 75% hourly data availability. Hourly $PM_{2.5}$ concentrations over 450 μg/m^3, accounting for around 0.00047% of all data, were considered outliers and removed before averaging. To ensure annual representativeness, AQM data from a station in a given year were omitted if the annual coverage of daily $PM_{2.5}$ samples for that year was below 70%. Remaining daily samples with missing $PM_{2.5}$ values (~2.29% of total samples) were discarded. In total, 103,639 samples were retained. Summary statistics of the measured data are provided in Table S1.

2.2. Predictor variables

Six LightGBM models were developed, each incorporating a distinct set of key predictors while control predictors were held constant. Further details of the LightGBM development are provided in Section 2.3. A complete list of all predictors is summarized in Table 1, with details of their preparation outlined below.

2.2.1. WRF and CMAQ simulations

Meteorological variables were simulated using Weather Research and Forecasting (WRF) v.4.3 and provided as inputs to CMAQ and LightGBM models. CMAQ v5.3.3 was used to estimate $PM_{2.5}$ concentrations. The WRF-CMAQ used two nested domains: an Asia domain (D1) at a 45 km grid resolution and a Thailand domain (D2) at a 15 km grid resolution (Fig. S1). The simulated data from D2 were used in this study. The WRF-CMAQ model configurations (i.e., WRF physics schemes, CMAQ chemical mechanisms, and emission inputs) followed Chantaraprachoom et al. (2024), except for the BB emissions, which used Fire INventory from NCAR v.2.5. The WRF and CMAQ simulations were evaluated against ground-based measurements, and their performance metrics are provided in Table S2.

The CMAQ model was run under three emission scenarios: (C1) all emissions, (C2) AT emissions excluded in D2, and (C3) BB emissions excluded in D2. Outputs from these runs were used to decouple individual emission contributions, following the calculation methods and labels summarized in Table 2. Daily averages of baseline CMAQ predictors and three contributions (AT, BB, and others) are shown with observed $PM_{2.5}$ concentrations in Fig. S2.

2.2.2. AOD data

Due to a high proportion of missing satellite-retrieval AOD values in this region (Chen et al., 2024b), reanalysis AOD was employed. The M2I3NXGAS collection from MERRA-2, produced by NASA's Global Modeling and Assimilation Office using the Goddard Earth Observing System Model v5.12.4, was utilized. The dataset has a 3-h temporal resolution and was aggregated to daily values. Further details can be found in Randles et al. (2017).

2.2.3. Auxiliary data

Road network data were obtained from OpenStreetMap (OpenStreetMap contributors, 2019) and classified into five categories:

Table 3
Hyperparameter search space.

Hyperparameter	Ranges	Sampling distribution
Learning rate (learning_rate)	0.01-0.3	Log-uniform
Maximum depth (max_depth)	4-10	Uniform
Number of boosting iterations (n_estimators)	100-1000	Uniform
Maximum number of leaves in one tree (num_leaves)	16-128	Uniform
Minimal number of data in one leaf (min_child_samples)	50-200	Uniform
Bagging (subsample)	0.5-0.8	Uniform
Frequency of bagging (subsample_freq)	1-10	Uniform
L1 regularization (reg_alpha)	0.1-10	Log-uniform
L2 regularization (reg_lambda)	0.1-20	Log-uniform

trunk, motorway, primary, secondary, and tertiary roads. They were further grouped into major roads (trunk, motorway, and primary roads) and minor roads (secondary and tertiary roads). The data were rasterized to a 1 km grid resolution, where each grid cell represented road density in km/km^2.

Other auxiliary datasets included population density (PD), elevation, and land use variables (built-up area, forest area, and rice-harvested area). PD data were sourced from the WorldPop Research Program at the University of Southampton (WorldPop and Bondarenko, 2020), and elevation data were obtained from NASA's Shuttle Radar Topography Mission v4 (Jarvis et al., 2008). The built-up area dataset was acquired from Pesaresi (2022), which utilized a Sentinel-2 global image composite for the reference year 2018. Forest area and rice-harvested area data were obtained from Hansen et al. (2013) and Yu et al. (2020), respectively.

2.2.4. Spatial data alignment

All spatial datasets, including WRF simulations, CMAQ simulations, AOD, and auxiliary data, were reprojected to a common coordinate reference system and resampled to a uniform 1 km grid using a bilinear interpolation method. The datasets were also cropped to a consistent spatial extent to ensure spatial alignment. Additionally, a focal-sum operation with an inverse distance-weighted matrix was applied to the road density, population, and land-use variables to account for the distance decay effect (Araki et al., 2018).

2.3. LightGBM model development

2.3.1. LightGBM configuration and hyperparameter tuning

Six models were developed: one baseline and five proposed models, including the AOD, BB, AT, AT-BB, and AT-BB-AOD models, each incorporating its corresponding key predictor as listed in Table 1. All models were implemented in Python v3.11 using the LightGBM library (Ke et al., 2017). Hyperparameter (HP) tuning was performed using the baseline and control predictors. A Bayesian optimization algorithm from the scikit-optimize library (Head et al., 2024) was applied, with the HP search space defined in Table 3. Model performance was assessed using the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). In addition, the fitted regression slope between observed (x-axis) and predicted (y-axis) $PM_{2.5}$ concentrations was calculated to investigate overestimation and underestimation.

To comprehensively examine model robustness, three evaluation strategies were implemented: station-holdout, region-holdout, and sparse-station. All strategies employed the same six LightGBM model configurations, evaluation metrics, and HP search space, but differed primarily in their data partitioning.

2.3.2. Station-holdout evaluation

This evaluation assesses the performance of each model on unseen stations. Training and test stations were drawn from all regions and followed similar regional proportions, reflecting the application scenario in which models are trained using existing stations and applied to other locations within the same domain. Fig. 2 illustrates the station-holdout evaluation. The dataset was spatially split by stations into

