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 gratefully thank the editor for his precious time and wise suggestion. A response to the comment and an introduction to the adjustments in the manuscript follow.

Response: We have carefully read the three referee reports and feel particularly grateful to them. According to the comments from the reviewers, we made further adjustments to our manuscript, and all changes have been highlighted in cyan color in the manuscript. The major revisions include:

1) Adjustments in the article structure. We have separated the experimental results and arranged Section 2 to Section 4 of the manuscript in the order of "Data – Methods - Experiment results".

2) Specific introduction to the experimental data, especially the physical-deep learning FMF (Phy-DL FMF) dataset.

3) Clear expressions of our design experiments. Related tables and statements have been added.

4) Three additional experiments:

A. Time series analysis of $PM_{2.5}$ bias at Beijing and Beijing-CAMS sites.

B. Spatial distribution of the deviation between PMRS and RF-PMRS in North China.

C. Feature importance of the embedded RF.

5) Article expansion of the above experiments, including Section 5.3 and Appendix B: Figures.

Other minor problems have also been responded and revised, and grammatical errors have been corrected.

Response to Comments of Reviewer #1:

Comments:

The paper presents an optimized ML approach to estimate complex physical parameters using remote sensing data. The ML method and results are described well, and I recommend acceptance.

Response: Thank the reviewer for acknowledging our work, we will further explore the combination of atmospheric mechanism and machine learning on the $PM_{2.5}$ retrieval methods.

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.

General Comments:

1. It will be better to separate data and experiment results into two sections. I suggest the authors move the data section before the method section, as some variables or datasets are mentioned in the method section (e.g., Phy-DL FMF dataset). I think a better layout will be Data as the second section, Method as the third, and Results as the fourth.

Response: We gratefully thank the reviewer for his constructive suggestions on the writing structure. To show the logic of the paper more clearly, we have separated data and experiment results into two sections and reorganized the manuscript. Sections 2 to 4 are arranged in the order of "Data - Methods - Experiment results". The specific modifications are:

1) The general introduction of the article structure: Page 4 Lines 101-104.
2) Section 2. Data: from Page 4 Line 106 to Page 7 Line 173.
3) Section 3. Methods: from Page 7 Line 174 to Page 12 Line 304.
4) Section 4. Experiment results: from Page 12 Line 305 to Page 20 Line 447.
5) The Discussion (from Page 20 Line 448) and Conclusions (from Page 22 Line 494) parts respectively become the fifth and sixth sections of the article.

At the same time, the serial numbers of the corresponding equations and figures have also been changed, which are all highlighted in cyan color in the text.

2. There are many experiments, but they are not presented in a clear way. If I understand them correctly, there are 1) a 10-fold CV and hold-out test (not sure for which year) for VEf validation, 2) a hold-out test of 2017 for PM2.5 validation at Beijing and Beijing-CAMS sites, and 3) a generalization test for PM2.5

validation within North China (not sure for which year). In addition, is it correct that AERONET AOD is used for calculating PM2.5 concentration for the experiments of BJ and BC while MODIS AOD is used for North China data? It will be better to include a table or state these experiments clearly.

Response: Thank the reviewer for pointing out this problem and we feel sorry for the unclear description of the main experiments. According to the comment, we have added an experiment information table (Table. R1 below) in Section 4. It includes the validation object, study region, study period, and temporal scale of three main experiments. Some descriptive statements are also adjusted or added in appropriate places. The specific modifications to this comment include: (1) Page 12 Lines 306-307, Page 13 Line 308: added Table 2 and related introduction; (2) Page 12 Lines 288-289: abbreviation addition corresponding to Table 2; (3) Page 6 Lines 164-168: explicit statement to the experimental area and period; (4) Page 12 Lines 293-298: adjusted from the original data part to the method introduction part; (5) Page 17 Lines 381-382: added statement of the experiment information. And we hope the experiments are presented in a clear way now.

As for the AOD data usage mentioned in the comment (yellow marked), it may be that we have not clearly explained, in fact, the MODIS AOD used in both the two sites (BJ and BC) and the North China region to calculate $PM_{2.5}$ concentrations. In general, the steps for obtaining $PM_{2.5}$ are "VE_f truth value calculation - RF model construction - VE_f value estimation - $PM_{2.5}$ calculation by formula". We only used AERONET AOD to calculate the VE_f truth value as the output of model training, as shown in equations (5)-(7) (Page 8 Line 193,194,198). And the input variables for estimating the VE_f value are FMF and spatiotemporal factors, without AOD. Finally, when calculating $PM_{2.5}$, the MODIS AOD dataset is used (Page 8 Line 210: Equation (8)). On the other hand, as is known from Fig. 2 (Page 9 Line 219), $PM_{2.5}$ of the RF-PMRS method is derived from satellite AOD (i.e., MODIS AOD in our study). In response to this misunderstanding, we have clarified the purpose of use of AERONET data, please see Page 4 Lines 114-116.

