Improving prediction of $PM_{2.5}$ in Metro Manila using XGBoost with Optuna hyperparameter optimization

Roseanne V. Ramos [1,2], Alvin Christopher G. Varquez[1]

[1]Department of Transdisciplinary Science and Engineering, School of Environment and Society, Institute of Science Tokyo, Meguro-ku, Tokyo 152-8550, Japan - ramos.r.4bbb@m.isct.ac.jp, varquez.a.aa@m.titech.ac.jp
[2]Department of Geodetic Engineering, University of the Philippines, Diliman, Quezon City, Philippines 1101 - rvramos@up.edu.ph

Keywords: $PM_{2.5}$ model, Metro Manila, XGBoost, Optuna, Kriging

Abstract

Reporting air pollution levels in Metro Manila, Philippines remains dependent on records from few ground monitoring stations. For impact studies on human health, grid-based pollution levels datasets will enhance the assessment of the exposure and risk of the local population with a finer spatial resolution. This paper presents an enhanced model for fine particulate matter ($PM_{2.5}$) in Metro Manila using extreme gradient boosting (XGBoost) with predictor variables from daily satellite observations and climate reanalysis datasets. This study employs XGBoost with Optuna workflow, a Bayesian optimization algorithm, to determine optimal learning parameters achieving least mean absolute error (MAE), root mean squared error (RMSE) and highest coefficient of determination R^2 for both training and test datasets to obtain more accurate predictions of $PM_{2.5}$. Predictor variables assessed are Aerosol Optical Depth (AOD), emissivity, direct solar radiation (DSR), albedo, land surface temperature during daytime (LSTDT) and nighttime (LSTNT), Normalized Difference Vegetation Index (NDVI), wind speed, air temperature, pressure and total precipitation. The results indicate that using a Kriging-based interpolated $PM_{2.5}$ as target variable achieves the highest $R^2 = 0.73$ and lowest RMSE (1.08 µg/ m^3) in the test data using K-fold cross-validation approach (k = 5). The predictors in the best XGBoost model with relatively high feature importance in both gain scores and SHAP values are DSR, precipitation, LSTDT, air temperature and AOD in the 0.47 µm band. These outputs can be utilized in generating daily gridded $PM_{2.5}$ maps for long-term health impact studies and assessment of seasonal variations of $PM_{2.5}$ at regional level.

1. Introduction

Air pollution levels are extensively monitored and investigated using various measurement and modeling techniques due to their impacts to environment, economy and human health. Numerous studies have applied machine learning techniques to estimate fine particulate matter ($PM_{2.5}$) concentrations in urban areas, aiming to overcome operational challenges such as limited monitoring stations and the costly maintenance of sensors. These models also enhance the spatial and temporal resolution of $PM_{2.5}$ predictions to support air quality management, public health protection, and sustainable environmental planning (Kunjir et al., 2025).

Metro Manila experiences high volumes of vehicular traffic and industrial activity, which the Environmental Management Bureau (EMB) has identified as major sources contributing to elevated PM concentrations. Ground-level $PM_{2.5}$ is measured by Continuous Ambient Air Quality Monitoring Stations operated by EMB. However, only 15 stations are distributed across the region's 16 cities, limiting spatial coverage. As a result, recent studies have focused on estimating $PM_{2.5}$ at unsampled locations. Torres et al. (2023a) applied regression models, specifically multiple linear regression and extreme gradient boosting (XGBoost), to analyze spatial and temporal patterns of coarse particles (PM_{10}) and $PM_{2.5}$. Predictor variables included ground-based $PM_{2.5}$ data from 2017 to 2020 and aerosol optical depth (AOD) in green and blue bands derived from the Moderate Resolution Imaging Spectroradiometer (MODIS). In a related study, a Land Use Regression (LUR) model incorporated geographic predictors such as road networks, traffic counts, Normalized Difference Vegetation Index (NDVI), population density, and elevation (Torres et al., 2024). However, these studies did not include meteorological parameters, which are addressed in this work.

Other research has demonstrated the relevance of meteorological variables derived from satellite-based sources, such as relative humidity, wind speed, temperature, solar radiation, cloud cover, albedo, mean sea level pressure, and surface pressure. These variables were obtained from MERRA-2 (with a spatial resolution of ~69 km) and the ERA5-Land dataset provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Aman et al., 2025; Kunjir et al., 2025). ERA5-Land, a reanalysis dataset generated by replaying the land component of the ERA5 climate reanalysis (Muñoz, 2019), is widely used in environmental modeling as an alternative to limited ground-based weather observations. Since Metro Manila has only two operational weather stations, this study utilizes ERA5-Land along with MODIS products to address spatial data gaps across the region's relatively small land area of 636 km².

A separate study in the Philippines employed artificial neural networks using monthly concentrations of trace gases (NO_2, SO_2, CO, O_3, HCHO, and H_2O) obtained from the Tropospheric Monitoring Instrument (TROPOMI) aboard Sentinel-5P to estimate PM at the national scale, aggregated by regions (De Hitta et al., 2025). However, the coarse spatial and temporal resolution of this study limits its applicability to localized modeling enhancements like those proposed here. While deep learning approaches are planned for future work, this study focuses on refining an existing $PM_{2.5}$ model for Metro Manila using an improved XGBoost-based framework. Prior research consistently showed that tree-based models such as XGBoost and Random Forest achieve higher accuracy when meteorological variables are included in the model inputs (Kunjir et al., 2025).

https://doi.org/10.5194/isprs-annals-X-5-W4-2025-369-2026 | © Author(s) 2026. CC BY 4.0 License.