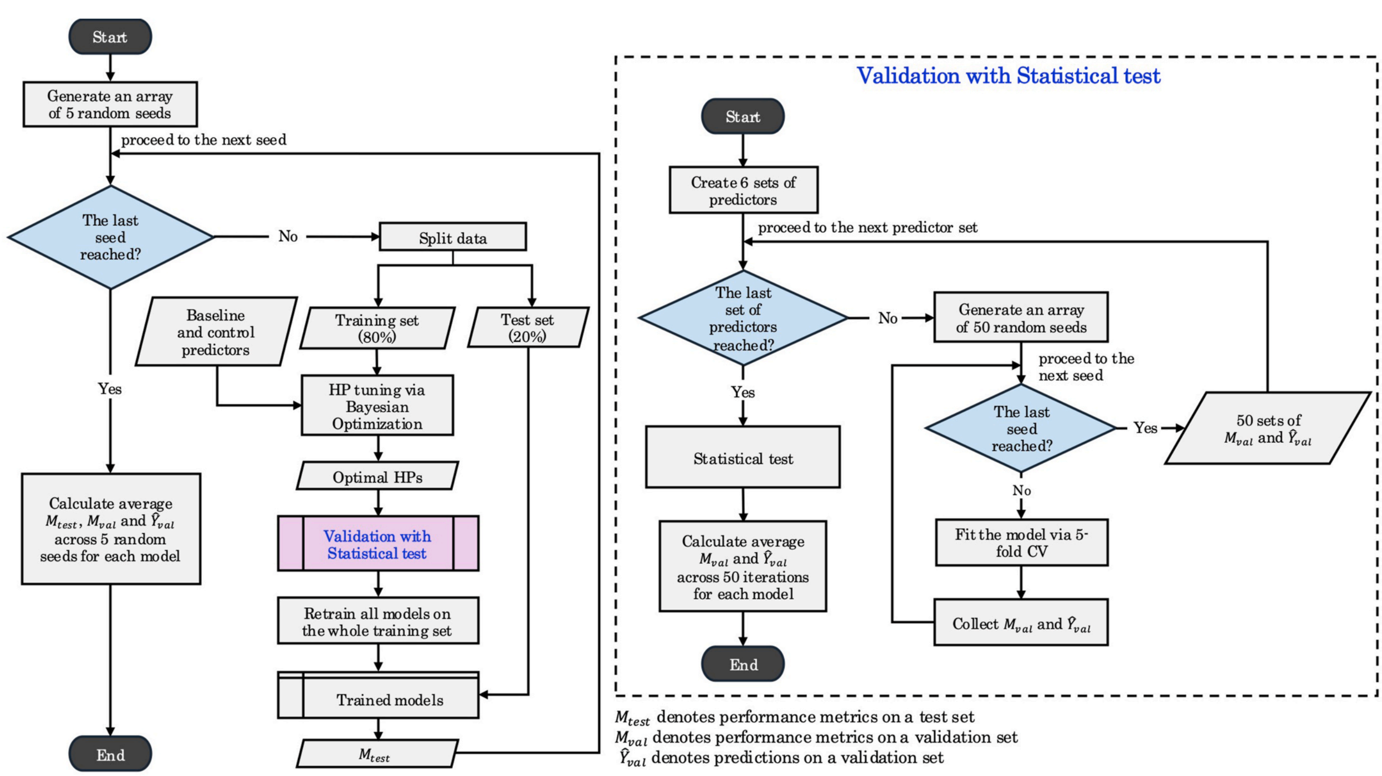


Fig. 2. Station-holdout evaluation.

Table 4
Validation performance of LightGBM models under the station-holdout evaluation across full-period, seasonal, and high-$PM_{2.5}$ perspectives.

Perspective	Model	Overall CV				Spatial CV			
		R^2	RMSE ($\mu g/m^3$)	MAE ($\mu g/m^3$)	Slope	R^2	RMSE ($\mu g/m^3$)	MAE ($\mu g/m^3$)	Slope
Full period	AOD	*0.810*	*8.18*	*5.18*	*0.817*	*0.703*	*10.24*	*6.41*	*0.724*
	Baseline	0.820	7.96	4.93	0.841	0.746	9.47	6.01	0.753
	AT	**0.829**	**7.77**	**4.78**	**0.848**	**0.753**	**9.34**	**5.91**	**0.756**
	BB	**0.838**	**7.56**	**4.66**	**0.852**	**0.762**	**9.17**	**5.80**	**0.767**
	AT-BB	**0.846**	**7.37**	**4.53**	**0.859**	**0.768**	**9.04**	**5.69**	**0.770**
	AT-BB-AOD	**0.860***	**7.02***	**4.38***	**0.863***	**0.779***	**8.83***	**5.58***	**0.778***
Wet season	AOD	*0.624*	*4.48*	*3.25*	*0.604*	*0.393*	*5.70*	*4.25*	*0.426*
	Baseline	0.677	4.15	3.01	0.660	0.459	5.38	4.00	0.486
	AT	**0.683**	**4.12**	**2.99**	**0.665***	0.464	5.35	3.99	**0.489***
	BB	**0.683**	**4.12**	**2.98**	*0.642*	0.458	5.38	3.99	*0.465*
	AT-BB	**0.688**	**4.08**	**2.96**	*0.647*	0.462	5.36	**3.98**	*0.467*
	AT-BB-AOD	**0.700***	**4.00***	**2.91***	*0.654*	**0.480***	**5.27***	**3.92***	*0.483*
Dry season	AOD	*0.767*	*10.04*	*6.57*	*0.787*	*0.636*	*12.54*	*7.98*	*0.681*
	Baseline	0.776	9.84	6.33	0.811	0.691	11.56	7.47	0.712
	AT	**0.788**	**9.58**	**6.08**	**0.820**	**0.700**	**11.39**	**7.30**	**0.715**
	BB	**0.801**	**9.29**	**5.89**	**0.823**	**0.713**	**11.14**	**7.11**	**0.726**
	AT-BB	**0.812**	**9.03**	**5.68**	**0.833**	**0.722**	**10.96**	**6.94**	**0.732**
	AT-BB-AOD	**0.830***	**8.57***	**5.45***	**0.839***	**0.735***	**10.70***	**6.78***	**0.741***
High $PM_{2.5}$	AOD	*0.506*	*15.54*	*10.87*	*0.754*	*0.185*	*19.96*	*13.67*	*0.568*
	Baseline	0.512	15.46	10.57	0.792	0.310	18.33	12.70	0.632
	AT	**0.537**	**15.06**	**10.12**	**0.803**	**0.331**	**18.06**	**12.36**	0.636
	BB	**0.564**	**14.61**	**9.80**	**0.796**	**0.361**	**17.64**	**11.98**	**0.639**
	AT-BB	**0.588**	**14.22**	**9.40**	**0.807**	**0.379**	**17.40**	**11.67**	**0.643**
	AT-BB-AOD	**0.631***	**13.45***	**9.02***	**0.813***	**0.407***	**16.99***	**11.39***	**0.651***

Bold values denote statistically significant improvements over the baseline model at the 95% CI ($p_c < 0.05$).
Italic values denote statistically significant worse performance than the baseline model at the 95% CI ($p_c < 0.05$).
A superscript star (*) indicates the best-performing model at the 95% CI ($p_c < 0.05$).

training (80%) and test (20%) sets using stratified random sampling to ensure regional balance (Fig. S3). This procedure was repeated with five random seeds to ensure reliable results. The test set was employed to assess models' predictive ability on unseen stations. Models were evaluated using spatial cross-validation (CV) with folds split by stations, as well as overall CV (sample-based splitting).

To reduce random effects, both CV approaches were repeated 50 times with different randomizations. This produced 50 sets of predictions and validation metrics for each model. These scores were statistically analyzed to answer two key questions: (1) whether each proposed model outperforms the baseline, and (2) which model among the six achieves the best performance. Statistical testing employed repeated measures analysis of variance and Tukey's honest significant difference test at the 95% Confidence Interval (CI). Because p-values across five seeds were not fully independent due to overlapping training data, they were aggregated into a single combined p-value (p_c) using the harmonic mean method (Wilson, 2019), defined as:

$$p_c = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{p_i}},$$

where w_i is the weight assigned to the p-value (p) from the seed i, with $\sum_{i=1}^{n} w_i = 1$. In this study, all weights were set equally to $1/n$, where n denotes the total number of seeds.

Final model predictions and validation metrics were averaged across all CV repetitions and five seeds. Model performance was evaluated from four perspectives: full-period, seasonal, high-$PM_{2.5}$, and regional. The full-period perspective used all predicted values. The seasonal perspective divided model predictions into wet (mid-May to mid-October) and dry (mid-October to mid-May) seasons. High-$PM_{2.5}$ perspective was defined as daily samples with $PM_{2.5}$ concentrations $\geq 37.6\ \mu g/m^3$, corresponding to the unhealthy category under the Thai Air Quality Index. The regional perspective, meanwhile, assessed model performance separately for the five regions defined in Section 2.1. Table S3 summarizes the number of samples for each perspective.

2.3.3. Region-holdout evaluation

This evaluation excludes an entire region from model training, e.g., all stations in the North are excluded. It provides a strict spatial generalization test to assess whether the proposed predictors remain effective in an unseen region, where $PM_{2.5}$ source contributions may differ. The dataset was partitioned by region, with one region reserved exclusively for testing and the remaining regions used for training. HPs were independently optimized for each train–test split to prevent spatial information leakage. Bayesian optimization was repeated with five random seeds, and performance metrics on the held-out region were averaged across seeds to obtain stable performance estimates.

2.3.4. Sparse-station evaluation

In this evaluation, the proportion of stations used for training was progressively reduced to simulate sparse station conditions. The dataset was spatially split by stations into training and test sets using four training ratios (10%, 25%, 50%, and 75% of stations). Stratified random sampling was applied to preserve regional balance within each split (Fig. S4). HPs were independently tuned for each train-test split. The entire procedure was repeated five times with different random seeds, and mean evaluation metrics on the test set were reported.

3. Results and discussion

3.1. Performance under the station-holdout evaluation

3.1.1. Full-period perspective

The AT-BB-AOD model achieved the highest performance under both overall and spatial CV, followed by the AT-BB model (Table 4). All proposed models, except the AOD model, significantly outperformed the baseline across both CV aspects. All models exhibited slopes of less than

Table 5
Regional validation performance of LightGBM models under the station-holdout evaluation.

