Evaluating different satellite-based Aerosol Optical Depth (AOD) in predicting inland daytime $PM_{2.5}$ using machine learning-based regression approach

Roseanne V. Ramos [1,2], Mark Joseph R. Calubad [3], 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
[3]Department of ICT Integrated Ocean Smart City Engineering, Dong-A University, Busan, South Korea – mjcalubad@donga.ac.kr

Keywords: satellite-based AOD, $PM_{2.5}$, daytime, machine learning, regression

Abstract

Aerosols play a critical role in the development of the boundary layer and build-up of air pollution in urban environments. Their presence in the atmosphere is quantified by Aerosol Optical Depth (AOD). Satellite sensors observe and retrieve AOD at varied spatial and temporal resolutions. In air quality monitoring, satellite-based AOD products are typically utilized to predict ground concentrations of fine particulate matter ($PM_{2.5}$) through various modelling approaches. This study evaluates AOD products observed by Moderate Resolution Imaging Spectroradiometer (MODIS), Visible Infrared Imaging Radiometer Suite (VIIRS) and the Advanced Himawari Imager (AHI) of Himawari-8 in predicting inland daytime $PM_{2.5}$ concentrations for test sites in Japan and South Korea. Prediction models were constructed using eXtreme Gradient Boosting (XGBoost) regression with input variables from observation datasets matched on ground $PM_{2.5}$ station locations. In addition to AOD, twelve (12) predictor variables representing topographic and meteorological parameters were considered. Prediction results were evaluated using Kruskal-Wallis test and effect size analysis to compare the absolute error distributions across AOD products. Statistical results indicate that while models utilizing MODIS AOD generalize better to new data, the overall difference in prediction accuracy is statistically negligible. These findings suggest that no single AOD product is significantly superior in predicting ground-level $PM_{2.5}$ concentrations, highlighting the potential of integrating AOD values from various sources. This work is essential for improving the accuracy of $PM_{2.5}$ estimates and for supporting more effective mapping of urban air pollution.

1. Introduction

Aerosols significantly influence the development of the atmospheric boundary layer and contribute to the air quality in urban environments (Jung et al., 2019). Numerous studies have investigated their sources, chemical composition and transport in the lower troposphere where the effects of the boundary layer are crucial. Beyond reducing visibility, high atmospheric aerosol concentrations pose severe health risks to populations exposed to inhalable particles, leading to respiratory diseases (Padimala and Matli, 2024). Monitoring these air pollutants is therefore essential in environmental and epidemiological studies.

Aerosol Optical Depth (AOD), which can be derived from either ground-based *in situ* sun photometers or from radiometers aboard satellite platform, serves as a key measure of the vertically integrated concentration of aerosols in the atmosphere. Compared to *in situ* observations, satellite-based AOD enables area-wide monitoring of tropospheric aerosols (Handschuh et al., 2022). These satellite-derived values are typically used to generate models for estimating ground-level fine particulate matter ($PM_{2.5}$), with their non-linear relationship established in previous studies (Padimala and Matli, 2024). The accuracy of these $PM_{2.5}$ models, however, depends on the input datasets and the geographical context in which they are applied. For instance, Xiao et al (2016) evaluated AOD products from Visible Infrared Imaging Radiometer Suite (VIIRS) Geostationary Ocean Color Imager (GOCI) and Moderate Resolution Imaging Spectroradiometer (MODIS) over East Asia, comparing them with AOD measurements from AErosol RObotic NETwork (AERONET) stations in 2012 and 2013. Despite these findings, to the best of our knowledge, there are no recent published assessments of these AOD products related to their current applicability in predicting inland daytime $PM_{2.5}$ concentrations across Japan and South Korea.

$PM_{2.5}$ concentrations close to the ground surface are estimated through various modelling approaches such as numerical, statistical and machine learning (ML) models (Vignesh et al., 2023). ML models have been widely used in predicting ground $PM_{2.5}$ concentrations because of their superior computational efficiency and their ability to capture complex nonlinearity between pollutant concentrations and predictor variables (Tang et al., 2024). Previous studies explored different ML algorithms in predicting $PM_{2.5}$ such as random forest, support vector regression, lasso regression, decision tree and gradient boosting (Kumar et al., 2024). Among these, gradient boosting methods are more advantageous for large datasets as they are scalable and can handle high-dimensional datasets (Wang et al., 2025). Hence, in this study, we employed ML-based regression modelling using gradient boosting to capture the complex, non-linear relationship between surface $PM_{2.5}$ and meteorological and topographic variables using observation datasets.

This work aims to evaluate different satellite-based AOD products with these specific objectives:

- to construct AOD-specific regression models from various AOD products and auxiliary observations in the prediction of inland daytime $PM_{2.5}$ concentrations, and
- to compare the accuracy of the model prediction across two regional sites.

https://doi.org/10.5194/isprs-archives-XLIX-B2-2026-1439-2026

2. Datasets and Methods

2.1 Datasets

Ground $PM_{2.5}$ measurements were obtained for two (2) regions: Kanto and Busan, which are located in Japan and South Korea, respectively (see Figure 1). The sampling period for this study is year 2021, with hourly $PM_{2.5}$ measurements for Kanto Region in Japan and Busan in South Korea obtained from the Atmospheric Environmental Regional Observation System (AEROS) operated by the National Institute for Environmental Studies (NIES) of Japan and the Korean Ministry of Environment, respectively. The stations in these regions (180 ground stations in Kanto region and 72 in Busan) are generally located within urban areas with a few stations near the coastline. In this study, we excluded roadside stations and water areas in the processing and analysis of the datasets.