Meteorology and topography play significant roles in the spatial and temporal changes of $PM_{2.5}$ in the atmosphere as discussed in previous studies. $PM_{2.5}$ concentrations are influenced by climate factors (Rusmili et al, 2023) and the particles' distribution is affected by ground features such as buildings and vegetation. To address these research gaps, this study aims to:

- include additional predictor variables representing topographical and meteorological conditions,
- explore interpolation models to generate gridded $PM_{2.5}$ to fill the spatial gaps in the station measurements,
- incorporate an efficient hyperparameter optimization workflow with XGBoost in improving predictions of ground $PM_{2.5}$ in the region, and
- evaluate prediction results using skill metrics and feature importance analysis.

2. Materials and Methods

2.1 Datasets

The input features to the prediction model are ground $PM_{2.5}$ as the target variable, and satellite-based observations and climate reanalysis data products as predictor variables. The data sources and specifications of these features are provided below.

2.1.1 Ground $PM_{2.5}$: Available records in the region from January 2017 to April 2023 were obtained from the EMB database in that underwent quality checks. EMB manages 15 monitoring stations located in various cities in the region and are situated mostly near urban centers and busy roads, as shown in Figure 1. In this study, we tested models using $PM_{2.5}$ records from 13 stations by excluding stations located at the extreme north (North Caloocan) and extreme south (Muntinlupa) of the region. The enhanced XGBoost model is limited to the central part of the region, bounded by the red dashed polygon.

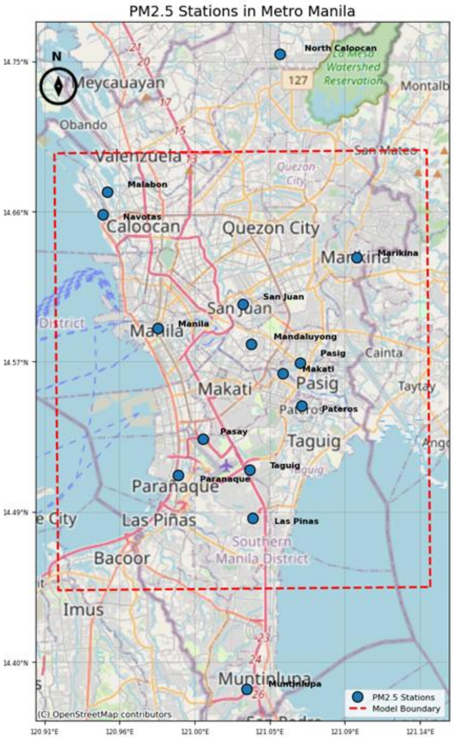


Figure 1. $PM_{2.5}$ stations in Metro Manila and model boundary (in red dashed polygon) for the enhanced XGBoost model.

2.1.2 Predictor variables: Temporally varying meteorological parameters, which impact the $PM_{2.5}$ distribution at least in a daily interval, are added in the prediction process. In addition to AOD, other factors that play a role in estimating $PM_{2.5}$ concentration in the atmosphere are topography and seasonally changing surface characterization (Gupta et.al., 2021). The scattering and absorption of fine particles in the atmosphere are affected by surface properties like land surface temperature (LST) and emissivity. Observations from satellite sensor MODIS and ERA-5 Land reanalysis datasets were used as predictors in calculating $PM_{2.5}$ concentrations in Metro Manila. These datasets are publicly available from the Earth Engine Data Catalog. Available images from 2017 to 2023 were extracted using R scripts utilizing rgee, a binding package for calling the Google Earth Engine API within R. As shown in Table 1, MODIS variables include AOD measured at blue and green bands (Lyapustin and Wang, 2022), direct solar radiation (Wang, 2024), black-sky albedo (Schaaf and Wang, 2021), LST in daytime/nighttime and emissivity (Wan et al., 2021), and NDVI derived from the MODIS/006/MOD09GA ground reflectance composite image provided by Google.

2.2 Methodology

Figure 2 illustrates the specific processes undertaken in this study to improve the predictions of $PM_{2.5}$ using XGBoost. The processing of the variables from ERA-5 Land includes bilinear interpolation to resample daily data from coarse (11 km) to fine resolution (1 km) and spatially align with MODIS observations for data consistency. Bilinear interpolation is based on distance weighted average of four nearest pixels.

Prediction tests cases are grouped by (a) values of input features at station-based locations and (b) values of input features matched on a grid with 1-km cell size within land areas of the dashed boundary shown in Figure 1. Specifically, inputs are (1) a dataframe consisting of daily average $PM_{2.5}$ and predictors matched at station locations (DF1) and (2) using a dataframe based on grid cells encompassing the region with a spatial resolution of 1km (DF2).

2.2.1 $PM_{2.5}$ outlier detection and filtering: Ground $PM_{2.5}$ values are measured in hourly basis but records from EMB show significant gaps and discontinuities in measurements, i.e. 60.39% missing values of hourly data in the period covered in the study. Figure 3 shows the distribution of unfiltered daily average values in the 13 stations. Outliers were detected using Interquartile Range (IQR) method to obtain a filtered dataframe for $PM_{2.5}$. All values below the first quartile and above third quartile were considered as outliers, comprising of 5.97% of the total samples of daily aggregated values. These outliers are attributed to extreme readings on the sensor on those dates with outlier values. Outlier detection techniques can be further improved in future work if data on local emission sources and weather sensors near those stations are available.