Region	Model	Overall CV				Spatial CV			
		R^2	RMSE (μg/m^3)	MAE (μg/m^3)	Slope	R^2	RMSE (μg/m^3)	MAE (μg/m^3)	Slope
Central	AOD	*0.796*	*7.07*	*4.96*	*0.766*	*0.705*	*8.51*	*6.14*	*0.690*
	Baseline	0.812	6.78	4.70	0.793	0.724	8.23	5.91	0.718
	AT	**0.829**	**6.46**	**4.49**	**0.808**	**0.741**	**7.96**	**5.74**	**0.733**
	BB	**0.838**	**6.29**	**4.39**	**0.812**	**0.749**	**7.84**	**5.66**	**0.738**
	AT-BB	**0.853**	**6.00**	**4.22**	**0.827**	**0.766**	**7.58**	**5.52**	**0.753**
	AT-BB-AOD	**0.859***	**5.86***	**4.13***	**0.830***	**0.774***	**7.45***	**5.42***	**0.759***
East	AOD	*0.770*	*6.14*	*4.35*	*0.735*	*0.719*	*6.78*	*4.87*	*0.681*
	Baseline	0.807	5.62	3.93	0.810	0.761	6.25	4.32	0.787
	AT	**0.813**	**5.53**	**3.85**	*0.805*	**0.766**	**6.18**	**4.25**	*0.767*
	BB	**0.815**	**5.49**	**3.84**	**0.821***	**0.764**	**6.21**	**4.29**	**0.809***
	AT-BB	**0.826**	**5.32**	**3.75**	**0.817**	**0.778**	**6.02**	**4.17**	**0.791**
	AT-BB-AOD	**0.831***	**5.24***	**3.70***	**0.815**	**0.785***	**5.92***	**4.12***	**0.787**
North	AOD	**0.825**	**12.80**	*7.01*	*0.873*	*0.688*	*17.06*	*8.95*	*0.745*
	Baseline	0.824	12.81	6.93	0.891	0.760	14.96	8.01	0.777
	AT	**0.826**	**12.77**	**6.88**	*0.890*	0.757	15.04	8.05	*0.768*
	BB	**0.836**	**12.38**	**6.55**	**0.894**	**0.768**	**14.69**	**7.71**	**0.785**
	AT-BB	**0.838**	**12.31**	**6.50**	**0.894**	**0.765**	**14.78**	**7.73**	0.777
	AT-BB-AOD	**0.862***	**11.34***	**6.05***	**0.901***	**0.779***	**14.34***	**7.46***	**0.786****
Northeast	AOD	*0.764*	*10.05*	*6.31*	*0.762*	*0.656*	*12.12*	*7.90*	*0.696*
	Baseline	0.779	9.70	6.06	0.769	0.677	11.74	7.46	0.699
	AT	**0.783**	**9.62**	**6.04**	**0.774**	*0.671*	*11.86*	*7.53*	0.699
	BB	**0.795**	**9.34**	**5.81**	**0.783**	**0.693**	**11.44**	**7.26**	**0.710**
	AT-BB	**0.801**	**9.21**	**5.75**	**0.785**	**0.690**	**11.50**	**7.27**	**0.706**
	AT-BB-AOD	**0.807***	**9.07***	**5.59***	**0.804***	**0.702***	**11.28***	**7.13***	**0.735***
South	AOD	*0.504*	*4.73*	*3.37*	*0.472*	*0.242*	*5.83*	*4.28*	*0.401*
	Baseline	0.560	4.46	3.23	0.549	0.413	5.14	3.85	0.492
	AT	*0.551*	*4.51*	*3.25*	*0.520*	0.392	5.23	3.88	*0.439*
	BB	*0.549*	*4.52*	*3.27*	**0.564**	*0.391*	*5.24*	3.89	**0.511***
	AT-BB	*0.544*	*4.54*	*3.27*	*0.534*	*0.365*	*5.35*	3.91	*0.445*
	AT-BB-AOD	**0.605***	**4.23***	**3.11***	**0.565***	0.421**	5.11**	**3.80****	*0.470*

Bold values denote statistically significant improvements over the baseline model at the 95% CI ($p_c < 0.05$).
Italic values denote statistically significant worse performance than the baseline model at the 95% CI ($p_c < 0.05$).
A superscript star (*) indicates the best-performing model at the 95% CI ($p_c < 0.05$). If the best-performing model cannot be clearly determined from the statistical tests, the double star superscript (**) indicate the model with the best score on average.

one with positive intercepts (Table S4), indicating underestimation at high $PM_{2.5}$ concentrations. However, slopes tended to increase as model complexity increased from the AOD model through the baseline, AT, BB, and AT-BB to AT-BB-AOD, suggesting that incorporating decoupled predictors helps reduce this underestimation.

The use of decoupled predictors, particularly AT-BB decoupled predictors, significantly improved model performance across the country. Source decoupling isolates contributions from individual emission sources, revealing distinct spatiotemporal patterns of $PM_{2.5}$ and providing a more informative representation of $PM_{2.5}$ variability. For example, BB emissions (and their BB-decoupled contributions) are typically high during the dry season, especially from March to April (Fig. S2). As a result, the LightGBM model receives clearer input signals, allowing it to learn the complex interplay between these signals (i.e., unique source contributions) and control predictors to effectively enhance predictive performance. Decoupled predictors with unique source contribution (CMAQ_PM25_BB and CMAQ_PM25_AT) consistently ranked among the top three predictors in the importance ranking for the decoupled models (AT, BB, AT-BB, and AT-BB-AOD models), further supporting this explanation (Fig. S5).

Adding AOD to the AT-BB model provided additional improvement, suggesting that AOD complements the decoupled predictors, which capture surface-level $PM_{2.5}$ concentrations, by adding total column aerosol information to the model. In the AT-BB model, the top three important predictors were CMAQ_PM25_BB, CMAQ_PM25_Othr, and CMAQ_PM25_AT. In the AT-BB-AOD model, AOD replaced CMAQ_PM25_Othr in the top three (Fig. S5), indicating that LightGBM learns to use AOD synergistically with decoupled predictors, leading to improved generalization. Nevertheless, meteorological variables also play an important role. The control WRF predictors were among the top five in the feature importance ranking, signifying the contribution of meteorological conditions to $PM_{2.5}$ estimations.

Chen et al., 2024b developed a stacked model combining a Convolutional Neural Network and LightGBM model to predict daily $PM_{2.5}$ in the Mekong River Basin (MRB), which includes Thailand. Their study reported better overall CV scores (R^2 of 0.92, an RMSE of 5.58 μg/m^3, and an MAE of 3.44 μg/m^3), but worse spatial CV scores (R^2 of 0.72, an RMSE of 10.21 μg/m^3, and an MAE of 6.74 μg/m^3) compared with this work. It is important to note that the overall CV is less constrained in its data splitting and tends to overestimate performance, whereas the spatial CV provides a more realistic model evaluation. This is because spatial CV evaluates the model on unseen stations. However, under the station-holdout framework, all regions remain represented during training; thus, spatial CV reflects generalization across stations within known regional structures. Within this evaluation, our models with the CMAQ-based predictors are expected to yield more reliable $PM_{2.5}$ estimates.

3.1.2. Seasonal and high-$PM_{2.5}$ perspectives

In the dry season, all proposed models, except the AOD model, significantly outperformed the baseline under both CV aspects. The AT-BB-AOD model achieved the highest performance, consistent with full-period results. This confirms that decoupled predictors are particularly effective in this season, which is characterized by stronger emission signals.

In contrast, inconsistent results were observed in the wet season.