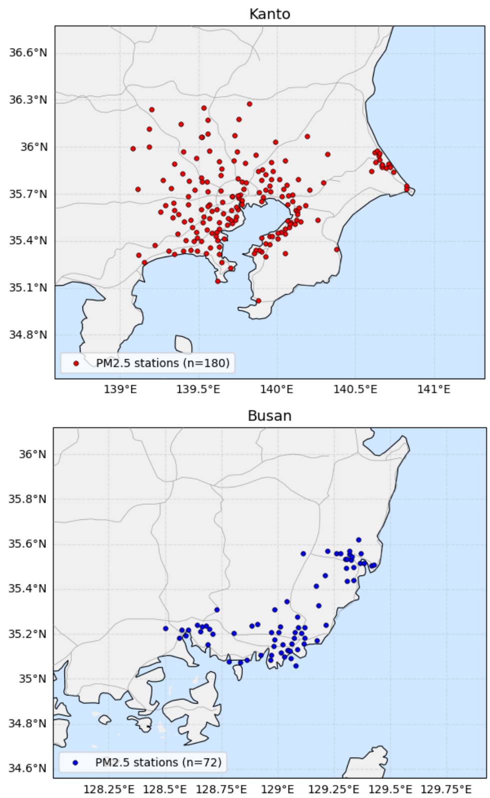


Figure 1. $PM_{2.5}$ stations in test sites Kanto region in Japan (top) and Busan in South Korea (bottom).

Figure 2 shows the data distribution of ground $PM_{2.5}$ concentrations per season in the Kanto region and Busan. In both sites, the winter season includes months of December-January-February (DJF), spring season includes months of March-April-May (MAM), summer season includes months of June-July-August (JJA) and autumn season includes months of September-October-November (SON). Table 1 provides the descriptive statistics of the hourly measurements in 2021 for both sites. The values show that the maximum $PM_{2.5}$ concentrations were recorded during the spring season with values 203 and 192 μg/m^3 in Busan and Kanto, respectively. The average concentrations in Busan are higher, with 13 to 53% of the 24-hour average values in 2021 exceeding the World Health Organization (WHO) guideline value (15 μg/m^3). In the study of Kang et al (2023), it was stated that the atmospheric environment of East Asian countries is significantly affected by Asian dust originating from deserts and plateaus of China and Mongolia and these dust emissions are confined to the lower troposphere. Although both test sites experience the effects of these dust emissions, each region is affected differently due to local emission characteristics, topography and meteorological factors that influence the spatial and temporal distribution of $PM_{2.5}$.

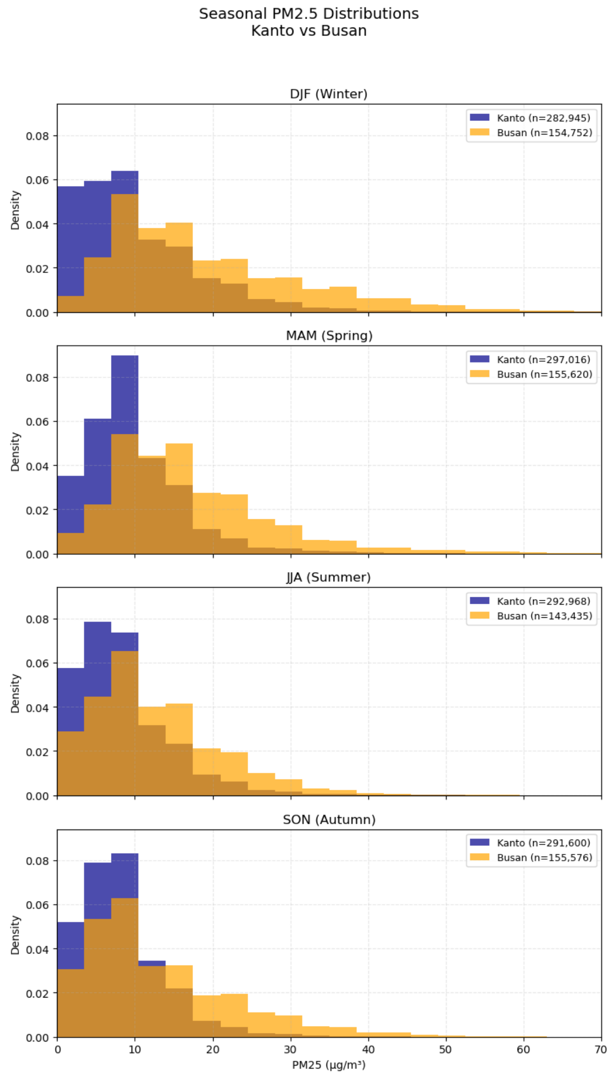


Figure 2. Seasonal data distribution (99th percentile) of ground $PM_{2.5}$ in 2021 for the Kanto region and Busan. The brown region indicates the distribution of $PM_{2.5}$ concentrations common to both sites.