2.2.2 Interpolation of ground $PM_{2.5}$ using Kriging: XGBoost performs well on large datasets, hence, the addition of interpolation scheme in pre-processing the ground $PM_{2.5}$ values. Gridded $PM_{2.5}$ values were generated using Ordinary Kriging (OK) interpolation method, an optimal spatial interpolation estimation method in which the value of a variable at an unsampled location is determined according to the linear combination of the known values of all the sampled locations (Rusmili et al, 2023). Similar studies explored widely used interpolation methods Kriging and Inverse Distance Weighting (IDW). Kriging generally provides more accurate estimates as it

Dataset/Code	Description	Spatial and temporal resolution
PM25mean	Ground $PM_{2.5}$ hourly values (in µg/ m^3) at 13 stations aggregated to daily average	1 km (interpolated), daily
MODIS		
AOD047/AOD55	AOD over land retrieved in the bands Blue (0.47 µm)/ Green (0.55 µm) of MCD19A2 V6.1 data product	1 km, daily
DSR	Total incident solar radiation (unit: W/ m^2) over land surfaces in the shortwave spectrum (300-4,000 nm)	
LSTDT/LSTNT	Daytime/ nighttime land surface temperature (in Kelvin) of MOD11A1 V6.1 data	
EMIS31/EMIS32	Surface emissivity values at band 31/32 of MOD11A1 V6.1 product	
BSAvis	Black-sky albedo in the visible band of MCD43A3 V6.1 Albedo Model dataset	
NDVI	Derived from near infrared and red bands of MOD09GA data product	
ERA5-Land		
WINDSPEED	Calculated using eastward (u) and northward (v) components of the 10m wind (in m/s)	1 km (interpolated), daily
TEMP	Air temperature (unit: Kelvin) at 2m above the surface of land, sea or in-land waters	
PRESSURE	Atmospheric pressure (in Pa) on the surface of land, sea and inland water	
PRECIP	Total accumulated liquid and frozen water (unit: m) from large-scale and convective precipitation reaching the earth's surface	

Table 1. Input features for the enhanced XGBoost model of $PM_{2.5}$ in Metro Manila.

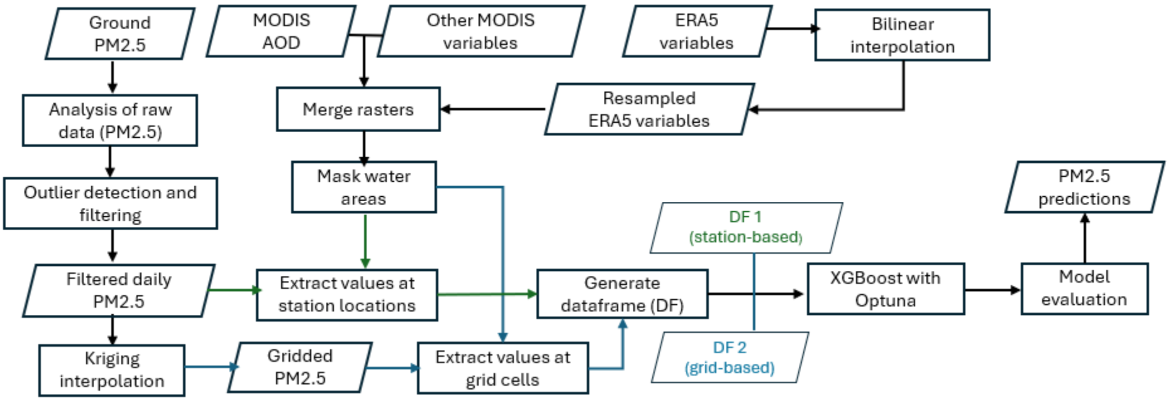


Figure 2. Methodological flowchart for the enhanced XGBoost model of $PM_{2.5}$ in Metro Manila.

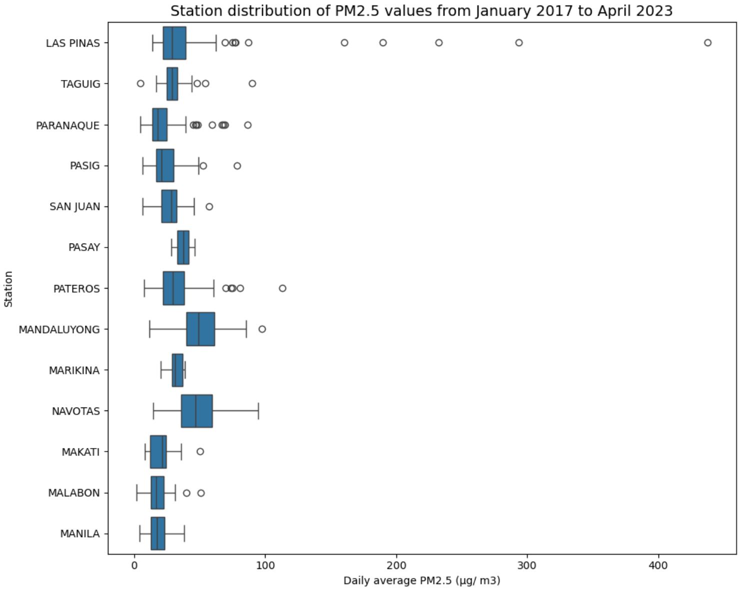


Figure 3. Daily average $PM_{2.5}$ values in 13 stations within Metro Manila from January 2017 to April 2023.

https://doi.org/10.5194/isprs-annals-X-5-W4-2025-369-2026 | © Author(s) 2026. CC BY 4.0 License.

accounts for spatial correlation of the measurements, which IDW often fails to capture as it does not sufficiently represent spatial variability of individual-level $PM_{2.5}$ concentrations within an area (Kim et.al., 2014). With Kriging interpolation, it is assumed that the variations of air pollutant parameters are almost always spatially dependent on some scale (Wong et al., 2004). This approach is combined with other regression techniques to predict air pollution gradients because of very limited sample measurements like in Metro Manila.

Kriging method is applied to predict $PM_{2.5}$ values on unmeasured stations given the characterized mean and variance structures (Kim et al., 2014). OK assumes a constant mean over space and uses variograms to perform unbiased and linear optimal estimates of spatial variables (Lin et al., 2018). Various OK interpolation tests were carried out using combinations of different variogram models (i.e., linear, power, gaussian, spherical and exponential) and *n* neighbors ranging from 2 to 7. Variograms are empirical models that describe the geographical variations of input values and quantifies the spatial correlation between two points (Rusmili et al, 2023). A variogram expresses the degree of similarity between two observations separated by a given distance and is used to calculate weights, which minimize the variance in the estimated value (Wong et al., 2004). In each set of sample points, predicted values are calculated using Equation 1 (Torres et al., 2023b):

$$\hat{Z}\,(x_0) = \sum_{i=1}^{N} \lambda i Z(x_i) \tag{1}$$

where $\hat{Z}$ = predicted value
x_0 = location of the predicted point
λ_i = weight for the measured value at sampled point
Z = measured value
x_i = location of measured value
N = number of measured values.