Overall CV indicated significant improvements for AT, BB, AT-BB and AT-BB-AOD models compared to the baseline, whereas spatial CV revealed significant improvements only for the AT-BB-AOD model and partially for the AT-BB model. Moreover, the performance drop from overall CV to spatial CV was more pronounced in the wet season. This can be attributed to three factors: First, the spatial CV is inherently more challenging, as discussed in Section 3.1.1. Second, during the wet season, the southwest monsoon brings widespread rainfall, which enhances wet deposition and reduces both the magnitude and spatial variability of $PM_{2.5}$ concentrations. This condition makes it more difficult for the model to learn distinct spatial patterns of $PM_{2.5}$. Third, the predictive power of decoupled predictors is diminished because emission variability is generally lower during the wet season (i.e., reduced BB activities), with this reduction reflected in the emission inventories.

During high-$PM_{2.5}$ episodes, decoupled models again outperformed the baseline, with AT-BB-AOD achieving the best performance. As high-$PM_{2.5}$ episodes are typically dominated by strong emission signals, incorporating decoupled predictors improves the model's ability to capture these high $PM_{2.5}$ concentrations. In contrast, the AOD model underperformed the baseline during these critical periods.

3.1.3. *Regional perspective and independent test performance*

The AOD model generally performed worse than the baseline in all regions under both CV aspects, except for R^2 and RMSE values under overall CV in the Northern region (Table 5). In contrast, the AT-BB-AOD model significantly outperformed the baseline in all regions under both CV aspects, except for spatial-CV R^2 and RMSE in the South. The AT-BB model also consistently outperformed the baseline in the Central, Eastern, Northern, and Northeastern regions.

In the Central region, where both AT and BB emissions contribute comparably to $PM_{2.5}$ concentrations (Fig. S6), applying decoupled predictors yielded the largest improvement in performance. In the Eastern region, where both emission sources contribute to $PM_{2.5}$ but BB is more dominant, model performance also improved when using decoupled predictors. These results demonstrate the effectiveness of source decoupling in regions with mixed emission sources.

In the Northern region, where BB dominates and AT contribution is minimal, no significant differences were found between the baseline and AT models, or between BB and AT-BB models under spatial CV. When one source overwhelmingly dominates, decoupling the weaker source adds limited predictive value.

The Northeastern region presents a notable contrast under the spatial CV. Despite having a contribution profile similar to the Eastern region, the BB model significantly outperformed the AT-BB model in terms of R^2 and RMSE. Moreover, the AT model even performed worse than the baseline. This may be attributed to data imbalance, as most training samples are concentrated in the Central region, leading the LightGBM model to learn AT-related patterns characteristic of that area. CMAQ_PM25_AT patterns in the Eastern and Northern regions correlate more strongly with the Central region (Pearson's $r = 0.77$ and 0.60, respectively), than CMAQ_PM25_AT patterns in the Northeastern region correlate with the Central region (Pearson's $r = 0.40$). Combined with a limited sample size ($N = 4220$), the model has insufficient information to learn AT patterns in the Northeast.

In the Southern region, the baseline model outperformed BB and AT-BB models in terms of R^2 and RMSE under spatial CV. The AT-BB-AOD model did not significantly differ from the baseline in spatial-CV R^2 and RMSE. R^2 values were substantially lower than other regions, consistent with the study by Chen et al., 2024b. These outcomes can be attributed to unique climatic and geographic characteristics. The southern peninsula experiences a prolonged wet season from May to December (Chokngamwong and Chiu, 2008). Unlike the rest of the country, which is primarily affected by the southwest monsoon (Section 3.1.2), the south receives additional rainfall from the northeast monsoon (Wangwongchai et al., 2005). Consequently, $PM_{2.5}$ concentrations are generally lower and more homogeneous across time and space. Contributions from both AT and BB emissions are minimal (Fig. S6), and under such conditions, incorporating decoupled predictors introduces unnecessary complexity, which potentially reduces predictive performance.

Table 6
Test performance of LightGBM models under the station-holdout evaluation.

Model	R^2	RMSE (μg/m^3)	MAE (μg/m^3)	Slope
AOD	0.715	9.85	6.17	0.750
Baseline	0.744	9.35	5.91	0.771
AT	0.749	9.25	5.83	0.780
BB	0.761	9.03	5.71	0.787
AT-BB	0.769	8.89	5.61	0.792
AT-BB-AOD	0.783	8.60	5.45	0.797

Table 7
Test performance of LightGBM models under the region-holdout evaluation.

Region	Model	R^2	RMSE (μg/m^3)	MAE (μg/m^3)	Slope
Central	AOD	**0.575**	**10.18**	**7.31**	*0.697*
	Baseline	0.568	10.26	7.65	0.780
	AT	*0.466*	*11.41*	*8.44*	**0.875***
	BB	**0.627***	**9.54***	**6.89***	*0.757*
	AT-BB	**0.577**	**10.16**	**7.47**	**0.839**
	AT-BB-AOD	**0.587**	**10.03**	**7.43**	**0.827**
East	AOD	*0.695*	*7.16*	*5.16*	*0.687*
	Baseline	0.742	6.59	4.65	0.775
	AT	*0.737*	*6.65*	**4.61**	*0.743*
	BB	*0.726*	*6.78*	*4.72*	**0.841***
	AT-BB	**0.746**	**6.53**	**4.60**	**0.794**
	AT-BB-AOD	**0.762***	**6.33***	**4.51***	**0.781**
North	AOD	*0.647*	*18.22*	*9.10*	*0.564*
	Baseline	0.688	17.13	8.73	0.639*
	AT	*0.684*	*17.22*	**8.64**	*0.616*
	BB	**0.689**	**17.08**	**8.59**	*0.636*
	AT-BB	**0.698**	**16.83**	**8.28**	*0.626*
	AT-BB-AOD	**0.709***	**16.53***	**8.02***	*0.628*
Northeast	AOD	*0.650*	*11.97*	*7.88*	**0.693***
	Baseline	0.677	11.50	7.44	0.667
	AT	*0.666*	*11.71*	*7.58*	*0.658*
	BB	**0.695***	**11.18***	**7.19**	**0.683**
	AT-BB	**0.688**	**11.31**	**7.28**	0.667
	AT-BB-AOD	**0.692**	**11.24**	**7.18***	**0.692**
South	AOD	*0.149*	*6.21*	*4.61*	*0.381*
	Baseline	0.331*	5.51*	4.20	0.485
	AT	*0.296*	*5.65*	*4.27*	*0.428*
	BB	*0.329*	*5.52*	**4.16***	**0.512***
	AT-BB	*0.248*	*5.84*	*4.38*	*0.446*
	AT-BB-AOD	*0.293*	*5.66*	*4.34*	*0.474*

Bold values denote better average performance than the baseline model.
Italic values denote worse average performance than the baseline model.
A superscript star (*) indicates the best-performing model based on average scores.

Independent test performance closely aligned with spatial CV results, suggesting good model generalizability across the study area (Table 6). However, generalization capability may be limited in regions with fewer training samples and more unique source characteristics, representing a form of feature distribution shift where local feature distributions differ from those in the training data. This issue is examined more rigorously under the region-holdout evaluation in Section 3.2. In general, the agreement between test and spatial CV results demonstrates the effectiveness of decoupled models within this evaluation strategy.

