Short-Term $PM_{2.5}$ Forecasting using Gradient Boosting on Geospatial Data

Nachat Jatusripitak
Department of Computer Science
Stanford University
Stanford, CA
nachatj@stanford.edu

***Abstract*—$PM_{2.5}$ (particulate matter smaller than 2.5 microns) is an important policy issue in Southeast Asia because of its adverse effect on human health. In this project, we developed and evaluated XGBoost models for short-term $PM_{2.5}$ forecasting over mainland Southeast Asia. Our models use engineered features to capture spatial and temporal dependencies. We constructed a spatiotemporal dataset by combining satellite-retrieved active fire detections, terrain data, and climate reanalysis variables into a regular grid. Our results show that XGBoost outperforms a persistence baseline, especially with the inclusion of spatiotemporal features. These findings are an early step towards developing more sophisticated architectures that can exploit the spatiotemporal structure of our data.**

***Index Terms*—geospatial analysis, machine learning, forecasting, air pollution**

I. Introduction

Air pollution remains a recurring public health and environmental problem in Southeast Asia. Each year, biomass burning releases particulate matter into the atmosphere, where seasonal meteorological conditions cause it to accumulate in urban areas. Particulate matter less than 2.5 microns in diameter ($PM_{2.5}$) in particular poses a significant threat because of its effect on human health. Exposure to $PM_{2.5}$ has been linked to a wide range of health problems, including respiratory illnesses, cardiovascular diseases, and neurological conditions [1], [2]. Curbing emissions and minimizing population exposure are thus key policy challenges in the region, complicated by the fact that policy interventions often conflict with livelihoods that depend on outdoor work.

Short-term spatiotemporal forecasting is an important tool for enabling more efficient and effective policy responses. In places like Southeast Asia where public resources are limited, this type of forecasting can identify where and when $PM_{2.5}$ pollution will be most severe. Using this approach, governments can issue timely warnings for vulnerable populations, impose targeted restrictions on high-emissions activities or deploy emergency services to high-risk communities. Unlike blanket measures, which may be costly and unpopular, targeted responses can mitigate public health risks while minimizing economic disruption. Developing fast, high-resolution spatiotemporal forecasting models thus plays an important role in supporting scalable, data-driven interventions.

However, forecasting $PM_{2.5}$ is challenging because pollution dynamics are complex. $PM_{2.5}$ concentrations are governed by emissions, meteorological conditions, and atmospheric chemistry. These interactions are nonlinear and operate at multiple spatial and temporal scales. For instance, $PM_{2.5}$ levels in one area may be affected by emissions transported by strong winds from distant sources. Or, high emissions could dissipate quickly under strong winds and high humidity, leading to a decrease in concentrations. Moreover, data scarcity—particularly the lack of high-resolution, high-quality spatiotemporal data—limits the development of more powerful models to capture these nuanced dependencies.

Numerous statistical approaches have been explored for better forecasting of $PM_{2.5}$ concentrations. Early attempts relied on traditional time series models such as ARMA, which were not powerful enough because of their linearity. In recent years, machine learning approaches have become more popular due to their ability to learn nonlinear interactions from data. Spatiotemporal deep learning models such as CNNs, RNNs, and hybrid architectures have gained traction but are especially data-hungry and computationally expensive. Existing methods are constrained by some combination of data availability, performance, or generalizability. The project presented here addresses this gap by exploring gradient boosting models, specifically XGBoost, with engineered spatiotemporal features as a flexible alternative for moderately-sized datasets.

In this project, we constructed a spatiotemporal dataset and benchmarked various XGBoost models against a naive baseline for next-day $PM_{2.5}$ forecasting. We integrated satellite-based active fire detection, topographic information, and climate reanalysis data into a structured grid over Northern Thailand. Using this dataset, we performed a statistical analysis to inform feature selection and modeling approaches. We then trained and evaluated XGBoost models with varying feature combinations to observe their impact on predictive performance.

Our results show that XGBoost benefits from exogenous variables and engineered spatiotemporal features. With hyperparameter tuning and thoughtful feature construction, XGBoost models outperform the naive baseline with minimal tradeoff in computational cost. Our results show that spatial context improves model performance, suggesting that there is potential for future work exploring hybrid architectures that can more effectively exploit the spatiotemporal structure of air pollution.

II. Related Work

In this section, we position our project relative to previous work in the $PM_{2.5}$ forecasting space.

A. Traditional Statistical Models

Early $PM_{2.5}$ forecasting used traditional time-series models. Autoregressive approaches such as ARMA fit a time series by estimating linear dependencies on past values and forecasting errors [3]. Seasonal, integrated models were shown to fit $PM_{2.5}$ time series data better than non-seasonal models due to seasonal variation and non-stationarity of the data [4]. These models are useful because of their interpretability, but they were shown to perform poorly during the smog season in China under conditions similar to those in Southeast Asia [5].