Kriging estimates are highly dependent on the daily average values, variogram model and number of neighbouring points used in the interpolation. The predictions are validated using Leave-One-Out Cross Validation (LOOCV) approach, wherein one station is excluded in the prediction, and the rest (*n* neighbours) are used to calculate $PM_{2.5}$. In each pair of variogram model and neighbour count, interpolation tests were evaluated using these statistical metrics discussed in previous studies (Rusmili et al., 2023) (Lin et al.,2018)(Kim et al., 2014): Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Coefficient of determination R^2 and Standardized Error (SE), which is the relative standard deviation of prediction errors.

2.2.3 Regression modeling of $PM_{2.5}$ using XGBoost with Optuna hyperparameter optimization: To enhance the existing $PM_{2.5}$ model developed for Metro Manila, this study employs an open source hyperparameter optimization software Optuna integrated into the XGBoost modeling workflow. Optimizing the learning parameters addresses the uncertainty and cost of manual tuning. Optuna allows users to dynamically construct the parameter search space and its architecture handles both small- and large-scale experiments with minimum setup requirements (Akiba et al, 2019). When a Bayesian optimization algorithm like Optuna is used with XGBoost, rather than traditional grid or random search method, it can significantly enhance prediction accuracy (Song et al., 2023). Optuna utilizes the Tree-structured Parzen estimator and formulates the hyperparameter optimization as a process of minimizing/maximizing an objective function that takes a set of hyperparameters as an input and returns its validation score (Akiba et al, 2019). It uses past prediction trials to search for better hyperparameters efficiently. The objective function defined in 50 prediction trials is minimizing MAE/MSE/RMSE and maximizing R^2 for both train dataset (70% of the samples) and test dataset (30% of the samples). Prediction test cases utilizing DF1 (station-based) are divided into these sub-groups with specific conditions:

- Case 1: Same predictors (AOD), stations (15) and period (2017 to 2020) used in previous study (Torres et al., 2023a)
- Case 2: Same predictors and stations but with extended period (2017 to 2023)
- Case 3: Same predictors and same period but measurements are selected at 13 stations
- Case 4: Same predictors, measurements at 13 stations and extended period

Prediction test cases utilizing DF2 (grid-based) are divided into these sub-groups:

- Case 5: Same predictors, gridded $PM_{2.5}$ based on 13 stations, extended period
- Case 6: Additional predictors (given in Table 1), gridded $PM_{2.5}$ based on 13 stations, extended period

The hyperparameters and range of values set in the prediction trials are summarized in Table 2. K-Fold cross-validation with 5 folds (k) is implemented in all test cases. In Cases 5 to 6, the best Kriging models (i.e., pair of variogram model and number of neighbors) with lowest RMSE, lowest MAE, highest R^2 and lowest SE were used to generate interpolated $PM_{2.5}$ values.

Hyperparameter	Values
n_estimators	100 to 2000 (step=100)
learning_rate	0.01 to 0.26
max_depth	3 to 20
colsample_bytree	0.5 to 1.0
subsample	0.5 to 1.0
lambda	0.0 to 5.0
alpha	0.0 to 5.0
gamma	0.0 to 0.5
tree_method	'hist'

Table 2. Range of values of hyperparameters optimized using Optuna framework applied in XGBoost.

2.2.4 Evaluation of XGBoost model results: Skill metrics for the train data and test data were calculated in the 6 test cases and were evaluated based on lowest RMSE and highest R^2. Feature importance was assessed in the models through F-scores by gain, which correspond to the average loss reduction caused by splits on a feature. To supplement the feature importance analysis, SHapley Additive exPlanations (SHAP) plots were also generated to check the impact of each feature on the predicted values. SHAP method was also applied in previous studies to understand the relative importance of the defined input features in explaining $PM_{2.5}$ predictions (Aman et al., 2025).

3. Results and Discussion

3.1 Comparison with previous XGBoost model

The study of Torres et al (2023a) employed XGBoost with RandomizedSearchCV approach in predicting $PM_{2.5}$ with AOD as predictor variables based on locations of 15 stations. Table 3 shows a comparison of the filtered dataframes containing daily average $PM_{2.5}$ values and quantile values. Previous study achieved low R^2 and low RMSE = 9.62 μg/ m^3 on dataset grouped by year. In this study, specific to Case 1 as shown in Table 4, all samples were used as input features in the enhanced XGBoost model and the results indicate lower accuracy, achieving relatively higher test RMSE = 10.136 μg/ m^3.

Value	Torres et al (2023a)	This study
Minimum	0	1.698
Maximum	50	55.367
1st quantile	15	14.876
3rd quantile	30	28.222

Table 3. Filtered $PM_{2.5}$ values compared with previous study.

Prediction case	Train RMSE	Train R^2	Test RMSE	Test R^2
Case 1	7.262	0.307	10.136	-0.227
Case 2	6.926	0.370	10.317	-0.271
Case 3	8.820	0.215	10.712	0.005
Case 4	8.387	0.291	11.061	-0.061

Table 4. Results of predictions tests for cases 1 to 4 using DF1

3.2 Kriging interpolation results for $PM_{2.5}$

The best performance statistics of each pair of variogram model and number of neighbours in obtaining gridded $PM_{2.5}$ values are summarized in Table 5. Interpolation results indicate three (3) model combinations that best estimates $PM_{2.5}$ at 1-km grid. The lowest RMSE is achieved using exponential variogram with n=3 (Exp3) while the lowest MAE is achieved using spherical variogram with n=3 (Sph3). The lowest SE and highest R^2 values were achieved using power variogram with n=5 (Pow5). Sample maps of these model pairs on common dates are provided in Figure 4.