3.2. *Performance under the region-holdout evaluation*

The region-holdout evaluation imposes a stricter spatial generalization test than the station-holdout, as the entire test region is excluded from training. Under this condition, the effectiveness of decoupled predictors became more conditional, driven by larger feature

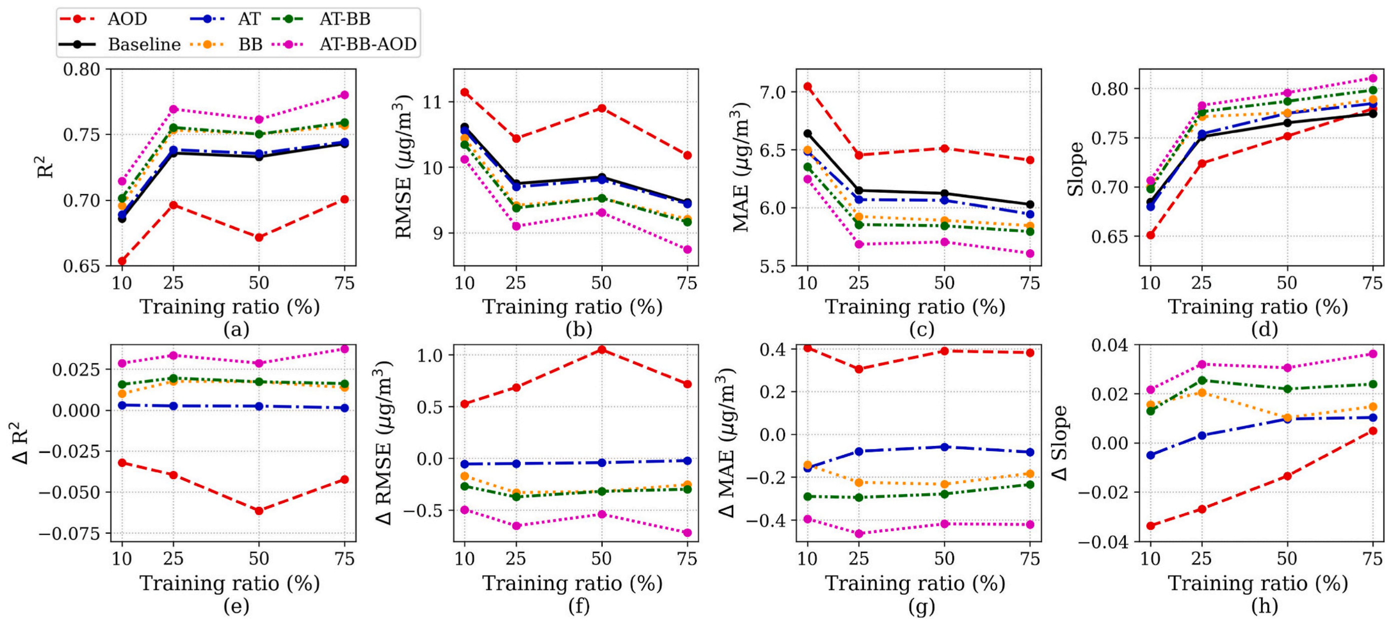


Fig. 3. (a–d) Performance metrics across training fractions. (e–h) Corresponding differences relative to the baseline (Proposed–Baseline).

distribution shifts between training and test regions. The AT model underperformed the baseline in terms of R^2 and RMSE across all holdouts, with the largest performance reduction observed when the Central region was excluded from training (Table 7). Because the Central region has the strongest AT contributions, removing it shifted the feature distribution between training and test sets (Fig. S7), leading to reduced performance. Even when correlations in CMAQ_PM25_AT between regions were high (i.e., Central and East), differences in magnitude distribution limited spatial generalization.

BB-decoupled predictors demonstrated stronger spatial robustness, remaining effective in the Central, Northern, and Northeastern holdouts. BB emissions exhibit strong and consistent seasonal patterns across several regions, supporting model learning and enhancing spatial generalizability. However, predictive ability declined in the South and East. Similar to the Central holdout, feature distribution shift contributed to the reduced performance of the BB model in both regions.

Furthermore, learned feature interactions can differ when the training data and HPs change, and may not generalize to a new region where features interact differently. Feature importance analysis showed that CMAQ_PM25_BB and WRF_Temp were the two most important predictors in the Eastern holdout, which differed from the Northeastern holdout, where the BB model performed best, with CMAQ_PM25_BB and CMAQ_PM25_No_BB jointly dominating.

Table 8
Computational cost and performance gain of CMAQ-based models.

Model	CMAQ run count	Simulation time (min/year)[a]	R^2 gain[b]	RMSE reduction ($\mu g/m^3$)[b]
Baseline	1	3903		
AT	2	7806	+0.007	−0.13
BB	2	7806	+0.016	−0.30
AT-BB	3	11710	+0.022	−0.43
AT-BB-AOD	3	11710	+0.033	−0.64

[a] An estimated time to produce one-year CMAQ simulation on a workstation with Xeon Gold 6248R (3.0 GHz, 24 core) x2, 192 GB RAM (2933 MHz DDR4 16GBx12). LightGBM training and inference time (~278–295 min/year on a workstation with 11th Gen Intel(R) Core(TM) i7-11700 @ 2.50 GHz, 16 GB RAM) is comparatively inexpensive. Thus, the overall computational cost is primarily determined by CMAQ model.

[b] Performance improvements were calculated based on spatial CV scores.

Because AT-decoupled predictors were ineffective under region-holdout evaluation, the effectiveness of AT-BB-decoupled predictors was also reduced in certain holdouts. In the Central holdout, for example, the AT-BB model even underperformed the BB model. Adding AOD improved performance in some regions; however, the BB model remained the best in the Central holdout. Findings demonstrate that the effectiveness of decoupled predictors depends on two key factors: (1) alignment of feature distributions between training and test regions, and (2) generalizability of learned feature interactions and potentially feature-response relationships.

3.3. *Performance under the sparse-station framework*

The AOD model underperformed the baseline across all training ratios, while the AT model showed minimal improvement (Fig. 3). The BB, AT-BB, and AT-BB-AOD models demonstrated larger improvements, though the performance gains of the BB and AT-BB models were nearly identical across all training ratios. Importantly, the magnitude of improvement over the baseline remained relatively consistent across training ratios for BB, AT-BB, and AT-BB-AOD models, suggesting that these models do not require large training samples to capture their key feature characteristics. Therefore, the decoupled models, particularly the BB, AT-BB and AT-BB-AOD models, remained robust under sparse station coverage.

3.4. *Comparison between AOD and CMAQ-based models*

Across all evaluation strategies, the AOD model generally failed to outperform the baseline. Its performance was particularly poor in the South, as indicated by degraded performance scores (Tables 5 and 7). Moreover, the AOD model tended to underestimate high $PM_{2.5}$ concentrations more than the other models (Table 4), when accurate estimation is critical for health-related applications.

MERRA-2 AOD is a model-derived product and thus carries biases from the modeling process, including assimilation technique and uncertainties in emission inventories (Ohno et al., 2022). Although AOD generally correlates with $PM_{2.5}$, the strength of this relationship varies by region and conditions. For example, the AOD–$PM_{2.5}$ correlation weakens during the wet season due to rainfall and hygroscopic growth (Ma et al., 2019). In addition, AOD represents a cumulative column measurement and includes both fine and coarse-mode particles,

