Reply to Reviewer #2 for Manuscript of “Estimating hourly ground-level aerosols using GEMS aerosol optical depth: A machine learning approach” by O et al.

The manuscript is based on the estimation of PM2.5 and PM10 from GEMS AOD. The main objective is to evaluate the effectiveness of GEMS AOD in estimating ground level PM concentrations. This study attempts to study how GEMS AOD can provide air quality estimates in a global scale, which is of great importance

⇨ *We thank the reviewer for recognizing the value of our study and for their thoughtful comments and efforts in reviewing our manuscript. Below, we provide point-by-point responses to the comments.*

⇨ *Here, we summarize the main revisions for the manuscript:*

- ***Data Update:*** *The dataset has been extended to December 2023, covering two years.*
- ***Additional Machine Learning Algorithm:*** *To enhance the robustness of our machine learning-based modeling, we have employed an additional algorithm, XGBoost (XGB).*
- ***Expanded Input Features:*** *We have considered additional chemical gas features, such as SO_2, NO_2, and O_3, in the modeling process.*
- ***The manuscript structure*** *is reorganized. Specifically:*
  - *Data and Methodology sections are separated.*
  - *Results with PM2.5 are reported in the main text.*

However, there are few concerns regarding the formulation of the study and the structure of the manuscript. Given below are my suggestions.

1. The overall paper lacks adequate explanations and citations to corroborate the objective of the study and how it differs from existing studies/novelty. (Ex: are there any ML based studies for estimation of PM concentrations? What are the advantages of this method over the existing?)

⇨ *We appreciate the reviewer’s comment and have clarified the study’s objectives and novelty. Our research is one of the first to validate GEMS AOD data, focusing on its utility for PM estimation rather than direct comparisons with ground-based AOD. Please note that we do not aim to develop best models for specific regions; instead, we evaluate GEMS AOD’s broader applicability, such as improving models with additional data (e.g., chemical data) or estimating PM in areas without ground observations (e.g., using neighboring stations). In this context, we use ML models that are flexible enough to integrate diverse input data and can be relatively easily implemented in other regions, unlike physics-based models that often require additional parameter calibrations.*

⇨ *These points have been added to the introduction and conclusions for better clarity. e.g. Here, we first evaluate GEMS AOD data through a direct comparison with ground-based AERONET observations over South Korea. However, we place greater emphasis on evaluating the utility of GEMS AOD for estimating ground-level PM concentrations, as it* *offers a unique opportunity to address aerosol data gaps in Asia. Moreover, South Korea has nationwide air quality monitoring stations, allowing us to obtain continuous and large data samples (PM10 and PM2.5) for validating the satellite data. To better utilize GEMS AOD for ground-level PM estimation, we employed machine learning models, which offer the advantage of experimenting with a wide range of input variables. For example, ground-level aerosols are influenced not only by meteorological conditions but also by precursor pollutants such as SO_2 and NO_2. Machine learning allows for the efficient integration and processing of these diverse datasets, enhancing the ability to utilize AOD for aerosol estimation.*

[...] We aim to employ practical ML models to assess the potential of GEMS AOD data for sub-daily PM estimation, without prioritizing to achieve the best possible model performance. More advanced ML models or different modelling approaches (e.g. chemical transport models) could improve performance, but this is beyond the scope of our study. Furthermore, to our knowledge, this is the first evaluation of GEMS AOD applications, providing baseline results for future model comparisons and development. This approach also applies to input selection. While we consider a relatively wide range of input variables, including both meteorological and chemical data, additional variables can be tested, and model performance can be compared. In this context, ML models are particularly advantageous, as they can incorporate a broad spectrum of variables, including those typically not used in process-based models. However, ML performance is highly dependent on the quality of the training data; therefore, careful attention to data quality is essential.

2. The introduction of the manuscript should include a brief description on the sections of the manuscript. Results and discussion should be a separate section from data and methodology. I suggest separating data and methodology as separate sections, as this manuscript lacks proper description on the methodology (there is too little information on the machine learning method (RF), selection criteria for input variables, ranges of the input variables.