Table R1. A brief information summary of the experiments conducted in our study.

Experiment	Object	Region	Period	Time scale
Model performance for training VE_f	VE_f	Global scale (Nine AERONET sites)	CV: Training period in Table 1 IV: Isolated-validation period in Table 1 (See Appendix A1)	Daily
Accuracy evaluation of PMRS/RF-PMRS	$PM_{2.5}$	Two AERONET Sites: Beijing, Beijing-CAMS	2017	Daily
Generalization performance of RF-PMRS	$PM_{2.5}$	North China region	2017	Daily

3. Validation selection: In section 3.2.2, the authors selected 2017 as the validation. I wonder why this year was selected as a validation year. Any characteristics? Also, was VEf based on RF obtained from the hold-out experiment (i.e., using data except or before 2017 at BJ and BC as training and 2017 at BJ and BC as testing) or 10-fold cross-validation? The experiment year of VEf and surface PM2.5 should be consistent.

Response: We gratefully thank the reviewer for this comment and we will answer the questions raised by the reviewer in three parts. In the early stage of the validation experiment at BC and BJ sites, we need to do some **accuracy comparisons** between AERONET data and our experimental data to ensure that our data match is correct. In view of the limited open data of AERONET in North China, and the complete data involved in the calculation of $PM_{2.5}$, we selected 2017, which has more data samples through comparison, and it is only a representative year. With the extension and disclosure of subsequent data years, we will select more years. For the sake of clarity, we have added corresponding statements in Page 12 Lines 296-297.

For the second question, VE_f is obtained by the machine learning training model (RF). In general, the machine learning model is established by 10-fold cross training and an optimal model is selected according to the statistics indicators. In this study, the isolated-validation period in Table 1 (Page 4 Lines 119-121) was not included in the experiment training. See Appendix A1 (Page 22 Lines 511-518) for the scope of use.

In this paper, we use two verification methods for VE_f. Among them, **10-fold cross validation** is to validate the **internal accuracy** of the model (recorded during training), and **isolated-validation** is to validate the temporal generalization of the model, that is, the **external accuracy** of the model.

After building the VE_F model, input the variables of the same year to get the VE_f of this year, and then deduce $PM_{2.5}$ through formula (8) (Page 8 Line 210), where the variables in the formula are also the same year. Therefore, the experimental years are all corresponding. First calculate VE_f, then derive $PM_{2.5}$. It can be seen that VE_f is the necessary step to calculate $PM_{2.5}$. For the problem of experiment year mentioned by the reviewer, machine learning relies on powerful data to **build a known model**, and then **estimates the model value of unknown range or time**. This experiment requires multiple VE_f data pairs (maybe not every day of the years) to build a universal RF model to derive unknown VE_f (every day in 2017). The $PM_{2.5}$ validation experiment only selects one of the representative years (i.e., 2017), so the experiment time here is not contradictory.

4. The temporal scale (daily or hourly?), study period, and study regions are not stated clearly. Maybe the authors could include this information along with the experiments I mentioned in comment #2.

Response: Thank the reviewer for pointing out this problem. According to the reviewer's comment, clear expressions of our design experiments have been added. First, an experiment information table is given in Section 4 (Page 12 Lines 306-307, Page 13 Line 308). It includes the validation object, study region, study period, and temporal scale of three main experiments. Also, some descriptive statements are adjusted in appropriate places. **The specific modifications are consistent with those listed in General Comments: Response #2.** We thank the reviewer again for the suggestions on the experimental statements.

Specific comments:

1. **Page 4, line 106: Please consider moving the method section after the data**

section.

Response: Thank the reviewer for this constructive suggestion. To show the logic of the paper more clearly, we have separated data and experiment results into two sections and reorganized the manuscript. Sections 2 to 4 are arranged in the order of "Data - Methods - Experiment results". The specific modifications are:

1) The general introduction of the article structure: Page 4 Lines 101-104.
2) Section 2. Data: from Page 4 Line 106 to Page 7 Line 173.
3) Section 3. Methods: from Page 7 Line 174 to Page 12 Line 304.
4) Section 4. Experiment results: from Page 12 Line 305 to Page 20 Line 447.
5) The Discussion (from Page 20 Line 448) and Conclusions (from Page 22 Line 494) parts respectively become the fifth and sixth sections of the article.