Variogram	RMSE	MAE	R^2	SE
Exponential (n =3)	9.78	8.56	0.85	1.34
Linear (n =3)	9.99	8.71	1.00	1.37
Spherical (n =3)	9.80	8.55	0.84	1.34
Gaussian (n = 2)	9.81	8.79	1.12	1.43
Power (n = 5)	13.39	10.73	0.80	1.32

Table 5. Average scores from Kriging interpolation

The filtered $PM_{2.5}$ dataset used in DF1 contains 763 samples, with 207 unique dates. Applying Kriging interpolation requires that the samples are spatially and temporally aligned. To ensure that there are enough $PM_{2.5}$ stations with non-zero values per day, each variogram model identified valid dates to consider in the interpolation process. Since there are a lot of gaps in the ground measurements in terms of temporal continuity, measurements on some days have low variance. Hence, gridded $PM_{2.5}$ values are only available on those valid dates. Both exponential and spherical models used 29 valid dates while power model only used 11.

The removal of outliers in the dataframes is crucial in the interpolation process. Initial tests include outliers, and the estimates were observed to be skewed on these extremely high values of $PM_{2.5}$. The statistical results of these tests with outliers showed lower accuracy for the different combinations of variogram model and neighbour count. Comparing scores with those provided in Table 5, initial interpolation tests generated these average scores with poor accuracy: RMSE ranges from 16.65 to 67.10, MAE ranges from 12.99 to 49.59, R^2 ranges from 0.38 to 4.79 and SE ranges from 0.40 to 5.70.

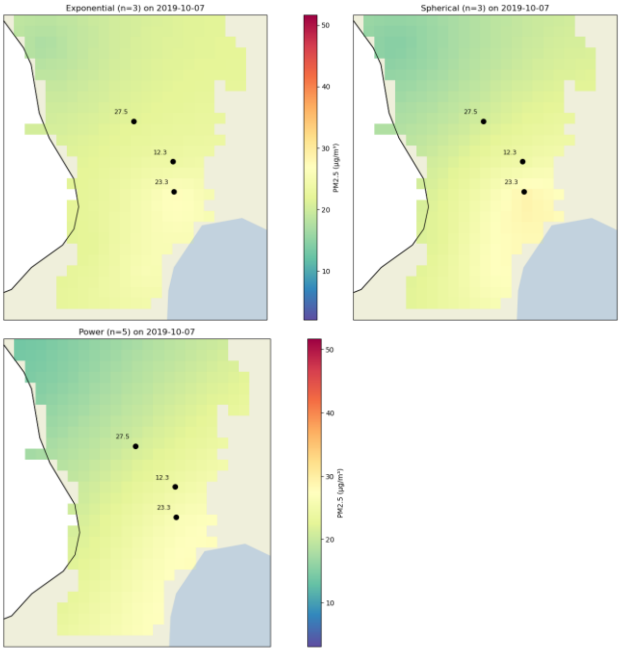

Exponential (n=3) on 2019-10-07
Spherical (n=3) on 2019-10-07
Power (n=5) on 2019-10-07
27.5
12.3
23.3
PM2.5 (μg/m³)
50
40
30
20
10

Figure 4. Interpolated $PM_{2.5}$ values on sample date (2019-10-07) common to the Kriging models.

3.3 Prediction results of the enhanced XGBoost model combined with gridded $PM_{2.5}$

3.3.1 Performance metrics: Table 6 shows the skill metrics of the enhanced XGBoost model combined with the gridded or interpolated $PM_{2.5}$ values based on the three Kriging models. Case 5 includes AOD only as predictors while Case 6 involves 13 predictors, with data distribution plots provided in Appendix A. The predictive accuracy of these models was assessed using these metrics to ensure robustness and generalization. CV metrics, based on average values across folds, provide the model's generalization performance while Test Set metrics indicates model accuracy. Checking overfitting in the model is done by comparing these two sets of metrics.

In 50 prediction trials using k-Fold CV (k=5), the spherical model (n=3) achieves the highest accuracy evaluated on the test data for both Cases 5 and 6. Adding predictor variables and adding years of samples for the spherical + XGBoost model resulted to higher test R^2 (from 0.2000 to 0.7279) and lower test RMSE (from 1.8454 to 1.0762). The final evaluation on the test set supports the cross-validation results, showing consistent accuracy and generalization. Training set and full-data evaluations show similar performance, indicating minimal overfitting. Additionally, close values between the training and test metrics indicate that the model generalizes the predictions well and is not overfitting.

Looking at the test R^2 and test RMSE in all cases, the model is sensitive to the number of samples. The samples are increasing from Case 1 to 6, and it is observed that these metrics improve as we add more predictors and samples. Like previous studies, XGBoost works well on large datasets and achieves higher accuracy with more predictors and samples in the training process. Results also show that gridded values, presented in Cases 5 and 6, addressed overfitting as metrics in the training, test and full datasets have smaller deviations compared to the metrics in Cases 1 to 4.