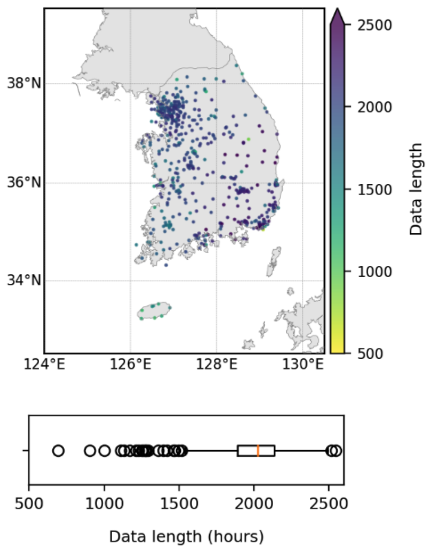
⇨ *Thank you for the suggestion. First, we have added the description on the sections in the end of the introduction: In the following sections, we first describe the data and its preprocessing in Section 2. Section 3 details the methodology, including the machine learning models employed for PM estimation. In Section 4, we present the results and discuss their implications, followed by conclusions and future research directions in Section 5.*

⇨ *Second, we have separated the data and methods, and we have added more description on the model in Methods: RF operates by constructing multiple decision trees during training and aggregating their predictions to enhance accuracy and avoid overfitting. It is widely recognized for its ability to efficiently handle non-linear relationships in data and is often used for estimating PM concentrations. We also use XGBoost, which is similarly based on decision trees. However, XGBoost builds trees sequentially, allowing each tree to learn from the errors of the previous one, and is generally considered to outperform RF.*

⇨ *Lastly, we have also added more description on the input variables: Those input variables are selected with reference to previous studies (Yang et al., 2020; Handschuh et al. 2022), including Seo et al. (2015), which conducted experiments in South Korea. [...] We obtain input data from reanalysis datasets, which are readily available across all* *areas within the GEMS satellite observation coverage. This ensures that the experiment conducted in this study can be easily extended to other locations, including other Asian countries, particularly in areas where meteorological measurements are unavailable. Additionally, reanalysis datasets provide consistent and reliable data updates over space and time. Nonetheless, it is well known that gases such as CO, NO_2 , and SO_2 can influence PM formation mechanisms either directly or indirectly. Therefore, we also incorporate chemical data measured at the AirKorea stations as additional input features. In this way, we can evaluate the potential improvements in PM estimation using AOD when supplemented with additional information, and we report the corresponding results.*

3. What is the sample size of the data used in RF?

⇨ *Since the models are applied to individual station points, the size of the training data varies for each model (station). To illustrate this, we have added the following figure as a Supplementary Material, showing the distribution of training data sizes across all stations.*

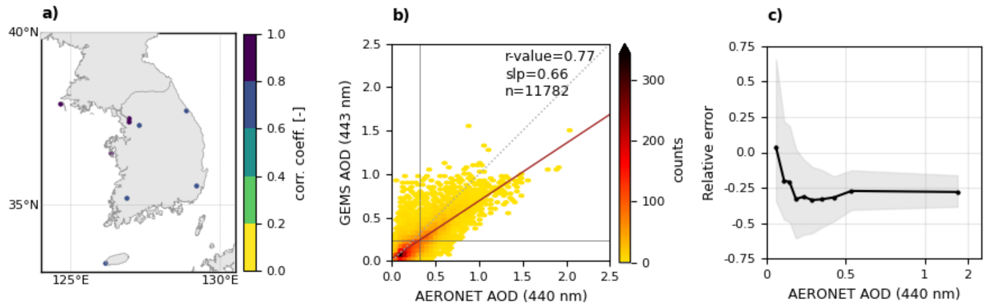


4. RF was selected to estimate PM concentrations out of some other ML methods. How do you evaluate the model effectiveness in this work? Model performance can also affect the conclusions you draw regarding the ability of GEMS AOD to accurately provide PM concentrations.