At the same time, the serial numbers of the corresponding equations and figures have also been changed, which are all highlighted in cyan color in the text.

2. Page 8, Table 1: This table should be with the data section of AERONET; the data section should be presented before the method section.

Response: We gratefully thank the reviewer for the suggestions. According to the suggestion, we have changed the position of Table 1 and moved it to the data section (Page 4 Lines 119-121). At the same time, the relevant descriptions of table 1 have been modified to make the presentation clear, including: 1) Page 4 Lines 116-117 in the data section; 2) Page 10 Lines 253-256 in the methods section.

Meanwhile, the structure of the article is adjusted and the revised contents are the same as the modifications in Specific comments: Response #1.

3. Page 9, line 216: How does the difference between station FMF and Phy-DL FMF influence surface PM2.5 estimation?

Response: Thank the reviewer for his comment and we will explain this question from the following two aspects. 1) Space coverage: ground FMF (S-FMF) is point data with a sparse distribution. Phy-DL FMF is areal data, which is applicable to a wider range of experiments. 2) Accuracy: S-FMF is a true value with high accuracy. Experiments

show that the precision of Phy-DL FMF is equivalent to that of S-FMF (Yan et al., 2022). In order to expand the applicability of VE_f model, this study uses the ground values as the truth to train the RF model and uses the planar FMF (Phy DL FMF) to estimate $PM_{2.5}$ in a wider range and more locations.

References:

Yan, X., Zang, Z., Li, Z., Luo, N., Zuo, C., Jiang, Y., Li, D., Guo, Y., Zhao, W., Shi, W., and Cribb, M.: A global land aerosol fine-mode fraction dataset (2001--2020) retrieved from MODIS using hybrid physical and deep learning approaches, *Earth Syst. Sci. Data*, 14, 1193-1213, https://doi.org/10.5194/essd-14-1193-2022, 2022.

4. Page 9, line 232: Please consider separate results from this section.

Response: Thank the reviewer for the suggestion. We have separated data and experiment results into two sections and reorganized the manuscript. The specific modifications are consistent with those listed in Specific comments: Response #1. We thank the reviewer again for the suggestions on the article structure.

5. Page 10, line 247: Please include more information about the Phy-DL FMF dataset, as it is one of the important components of this paper. How did you calculate or derive FMF in this dataset? What are the differences between FMF in this dataset and at the AERONET sites?

Response: We gratefully thank the reviewer for the suggestions. Detailed descriptions have been supplemented in the manuscript, including its generation process and performance (Page 5 Lines 139-146, Page 6 Lines 147-148, Page 10 Lines 243-245, Page 30 Lines 726-729). Also, for the convenience of statements, we have adjusted the order of data introduction (Page 4 Line 107 to Page 6 Line 150, Section 2.1 to section 2.4). Other detailed modifications are: 1) abbreviation explanation, Page 5 Line 124; 2) definition of meteorological data types, Page 6 Lines 151-152.

The Phy-DL FMF dataset is published in Geotiff format. We obtained the FMF value by reading the value on the image files. Specifically, it selects the FMF data obtained by a physical method (i.e., Look-Up-Table-based Spectral Deconvolution Algorithm, LUT-SDA) as the optimization target. Then it combines the Phy-based FMF into a

deep-learning model along with multiple auxiliary data such as satellite observations for the final Phy-DL results. Note that the process is trained with AERONET data as the ground truth. In the comparison experiment against the ground FMF, Phy-DL FMF shows a high accuracy (R = 0.78, RMSE = 0.100).

6. Page 10, line 257: It seems like the spatial resolutions of AOD, FMF, and ERA5 meteorology are different. How do different spatial resolutions affect PM2.5 estimation? Please elaborate the uncertainties of various resolutions of the input data.

Response: We gratefully appreciate the reviewer for his comment. In most experiments, the lowest resolution of all data will be taken as the unified resolution when taking values. The average values of different data may lose some spatial details during the upsampling/downsampling process, which shows the accuracy and uncertainty in the estimation results. In this study, there is no such uncertainty problem. We set 1 ° as the unified spatial unit, and take the longitude and latitude of each cell's center as the reference longitude and latitude. **The variables in the data section are spatially matched to ground sites at their respective resolutions.** And we have described this space-time matching method in the methods section (see Page 12 Lines 293-295). So, all kinds of data uncertainties only exist in instrument measurement or statistical release. Please refer to the website and the official instructions for each data (All websites are listed in the Code and data availability section on Pages 24-25 Lines 549-558).

