Response to Comments on the Manuscript:

"An optimized semi-empirical physical approach for satellite-based PM$_{2.5}$ retrieval: embedding machine learning to simulate complex physical parameters"

Response to Comments of the Editor:

We would like to thank the editor for his precious time. A response to the comments and an introduction to the adjustments in the manuscript follows.

Response: We have carefully read the referee's comments and feel particularly grateful to him/her for wise suggestions. According to the comments, we made further adjustments to our manuscript, especially in the discussion part. And all changes have been highlighted in yellow color in the manuscript. The major revisions include:

1) **Describe where different data categories (site and surface data) are applied in the experimental steps.** The introduction in Section 5.4.3 can clarify the reviewer's understanding of the different types of AOD and FMF data used in this paper.

2) **Add the detailed process of RF model construction.** The additions to sections 5.4.1 and 4.1 can explain the applicability of our model in North China and experimental consistency in 2017.

3) **Analyze the overall performance of the RF-PMRS method.** Section 5.4 is added to this document, which analyzes:

A. the universality of RF-PMRS to answer why it applies to North China (Section 5.4.1);

B. limitations of the validation experiments and reasons (Section 5.4.2);

C. variable uncertainty and it is carried out from five aspects (Section 5.4.3).

Other minor problems have also been responded and revised, including **a clear definition of the relationship between regression and machine learning**, and **the difference between AERONET sites and ground PM$_{2.5}$ stations**.

Response to Comments of Reviewer #2:

We would like to take this opportunity to gratefully thank the reviewer for his/her constructive suggestions for improving the paper. An item-by-item response to the comments raised by the reviewer follows.

Comments:

1. Page 2, line 62-65: "Machine learning….. between multiple variables (Irrgang et al., 2021). But the regression is ….. ground stations (Gupta and Christopher, 2009; Li et al., 2017)." There is no apparent connection between the two sentences. The first part of the sentence is about machine learning, but the second sentence jumps to the regression. Please consider rewriting the sentence.

Response: Thank the reviewer for pointing out this problem and we feel sorry for the unclear description of the relationship between regression and machine learning (ML). This paragraph focuses on univariate/multivariate regression methods used to retrieve $PM_{2.5}$ concentrations (Page 2 Lines 60-66). **Some algorithms of ML are useful tools for regression, which can be used to establish high-precision regression models.** This is also the connection between ML and such regression methods.

For clarity, we have changed the statement of the article (Page 2 Lines 61-64) to "This kind of data-driven method establishes a statistical model between AOD, auxiliary variables, and ground PM2.5 observations. Machine learning is a common tool for such regression methods due to its powerful nonlinear fitting ability between multiple variables (Irrgang et al., 2021). But the regression algorithms in machine learning are affected by the distribution and density of ground stations." We hope your doubts are now resolved.

2. Page 7, Fig. 1: Please mark the BJ and BC sites with triangles or stars.

Response: Since BJ and BC sites do not belong to the site category shown in Fig. 1, they cannot be labeled in the figure. We apologize for your misunderstanding of the types of BJ and BC sites due to our description. Fig. 1 shows the location of $PM_{2.5}$

ground monitoring stations in the NC region, mainly used to validate the accuracy of $PM_{2.5}$ estimation results. However, the BJ and BC sites are AERONET sites. Therefore, AERONET sites including BC and BJ cannot be labeled in Fig. 1.

In fact, step 1 of Fig. 3 marks the nine AERONET site locations used in the experiment (Page 11 Line 261). According to your previous suggestion, we have highlighted the locations of BJ and BC sites in this figure with two yellow quadrangles in the zoom-in view (Subgraph 1 "VE_f calculation" of Fig. 3) and explained accordingly (Page 11 Lines 263-265). Meanwhile, to prevent misunderstandings about the type of ground sites, we have modified the figure title and annotations of Fig. 1 (Page 7 Lines 170-172).

3. Page 10, line 252: As the authors mentioned in comment #2, the AERONET AOD is used in step 1, while MODIS AOD is used in step 4. How does the AOD from two sources affect the PM2.5 estimation, considering the uncertainties of the two sources?