Site	Season	Mean ± SD	IQR	Min-Max	%>15
Kanto	Winter	10.0 ± 8.0	10	0–149	2.7
	Summer	8.2 ± 6.1	7	0–94	7.6
	Spring	9.5 ± 6.2	7	0–192	7.6
	Autumn	8.0 ± 5.5	7	0–113	3.3
Busan	Winter	18.7 ± 12.3	15	0–115	13.4
	Summer	12.4 ± 8.1	11	0–87	28.3
	Spring	17.6 ± 14.3	12	0–203	53.3
	Autumn	13.0 ± 9.6	12	0–94	28.6

Table 1. Hourly measured $PM_{2.5}$ statistics by season in 2021 (%>15 = percentage of samples exceeding 24-hour average WHO guideline value of 15 μg/m^3).

Considering both meteorology and topographical parameters, predictor variables were defined according to their direct and indirect contributions to the formation and transport of atmospheric $PM_{2.5}$ and their availability. Observations from different satellite sensors (i.e., MODIS, VIIRS onboard NOAA-20 and Advanced Himawari Imager (AHI) of Himawari-8) and climate reanalysis datasets (i.e., ERA5 and ERA5-Land) were extracted from publicly accessible databases. The observation datasets with their varying spatiotemporal resolutions are shown in Table 2.

Predictor	Data Source (Resolution)
[1] AOD550 or [2] AOT_Land or [3] AOT_Merged	[1] MODIS MAIAC AOD measured at 550nm (1 km, daily at ~10AM); [2] VIIRS L2 AOD (6 km, daily at ~1PM); [3] AHI Himawari-8 L3 AOD measured at 500nm (5 km, hourly at daytime)
relative humidity, wind speed, temperature, surface pressure and total precipitation	ERA5-Land climate reanalysis data (11.13 km, hourly)
mean surface downward shortwave radiation flux, albedo from direct and diffuse radiation in the visible bands, boundary layer height	ERA5 climate reanalysis data (27.83 km, hourly)
vegetation fraction	MODIS (1 km, daily) derived from Normalized Difference Vegetation Index (NDVI)
leaf area index	VIIRS (5.56 km, daily)
distance from the coastline	NASA (invariant)

Table 2. Predictor variables in estimating inland daytime $PM_{2.5}$

2.1.1 Satellite AOD products: The AOD product observed by MODIS is a Level 2 (L2) gridded data over land surfaces with correction applied using the Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm. MODIS MAIAC AOD is derived using the global measurements from both Terra and Aqua sensors and produced daily at 1km resolution (Lyapustin and Wang, 2018). In this study, we utilized the AOD values measured at 550nm wavelength (denoted in this paper as AOD550). The MODIS MAIAC dataset is publicly distributed by NASA (Wang, D., 2024). VIIRS produced L2 aerosol products from NOAA-20, formerly known as the Joint Polar Satellite System (JPSS-1) satellite, specifically the 6 km Version 2.0 Level 2 aerosol product AERDB_L2_VIIRS_NOAA-20. This AOD product is retrieved using Deep Blue algorithm (Lee et al., 2024) and is distributed by NASA Level-1 and Atmosphere Archive & Distribution System (LAADS) Distributed Active Archive Center (DAAC) with daily global coverage. In this study, the *Aerosol_Optical_Thickness_550_Land_Best_Estimate* dataset (denoted in this paper as *AOT_Land*) is used to represent daily AOD observed by VIIRS, with orbital pass at approximately 13:30 local time. For the AHI Himawari-8 AOD product, hourly daytime Level 3 (L3) gridded product at 500nm was supplied by the P-Tree System of the Japan Aerospace Exploration Agency (JAXA) at 0.05 degree (approximately 5-km) spatial resolution. L3 aerosol products are aggregated hourly based on L2 AOD datasets. We evaluated the *AOT_Merged* values, the spatial and temporal optimum interpolation of *AOT_Pure* within an hour, which were retrieved using the methodology of Kikuchi et al (2018) combining aerosol optical thickness with strict cloud-screening. Tan et al (2022) evaluated hourly AOD products of Himawari-8 over Japan through comparison with different ground monitoring networks recording AOD measurements and found that the *AOT_Merged* variable had the best performance.

2.1.2 Climate reanalysis datasets: Meteorological parameters were extracted from ERA5 hourly data on single levels and ERA5-Land hourly datasets derived from climate reanalysis datasets. ERA5, the fifth generation European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis for the global climate and weather for the past 8 decades, provides hourly estimates of atmospheric quantities on single levels with a spatial horizontal resolution of 0.25° x 0.25° or approximately 27.8 km (Hersbach et al., 2023). ERA5-Land, on the other hand, provides a consistent view of the evolution of land variables over several decades at an enhanced resolution compared to ERA5 (Muñoz Sabater, 2019). ERA5-Land variables are available at a horizontal resolution of 0.1° x 0.1° or approximately 11.1 km. Wind speed was calculated from 10m u and v-components of wind data while relative humidity was derived from temperature and dewpoint temperature at 2m.