⇨ *The model effectiveness is evaluated through the k-fold cross validation, as explained in Methods; "The main analysis is based on model predictions obtained through five-fold cross-validation." Model correlations, linear regression slopes, and relative errors (which is newly added) are all based on the five-fold cross-validation. Please note that additionally we have employed another machine learning method, XGBoost, to compare the performance of different models and further validate our approach.*

5. The first part of the results should be to validate the GEMS AOD retrievals

⇨ *We have reorganized the section to present the comparison between GEMS and AERONET AOD as the first subsection, with the following figure now labeled as Fig. 1.*



⇨ *The added text is as follows: First, we directly compare the GEMS AOD data with ground-based AOD measurements from the AERONET. As shown in Fig.1a, the temporal variations of AODs at each station exhibit overall good correlations (Pearson's r), ranging from 0.68 to 0.89. When the entire time series from all AERONET sites are compared, the correlation remains strong (r = 0.77), although GEMS tends to underestimate AOD compared to the ground-based AOD measurements, as indicated by the linear regression slope (0.66) in Fig.1b. Furthermore, Fig. 1c demonstrates that this underestimation is consistent across most AERONET AOD ranges, with overestimation can also occur at very low AOD values. A study on the early version of GEMS L2 algorithm prior to the launch of GEMS also reported high correlation but slight underestimation of GEMS AOD relative to AERONET (Kim et al., 2020). Recent studies using GEMS L2 data in Asia regions have reported similar findings (e.g. Cho et al. 2024; Jang et al. 2024).*

6. The labeling of PM measurements used in RF, and the PM estimations, is vague. Make it more distinct.

⇨ *We have reviewed and revised the labeling throughout the paper to ensure a clear distinction between PM measurements and PM estimations.*

7. The use of mean vs error plots would be a better way of understanding the model performance rather than comparing the correlation coefficients. (Refer, Bland-Altman analysis)

⇨ *Thank you for the suggestion. We have conducted additional analysis on the model error structure using relative errors, defined as the difference between estimated and measured PM divided by the measured values, at each 10 percentile of the measured PM. The result is now reported in Fig. 3, as shown below.*

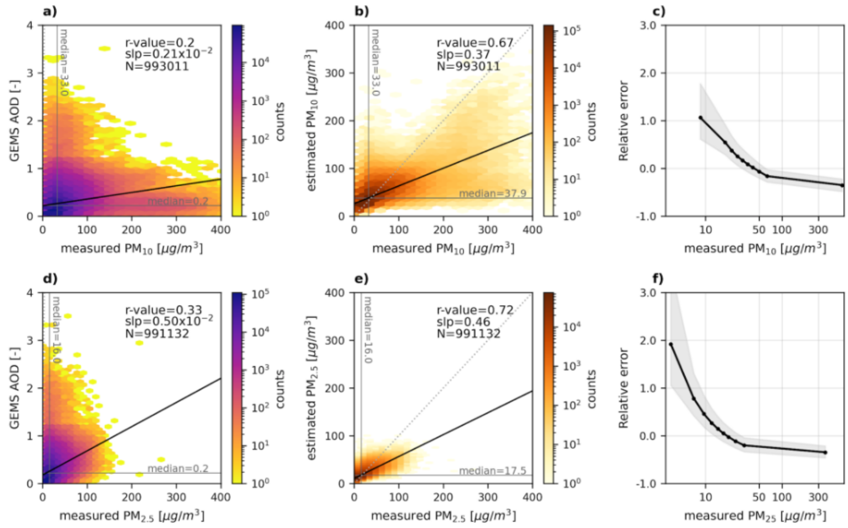


Fig. 3 Performance of RF models in estimating ground-level PM concentrations. *(a) Density scatter plot between measured PM10 and GEMS AOD across all stations. (b) Density scatter plot between measured PM10 and model-estimated PM10. The vertical and horizontal lines represent the corresponding median values. The thick solid line is the regression line, and the dotted diagonal line is the one-to-one. (c) Relative errors, defined as the difference between estimated and measured PM divided by the measured values, are shown at each 10th percentile of measured PM10. Panels (d), (e), and (f) correspond to (a), (b), and (c), respectively, but for PM25.*