B. Tree-Based Models

Tree-based models are a subclass of machine learning methods that recursively split the data into subsets based on feature values and use the resulting "decision tree" to generate predictions. Unlike traditional methods, which typically assume linear relationships between covariates, tree-based models can capture nonlinear dependencies more effectively. A comparison between machine learning approaches for gap-filling $PM_{2.5}$ data over Tehran, Iran found that tree-based XGBoost and random forest models achieved lower estimation error than a neural network [6]. XGBoost was found to be superior to the physics-based WRF-Chem model at daily forecasting in Shanghai, China [7]. In another study, XGBoost performed better than multiple linear regression, GRUs, and RNNs in tandem with a CNN encoder [8].

C. Deep Learning Models

With increasing computational power, deep learning methods have become more viable for $PM_{2.5}$ forecasting. There has been much work on applying architectures specialized for handling time-series and gridded data to this task. Interpolation techniques are sometimes used to transform $PM_{2.5}$ data into a regular grid, upon which a standard CNN can be trained [9]. Studies have shown that hybrid CNN-LSTM models can achieve lower estimation error than other deep learning models. These hybrid models benefit greatly from large data resources [10], [11].

D. Our Work

This project distinguishes itself from the previous work discussed above in three ways. First, our analysis combines remote sensing and climate reanalysis data sources into a born-regular (rather than interpolated) grid, reducing interpolation error. Second, we draw from traditional time series principles to construct a cross-validation pipeline that is more robust to spatial and temporal data leakage. Third, we evaluate the sensitivity of XGBoost models to various feature combinations rather than optimizing for performance directly.

III. Dataset

We processed our data in Google Earth Engine (GEE), a cloud-based platform with parallelized operations for geospatial analysis [12]. We retrieved the data from the Earth Engine Data Catalog and the Awesome GEE Community Catalog [13].

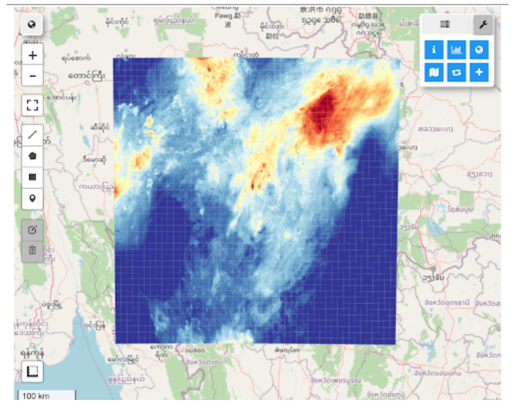

Fig. 1. False-color visualization of $PM_{2.5}$ concentration over Northern Thailand and surrounding regions on a high-pollution day. The gridded overlay shows the 11.1 km × 11.1 km resolution used in the study. Warmer colors (red/yellow) indicate higher $PM_{2.5}$ concentrations, while cooler colors (blue) indicate lower concentrations. Base map shows regional context, including parts of Myanmar, Laos, and Thailand.

The study region is a $\sim 500 \times 500$ km area over mainland Southeast Asia spanning Northern Thailand, Myanmar, and Laos. Fig. 1 shows a representative image of the region.

TABLE I
Summary of Dataset Sources

Dataset	Resolution	Variables Used	Citation
$PM_{2.5}$ (GHAP)	1 km	$PM_{2.5}$ (μg/m^3)	[14]
Weather (ERA5-Land)	11.1 km	u/v wind (m/s), dew point (K), temperature (K), pressure (Pa), precipitation (m)	[15]
Fires (VIIRS VNP14)	1 km	Max. FRP (MW), high-confidence only	[16]
Terrain (SRTM DEM)	90 m	Elevation (m)	[17]

We compiled data from 2018-07-04 to 2022-12-30 with daily frequency. We loaded and preprocessed images on each day and stacked them into a single multi-band image. The precipitation data was truncated to exclude negative values and the fire radiative power (FRP) data was filtered for medium-to-high confidence detections.

The images were gridded at the coarsest resolution of our data, in this case the 11.1km climate data, resulting in 44 x 44 pixel images. We aggregated the other data by averaging the $PM_{2.5}$, wind, dew point, temperature, pressure, and elevation data, as well as by summing the precipitation and FRP values. These multi-band images were exported from GEE as a collection of GeoTIFF files using the EPSG:32647 coordinate reference system before being converted to a tabular format for compatibility with downstream processing steps. The final tabular dataset consists of daily observations for each 11.1km x 11.1km grid cell in the study region, totaling roughly 3.8 million observations across the full time span.

IV. Exploratory Data Analysis

We conducted an exploratory data analysis to understand patterns in the data and guide the experimental protocol. We computed (1) summary statistics, (2) time series plots to observe temporal trends, (3) a correlation matrix to identify linear relationships, (4) plotted $PM_{2.5,t+1}$ versus $PM_{2.5,t}$, and (5) visualized spatial distribution of $PM_{2.5}$ using heatmaps

Table II displays the summary statistics across the bands of the dataset. Most variables have reasonable distributions, but a select few are heavily right-skewed. $PM_{2.5}$, precipitation, and FRP have medians of 12 µg/m³, 8.5×10^{-4} m, and 0 MW but maximums of 266 µg/m³, 0.2034 m, and 12,930 MW, respectively. These values are extreme but not anomalous.