2.1.3 Other variables: Topographical parameters were considered using vegetation fraction, leaf area index (LAI) and distance from the coastline. The distance values from the nearest coastline, expressed in kilometers, are based on the global interpolated 0.01-degree coastline dataset released by the National Aeronautics and Space Administration (NASA, 2019). Vegetation fraction is extracted from MODIS datasets while LAI, a biophysical variable for defining vegetation properties to evaluate stress and yields, is obtained from observations of VIIRS (Claverie et al., 2023).

2.2 Methodology

Building on our previous work (Ramos and Varquez, 2026), we conducted data processing and prediction tasks using the eXtreme Gradient Boosting (XGBoost) algorithm. This study is focused on evaluating different AOD input predictors. Each prediction pipeline follows the same workflow, differentiated only by the AOD data in defining the predictor variables. Since the satellite-based AOD products vary by time of capture due to the satellites' orbital pass, these three (3) AOD-specific models are evaluated in this study:
(1) daily at 10AM JST for MODIS MAIAC AOD (AOD550),
(2) daily at 1PM JST for VIIRS AOD (AOT_Land), and
(3) daytime hours 10AM and 1PM JST for AHI Himawari-8 AOD (AOT_Merged).

2.2.1 Dataframe generation: The data pre-processing workflow in creating ML-ready inputs called “dataframes” are described in these general steps: (1) conversion of input datasets with various formats like NetCDF, GeoTIFF and CSV into ML-ready dataframes, (2) resampling of raster (.tif format) layers to a common spatial grid, (3) matching of all variables based on station locations and timestamps and (4) filtering of negative values. We adopted the finest resolution from MODIS at 1km to resample observation datasets of predictor variables, except AOD, to the common spatial grid using bilinear interpolation. Station-based matching is done using the nearest neighbor function.

2.2.2 ML-based regression modeling: The regression model was constructed using the XGBoost package accessible from an open-source machine learning library in Python called scikit-learn. XGBoost can handle missing values and mixed types of input features that are either numerical or categorical. In a related work that evaluated tree-based ML algorithms (Gianquintieri et al., 2025), XGBoost consistently achieved high accuracy in structured datasets.

The XGBoost model with predictor variables composed of a specific AOD product and other variables listed in Table 2 was trained using the generated dataframes for Kanto region and Busan sites. Since the data distributions in these test sites differ, separate models were generated. Each prediction pipeline was carried out with these general processes: (1) definition of predictor variables, (2) hyperparameter optimization, (3) training the models using cross-validation approaches with a defined objective function, and (4) evaluation of the prediction skill metrics. The objective function defined in the models is minimizing RMSE in all trials (n=20). Hyperparameters, the learning parameters of the ML model to implement the predictions, were optimized using a Bayesian algorithm developed by Akiba et al. (2019). Table 3 shows the range of the hyperparameters in the optimization process. In this study, the same settings for the hyperparameters were applied in training the AOD-specific models for a fair comparison in the evaluation of the prediction performance of each model. Fine-tuning these hyperparameters are not covered in this study and will be explored in future work.

Hyperparameter	Values
max_depth	2 to 4
learning rate	0.01 to 0.1
n_estimators	300 to 600
subsample	0.4 to 0.6
colsample_bytree	0.6 to 0.9
min_child_weight	20 to 40
gamma	0.5 to 20.0
reg_alpha	5 to 20.0
reg_lambda	5 to 20.0

Table 3. Values considered in the optimization process of the hyperparameters of the XGBoost models.

After applying K-Fold (k=5) cross-validation (CV) tests with random seed values, the out-of-fold predictions results were assessed based on the coefficient of determination (R^2), root mean square error (RMSE), mean and absolute error (MAE). These metrics describe the predictive accuracy of the AOD-specific models. To determine if the prediction errors obtained from these AOD-specific models have statistically significant differences, we employed the Kruskal-Wallis test. It is a non-parametric statistical method and a rank-based test suitable for datasets that do not follow a normal distribution (Haseeb et al., 2024). The Kruskal-Wallis statistic (H) is calculated using Equation 1

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3\,(N+1) \tag{1}$$

where N is the total number of rows across all groups, k is the number of groups, R_i is the sum of ranks for group *i* and n_i is the number of rows in group *i* (Ostertagová et al., 2014). To quantify the effect of the sample density (i.e., total number of valid AOD values) on the predictions, the effect size (η^2) was calculated using Equation 2

$$\eta^2 = \frac{H - k + 1}{n - k} \tag{2}$$

which uses the same parameters described in Equation 1.