Response: We gratefully thank the reviewer for this comment. In experimental comparisons, uncertainty is rarely used as a quantitative indicator, which we understand as the deviation or error between the estimated results and the true values. **In order to obtain an accurate RF model, high-precision point-scale AERONET AOD data is used for modeling, and when generalized to surface-scale, only satellite remote sensing AOD can be used, and our $PM_{2.5}$ estimation results show that our generalization is feasible and the accuracy is reliable.** Thus, these two types of AOD are used for different experimental steps, and in the point-to-surface extension experiments, there is no replacement of AERONET AOD with MODIS AOD. So, there is no error caused by the AOD category replacement on the $PM_{2.5}$ calculation. Equation R1 shows how the RF-PMRS method works.

$$PM_{2.5} = AOD\frac{FMF \cdot VE_f \cdot \rho_{f,dry}}{PBLH \cdot f_0(RH)} \tag{R1}$$

AERONET AOD is applied to calculate the true values of VE_f for establishing the RF simulation model. And the RF model construction is a step of $PM_{2.5}$ estimation (as

VE_f variable in equation (R1)). MODIS AOD is satellite AOD data, which is the most commonly used remote sensing data for large-scale retrieval of $PM_{2.5}$. It is an important variable for $PM_{2.5}$ estimation in RF-PMRS (as AOD variable in equation (R1)).

As for uncertainty, AERONET AOD provides truth values for calculating VE_f, which theoretically has negligible uncertainty, and the simulation accuracy of VE_f represents its influence on estimating $PM_{2.5}$ to a certain extent. And it is generally considered that MODIS AOD has guaranteed quality and sufficient accuracy to be used directly. Few $PM_{2.5}$ estimation articles specifically discuss the error introduced by MODIS AOD. And we think that uncertainty is relative. In the future, if a better surface AOD product appears, we will bring it into the estimation model and compare the corresponding deviation (uncertainty) with the $PM_{2.5}$ results of the MODIS AOD.

Based on the reviewer's suggestion, we have added application interpretation and uncertainty analysis of AOD data categories in Section 5.4.3 (Page 24 Lines 556-568: "AERONET AOD vs. MODIS AOD").

4. Page 11, line 274: In experiment 3, the authors applied the RF to estimate VEf and PM2.5 concentration over North China. The authors used sites worldwide (Table 1) to train the RF for estimating VEf, and the relationships learned by RF are based on the training data. How can RF represent the relationships within North China based on only two sites in this region in the training data? The authors should at least include this issue and the associated uncertainties in the discussion. I think this comment was also raised by Reviewer #3, but it was not fully addressed in the response.

Response: Thank the reviewer for the comments and constructive suggestions. **Firstly,** training a universal model based on global data and applying the model to local areas is a commonly used method for estimating atmospheric $PM_{2.5}$ (Zhang and Li, 2015; Li et al., 2016). We construct the VE_f model based on RF using high-precision point data and extend it to surface data for $PM_{2.5}$ estimations. **Secondly,** our ultimate goal is to estimate $PM_{2.5}$, and from the results, although there is uncertainty, $PM_{2.5}$ estimation results are still quite good. The experimental results demonstrate the $PM_{2.5}$ accuracy in North China (Sections 4.2 to 4.3), showing that the method has certain universality from point scale to surface scale. Meanwhile, relevant statistical indicators are common criteria for evaluating model accuracy (uncertainty), and the results of our study can fully show the applicability of the RF model in North China. **Thirdly,** if more AERONET sites can be found in the local area in the future, it will more effectively promote the accuracy of our method, and we are very grateful for the suggestions of the reviewer.

According to the reviewer's comments, we have added a new section to the article explaining the universality and overall performance of the approach (Section 5.4.1: Pages 22-23 Lines 509-539). It answers why RF-PMRS applies to the NC region. The reasons that this paper only validated $PM_{2.5}$ estimations in North China and variable uncertainty analysis are also added in Section 5.4.2 (Pages 23-24 Lines 541-551) and Section 5.4.3 (Pages 24-25 Lines 553-604). The universality of this method is analyzed from the following two specific aspects.

1) The overall performance of the model is high. We use the ground data of 9 AERONET sites around the world to train the RF model and simulate the VE_f values, the site distribution is relatively uniform and the amount of training data is sufficient. Table 1 shows a total of 6463 data matching pairs in the training period, which is enough to establish a credible RF model. Table 3 results show that in IV experiments, the accuracy of the model is well and can be generalized in different periods. For VE_f, the model shows both high internal accuracy (CV) and external accuracy (IV), so it can be generalized in regions with different aerosol types.