TABLE II
Summary Statistics for Dataset Bands

Band	Mean	Median	Min	Max
$PM_{2.5}$ (t)	19.56	12.00	4.27	266.00
U wind (t)	0.091	0.0586	-4.00	4.28
V wind (t)	0.383	0.3674	-3.18	3.88
Dew point (t) [K]	291.7	293.0	270.4	299.4
Air temp (t) [K]	296.9	297.0	279.2	309.3
Surface pressure (t) [Pa]	93,820	93,850	83,740	102,700
Precipitation (t) [m]	0.0045	0.00085	0.0000	0.2034
Fire radiative power (t)	1.671	0.0000	0.0000	12,930
Elevation [m]	646.9	638.0	9.0	1,743
$\Delta PM_{2.5,t+1}$ (target)	0.0425	0.0107	-213.1	219.1

Fig. 2 shows that daily mean $PM_{2.5}$ levels in the study region begin to rise in November, with a sharp peak in mid-March, before decreasing precipitously in April and May. Note that during the burning season from February to April, daily mean $PM_{2.5}$ levels greatly exceed the World Health Organization's Air Quality Guideline of 15 µg/m³ [18]. This pattern is consistent across the five years of data, indicating a seasonal component.

A. Temporal autocorrelation

We quantified temporal autocorrelation using autocorrelation and partial autocorrelation (ACF/PACF) plots. We averaged ΔPM2.5 over all grid cells on each day to generate a time series. An augmented Dickey-Fuller (ADF) test showed that the region-averaged ΔPM2.5 time series is stationary, with $p < 10^{-6}$. The region-wide ACF/PACF plot shown in Fig. 3 shows modest autocorrelation at lags $t = 1, \ldots, 10$, with noisy correlations beyond those lags possibly due to the seasonal component in the data.

We confirmed these results with ACF/PACF plots for location-specific ΔPM2.5 time series. We used k-means clustering with $k = 6$ to group grid cells by their summary statistics, then computed ACF/PACF plots for the medoid points. These medoid plots reflected the region-wide plots.

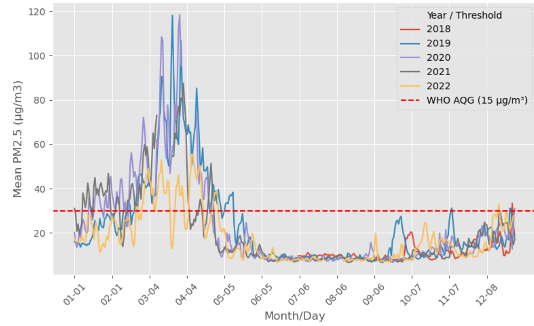

Fig. 2. Daily mean $PM_{2.5}$ concentrations across five years (2018–2022) averaged over the study region. A clear seasonal trend is visible, with pollution levels peaking between February and April each year—corresponding to the regional burning season. The red dashed line represents the World Health Organization (WHO) Air Quality Guideline (AQG) for $PM_{2.5}$ (15 µg/m³), which is consistently exceeded during the high-pollution months.

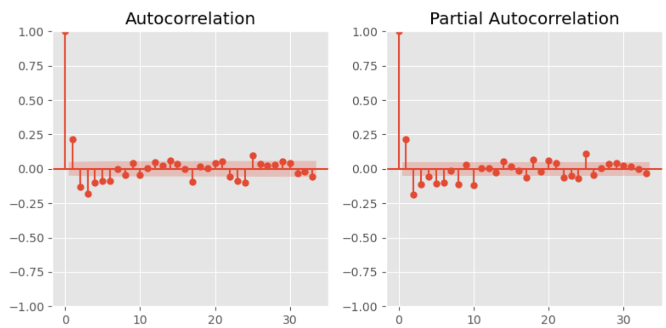

Fig. 3. Autocorrelation and partial autocorrelation plots of the region-averaged daily $\Delta PM_{2.5}$ time series. Significant autocorrelation is observed at short lags (1–3 days), indicating temporal dependencies that can be exploited for forecasting. The rapid decay in both ACF and PACF supports the use of short-term lag and moving average features.

B. Spatial autocorrelation

We quantified spatial autocorrelation using empirical variograms. We randomly sampled 50 days from our date range without replacement and fitted variograms on the data from each date up to a maximum range of 22 pixels (half the image size). We assumed a Gaussian model and used the Cressie-Hawkins estimator, which is more robust against outliers. Fig. 4 is a representative example of such a variogram from 2022-10-23, which shows a range of approximately 21 pixels. The variograms indicated significant spatial autocorrelation in the data on many dates.

In sum, we found spatiotemporal autocorrelation in the data, which has important implications for model evaluation. Temporal autocorrelation means that a naive train-test split would have leakage of information from consecutive days. Spatial autocorrelation means that grouping data points by day is necessary to avoid leakage of information from neighboring locations. In the next section, we discuss how our experimental protocol respects these constraints.