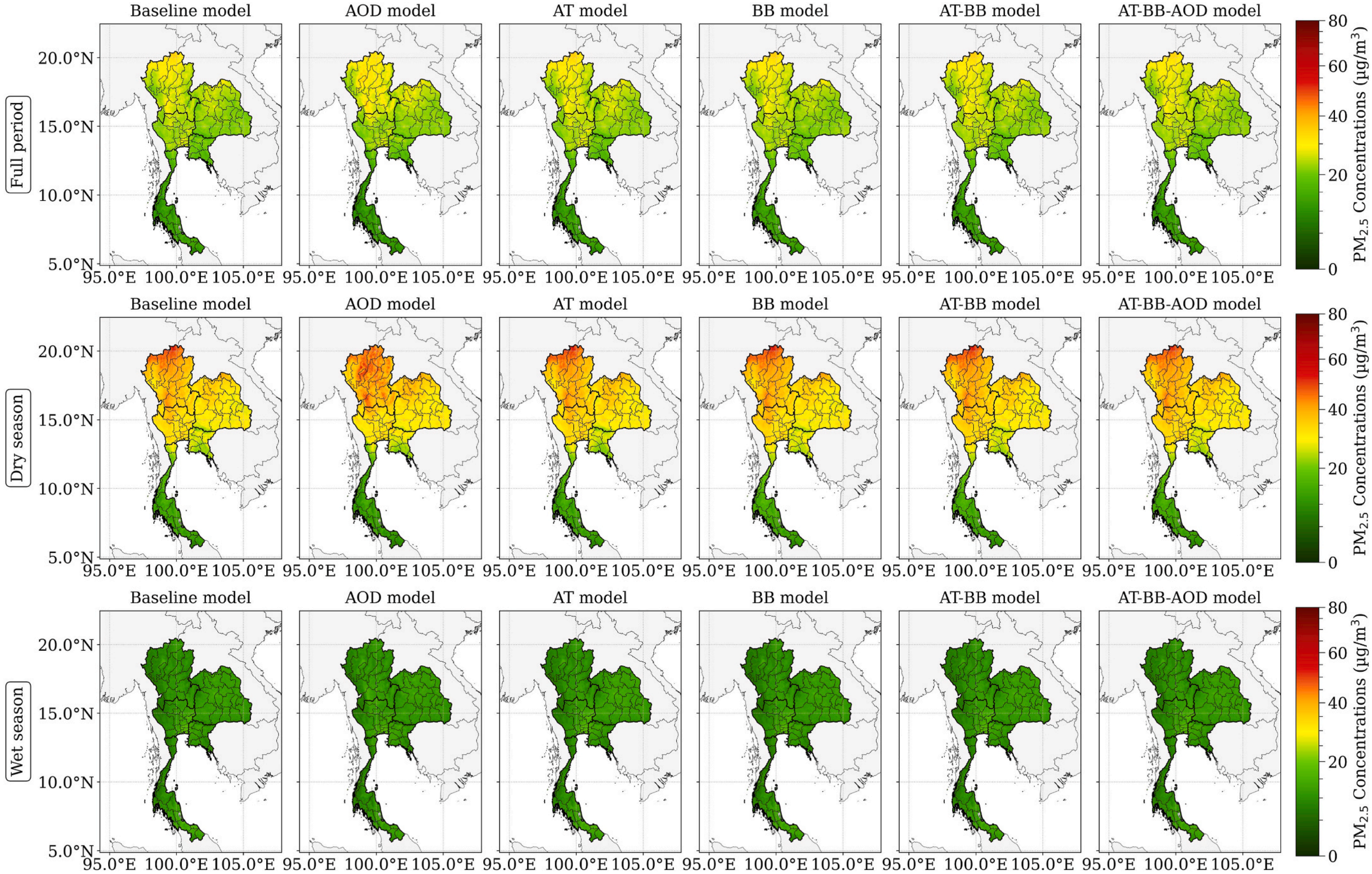


Fig. 4. Average $PM_{2.5}$ predictions of the five models in the full period, dry and wet seasons.

reducing its correlation with ground-level $PM_{2.5}$ compared to CMAQ-simulated $PM_{2.5}$ concentrations (Fig. S8). The coarse spatial resolution of MERRA-2 AOD limits its ability to capture local-scale pollution variability. All of these factors contribute to the degraded performance of the AOD model.

Nevertheless, the AOD model achieved performance comparable to the baseline on average. It remains a practical alternative when CMAQ simulations are unavailable and moderate accuracy is acceptable (e.g., long-term $PM_{2.5}$ trend analysis). For applications demanding highly accurate $PM_{2.5}$ estimates or more reliable peak estimation, e.g., epidemiological studies and health-impact assessments, CTM-based predictors remain preferable. Further accuracy gains can be achieved by supplementing the AT-BB model with AOD, though this comes with a trade-off between computational demand and accuracy. Table 8 summarizes the computational costs for each model configuration. Incorporating decoupled predictors increases computational cost due to additional CMAQ runs.

3.5. *High-resolution $PM_{2.5}$ prediction mapping*

LightGBM models trained on the full training set generated high-resolution daily $PM_{2.5}$ maps, which were averaged over full-period and seasonal scales (Fig. 4). Prediction differences were computed by subtracting the baseline output from those of each proposed model (Proposed – Baseline), where positive values indicated higher $PM_{2.5}$ concentrations (Fig. 5). Across the full period, all models exhibited similar spatial patterns, with lower $PM_{2.5}$ concentrations in the South and progressively higher levels toward the North, where the highest concentrations occurred (Fig. 4).

Among proposed models, the AOD model showed the largest deviation from the baseline (Fig. 5). The AT model predicted slightly higher $PM_{2.5}$ concentrations in the Central region, likely associated with strong AT emissions from dense population. The BB model predicted higher $PM_{2.5}$ concentrations along the Thailand-Myanmar border, covering parts of the Central and Southern regions. Stations used in the analysis are not located along these border areas, making it difficult to verify the accuracy of these higher predictions. Nevertheless, the spatial pattern of higher predicted values aligns well with CMAQ_PM25_BB and fire-density distributions (Fig. S9), suggesting that the elevated concentrations are likely associated with BB emissions. It is also consistent with the extensive forest cover, where forest fires are a major source of BB emissions besides agricultural burning (Punsompong et al., 2021). The AT-BB and AT-BB-AOD models generally predicted slightly higher $PM_{2.5}$ concentrations across most regions, particularly in the Central region.

In the dry season, the spatial distribution of predicted $PM_{2.5}$ closely resembled the full-period distribution but with more pronounced patterns, reflecting the stronger influence of emission sources during this period. In contrast, wet season predictions were largely uniform across models, with minimal divergence from the baseline (Figs. 4 and 5).

4. Conclusion

This study evaluated CMAQ-based and AOD-based variables as key predictors for daily $PM_{2.5}$ predictions in Thailand using a LightGBM model. CMAQ-based predictors were further examined through a source-decoupling approach that separated AT and BB emissions. Six key predictor sets were constructed: (1) Baseline predictor with total emission sources, (2) AT-decoupled, (3) BB-decoupled, (4) AT-BB decoupled, (5) AOD, and (6) hybrid AT-BB-AOD. Corresponding models were developed and evaluated under three strategies: station-holdout, region-holdout and sparse-station.