3. Results and Discussion

3.1 Spatiotemporal distribution of AOD products

To fairly evaluate the influence of the AOD products on the predicted values of $PM_{2.5}$ concentrations, any feature modifications like grid resampling or gap-filling techniques on the AOD values were not applied. In both regions, MODIS MAIAC AOD has complete coverage while VIIRS AOD and AHI Himawari-8 AOD have very poor coverage in 2021 (see Table 4). The distribution of AOD products vary because of spatial gaps attributed to the presence of clouds in the regions. These gaps are quantified in the AOD data density maps shown in Figure 3 (for VIIRS) and Figures 4 to 5 (for AHI datasets). Refer to Appendix A and B for the full-year distribution of the AOD values extracted on $PM_{2.5}$ station locations in Kanto region and Busan, respectively.

AOD source	Valid data (%)	Missing data (%)
Kanto	***Station n = 182***	
MODIS	100	0
VIIRS	4.15	95.85
AHI	2.27	97.73
Busan	***Station n = 72***	
MODIS	100	0
VIIRS	8	92
AHI	0.48	99.52

Table 4. Sample density of AOD products in test sites for 2021. Valid data include non-negative AOD values.

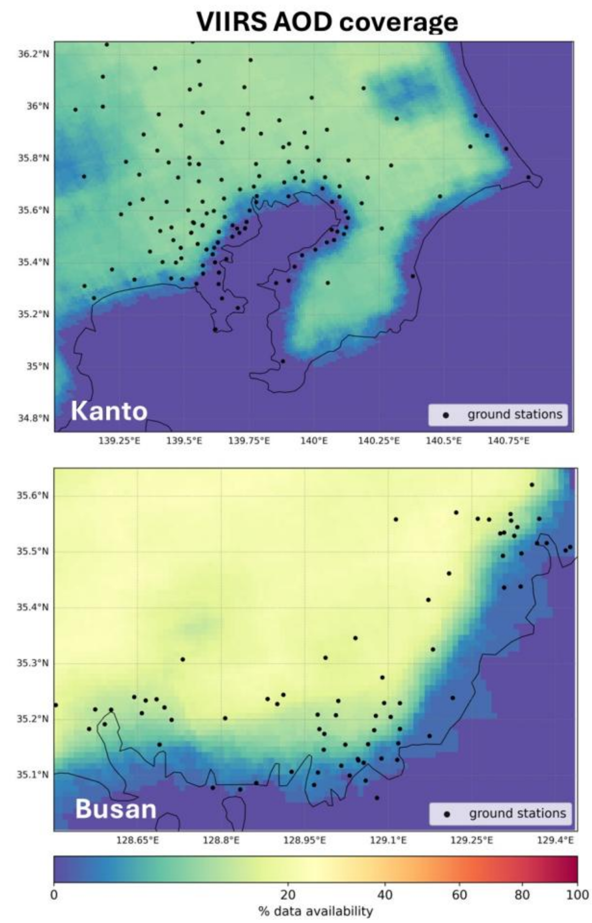


Figure 3. Availability (in %) of VIIRS AOD data in 2021 with observation time at 1PM JST. Black dots and lines refer to *in situ* ground stations and coastlines, respectively.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIX-B2-2026-1439-2026

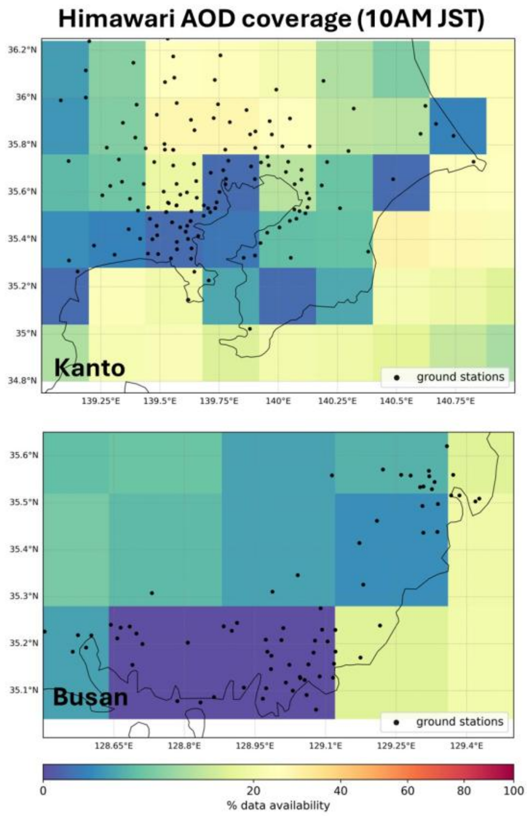


Figure 4. Availability (in %) of AHI Himawari-8 AOD data in 2021 with observation time at 10AM JST.