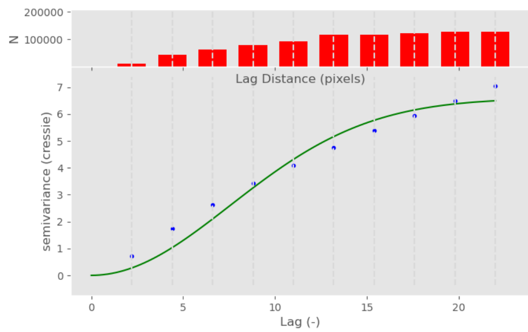

Fig. 4. Empirical semivariogram for a representative day (October 23, 2022), illustrating spatial autocorrelation in $PM_{2.5}$ across the study region. The semivariance (Cressie-Hawkins estimator) increases with lag distance, reaching a plateau around 21 pixels—indicating a spatial correlation range of approximately 230 km. The fitted Gaussian model captures the spatial dependence structure, supporting the inclusion of spatial features in the forecasting model.

V. Methodology

We trained models using the Scikit-learn estimator interface to predict $\Delta PM_{2.5}$ for each grid cell [19]. Unlike many studies in the literature that predict $PM_{2.5}$ levels directly, we chose to predict $\Delta PM_{2.5}$ so that the models learn meaningful dynamics. We chose gradient-boosted decision tree models (XGBoost) because they are quick to train, are robust against overfitting, and have shown strong performance in similar applications. While other lightweight forecasting methods such as linear regression, ridge regression, or ARIMA could be considered, we focused on XGBoost due to its ability to handle nonlinearity and interactions between features.

We conducted three main feature experiments to assess the effect of temporal, spatial, and seasonality features on model performance. These feature experiments are not only self-contained studies, but also integrated steps towards developing the final model. In each experiment, we trained models with different feature combinations, holding hyperparameters fixed. The models were trained using the Pseudo-Huber loss function for robustness against extreme values and evaluated using root mean squared error (RMSE). While mean absolute error (MAE) is more conventional given the absolute error loss function, RMSE is more useful for punishing large errors on high-pollution days. The most effective feature combinations are retained from each experiment; each experiment adds more features incrementally. We evaluated the final model with complete features on a held-out test set from the year 2022.

We used rolling-window cross-validation with spatial blocking to mitigate the spatiotemporal autocorrelation that we found in our exploratory data analysis. In each window, we trained the model on 730 consecutive days of data, left a purge gap of 21 days, then evaluated the model on the next 49 days of data. We advanced the window by 49 days and repeated the process until the dataset is exhausted. Fig. 5 visualizes the rolling-window aspect of the procedure.

Our procedure respects temporal order since no models are trained on future observations, and the purge gap allows

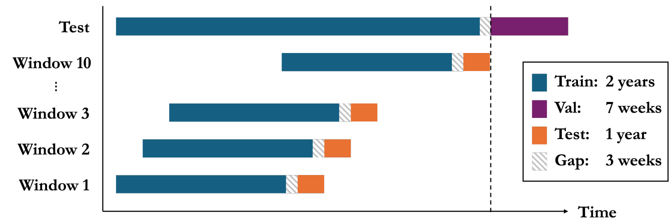

Fig. 5. Visualization of the rolling-window cross-validation scheme with spatial blocking used for model evaluation. Each training window spans two years, followed by a 3-week purge gap to prevent temporal leakage, and a 7-week validation period. This process is repeated across multiple windows until the dataset is exhausted. A separate, held-out test set from the year 2022 is reserved for final evaluation. This design ensures that temporal order is respected and spatial autocorrelation is controlled.

residual temporal autocorrelation to ”die off.”At the same time, we avoided spatial autocorrelation by grouping data points from the same day together so that no spatial leakage occurred between training and evaluation.

We recorded RMSE for each evaluation window and computed mean RMSE, which was used to compare models. The experimental models were compared to each other, as well as against a base model with basic features and a naive persistence model. Here, “persistence” means predicting $\Delta PM2.5 = 0$, which is strong because $PM_{2.5}$ is highly correlated between days, thus predicting zero change captures much of the variance.

A. Temporal Features

To capture temporal dependencies, we engineered moving average and time-lagged features for $\Delta PM_{2.5}$.

Moving averages were used to smooth short-term fluctuations and capture recent trends. The n-day moving average at location (x, y) is given by

$$\mathrm{MA}_n(x, y) = \frac{1}{n} \sum_{i=0}^{n-1} \Delta \mathrm{PM}_{2.5,t-i}^{(x,y)}$$

where $\Delta \mathrm{PM}_{2.5,t}^{(x,y)}$ is the change in $PM_{2.5}$ at time t at position (x, y). We experimented with multiple window sizes ($n = 3, 5, 7, 9, 11$) to assess how short and medium-term trends influenced predictive performance. Starting with $n = 3$, we cumulatively added additional moving averages up to $n = 11$.

Time-lags were used to leverage temporal autocorrelation and capture recent increases or decreases in concentration levels, which often signal upcoming shifts in air quality. We experimented with lags of up to 7 days in a cumulative manner, similar to our approach to moving averages.