Under station-holdout and sparse-station evaluations, where stations

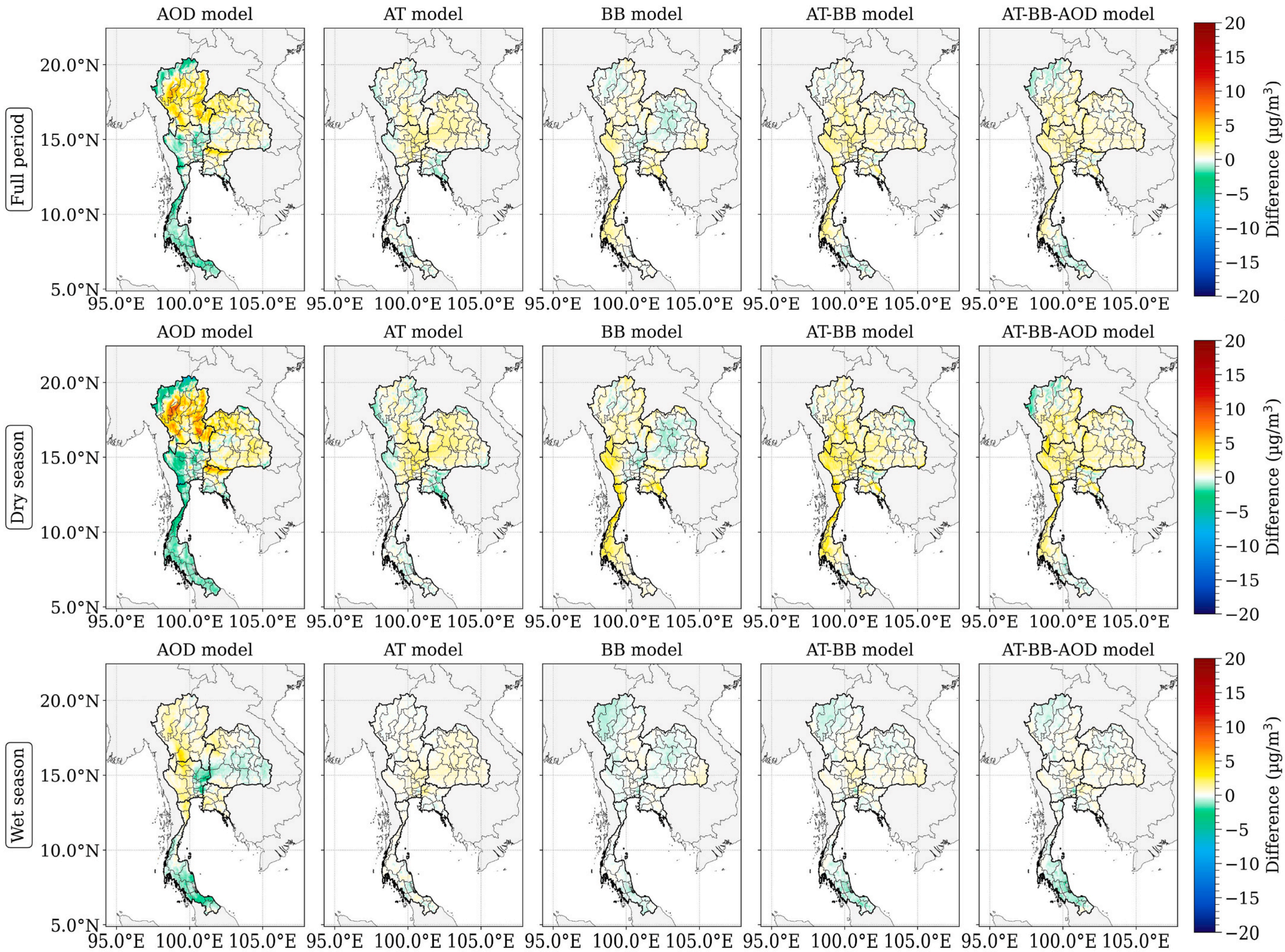


Fig. 5. The difference between baseline and proposed models in the full period, dry and wet seasons.

in all regions were represented during training, the AT-BB model mostly outperformed the baseline, with further improvement when combined with AOD. The AT-BB decoupled predictors were particularly effective during the dry season, high-$PM_{2.5}$ periods, and in regions with mixed and balanced source contributions. Their benefit persisted even under substantially reduced training station coverage. However, in regions dominated by a single source, decoupling the less dominant source was ineffective. In the South, where $PM_{2.5}$ concentrations were low and spatially homogeneous, most decoupled predictors did not improve performance. Under the stricter region-holdout evaluation, however, the effectiveness of decoupled predictors in completely unseen regions depended on the alignment of feature distributions between training and test regions, and the generalizability of learned feature interactions.

From a deployment perspective, decoupled predictors are most effective in within-domain applications, where models are applied to new locations within known regions. In this setting, AT-BB and AT-BB-AOD configurations provide robust improvements. For out-of-domain applications, it is recommended to incorporate training data from regions with diverse emission and meteorological characteristics to fully exploit the benefits of decoupled predictors.

The AOD model demonstrated comparable predictive ability to the baseline, making it a viable alternative when CMAQ simulations are not available. However, its greater tendency to underestimate $PM_{2.5}$ at elevated concentrations and its overall lower accuracy compared to CMAQ-based models, limits its suitability for applications demanding highly accurate $PM_{2.5}$ estimates. In such cases, CMAQ-based predictors with source decoupling are preferred. Nonetheless, the final decision on predictor configuration involves a trade-off between predictive accuracy and computational cost.

CRediT authorship contribution statement

Supitcha Sukprasert: Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Pongpisit Thanasutives:** Methodology, Writing – review & editing. **Shin Araki:** Conceptualization, Writing – review & editing. **Katsushige Uranishi:** Data curation, Writing – review & editing. **Tomohito Matsuo:** Writing – review & editing. **Hikari Shimadera:** Conceptualization, Data curation, Funding acquisition, Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

5 Acknowledgement

This work was supported by the Environment Research and Technology Development Fund of the Environmental Restoration and Conservation Agency of Japan [grant number JPMEERF20215005], and the

JSPS KAKENHI, Japan [grant number 23K25011].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atmosenv.2026.122006.

Data availability

Data will be made available on request.

References

Ahmad, M., Manjantrarat, T., Rattanawongsa, W., Muensri, P., Saenmuangchin, R., Klamchuen, A., Aueviriyavit, S., Sukrak, K., Kangwansupamonkon, W., Panyametheekul, S., 2022. Chemical composition, sources, and health risk assessment of PM2.5 and PM10 in urban sites of Bangkok, Thailand. Int. J. Environ. Res. Publ. Health 19, 14281. https://doi.org/10.3390/ijerph192114281.

Araki, S., Shima, M., Yamamoto, K., 2018. Spatiotemporal land use random forest model for estimating metropolitan NO2 exposure in Japan. Sci. Total Environ. 634, 1269–1277. https://doi.org/10.1016/j.scitotenv.2018.03.324.

Bi, J., Knowland, K.E., Keller, C.A., Liu, Y., 2022. Combining machine learning and numerical simulation for high-resolution $PM_{2.5}$ concentration forecast. Environ. Sci. Technol. 56, 1544–1556. https://doi.org/10.1021/acs.est.1c05578.

Bran, S.H., Macatangay, R., Surapipith, V., Chotamonsak, C., Chantara, S., Han, Z., Li, J., 2022. Surface PM2.5 mass concentrations during the dry season over northern Thailand: sensitivity to model aerosol chemical schemes and the effects on regional meteorology. Atmos. Res. 277, 106303. https://doi.org/10.1016/j.atmosres.2022.106303.

Chantaraprachoom, N., Shimadera, H., Uranishi, K., Mui, L.V., Matsuo, T., Kondo, A., 2024. A nation-by-nation assessment of the contribution of Southeast Asian open biomass burning to PM2.5 in Thailand using the community multiscale air quality-integrated source apportionment method model. Atmosphere 15, 1358. https://doi.org/10.3390/atmos15111358.

Chen, C.-Y., Huang, K.-Y., Chen, C.-C., Chang, Y.-H., Li, H.-J., Wang, T.-H., Yang, P.-C., 2024a. The role of PM2.5 exposure in lung cancer: mechanisms, genetic factors, and clinical implications. EMBO Mol. Med. 17, 31–40. https://doi.org/10.1038/s44321-024-00175-2.

Chen, D., Gu, X., Guo, H., Cheng, T., Yang, J., Zhan, Y., Fu, Q., 2024b. Spatiotemporally continuous PM2.5 dataset in the Mekong River Basin from 2015 to 2022 using a stacking model. Sci. Total Environ. 914, 169801. https://doi.org/10.1016/j.scitotenv.2023.169801.

Chen, G., Li, Y., Zhou, Y., Shi, C., Guo, Y., Liu, Y., 2021. The comparison of AOD-based and non-AOD prediction models for daily PM2.5 estimation in Guangdong province, China with poor AOD coverage. Environ. Res. 195, 110735. https://doi.org/10.1016/j.envres.2021.110735.

Chen, L., Gao, Y., Ma, M., Wang, L., Wang, Q., Guan, S., Yao, X., Gao, H., 2023. Striking impacts of biomass burning on PM2.5 concentrations in Northeast China through the emission inventory improvement. Environ. Pollut. 318, 120835. https://doi.org/10.1016/j.envpol.2022.120835.

Chen, Z.-Y., Jin, J.-Q., Zhang, R., Zhang, T.-H., Chen, J.-J., Yang, J., Ou, C.-Q., Guo, Y., 2020. Comparison of different missing-imputation methods for MAIAC (Multiangle implementation of atmospheric correction) AOD in estimating daily PM2.5 levels. Remote Sens. 12, 3008. https://doi.org/10.3390/rs12183008.

Chokngamwong, R., Chiu, L.S., 2008. Thailand daily rainfall and comparison with TRMM products. J. Hydrometeorol. 9, 256–266. https://doi.org/10.1175/2007JHM876.1.

Feng, S., Huang, F., Zhang, Yuqi, Feng, Y., Zhang, Ying, Cao, Y., Wang, X., 2023. The pathophysiological and molecular mechanisms of atmospheric PM2.5 affecting cardiovascular health: a review. Ecotoxicol. Environ. Saf. 249, 114444. https://doi.org/10.1016/j.ecoenv.2022.114444.

Gueymard, C.A., Yang, D., 2020. Worldwide validation of CAMS and MERRA-2 reanalysis aerosol optical depth products using 15 years of AERONET observations. Atmos. Environ. 225, 117216. https://doi.org/10.1016/j.atmosenv.2019.117216.