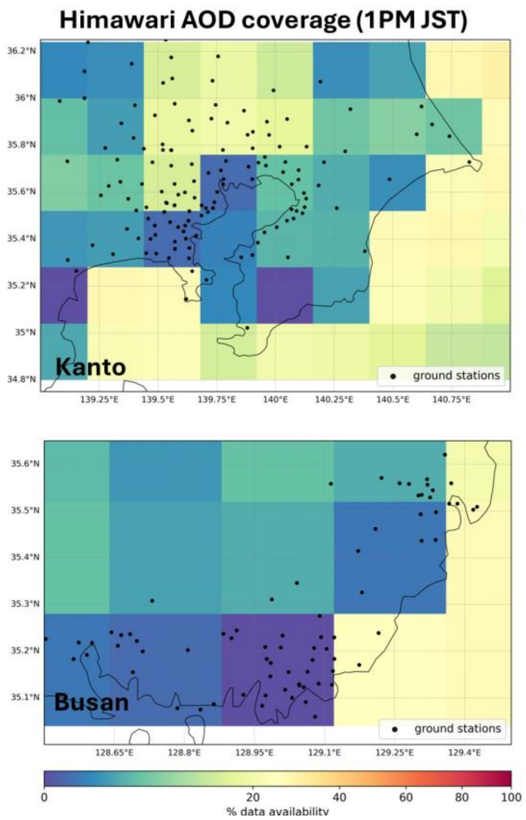


Figure 5. Availability (in %) of AHI Himawari-8 AOD data in 2021 with observation time at 1PM JST.

3.2 Effect of AOD products on $PM_{2.5}$ predictions

The $PM_{2.5}$ prediction results of the AOD-specific models were evaluated on the validation sets or the “out-of-fold” predictions (refer to Table 5). To quantify how the model performs on unseen data compared to training datasets, the differences in RMSE values were obtained and provided in Table 5 as %gap. Low %gap indicates high generalizability of the prediction model. These results suggest that the model with MODIS AOD (with RMSE gap equivalent to 10.49% and 15.08% for Kanto and Busan, respectively) generalizes better than the models with VIIRS and AHI AOD on both regions. However, the predictive accuracy varies across sites and AOD products. In Kanto region, VIIRS AOD achieved the highest accuracy (R^2=0.58, MAE=2.33 μg/m^3, RMSE=3.14 μg/m^3) but has the largest RMSE gap (30.52%) which indicates poor generalization. In Busan, MODIS AOD performed best in terms of predictive accuracy (R^2=0.58, MAE=5.79 μg/m^3, RMSE=7.77 μg/m^3) and model generalization (RMSE gap=15.08%).

AOD source	N	R^2	MAE	RMSE	%gap
Kanto					
MODIS	19906	0.53	3.24	4.47	10.49
VIIRS	2143	0.58	2.33	3.14	30.52
AHI	7730	0.58	3.11	4.27	18.46
Busan					
MODIS	8529	0.58	5.79	7.77	15.08
VIIRS	2093	0.57	5.97	8.08	29.95
AHI	651	0.47	6.88	9.54	30.84

Table 5. Average CV metrics of prediction runs per AOD-specific model in each region where N = total variable samples.

The Kruskal-Wallis test revealed statistically significant differences in the prediction errors obtained by the AOD-specific models in Kanto ($H = 11.567$, $p < 0.05$) as shown in Table 6. For Busan, the differences in the prediction errors are not statistically significant ($p > 0.05$). To determine which AOD products perform differently in predicting $PM_{2.5}$ concentrations in the Kanto region, post hoc pairwise comparisons were conducted using Dunn’s test with Bonferroni correction. Dunn’s test results indicate that AOD values from MODIS and VIIRS have significant difference (p-value=0.0059).

Region	H	p-value	η^2
Kanto	11.567	0.003	0.00032
Busan	2.984	0.225	0.00087

Table 6. Kruskall-Wallis statistical results and effect size values in both regions.

The effect size η^2 indicates how much of the variation in the $PM_{2.5}$ prediction metrics is explained by the models utilizing different AOD products. The results show extremely small values of η^2 (0.00032 and 0.00087 for Kanto and Busan, respectively) suggesting that the effect of choosing one AOD product over another in predicting $PM_{2.5}$ is negligible. With the current configuration and performance of the XGBoost model, given the set of predictor variables listed in Table 2, the AOD values from the different satellite sensors have minimal impact as quantified by the effect size. In future work, combining these three AOD values in predicting inland daytime $PM_{2.5}$ concentrations may be necessary and will be considered in improving the prediction workflow.

4. Summary and Conclusion

This work evaluated the predictive performance of an XGBoost regression model utilizing AOD products from three satellite sensors (MODIS, VIIRS and AHI Himawari-8) across two distinct geographic sites: Kanto region in Japan and Busan in South Korea. The key findings are as follows:

- MODIS MAIAC AOD achieved the lowest %RMSE gap (10.49%-15.08%) across both study sites. This demonstrates that models utilizing MODIS AOD are less prone to overfitting and exhibit higher generalizability when applied to unseen datasets compared to those using VIIRS and AHI AOD.

- While the Kruskal-Wallis test identified statistically significant differences in prediction errors within the Kanto region ($p < 0.05$), the corresponding effect sizes were notably low ($\eta^2 < 0.001$). This indicates that the observed statistical differences are negligible.

In conclusion, while MODIS offers high sample density of AOD values and high generalizability of the XGBoost model, the negligible effect sizes suggest that no single AOD product provides a definite advantage over the others. Hence, rather than selecting a single AOD source, future studies will explore developing multi-sensor fusion techniques of AOD datasets. Additionally, we will implement advanced feature engineering and incorporate additional predictor variables to reduce overfitting and enhance the estimation of ground $PM_{2.5}$ concentrations.