B. Spatial Features

To incorporate spatial context, we constructed features based on values from neighboring grid cells. For each grid cell (x, y), we computed the mean and maximum $PM_{2.5}$ (or FRP) within a local square neighborhood of size $(2n+1)\times(2n+1)$. For a variable v at time t, the spatial mean and maximum of neighborhood size $2n + 1$ centered at (x, y) is given by

$$\text{Mean}_n(x,y) = \frac{1}{(2n+1)^2} \sum_{i=-n}^{n} \sum_{j=-n}^{n} \text{PM}_{2.5,t}^{(x+i,y+j)}$$

$$\text{Max}_n(x,y) = \max_{-n \leq i,j \leq n} \text{PM}_{2.5,t}^{(x+i,y+j)}$$

We experimented with $n = 1, 2, 3, 4$ (i.e., 3×3, 5×5, 7×7, and 9×9 neighborhoods). These spatial features correspond to increasing spatial windows of up to approximately 100km in diameter, which helps capture diffusion and transport patterns that arise from regional fire activity and wind conditions.

C. Seasonal Features

As noted earlier, $\Delta\text{PM}_{2.5}$ has a strong seasonal component. These seasonal cycles were encoded using calendar-based features. We included the following features:

- Day of the week (as both one-hot and numerical values from 0 to 6),
- A binary weekend indicator,
- Month of the year (as both one-hot and numerical values from 1 to 12).

These features signal periodic patterns in emissions such as agricultural burning, wildfire seasons, and human activity that can influence $\text{PM}_{2.5}$ levels. In particular, monthly features allow the models to distinguish between different pollution regimes—low pollution in the summer and autumn and high pollution in the winter and spring.

VI. Results and Discussion

A. Temporal features

The moving average $\Delta\text{PM}_{2.5}$ features surprisingly worsened performance. No moving average models outperformed the base model, so these features were excluded from following experiments. Given the short-term fluctuations in $\Delta\text{PM}_{2.5}$, it is possible that these moving averages over-smoothed these trends. In contrast, time-lagged $\Delta\text{PM}_{2.5}$ improved performance across the board, minus 1-4 day time lags. Table III shows that adding 1-3 day time lags led to the greatest reduction in mean RMSE and MAE. These reductions were modest but noticeable, on the order of 1%. These results align with earlier observations from the ACF/PACF plots.

TABLE III
Mean RMSE and MAE for $\Delta\text{PM}_{2.5}$ Time Lag Experiments

Experiment	Mean RMSE (µg/m³)	Mean MAE (µg/m³)
Lags 1–3	**4.335**	**2.945**
Lags 1–7	4.344	2.947
Lags 1–5	4.351	2.953
Lags 1–6	4.356	2.957
Lags 1–2	4.358	2.958
Lag 1	4.359	2.962
Base	4.362	2.965
Lags 1–4	4.367	2.958

B. Spatial features

Spatial features were effective at reducing forecasting error. Table IV shows that performance increases with additional spatial context for $\Delta\text{PM}_{2.5}$. While the $n = 1$ (3x3 window) worsened performance relative to the base model, there were increasing performance gains from adding higher orders of n. Moreover, adding spatial context for $\text{PM}_{2.5}$ also improved performance, as it did for $\Delta\text{PM}_{2.5}$. Table V shows similar gains from including spatial averages of $\text{PM}_{2.5}$. These increases in performance do not seem to diminish with increasing n, suggesting that even wider spatial windows should be considered. Our results show that spatial averages of $\Delta\text{PM}_{2.5}$ and $\text{PM}_{2.5}$ are both effective at reducing mean RMSE and MAE relative to the base model.

TABLE IV
Mean RMSE and MAE for $\Delta\text{PM}_{2.5}$ Spatial Mean Experiments

Experiment	Mean RMSE (µg/m³)	Mean MAE (µg/m³)
$\mathbf{n = 1–n = 4}$	**4.315**	**2.920**
$n = 1$–$n = 2$	4.326	2.932
$n = 1$–$n = 3$	4.332	2.930
Base	4.350	2.955
$n = 1$	4.358	2.952

TABLE V
Mean RMSE and MAE for $\text{PM}_{2.5}$ Spatial Mean Experiments

Experiment	Mean RMSE (µg/m³)	Mean MAE (µg/m³)
$\mathbf{n = 1–n = 4}$	**4.318**	**2.929**
$n = 1$–$n = 2$	4.340	2.945
$n = 1$–$n = 3$	4.344	2.943
Base	4.350	2.955
$n = 1$	4.367	2.954

The addition of spatial features for fire activity and pollution in neighboring grid cells yielded modest but consistent improvements. These improvements are supported by real-world pollution dynamics, which are influenced not only by local conditions, but also by surrounding activity. Emissions from fires can be transported over long distances and contribute to distant pollution concentrations elsewhere. The model can capture medium-range dispersion and transport effects through these spatial features. However, these gains remain incremental, possibly because of the coarse 11km x 11km resolution that may obscure finer transport dynamics.

C. Seasonal features

Only month indicators were effective at reducing forecasting error. As shown in Table VI, both one-hot and categorical month encodings improved model performance relative to the base model, with one-hot encodings showing the most benefit. These monthly indicators may help the model distinguish between different pollution regimes depending on the time of year. In contrast, the binary weekend flag did not show any significant benefit and day-of-week features worsened performance relative to the base model.