Metrics		Exp3	Sph3	Pow5
Case 5: AOD047 and AOD55, 4 years				
CV Train	RMSE	1.9696	1.8392	5.7542
	MAE	1.6388	1.6768	4.8735
	R^2	0.1621	0.2005	0.0461
CV Test	RMSE	1.9744	1.8612	5.7540
	MAE	1.6471	1.7029	4.8750
	R^2	0.1520	0.1735	0.0457
Train Set	RMSE	1.9701	1.8398	5.7543
	MAE	1.6400	1.6785	4.8793
	R^2	0.1620	0.2005	0.0461
Test Set	RMSE	1.9014	1.8454	5.7619
	MAE	1.6280	1.6857	4.8337
	R^2	0.1486	0.2000	0.0555
Full Data	RMSE	1.9498	1.8415	5.7566
	MAE	1.6364	1.6806	4.8656
	R^2	0.1582	0.2003	0.0490
Case 6: All predictors, 6 years				
CV Train	RMSE	1.2749	1.0436	4.7682
	MAE	0.6273	0.5708	2.5267
	R^2	0.6489	0.7425	0.5996
CV Test	RMSE	1.3008	1.0748	4.7957
	MAE	0.6571	0.6092	2.5737
	R^2	0.6318	0.7240	0.5945
Train Set	RMSE	1.2750	1.0436	4.7677
	MAE	0.6258	0.5696	2.5187
	R^2	0.6490	0.7427	0.5998
Test Set	RMSE	1.2165	1.0762	4.7238
	MAE	0.6027	0.6108	2.4864
	R^2	0.6515	0.7279	0.6027
Full Data	RMSE	1.2578	1.0535	4.7546
	MAE	0.6189	0.5820	2.5090
	R^2	0.6497	0.7383	0.6006

Table 6. Skill metrics of enhanced XGBoost models trained using K-Fold CV approach (k = 5) and gridded $PM_{2.5}$ values.

3.3.2 Optimized hyperparameters: Table 7 shows the optimized hyperparameters for each model in Case 6. All models utilized high n_estimators and max_depth and low learning_rate values in the prediction tests. Regularization parameter values resulting in each model differ largely in the lambda (L2 or ridge regularization) which helps prevent overfitting.

Hyperparameter	Exp3	Sph3	Pow5
n_estimators	1600	1500	1000
learning_rate	0.13465	0.03531	0.04721
max_depth	11	15	20
colsample_bytree	0.69046	0.75223	0.63001
subsample	0.870501	0.67266	0.94684
lambda	0.29375	4.91975	1.60550
alpha	0.04793	0.07173	0.38186
gamma	0.15961	0.08927	0.25431

Table 7. Optimized hyperparameters with gridded inputs.

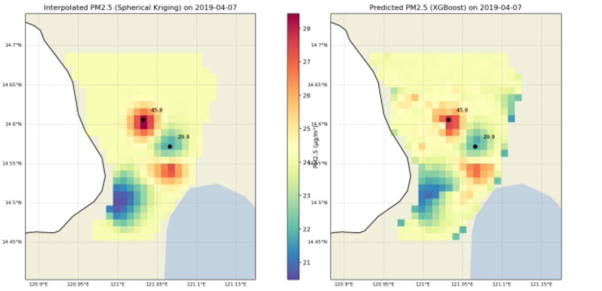


Figure 5. Sample interpolated $PM_{2.5}$ and XGBoost predictions.

Results showing the observed (interpolated values using Spherical model) and predicted $PM_{2.5}$ on a sample date is illustrated in Figure 5. Point stations with measurements on that day is overlaid for comparison with the gridded values. Clusters of high $PM_{2.5\ v}$ values are located near these stations while clusters of low values mostly appear on the southwest part of the region.

3.4 Feature Importance Analysis

The importance of the predictor variables in the best model (spherical, n=3) was evaluated using XGBoost F-Score by Gain and SHAP plots. For the Exponential model, the top 5 variables with relatively high gain scores or the features that caused the largest average reduction in prediction errors are EMIS31, DSR, LSTDT, AOD047 and temp. For the Spherical model. The top 5 important features are precipitation, DSR, LSTDT, air temperature and EMIS31. Table 8 shows the F-scores obtained for Exponential and Spherical models. For Power model, the F-scores obtained are skewed on the precipitation variable and other variables with relatively high F-scores are pressure, temperature and wind speed.

Predictor	F-Score (Exp3)	F-Score (Sph3)
windspeed	1.4	1.3
temp	6.9	4.5
pressure	3.7	3.6
precip	5.5	19.3
NDVI	3.8	2.8
LSTNT	1.3	0.9
LSTDT	9.2	7.2
EMIS32	0.7	0.6
EMIS31	83.2	4.4
DSR	13.2	16.2
BSAvis	4.2	1.6
AOD055	3.7	2.8
AOD047	7.9	3.1

Table 8. XGBoost feature importance plot showing F-scores by Gain for Exponential and Spherical models

Additional insights on the feature importance for the Spherical model are illustrated in the SHAP plot in Figure 6. These plots indicate the impact of each feature on the $PM_{2.5}$ predictions based on SHAP values and are ranked from most to least important. Each dot on the plot is a single prediction and are classified as high (if it increases the prediction) and low (if it decreases the prediction).

The interpretation from these results based on the SHAP values of the top features shown in Figure 6 are summarized below:
(1) DSR points with high values have positive SHAP value, indicating that high DSR tends to increase the predicted value of $PM_{2.5,}$
(2) High precipitation values lead to negative SHAP values, indicating that higher precipitation value in the area decreases the concentration of $PM_{2.5,}$
(3) Although NDVI points with negative SHAP values are mixed of highs and lows, the plot suggests that most values tend to decrease the predicted value of $PM_{2.5,}$
(4) LSTDT exhibits a similar pattern with DSR, where high values are associated with positive SHAP values. This indicates that high daytime temperatures tend to increase the predicted $PM_{2.5}$ concentrations. On the other hand, low LSTDT values cluster around SHAP values near zero, suggesting that low daytime temperatures have minimal impact on the model's $PM_{2.5}$ prediction.

https://doi.org/10.5194/isprs-annals-X-5-W4-2025-369-2026 | © Author(s) 2026. CC BY 4.0 License.