Acknowledgements

The authors are grateful for the Doctoral program scholarship funding provided by the Department of Science and Technology -Engineering Research and Development for Technology (DOST-ERDT) of the Philippines and the financial support from the Environment Research and Technology Development Fund (JPMEERF2024RA1) of the Environmental Restoration and Conservation Agency provided by the Ministry of the Environment of Japan. We also thank the agencies NIES of Japan and the Korean Ministry of Environment for sharing the $PM_{2.5}$ data records of their ground monitoring stations. We also acknowledge the technical support from Dr. 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.

Claverie, M., Vermote, E., Justice, C., Csiszar, I., Myneni, R., Baret, F., Masuoka, E., Wolfe, R., Ray, J., 2023. NOAA Climate Data Record (CDR) of VIIRS Leaf Area Index (LAI) and Fraction of Absorbed Photosynthetically Active Radiation (FAPAR), Version 1. NOAA National Centers for Environmental Information. doi.org/10.25921/9x3m-0e02.

Gianquintieri, L., Oxoli, D., Caiani, E. G., Brovelli, M. A., 2025. State-of-art in modelling particulate matter (PM) concentration: a scoping review of aims and methods. *Environment, Development and Sustainability* vol. 27 25889–25911. doi.org/10.1007/s10668-024-04781-5.

Handschuh, J., Erbertseder, T., Schaap, M., Baier, F., 2022. Estimating PM2.5 surface concentrations from AOD: A combination of SLSTR and MODIS. *Remote Sensing Applications: Society and Environment*, Volume 26, 100716, ISSN 2352-9385. doi.org/10.1016/j.rsase.2022.100716.

Haseeb, M., Tahir, Z., Mahmood, S. A., Arif, H., Almutairi, K. F., Soufan, W., Tariq, A., 2024. Comparative analysis of machine learning models for predicting $PM_{2.5}$ concentrations using meteorological and chemical indicators. *Journal of Atmospheric and Solar-Terrestrial Physics*, Volume 263, 106338. doi.org/10.1016/j.jastp.2024.106338.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J-N., 2023. ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). doi.org/10.24381/cds.adbb2d47.

Jung, J., Souri, A. H., Wong, D. C., Lee, S., Jeon, W., Kim, J., Choi, Y., 2019. The impact of the direct effect of aerosols on meteorology and air quality using aerosol optical depth assimilation during the KORUS-AQ campaign. *Journal of Geophysical Research: Atmospheres*, 124, 8303–8319. doi.org/10.1029/2019JD030641.

Kang, D.-D., Goo, T.-Y., Lee, D.-I., 2023. Classification of High-Concentration Aerosol Phenomena Using Their Physical Properties in Busan, South Korea. *Applied Sciences*, *13*(1), 355. doi.org/10.3390/app13010355.

Kikuchi, M., Murakami, H., Suzuki, K., Nagao, T., Higurashi, A., 2018. Improved Hourly Estimates of Aerosol Optical Thickness using Spatiotemporal Variability Derived from Himawari-8 Geostationary Satellite, *IEEE Trans. Geosci. Remote Sensing*. doi: 10.1109/TGRS.2018.2800060.

Kumar, R. P., Prakash, A., Singh, R., Kumar, P., 2024. Machine learning-based prediction of hazards fine PM2.5 concentrations: a case study of Delhi, India. *Discover Geoscience 2*, 34. doi.org/10.1007/s44288-024-00043-z.

Lee, J., Hsu, N. C., Kim, W. V., Sayer, A. M., Tsay, S. C., 2024. VIIRS Version 2 Deep Blue Aerosol Products. *J. Geophys. Res.-Atmos., 129*, e2023JD040082, doi.org/10.1029/2023JD040082.

Lyapustin, A., Wang, Y., 2018. MCD19A2 MODIS/Terra+Aqua Land Aerosol Optical Depth Daily L2G Global 1km SIN Grid V006 [Data set]. NASA Land Processes Distributed Active Archive Center. doi.org/10.5067/MODIS/MCD19A2.006.

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.

NASA Ocean Biology Processing Group, 2019. Distance to the Nearest Coast [Dataset]. Retrieved online from https://oceancolor.gsfc.nasa.gov/resources/docs/distfromcoast/.

Ostertagová E., Ostertag, O., Kováč, J., 2014. Methodology and application of the Kruskal-Wallis test. In Applied Mechanics and Materials. Trans Tech Publications Ltd 611:115-120. doi.org/10.4028/www.scientific.net/AMM.611.115.

Padimala, S. S. K., Matli, C. S., 2024. A comprehensive review delineates advancements in retrieving particulate matter utilising satellite aerosol optical depth: Parameter consideration, data processing, models development and future perspectives. *Atmospheric Research*, Volume 308, pp. 107514. doi.org/10.1016/j.atmosres.2024.107514.

Ramos, R. V. and Varquez, A. C. G., 2026. Improving prediction of $PM_{2.5}$ in Metro Manila using XGBoost with Optuna hyperparameter optimization, *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.,* X-5/W4-2025, 369–376. doi.org/10.5194/isprs-annals-X-5-W4-2025-369-2026.