2) In the subsequent $PM_{2.5}$ estimation, the model displays high applicability in North China. From the perspective of the model, the four aerosol types are the classification basis of the training data, and comprehensive modeling can improve the generalization performance. Also, the addition of spatiotemporal variables can increase the model applicability in North China. On the other hand, the number of stations used in an area does not determine the regional accuracy of the established model, which can be derived from our results. Compared with the $PM_{2.5}$ ground measurements in the NC

region, the relative deviation of the RF-PMRS $PM_{2.5}$ is only 2.31 μg/m^3, which confirms that RF can represent the relationships within North China.

References:
Zhang, Y., and Li, Z.: Remote sensing of atmospheric fine particulate matter (PM2.5) mass concentration near the ground from satellite observation, Remote Sens Environ, 160, 252-262, https://doi.org/10.1016/j.rse.2015.02.005, 2015.

Li, Z., Zhang, Y., Shao, J., Li, B., Hong, J., Liu, D., Li, D., Wei, P., Li, W., Li, L., Zhang, F., Guo, J., Deng, Q., Wang, B., Cui, C., Zhang, W., Wang, Z., Lv, Y., Xu, H., Chen, X., Li, L., and Qie, L.: Remote sensing of atmospheric particulate mass of dry PM2.5 near the ground: Method validation using ground-based measurements, Remote Sens Environ, 173, 59-68, https://doi.org/10.1016/j.rse.2015.11.019, 2016.

5. Page 12, line 283: Is station FMF calculated from equation 10?

Response: Thank the reviewer for pointing out this problem. The station FMF (S-FMF) here is obtained directly from the AERONET monitoring site and does not require a calculation step. Equation 10 is the formula to calculate the VE_f value used by the original method (PMRS). The starting point of our improvement (RF-PMRS) is to replace this simple polynomial with the RF model to optimize the expression of VE_f. Finally, the accuracy of $PM_{2.5}$ obtained by RF-PMRS is improved.

To avoid misunderstandings, we have modified the statement to "Note that the station FMF values (S-FMF) from AERONET sites are used when training" (Page 12 Line 284). At the same time, the article has added a new section to describe the application of S-FMF and how it differs from Phy-DL FMF (Page 24 Lines 569-577: "S-FMF vs. Phy-DL FMF"). Hope it is clear now.

6. Page 13, line 323: If I understand correctly, experiment 1 in Table 2 is just for model evaluation (internal and external accuracy). In experiments 2 and 3, did the authors use data except for 2017 to retrain the model and get the estimation of VEf of 2017 for PM2.5 calculation? Or the authors used the VEf of 2017 from results from the 10-fold CV? If it's the second one, I doubt the consistency of the experiment, as mentioned before. The authors should clearly describe how they

obtained the VEf for experiment 2 (3) at the beginning of Section 4.2 (4.3).

Response: Thank the reviewer for the comments and constructive suggestions. The VE_f of 2017 is obtained by the RF model trained on all data pairs (including 2017) after 10-fold CV tuning. Specifically, the 10-fold CV result is used to determine the optimal combination of parameters for the model, and see Appendix A3 for the adjustment of the model parameters. Considering that the completeness of the training data will optimize the generalization performance of the model, the experiment fine-tunes the model based on all the original datasets (the training period of Table 1) under the optimal parameters, then the final RF model is constructed. This is also the most common method for ML model construction. The 10-fold CV can evaluate the internal accuracy of the model and the IV experiment provides independent time validation of the final model. According to the comment, we have added the detailed process explanation of the RF model (Page 13 Lines 312-319). At the same time, the role of 10-fold CV has been clarified (Page 13 Lines 312-314; Page 26 Lines 628-629).

Next, there is an explanation of the experimental consistency mentioned by the reviewer. Although the VE_f training period includes 2017, it does not affect the universality of experimental validation. **1)** 2017 is only a representative year. This experiment requires multiple point-scale VE_f data pairs (not every day of the year) to build a universal RF model to derive unknown VE_f (every day in 2017 when Phy-DL FMF is available). The model captures the spatiotemporal characteristics of 2017, and the generalization results are also applicable to 2017. The validation in North China shows that the model has excellent spatiotemporal generalization performance (from point to the surface). **2)** The types of results evaluated are not the same. The accuracy of the estimated $PM_{2.5}$ values is not affected by the selected year, since VE_f is obtained by introducing the Phy-DL FMF datasets (surface data) to the final RF model. Comparing the estimated results of $PM_{2.5}$ with ground values can demonstrate the superiority of the RF-PMRS method. **3)** As for why 2017 is chosen as a representative year, it is because there are more data samples for 2017 in view of the limited open data of AERONET in North China, and the complete data involved in the calculation of $PM_{2.5}$ (Page 12 Lines 297-299). The additions to section 5.4.1 (Pages 22-23 Lines

509-539) can explain the applicability of our model in North China and experimental consistency in 2017.