TABLE VI
MEAN RMSE AND MAE FOR TEMPORAL FEATURE EXPERIMENTS

Experiment	Mean RMSE (μg/m³)	Mean MAE (μg/m³)
Month (one-hot)	**4.222**	**2.862**
Month (numerical)	4.231	2.868
Weekend (binary)	4.248	2.874
Base	4.248	2.882
Day-of-week (numerical)	4.259	2.882
Day-of-week (one-hot)	4.296	2.898

D. Final model

On the held-out 2022 test set, the XGBoost model with complete feature set moderately outperformed both the persistence baseline and a basic XGBoost model with only the features from the dataset. As shown in Table VII, the complete XGBoost model has lower RMSE and MAE (5.057 μg/m³ and 3.040 μg/m³) compared to the persistence baseline (5.622 μg/m³ and 3.327 μg/m³)and the basic XGBoost model (5.240 μg/m³ and 3.160 μg/m³). These performance gains amount to 10% and 8.6% reductions in RMSE and MAE over the persistence baseline, and 3.5% and 3.8% over the basic XGBoost model. While our results indicate that XGBoost is able to

TABLE VII
TEST SET RMSE AND MAE FOR FINAL MODEL EVALUATION

Model	RMSE (μg/m³)	MAE (μg/m³)
Persistence	5.622	3.327
Basic XGBoost	5.240	3.160
Final XGBoost	**5.057**	**3.040**

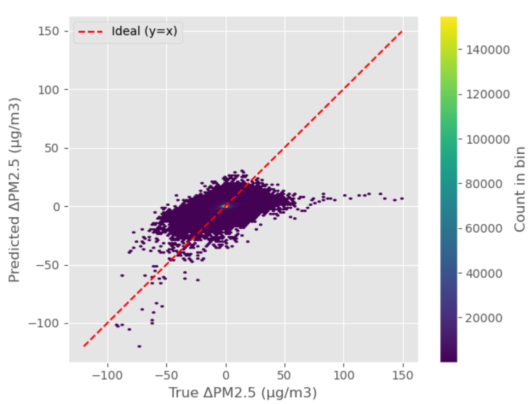

Fig. 6. Hexbin plot comparing predicted versus true daily changes in $PM_{2.5}$ ($\Delta PM_{2.5}$) on the 2022 test set. The red dashed line represents the ideal prediction line (y = x). The model performs well for small to moderate changes, but underestimates large positive changes and tends to overpredict large decreases. The skewed spread and clustering around zero highlight the model's bias toward more frequently occurring low-magnitude events.

capture some underlying pollution dynamics, there is much room for improvement in the modeling of emissions. Fig. 6 compares model predictions to the true values and illustrates three important patterns in the results.

First, the model is relatively good at predictions concentrated around the center. These are mild to moderate changes in $PM_{2.5}$ that are more frequently observed in the data.

Second, the model is quite good at predicting *decreases* in $PM_{2.5}$, even if these decreases are extreme. This result suggests that time-lagged variables were effective at signaling these reductions; if $PM_{2.5}$ levels are high today, they are more likely to decrease tomorrow than to increase.

Third, the model mostly fails to predict large changes in $PM_{2.5}$ above 50 μg/m³, indicated by the mostly horizontal, sparse line of predictions extending from the main cluster. Fig. 7 overlays the true $PM_{2.5}$ time series at a particular location with the predicted $PM_{2.5}$. The bottom plot shows the

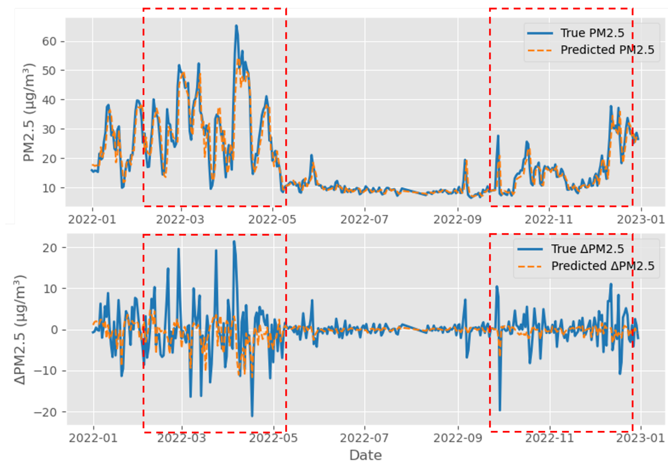

Fig. 7. Time series plots of true and predicted $PM_{2.5}$ values (top) and daily changes in $PM_{2.5}$ ($\Delta PM_{2.5}$, bottom) for a representative location in 2022. The model closely tracks moderate fluctuations but underestimates sharp increases in $PM_{2.5}$, particularly during the peak burning seasons (highlighted in red boxes). While the model captures many sharp decreases, its predictions during high-variability periods remain conservative, reflecting a bias toward stable or declining pollution levels.

direct $\Delta PM_{2.5}$ predictions whereas the top plot reconstructs the predicted $PM_{2.5}$ time series by adding

$$PM_{2.5,t+1}^{\text{pred.}} = PM_{2.5,t} + \Delta PM_{2.5,t+1}^{\text{pred.}}.$$

The overlaid time series confirms that the model is better at predicting decreases than increases. There is a major discrepancy during the peak burning season where increases are more extreme. Here, the model fails to predict many increases but is relatively good at predicting the decreases that follow.