8. Add more details description on SHAP analysis.

⇨ *We have added the following sentences in Methods. However, the primary disadvantage of machine learning is its 'black-box' nature, meaning we cannot fully understand why it produces certain estimations. To address this limitation and examine the role of the input features, we further use SHapley Additive exPlanations (SHAP) and quantify the relative importance of the considered input features on the model's predictions. SHAP is an explainable machine learning method based on Shapley values, which measure the marginal contribution of each predictor to the model's output or prediction across all the possible predictor combinations.*

⇨ *We have added more detailed descriptions of the SHAP analysis results, as follows: Furthermore, we use SHAP to examine the relative importance of the considered input features on the model's estimations (see Methods). As SHAP is computed for individual observations, we take the mean of absolute SHAP values for each input variable across all the estimations to explain its global feature contributions (Fig. 4). For both PM10 and PM25 estimations, AOD has a greater influence on model performance than most meteorological variables, as expected from its relatively strong correlation with ground aerosols (Fig. 2a), demonstrating the effectiveness of satellite information for ground aerosol simulations. [...] Among other meteorological variables, relative humidity (RH) emerges as one of the most important predictors for PM10 estimations. RH influences wet deposition of PM and also characterises seasonal variations in PM concentrations. For instance, in Korea, PM10 tends to increase due to wildfire emissions during dry seasons or from relatively coarse particles transported by Asian dust in the spring season. On the other hand, BLH has a greater importance for PM25. As aerosols are* *primarily confined to the planetary boundary layer, BLH is a good proxy for estimating the height of the aerosol layer (Lee et al., 2024) and can help relate columnar satellite data to surface aerosol values (Handschuh et al., 2022; Gupta and Christopher, 2009a). The stronger importance of BLH for PM25 suggests its effectiveness in capturing the vertical distribution of finer aerosols. Previous studies have suggested a strong relationship between BLH and PM25 due to well-recognized positive feedbacks (e.g. Li et al., 2017; Su et al., 2022), which further underscores the importance of BLH. The significant influence of both BLH and RH on the relationship between AOD and PM has also been reported by Zheng et al., 2017.*

9. L 59-61 Include more details about GEMS instrument (uncertainties, wavelength channels). Do you perform any pixel averaging?

⇨ *We have added the following explanation about the GEMS instrument: The GEMS measures radiance in the 300–500 nm range with a spectral resolution of 0.6 nm and retrieves aerosol properties. The GEMS aerosol retrieval algorithm (AERAOD) uses the optimal estimation (OE) method, which integrates satellite-observed radiances with initial estimates of aerosol properties, including AOD, derived using the two-channel inversion approach employed by the OMAERUV algorithm (Torres et al., 2007).*

⇨ *No, we do not perform any pixel averaging. GEMS Level 2 data is provided on a pixel basis, and we select the pixel closest to the ground station at each time step, resulting in average distances ranging from 1.5 to 2.7 km. While gridding or interpolation could be applied to align the spatial resolution of GEMS with ground station data, such preprocessing may introduce additional uncertainties. This is especially relevant given that GEMS data contain many spatial gaps due to cloud contamination and other factors.*

⇨ *We discuss the potential uncertainty from the spatial mismatch of the data in the main text as follows: While GEMS data could be collocated to the exact location of ground stations through interpolation, additional preprocessing may introduce further uncertainties. Furthermore, we confirm that the spatial distances between GEMS pixels and ground stations do not significantly affect PM estimation performance (see Fig. S2).*