At the same time, we have added the descriptions of how VE_f values are obtained for experiments 2 and 3 at the beginning of Section 4.2 (Page 14 Lines 333-336). The application time of Phy-DL FMF in $PM_{2.5}$ estimation is also clarified (Page 14 Lines 335-338). To be specific, Phy-DL FMF is introduced into the RF model to replace S-FMF, and the 2017 VE_f values are estimated. Besides, Phy-DL FMF data is applied to the $PM_{2.5}$ estimation steps (as FMF variable in equation (8)) for a wider range of validation experiments.

7. Page 13, line 324: Please specify when you applied Phy-DL FMF in the process. It wasn't very clear to me. The authors mentioned that station FMF was used in VEf true value calculation and RF training. It was not clear when Phy-DL was applied.

Response: Thank the reviewer for pointing out this problem and we apologize for this lack of clarity. We construct the VE_f model based on RF using high-precision point data and extend it to surface data for $PM_{2.5}$ estimations. S-FMF is obtained directly from the AERONET monitoring sites and is one of the variables of the RF model (as FMF variable in equation (11)). In the point-to-surface extension, **Phy-DL FMF is introduced into the RF model to replace S-FMF, and the 2017 VE_f values are obtained.** The basis of the above replacement is that the accuracy of Phy-DL FMF is relatively consistent with that of S-FMF. **Besides, Phy-DL FMF data is applied to the $PM_{2.5}$ estimation steps (as FMF variable in equation (8)) for a wider range of validation experiments.** The results show that the $PM_{2.5}$ concentration estimated by RF-PMRS has high accuracy, proving the credibility of Phy-DL FMF.

According to the reviewer’s comment, **we have added the application time of Phy-DL FMF in two places in the article**. One addition is in Section 4.2 to illustrate how VE_f and $PM_{2.5}$ of experiments 2 (3) are obtained (Page 14 Lines 333-338). The other addition exists in Section 5.4.3, which comprehensively compares S-FMF and Phy-DL FMF (Page 24 Lines 569-577).

8. Page 17, line 380: In experiment 3, I guess the author also used Phy DL FMF to replace S-FMF in the RF to derive the VEf since there is no S-FMF data for each site in North China. In this case, how does the difference between station FMF and Phy-DL affect the VEf and PM2.5 estimation?

Response: Thank the reviewer for the comment. We construct the VE_f model based on RF using high-precision point data and extend it to surface data for $PM_{2.5}$ estimations. S-FMF is obtained directly from the AERONET monitoring sites and is one of the variables of the RF model (as FMF variable in equation (11)). In the point-to-surface extension, Phy-DL FMF is introduced into the RF model to replace S-FMF, and the 2017 VE_f values are obtained. The basis of the above replacement is that the accuracy of Phy-DL FMF is relatively consistent with that of S-FMF. Besides, Phy-DL FMF data is applied to the $PM_{2.5}$ estimation steps (as FMF variable in equation (8)) for a wider range of validation experiments. The results show that the $PM_{2.5}$ concentration estimated by RF-PMRS has high accuracy, proving the credibility of Phy-DL FMF. **Accordingly, we have added Section 5.4.3, which comprehensively compares S-FMF and Phy-DL FMF (Page 24 Lines 569-577).**

There is indeed a difference between S-FMF and Phy-DL FMF, but this experiment requires replacing point data (S-FMF) with surface data (Phy-DL FMF) to achieve VE_f generalization. Different surface FMFs affect the estimation accuracy of VE_f and $PM_{2.5}$. When selecting the surface FMF data, Phy-DL FMF is compared with MODIS FMF, and it is found that Phy-DL FMF has a lower bias (uncertainty) on the $PM_{2.5}$ estimations (Section 5.1). So Phy-DL FMF is chosen for the replacement. In the future, if a better surface FMF product appears, we will bring it into the RF-PMRS method to validate whether the $PM_{2.5}$ accuracy is consistent with that obtained by Phy-DL FMF.