We stratified test-set errors by the magnitude of $\Delta PM_{2.5}$ to better understand differences in model performance across various pollution dynamics, as shown in Table VIII. As expected, the persistence baseline performed well in stable or low-pollution conditions, achieving the lowest MAE (1.56 μg/m³) and RMSE (2.03 μg/m³). However, the performance of the persistence baseline deteriorated in high-change regimes, with RMSE of 29.43 μg/m³ and 27.74 μg/m³ and MAE of 28.14 μg/m³ and 26.41 μg/m³ in the "Large Decrease" and "Large

Increase" bins, respectively. Both the basic and complete XGBoost models reduced errors in higher-change regimes. The complete XGBoost reduced RMSE by 29.4% and 7% in, respectively, the "Large Decrease" and "Large Increase" bins compared to the persistence model. The models were able to learn to predict decreases effectively, whereas the addition of spatial and temporal features in the complete XGBoost model somewhat improved its ability to predict increases in $PM_{2.5}$ over the basic XGBoost model. Interestingly, while the XGBoost models slightly underperformed against the persistence model in low-change bins, they consistently achieved lower RMSE and MAE in high-change bins. These results suggest that XGBoost can capture some nonlinear dynamics, particularly during large decreases, at the expense of some bias during stable regimes.

These improvements depend on reliable and noiseless data, which was a concern in this study. First, inherent limitations of satellite-retrieved data introduce uncertainty into fire detections. VIIRS can miss fires that are small or brief, like many agricultural fires. Second, the weather data is less informative because of its coarse spatial resolution - at 11.1km resolution, variables such as wind direction and temperature may be less meaningful, as gradients are lost in the aggregation process. Third, the $PM_{2.5}$ ground truth data itself contains noise. The data was estimated by a machine learning model and validated at weather stations in the area. The $PM_{2.5}$ data itself are noisy and potentially less reliable due to fewer weather stations in Southeast Asia. These limitations in the data make constructing a reliable model challenging.

VII. Conclusion and Future Work

This project explored whether lightweight machine learning methods–specifically gradient boosting with engineered spatiotemporal features–could provide meaningful improvements in short-term $PM_{2.5}$ forecasting over mainland Southeast Asia. We constructed a dataset of daily 44 x 44 grids using satellite, topographic, and climate reanalysis data. Spatiotemporal autocorrelation analysis revealed strong spatial and temporal dependencies, which guided our evaluation protocol.

Through a series of feature experiments, we found that XGBoost could predict changes in $PM_{2.5}$ with modest effectiveness, with one configuration achieving a 10% reduction in RMSE compared to a naive persistence baseline. These gains, while incremental, show that tree-based models can partially capture some of the underlying pollution dynamics. However, the models struggled to predict sharp increases during the burning season, limited by data quality and the expressiveness of the existing feature set.

Future work could explore:

- Spatiotemporal architectures such as CNNs and LSTMs, which may better exploit spatiotemporal dependencies.
- Hybrid models using shallow feature extractors and XGBoost to balance performance on smaller datasets.
- Ground-based $PM_{2.5}$ measurements, which could eliminate estimation noise from the current synthetic dataset.
- Discretization or transformation of the prediction space to improve robustness for extreme pollution events.

Overall, this work is an early step in developing scalable air quality forecasting tools using large-scale geospatial data.

Acknowledgment

I am deeply grateful to the many individuals who have contributed to this senior research project.

First and foremost, I would like to extend my deepest gratitude to Dan Iancu, Associate Professor of Operations, Information, and Technology at the Graduate School of Business. He set a vision for our work that guided my research direction, and provided constructive feedback and criticism that shaped my research outcomes for the better.

Second, I would like to thank Mykel Kochenderfer, Associate Professor of Aeronautics and Astronautics and, by courtesy, of Computer Science. Mykel was supportive of my research process and provided valuable advice in principled experimental design.

Third, I would like to thank Mary McDevitt, Director of the Technical Communication Program. Mary gave me useful technical writing advice that has made me a better technical writer.

References

[1] S. Feng, D. Gao, F. Liao, F. Zhou, and X. Wang, "The health effects of ambient pm2. 5 and potential mechanisms," *Ecotoxicology and environmental safety*, vol. 128, pp. 67–74, 2016.

[2] P. Fu, X. Guo, F. M. H. Cheung, and K. K. L. Yung, "The association between pm2. 5 exposure and neurological disorders: a systematic review and meta-analysis," *Science of the Total Environment*, vol. 655, pp. 1240–1248, 2019.

[3] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[4] W. Wang and Y. Guo, "Air pollution pm2. 5 data analysis in los angeles long beach with seasonal arima model," in *2009 international conference on energy and environment technology*, vol. 3. IEEE, 2009, pp. 7–10.