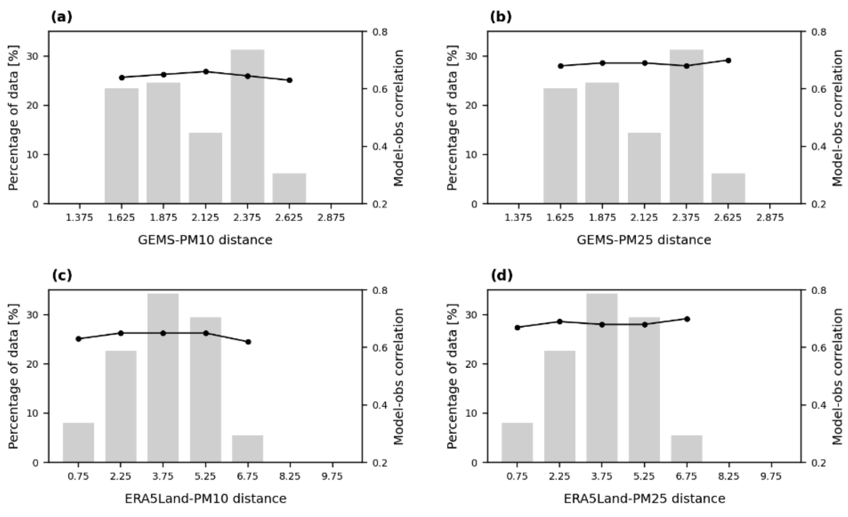


Figure S2. *PM estimation performance by distance between data sources and measurement stations. (a) and (b) show the distribution of distances between GEMS AOD observation pixels and PM measurement stations, along with* *average model performance (correlations) for PM10 and PM2.5, respectively. (c) and (d) depict the same, but for the distances between meteorological input data (i.e. ERA5-Land) and PM measurement stations.*

10. You need to add a description on GEMS AOD retrieval algorithm and explain possible uncertainties in AOD retrievals. The citation is not enough.

⇨ *Thank you for pointing this out. We have added a more explanation of the GEMS AOD retrieval algorithm (please see our response to Comment #9). Additionally, we have expanded on the results of the comparison between GEMS and AERONET AOD, as addressed in our reply to Comment #5.*

⇨ *However, we would like to emphasize that the primary focus of our study is on the usefulness of GEMS AOD in estimating PM concentrations rather than a detailed analysis of the AOD retrieval process itself (see our response to Comment #1).*

⇨ *Nonetheless, to address the Reviewer's concern, we have included a discussion of potential sources of bias in GEMS AOD retrievals, drawing from recent studies; Cho et al. (2024) specifically compared GEMS and AERONET AOD measurements in Asia and pointed out that the absence of region-specific aerosol type information in the GEMS aerosol model, as well as inaccuracies in cloud-masking processes, may negatively impact the accuracy of GEMS AOD data.*

11. L 74 – Do you perform any geolocation of data? How do you collocate reanalysis data? What is the maximum possible difference in the colocation of ground-based PM concentrations, AOD and reanalysis data?

⇨ *We initially selected the closest AOD and reanalysis data to the ground-based PM measurements. However, we have updated the method as follows: Both gridded datasets are interpolated to the locations of AirKorea stations using inverse distance weighting (IDW) based on the four closest grid points. If data are missing in the nearest grid points (e.g., over ocean areas), the corresponding locations are excluded from the analysis*.

⇨ *Please note that the PM estimation results remain consistent after changing the geolocation method. For details on the distance differences between the datasets, please refer to our response and the corresponding figure in Comment #9.*

12. L 115 – Is this something evident across all AOD values? Are there any differences seen for PM estimations under lower AOD and higher AOD values. Is there any detection limit?

⇨ *Thank you for your question. As detailed in the figure provided in our response to Comment #7, we observe a relatively strong overestimation at low AOD values and a weaker underestimation at high AOD values. We did not explicitly set any detection limits. If such limits exist, they are assumed to be inherently learned by the machine learning model during the training process.*

13. L 154 – AERONET data should be introduced under the data section. How do you collocate AERONET data? What do you mean by closest? You should specify the

distance limit. How do you average temporal data. Does the difference between AERONET AOD and GEMS AOD lies with AEROENT AOD uncertainty?