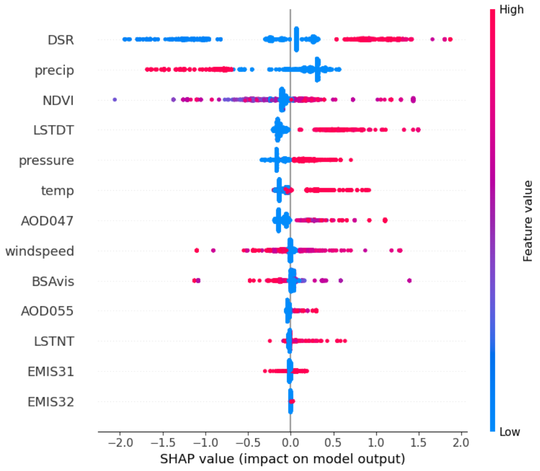


Figure 6. SHAP values of features from the Spherical (n=3) model trained in XGBoost.

4. Conclusion and Recommendations

This paper presents the improvements made on the existing model for $PM_{2.5}$ in Metro Manila, Philippines using XGBoost regression approach. The model developed in this study has these key enhancements in improving the predictions of $PM_{2.5}$: (1) inclusion of additional input features or predictor variables from MODIS observations and ERA-5 Land climate reanalysis datasets to represent topographical and meteorological parameters that can affect spatial and temporal distribution of $PM_{2.5}$ within the region and (2) use of hyperparameter optimization workflow Optuna to develop a model with better generalization performance and higher predictive accuracy. The enhanced XGBoost model, when combined with a gridded $PM_{2.5}$ generated by Kriging interpolation, achieves the highest R^2 (0.73) and lowest RMSE (1.08 μg/ m^3). To understand the impacts of the defined predictor variables, feature importance was evaluated through XGBoost F-scores and SHAP values. Results indicate that DSR, precipitation, LSTDT, air temperature and AOD in the 0.47 μm band have relatively high importance in estimating daily $PM_{2.5.}$ The methodological workflow and findings presented in this study can be extended in assessing the long-term trends and seasonal variations of $PM_{2.5}$.in the region. We recommend the inclusion of past sensor readings and most recent ground measurements of $PM_{2.5}$ in the training process to address the spatial and temporal gaps introduced by the existing stations. It is also recommended to explore imputation methods on missing data or gaps in the satellite observations utilized in the regression model, as well as other statistical metrics in the model evaluation. In applications related to human health exposure and risk, we also recommend that the model developed in this study is extended to hourly datasets to capture the diurnal variations of $PM_{2.5.}$ For future work, other remotely sensed and climate reanalysis datasets with hourly resolution that are publicly available will be considered in modelling $PM_{2.5.}$

Acknowledgements

The authors would like to acknowledge the Department of Science and Technology of the Philippines for the scholarship funding provided through the Engineering Research and Development for Technology. We are also grateful for the technical support of Mr. Ryoga Hiroki for the hyperparameter optimization workflow.

References

Akiba, T., Sano S., Yanase T., Ohta, T., Koyama, M., 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. doi.org/10.48550/arXiv.1907.10902.

Aman, N., Panyametheekul, S., Pawarmart, I., Sudhibrabha, S., Manomaiphiboon, K., 2025. A Visibility-Based Historical $PM_{2.5}$ Estimation for Four Decades (1981–2022) Using Machine Learning in Thailand: Trends, Meteorological Normalization, and Influencing Factors Using SHAP Analysis. *Aerosol Air Qual. Res. 25*, 4. doi.org/10.1007/s44408-025-00007-z.

De Hitta, M.A.S., Castro, E.C., Blanco, A.C., Felix, M. J., 2025. Estimation of PM concentrations in the Philippines using Sentinel-5P and artificial neural network, *Proc. SPIE 13262*, *Remote Sensing of the Atmosphere, Clouds, and Precipitation VIII*, 132620R. doi.org/10.1117/12.3042769.

Gupta, P., Zhan, S., Mishra, V., Aekakkararungroj, A., Markert, A., Paibong, S., Chishtie, F., 2021. Machine Learning Algorithm for Estimating Surface PM2.5 in Thailand. *Aerosol and Air Quality Research.* doi.org/10.4209/aaqr.210105.

Kim, Y., Yi, J., Eum, Y. S., Choi, J., Shin, H., Ryou, H. G., & Kim, H., 2014. Ordinary kriging approach to predicting long-term particulate matter concentrations in seven major Korean cities. *Environmental Health and Toxicology,* 29, e2014012. doi.org/10.5620/eht.e2014012.

Kunjir, G.M., Tikle, S., Das, S., Karim, M. Roy, S.K., Chatterjee, U., 2025. Assessing particulate matter ($PM_{2.5}$) concentrations and variability across Maharashtra using satellite data and machine learning techniques. *Discov Sustain 6*, 238. doi.org/10.1007/s43621-025-01082-3.

Lin, J., Zhang, A., Chen, W., & Lin, M., 2018. Estimates of Daily PM2.5 Exposure in Beijing Using Spatio-Temporal Kriging Model. *Sustainability*, *10*(8), 2772. doi.org/10.3390/su10082772.

Lyapustin, A., Wang, Y., 2022. MODIS/Terra+Aqua Land Aerosol Optical Depth Daily L2G Global 1km SIN Grid V061 [Data set]. NASA EOSDIS Land Processes Distributed Active Archive Center. doi.org/10.5067/MODIS/MCD19A2.061.

Muñoz Sabater, J., 2019: ERA5-Land hourly data from 1950 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). doi.org/10.24381/cds.e2161bac.

Rusmili, S. H. A., Mohamad Hamzah, F., Choy, L. K., Azizah, R., Sulistyorini, L., Yudhastuti, R., Chandraning Diyanah, K., Adriyani, R., & Latif, M. T., 2023. Ground-Level Particulate Matter ($PM_{2.5}$) Concentration Mapping in the Central and South Zones of Peninsular Malaysia Using a Geostatistical Approach. *Sustainability*, *15*(23), 16169. doi.org/10.3390/su152316169.

Schaaf, C., Wang, Z., 2021. MODIS/Terra+Aqua BRDF/Albedo Daily L3 Global - 500m V061 [Data set]. NASA EOSDIS Land Processes Distributed Active Archive Center. doi.org/10.5067/MODIS/MCD43A3.061.