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G., 2013. High-Resolution global maps of 21st-Century Forest cover change. Science 342, 850–853. https://doi.org/10.1126/science.1244693.

Hassan Bran, S., Macatangay, R., Chotamonsak, C., Chantara, S., Surapipith, V., 2024. Understanding the seasonal dynamics of surface PM2.5 mass distribution and source contributions over Thailand. Atmos. Environ. 331, 120613. https://doi.org/10.1016/j.atmosenv.2024.120613.

Head, T., Kumar, M., Nahrstaedt, H., Louppe, G., Shcherbatyi, I., 2024. holgern/scikit-optimize. https://doi.org/10.5281/ZENODO.11478072.

Jarernwong, K., Gheewala, S.H., Sampattagul, S., 2023. Health impact related to ambient particulate matter exposure as a spatial health risk map case Study in Chiang Mai, Thailand. Atmosphere 14, 261. https://doi.org/10.3390/atmos14020261.

Jarvis, A., Reuter, H., Nelson, A., Guevara, E., 2008. Hole-Filled Seamless SRTM Data V4. Tech. Rep. International Centre for Tropical Agriculture (CIAT), Cali Columbia.

Karimian, H., Li, Y., Chen, Y., Wang, Z., 2023. Evaluation of different machine learning approaches and aerosol optical depth in PM2.5 prediction. Environ. Res. 216, 114465. https://doi.org/10.1016/j.envres.2022.114465.

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

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17. Curran Associates Inc., Red Hook, NY, USA, pp. 3149–3157.

Liu, Y., Paciorek, C.J., Koutrakis, P., 2009. Estimating regional spatial and temporal variability of $PM_{2.5}$ concentrations using satellite data, meteorology, and land use information. Environ. Health Perspect. 117, 886–892. https://doi.org/10.1289/ehp.0800123.

Ma, Yining, Xin, J., Zhang, W., Liu, Z., Ma, Yongjing, Kong, L., Wang, Y., Deng, Y., Lin, S., He, Z., 2019. Long-term variations of the PM2.5 concentration identified by MODIS in the tropical rain forest, Southeast Asia. Atmos. Res. 219, 140–152. https://doi.org/10.1016/j.atmosres.2018.12.024.

Mallet, V., Sportisse, B., 2006. Uncertainty in a chemistry-transport model due to physical parameterizations and numerical approximations: an ensemble approach applied to ozone modeling. J. Geophys. Res. Atmospheres 111. https://doi.org/10.1029/2005JD006149, 2005JD006149.

Ohno, T., Irie, H., Momoi, M., Da Silva, A.M., 2022. Quantitative evaluation of mixed biomass burning and anthropogenic aerosols over the Indochina Peninsula using MERRA-2 reanalysis products validated by sky radiometer and MAX-DOAS observations. Prog. Earth Planet. Sci. 9, 61. https://doi.org/10.1186/s40645-022-00520-4.

OpenStreetMap contributors, 2019. Openstreetmap Road Network for Thailand.

Pesaresi, M., 2022. GHS-BUILT-S R2022A - GHS built-up surface grid, derived from Sentinel 2 composite and Landsat, multitemporal (1975-2030). https://doi.org/10.2905/D07D81B4-7680-4D28-B896-583745C27085.

Pollution Control Department, 2025. Thailand State of Pollution Report 2024. AP CONNEX Co.,Ltd.

Punsompong, P., Pani, S.K., Wang, S.-H., Bich Pham, T.T., 2021. Assessment of biomass-burning types and transport over Thailand and the associated health risks. Atmos. Environ. 247, 118176. https://doi.org/10.1016/j.atmosenv.2020.118176.

Randles, C.A., Da Silva, A.M., Buchard, V., Colarco, P.R., Darmenov, A., Govindaraju, R., Smirnov, A., Holben, B., Ferrare, R., Hair, J., Shinozuka, Y., Flynn, C.J., 2017. The MERRA-2 aerosol reanalysis, 1980 onward. Part I: system description and data assimilation evaluation. J. Clim. 30, 6823–6850. https://doi.org/10.1175/JCLI-D-16-0609.1.

Shi, H., Xiao, Z., Zhan, X., Ma, H., Tian, X., 2019. Evaluation of MODIS and two reanalysis aerosol optical depth products over AERONET sites. Atmos. Res. 220, 75–80. https://doi.org/10.1016/j.atmosres.2019.01.009.

Solazzo, E., Galmarini, S., 2016. Error apportionment for atmospheric chemistry-transport models – a new approach to model evaluation. Atmos. Chem. Phys. 16, 6263–6283. https://doi.org/10.5194/acp-16-6263-2016.

Tang, B., Stanier, C.O., Carmichael, G.R., Gao, M., 2024. Ozone, nitrogen dioxide, and PM2.5 estimation from observation-model machine learning fusion over S. Korea: influence of observation density, chemical transport model resolution, and geostationary remotely sensed AOD. Atmos. Environ. 331, 120603. https://doi.org/10.1016/j.atmosenv.2024.120603.

Thongsame, W., Henze, D.K., Barth, M., Pfister, G., Kumar, R., Macatangay, R., Hassan Bran, S., 2025. Source attribution and health burden of $PM_{2.5}$ in mainland Thailand. GeoHealth 9. https://doi.org/10.1029/2024GH001315 e2024GH001315.

Thongsame, W., Henze, D.K., Kumar, R., Barth, M., Pfister, G., 2024. Evaluation of WRF-Chem PM2.5 simulations in Thailand with different anthropogenic and biomass-burning emissions. Atmos. Environ. X 23, 100282. https://doi.org/10.1016/j.aeaoa.2024.100282.

Thongthammachart, T., Araki, S., Shimadera, H., Eto, S., Matsuo, T., Kondo, A., 2021. An integrated model combining random forests and WRF/CMAQ model for high accuracy spatiotemporal PM2.5 predictions in the Kansai region of Japan. Atmos. Environ. 262, 118620. https://doi.org/10.1016/j.atmosenv.2021.118620.

Thongthammachart, T., Shimadera, H., Araki, S., Matsuo, T., Kondo, A., 2023. Land use regression model established using Light Gradient Boosting Machine incorporating the WRF/CMAQ model for highly accurate spatiotemporal PM2.5 estimation in the central region of Thailand. Atmos. Environ. 297, 119595. https://doi.org/10.1016/j.atmosenv.2023.119595.

Wang, M., Kim, R.Y., Kohonen-Corish, M.R.J., Chen, H., Donovan, C., Oliver, B.G., 2025. Particulate matter air pollution as a cause of lung cancer: epidemiological and experimental evidence. Br. J. Cancer 132, 986–996. https://doi.org/10.1038/s41416-025-02999-2.

Wangwongchai, A., Sixiong, Z., Qingcun, Z., 2005. A case study on a strong tropical disturbance and record heavy rainfall in Hat Yai, Thailand during the winter monsoon. Adv. Atmos. Sci. 22, 436–450. https://doi.org/10.1007/BF02918757.
Wilson, D.J., 2019. The harmonic mean p-value for combining dependent tests. Proc. Natl. Acad. Sci. 116, 1195–1200. https://doi.org/10.1073/pnas.1814092116.
WorldPop, Bondarenko, M., 2020. Individual countries 1km UN adjusted population density (2000-2020). https://doi.org/10.5258/SOTON/WP00675.
Xu, M., Jin, J., Wang, G., Segers, A., Deng, T., Lin, H.X., 2021. Machine learning based bias correction for numerical chemical transport models. Atmos. Environ. 248, 118022. https://doi.org/10.1016/j.atmosenv.2020.118022.
Yu, Q., You, L., Wood-Sichra, U., Ru, Y., Joglekar, A.K.B., Fritz, S., Xiong, W., Lu, M., Wu, W., Yang, P., 2020. A cultivated planet in 2010: 2. the global gridded agricultural production maps. Earth Syst. Sci. Data Discuss. 1–40. https://doi.org/10.5194/essd-2020-11.