[5] L. Zhang, J. Lin, R. Qiu, X. Hu, H. Zhang, Q. Chen, H. Tan, D. Lin, and J. Wang, "Trend analysis and forecast of pm2. 5 in fuzhou, china using the arima model," *Ecological indicators*, vol. 95, pp. 702–710, 2018.

[6] M. Zamani Joharestani, C. Cao, X. Ni, B. Bashir, and S. Talebiesfandarani, "Pm2. 5 prediction based on random forest, xgboost, and deep learning using multisource remote sensing data," *Atmosphere*, vol. 10, no. 7, p. 373, 2019.

[7] J. Ma, Z. Yu, Y. Qu, J. Xu, and Y. Cao, "Application of the xgboost machine learning method in pm2. 5 prediction: A case study of shanghai," *Aerosol and Air Quality Research*, vol. 20, no. 1, pp. 128–138, 2020.

[8] A. Usha Ruby, J. G. C. Chandran, P. Theerthagiri, R. Patil, B. Chaithanya, and T. S. Jain, "Forecasting pm2. 5 concentration using gradient-boosted regression tree with cnn learning model," *Optical Memory and Neural Networks*, vol. 33, no. 1, pp. 86–96, 2024.

[9] S. Chae, J. Shin, S. Kwon, S. Lee, S. Kang, and D. Lee, "Pm10 and pm2. 5 real-time prediction models using an interpolated convolutional neural network," *Scientific Reports*, vol. 11, no. 1, p. 11952, 2021.

[10] C.-J. Huang and P.-H. Kuo, "A deep cnn-lstm model for particulate matter (pm2. 5) forecasting in smart cities," *Sensors*, vol. 18, no. 7, p. 2220, 2018.

[11] T. Li, M. Hua, and X. Wu, "A hybrid cnn-lstm model for forecasting particulate matter (pm2. 5)," *Ieee Access*, vol. 8, pp. 26 933–26 940, 2020.

[12] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google earth engine: Planetary-scale geospatial analysis for everyone," *Remote Sensing of Environment*, 2017. [Online]. Available: https://doi.org/10.1016/j.rse.2017.06.031

TABLE VIII
STRATIFIED MAE AND RMSE BY $PM_{2.5}$ CHANGE LEVEL

Bin	**Count**	**MAE P (µg/m³)**	**MAE B (µg/m³)**	**MAE C (µg/m³)**	**RMSE P (µg/m³)**	**RMSE B (µg/m³)**	**RMSE C (µg/m³)**
Large Decrease	4,513	28.14	**19.82**	18.87	29.43	21.47	**20.79**
Moderate Decrease	64,550	9.01	6.89	**6.85**	9.66	**7.83**	7.93
Stable	534,755	**1.56**	1.66	1.64	**2.03**	2.30	2.35
Moderate Increase	67,756	8.82	8.99	**8.11**	9.47	9.86	**9.10**
Large Increase	4,090	26.41	26.05	**23.98**	27.74	27.51	**25.69**

Bin thresholds: **Large Decrease** ($\Delta\mathrm{PM}_{2.5} < -20$), **Moderate Decrease** ($-20 \leq \Delta\mathrm{PM}_{2.5} < -5$), **Stable** ($|\Delta\mathrm{PM}_{2.5}| < 5$), **Moderate Increase** ($5 \leq \Delta\mathrm{PM}_{2.5} < 20$), **Large Increase** ($\Delta\mathrm{PM}_{2.5} \geq 20$). **P** = Persistence, **B** = Basic XGBoost, **C** = Complete XGBoost.

[13] S. Roy, E. Jensen, A. Saah, and T. Swetnam, "samapriya/awesome-gee-community-datasets: Community catalog (3.5.0)," 2025. [Online]. Available: https://doi.org/10.5281/zenodo.15587604

[14] J. Wei, Z. Li, A. Lyapustin, J. Wang, O. Dubovik, J. Schwartz, L. Sun, C. Li, S. Liu, and T. Zhu, "First close insight into global daily gapless 1 km pm2. 5 pollution, variability, and health impact," *Nature communications*, vol. 14, no. 1, p. 8349, 2023.

[15] J. Muñoz-Sabater, E. Dutra, A. Agustí-Panareda, C. Albergel, G. Arduini, G. Balsamo, S. Boussetta, M. Choulga, S. Harrigan, H. Hersbach *et al.*, "Era5-land: A state-of-the-art global reanalysis dataset for land applications," *Earth system science data*, vol. 13, no. 9, pp. 4349–4383, 2021.

[16] L. Giglio, "Viirs/jpss1 thermal anomalies and fire daily l3 global 1km sin grid v002," *NASA EOSDIS Land Processes Distributed Active Archive Center (DAAC) data set*, pp. VJ114A1–002, 2023.

[17] A. Jarvis, H. I. Reuter, A. Nelson, E. Guevara *et al.*, "Hole-filled srtm for the globe version 4," *available from the CGIAR-CSI SRTM 90m Database (http://srtm. csi. cgiar. org)*, vol. 15, no. 25-54, p. 5, 2008.

[18] W. H. Organization *et al.*, *WHO global air quality guidelines: particulate matter (PM2. 5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. World Health Organization, 2021.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.