Song, T., Zhang, C., Jin, X., Zhao,X., Huang, W., Sun, X., Yang, Z., Wang, s., 2023. Spatial prediction of PM2.5 concentration using hyper-parameter optimization XGBoost model in China, *Environmental Technology & Innovation*, Volume 32, 103272, ISSN 2352-1864. doi.org/10.1016/j.eti.2023.103272.

Torres, R. A. B., Ramos, R. V., Recto, B. A. B., Tamondong, A. M., Jiao, B. J. D., and Cayetano, M. G., 2023a: Estimating airborne particulate matter in the National Capital Region, Philippines using Multiple Linear Regression and Gradient Boosting Algorithm on MODIS MAIAC Aerosol Optical Depth. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, X-4/W1-2022, 729–736. doi.org/10.5194/isprs-annals-X-4-W1-2022-729-2023.

Torres, R.A., Ramos, R., Recto, B.A., Tamondong, A., 2023b. Analysis of Kriging Interpolation Models for Particulate Matter Levels in the National Capital Region, Philippines. *2023 Asian Conference on Remote Sensing (ACRS2023) Proceedings by the Asian Association on Remote Sensing*.

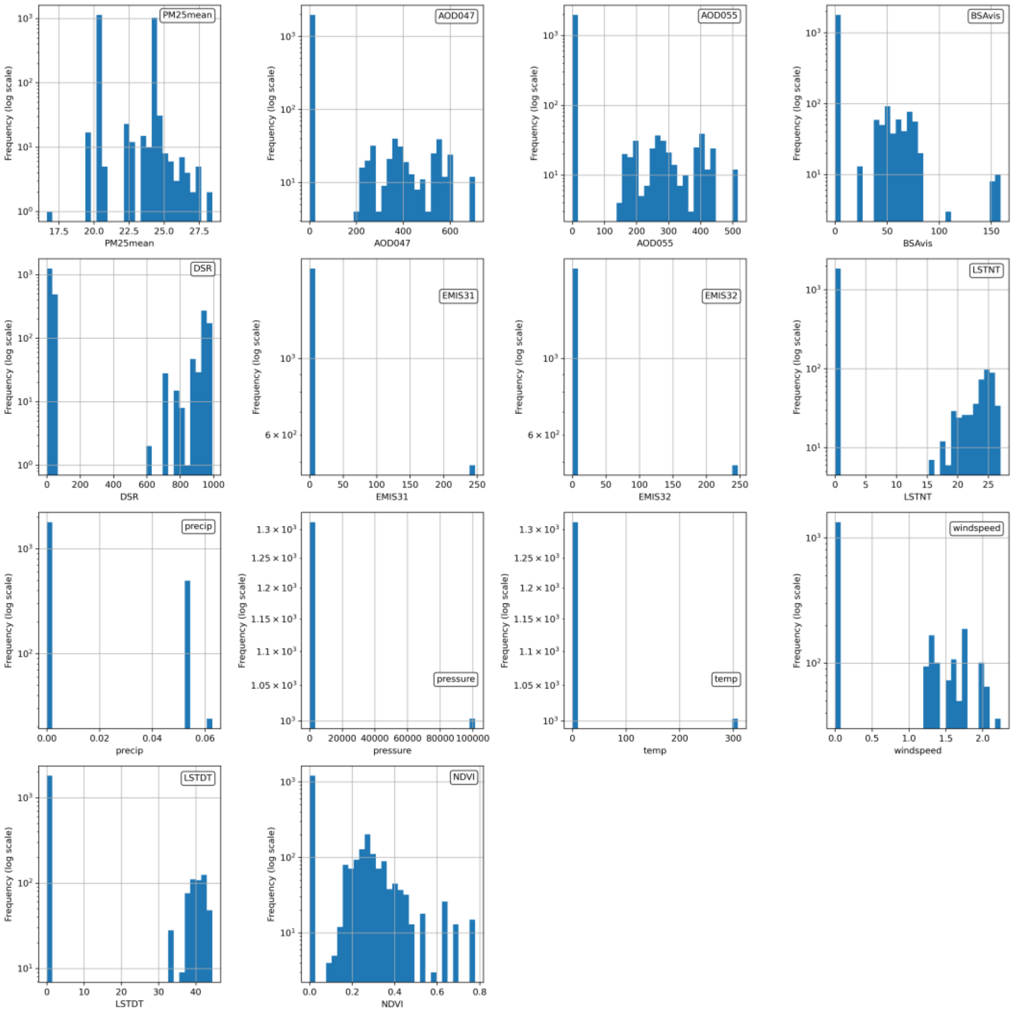
Torres, R. A. B., Ramos, R. V., Recto, B. A. B., and Tamondong, A. M., 2024: Development of Land-Use Regression Models for particulate matter estimation in National Capital Region, Philippines, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVIII-4/W8-2023, 437–444. doi.org/10.5194/isprs-archives-XLVIII-4-W8-2023-437-2024.

Wan, Z., Hook, S., Hulley, G., 2021. MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V061 [Data set]. NASA EOSDIS Land Processes Distributed Active Archive Center. doi.org/10.5067/MODIS/MOD11A1.061.

Wang, D., 2024. MODIS/Terra+Aqua Surface Radiation Daily/3-Hour L3 Global 1km SIN Grid V062 [Data set]. NASA EOSDIS Land Processes Distributed Active Archive Center. doi.org/10.5067/MODIS/MCD18A1.062.

Wong, D. W., Yuan, L., & Perlin, S. A., 2004. Comparison of spatial interpolation methods for the estimation of air quality data. *Journal of Exposure Science & Environmental Epidemiology*, *14*(5), 404-415. doi.org/10.1038/sj.jea.7500338.

Appendix.



Appendix A. Log-scaled data distribution plots of the input features with gridded $PM_{2.5}$.