⇨ *We have introduced the AERONET data in the data section and also explain how we prepared the data, as follows: Finally, for direct comparison, we obtain ground-based AOD measurements from AERONET sites in South Korea. A total of nine stations are selected, where data are available during the study period. AERONET provides highly accurate AOD measurements using Cimel Electronique Sun–sky radiometers, with an uncertainty of approximately 0.01-0.02. For this study, we use the version 3, level 2.0 quality-assured AOD at 440nm. For the comparison, GEMS AOD data within a 5 km radius of the AERONET sites are considered, and sub-hourly AERONET data are averaged within a temporal window of ±20 minutes around the GEMS observation time.*

⇨ *Please note that, as shown in the figure provided in our response to Comment #5, the discrepancy exceeds AERONET AOD uncertainty. As explained in our response to Comment #10, this is likely due to additional factors, such as characteristics of the GEMS algorithm.*

14. L 63 – What is ARA?

⇨ *Aerosol Retrieval Algorithm. The text is revised.*

15. L 69 – What is ERA5?

⇨ *ECMWF Reanalysis v5 (ERA5) data. The text is revised.*

16. L 75 – How did you perform the AOD-PM simulations? Or do you mean estimations?

⇨ *We meant the PM estimation from the model simulations. The text is revised.*

17. L 83 – 89 This paragraph should go under the data section

⇨ *Thank you for the suggestion. We have moved the part of the paragraph to the data section.*

18. L 157 – Has it been observed for low AOD or high AOD?

⇨ *The underestimation is strong in the high AOD, but observed across all ranges. Please refer to the figure in our response to Comment #5,*

19. L 164 -165 Does the GEMS AOD algorithm consider any non-sphericity dust? You should add a description about the AOD algorithm

⇨ *The current GEMS aerosol retrieval algorithm primarily assumes spherical particles in its calculations. We have added a more detailed description of the algorithm and included references to a relevant recent study (see our response to Comment #6).*

20. L 176 – Fig 6. What does n=1,2,… stand for?

⇨ *Thank you for pointing this out. It refers to the number of neighboring stations providing the training dataset. We have updated the figure caption to clarify this. Additionally, please note that we have slightly modified the experimental setup to demonstrate the potential of satellite-derived AOD and machine learning for estimating PM concentrations at ungauged locations.*

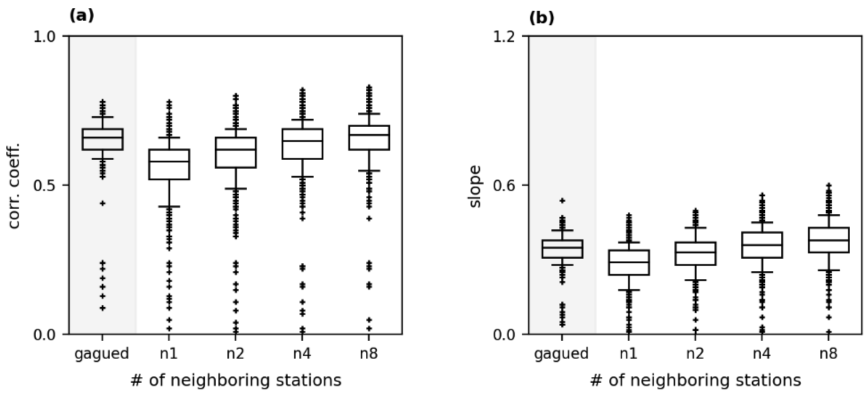


Figure 7. Potential of satellite data in PM estimation at ungauged locations. *(a) Correlation and (b) slope of the linear regression between the measured and estimated PM10 concentrations at each station. Data from the n closest neighboring sites are used to train the RF models, and the model performance is evaluated at the target station, where the training data is deliberately excluded for this experiment. The term 'gauged' indicates that the model is trained and tested at the same station (as shown in the main analysis in Fig. 2).*