9. Page 20, line 453 (Table 5): Maybe the comparison would be valid when the PMRS with MODIS and Phy-DL FMF were sampled on the same days. It seems like the statistics were based on two sets of days.

Response: Thank the reviewer for the comment. We apologize for any confusion regarding the results and conclusions presented in Table 5. In fact, the different days exactly indicate that the two types of FMF have different amounts of data available for $PM_{2.5}$ estimations, showing the superiority of Phy-DL FMF. The specific explanation is as follows.

It is believed that different surface data sources may affect the $PM_{2.5}$ results, introducing some uncertainty. Table 5 compares the $PM_{2.5}$ accuracy using two FMF data in 2017. The data missing time for MODIS FMF and Phy-DL FMF in North China are different, which can be found in the statistics on their respective available days (refer to valid DOY). There are far more valid days based on Phy-DL FMF than MODIS FMF (143 and 31 days), demonstrating the superiority of Phy-DL FMF. Although the specific validation time of two FMF varies, the overall accuracy of the $PM_{2.5}$ estimation (which can be regarded as the average accuracy over the year) shows that the Phy-DL FMF increases R to 0.68 (MODIS FMF: 0.38) with low uncertainty. Therefore, it is feasible and reasonable to use two sets of days in this experiment for $PM_{2.5}$ accuracy comparison. **At the same time, we have added a description of this issue and summarized it as an analysis of the uncertainty (error) of the $PM_{2.5}$ results based on two FMF datasets (Pages 24-25 Lines 578-587).**

10. Page 20, Section 5: After reading the revised manuscript, I now understand the flow and structure of this study better. It seems to me the (same) parameters from different sources are often used to estimate surface PM2.5 (e.g., S-FMF vs. Phy-DL FMF and AERONET AOD vs. MODIS AOD). I am confused about when and where they are used. Also, the differences in data used in training and generalization (e.g., S-FMF used in training vs. Phy-DL FMF used in generalization (North China) for VEf prediction) and the associated biases would propagate to PM2.5 estimation.

How do you quantify the uncertainties? It is critical to include this issue in the discussion.

Response: Thank the reviewer for the comments and constructive suggestions. In experimental comparisons, uncertainty is rarely used as a quantitative indicator, which we understand as the deviation or error between the estimated results and the true values. Meanwhile, relevant statistical indicators are common criteria for quantifying model accuracy (uncertainty). In this study, the statistical indicators used for evaluation include correlation coefficient (R), mean bias (MB), relative mean bias (RMB), root mean square error (RMSE), and mean absolute error (MAE). In addition, relative predictive error (RPE) is added to validate the accuracy of the RF-based VE_f model. See Appendix A2 for the specific information on these indicators. Regarding the uncertainty of different variables on the experimental results, the reviewers may be referring to the effect of the propagation of VE_f error on $PM_{2.5}$ estimations. This paper focuses on method optimization, and the point data used in VE_f modeling is recognized as truth values. The extension of the model to the surface data brings certain uncertainty, so the estimation accuracy of VE_f and $PM_{2.5}$ is quantitatively evaluated by experiments (Section 4).

$PM_{2.5}$ results are affected by the combination of multiple variables, and uncertainties may lie in many aspects including the instrument measurement of acquiring data or the resolution of surface data. Therefore, a comprehensive analysis of possible uncertainty in the results has been added in Section 5.4.3 (Pages 24-25 Lines 553-608). The analysis is carried out from five aspects: AOD data category, FMF data category, the uncertainty of Phy-DL FMF and MODIS FMF, $\rho_{f,dry}$ value, and unified method of variable resolutions.

Meanwhile, we describe the steps for different data categories (site and surface data) for experiments. The introduction in Section 5.4.3 can clarify the reviewer's understanding of the different types of AOD and FMF data used in this paper (Pages 24-25 Lines 553-577).

Overall, RF-PMRS shows excellent estimation performance in North China, and the accuracy of surface $PM_{2.5}$ estimation based on remote sensing data is guaranteed. In the future, with the improvement of related experimental data, we will verify our proposed method in a broader range and continuously optimize it from all aspects.

Thanks again to the reviewer for giving us a chance to revise and improve the quality of our article. In all, we find these comments quite helpful. We wish this revision will be acceptable.

Thanks to the editor for his consideration. If you still have any questions about our study, don't hesitate to contact us.