Tan, Y., Wang, Q., Zhang, Z., 2022. Assessing spatiotemporal variations of AOD in Japan based on Himawari-8 L3 V31 aerosol products: Validations and applications. *Atmospheric Pollution Research,* Volume 13, Issue 6, 101439, ISSN 1309-1042. doi.org/10.1016/j.apr.2022.101439.

Tang, D., Zhan, Y., Yang, F, 2024. A review of machine learning for modeling air quality: Overlooked but important issues. *Atmospheric Research* vol. 300. doi.org/10.1016/j.atmosres.2024.107261.

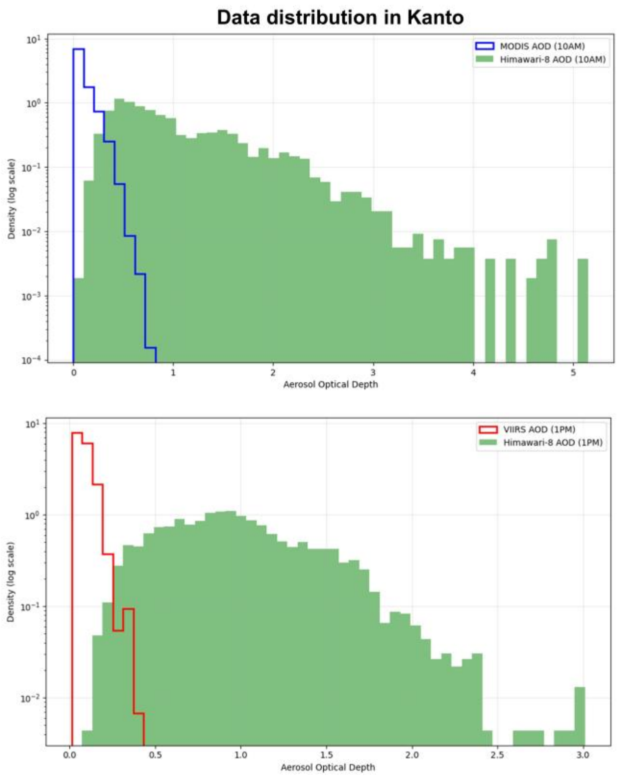
Vignesh, P. P., Jiang, J. H., Kishore, P., 2023. Predicting PM2.5 Concentrations Across USA Using Machine Learning. *Earth and Space Science*, 10(10). doi.org/10.1029/2023EA002911.

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.

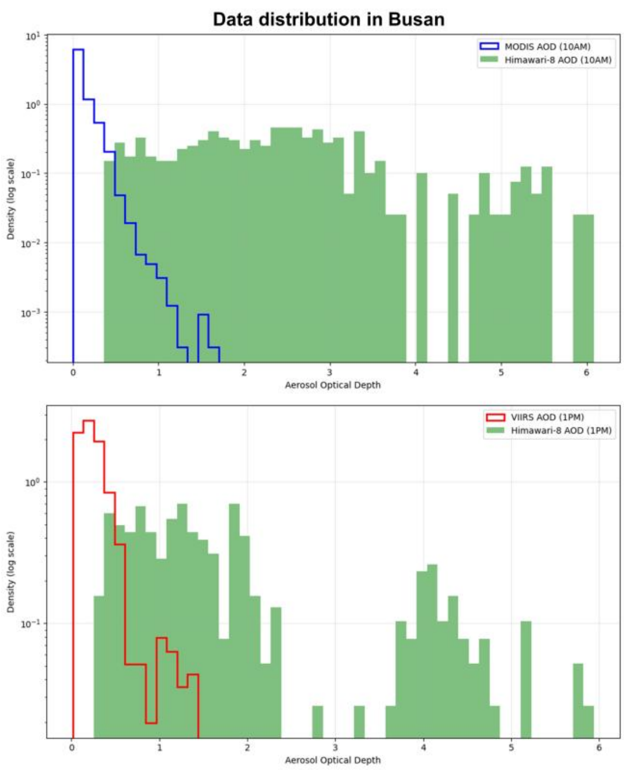
Wang, Y., Xu, W., Zhong, H., Chen, B., Yang, L., Wang, L., Ren, J., Duan, J., Lin, C., Huang, R., 2025. Advancing Aerosol Chemistry with Machine Learning: A Short Review. *ACS ES&T Air 2*, 2323–2341. doi.org/10.1021/acsestair.5c00001.

Xiao, Q., Zhang, H., Choi, M., Li, S., Kondragunta, S., Kim, J., Holben, B., Levy, R. C., Liu, Y., 2016. Evaluation of VIIRS, GOCI, and MODIS Collection 6 AOD retrievals against ground sunphotometer observations over East Asia. *Atmospheric Chemistry and Physics*, Volume 16 (3), pp.1255-1269. doi.org/10.5194/acp-16-1255-2016.

Appendix



Appendix A. Distribution of satellite-based AOD values extracted in the locations of $PM_{2.5}$ ground stations in Kanto region.



Appendix B. Distribution of satellite-based AOD values extracted in the locations of $PM_{2.5}$ ground stations in Busan.