7. Page 11, line 270: Is AERONET AOD used for calculating PM2.5 concentration for the experiments of BJ and BC, while MODIS AOD is used for North China? If so, how do the differences between two AOD products affect PM2.5 estimation? Suppose this approach would be applied to regions where AERONET is not available (the most likely scenario); it is important to evaluate the biases caused by different AOD products, particularly the input variables of RF are based on AERONET data.

Response: Thank the reviewer for his comment and we are sorry for the

misunderstanding of data usage. As for the AOD data usage mentioned in the comment, it may be that we have not clearly explained, in fact, the MODIS AOD used in both the two sites (BJ and BC) and the North China region to calculate $PM_{2.5}$ concentrations. In general, the steps for obtaining $PM_{2.5}$ are "VE_f truth value calculation - RF model construction - VE_f value estimation - $PM_{2.5}$ calculation by formula". We only used AERONET AOD to calculate the VE_f truth value as the output of model training, as shown in equations (5)-(7) (Page 8 Line 193,194,198). And the input variables for estimating the VE_f value are FMF and spatiotemporal factors, without AOD. Finally, when calculating $PM_{2.5}$, the MODIS AOD dataset is used (Page 8 Line 210: Equation (8)). On the other hand, as is known from Fig. 2 (Page 9 Line 219), $PM_{2.5}$ of the RF-PMRS method is derived from satellite AOD (i.e., MODIS AOD in our study). In response to this misunderstanding, we have clarified the purpose of use of AERONET data, please see Page 4 Lines 114-116.

8. Page 12, Fig. 3: Please mark the AERONET sites (Beijing and Beijing-CAMS) on the map (use different colors and shapes).

Response: Thank the reviewer for pointing out this problem. To highlight Beijing and Beijing-CAMS sites, we have marked them in bold font in Table 1 (Page 4 Lines 119-121). Referring to the suggestion, we use the yellow quadrangles to highlight the positions of the two sites (BJ and BC) (Step 1 zoom-in view map in Fig. 3, Page 11 Line 260), which are different from the other seven sites (red points). At the same time, we have described the prominent sites in the article, please see Page 10 Lines 256-258, and Page 11 Lines 264-265.

9. Page 12, line 288: The experiment period is a little bit confusing. The surface PM2.5 validation is conducted for 2017, while the VEf validation is based on the 10-fold CV and different hold-out periods. Also, please justify the test selection for 2017.

Response: We gratefully appreciate the reviewer for his comment. In our study, VE_f is obtained by a machine learning training model (RF). In general, we use two verification

methods for VE_f. Among them, 10-fold cross-validation is to validate the internal accuracy of the model (recorded during training), and isolated-validation is to validate the temporal generalization of the model, that is, the external accuracy of the model. In principle, the trained model applies to all periods, including 2017. The $PM_{2.5}$ validation experiment only selects one of the representative years (i.e., 2017), the main reason is that $PM_{2.5}$ calculation also involves many other variables (formula (8), Page 8 Line 210), and the accuracy, validity and number pair matching of all variables need to be considered. Specifically, in the early stage of the validation experiment at BC and BJ sites, we need some accuracy comparisons between AERONET data and our experimental data to ensure that our data match is correct. In view of the limited open data of AERONET in North China, and the complete data involved in the calculation of $PM_{2.5}$, we selected 2017, which has more data samples through comparison, and it is only a representative year. With the extension and disclosure of subsequent data years, we will select more years. For the sake of clarity, we have added corresponding statements in Page 12 Lines 296-297.

10. Page 13, lines 308-309: Was the VEf based on RF derived from the hold-out experiment? Ideally, the VEf based on RF should be from test results (i.e., using data in Table 1 but excluding data at BJ and BC in 2017 as training and data at BJ and BC in 2017 as testing).

Response: We gratefully appreciate the reviewer for this comment. In general, the machine learning model is established by 10-fold cross training and the optimal model is selected according to the statistics indicators. In this study, VE_f is obtained by 10-fold cross training, and two verification methods are used for VE_f, that is, 10-fold cross-validation (recorded during training), and isolated-validation (i.e., hold-out experiment).

For the training period mentioned, the reviewer may have some misunderstanding about the data used in each step of our experiment. The steps for obtaining $PM_{2.5}$ are "VE_f truth value calculation - RF model construction - VE_f value estimation - $PM_{2.5}$ calculation by formula". We only used AERONET AOD to calculate the VE_f truth value as the output of model training, as shown in equations (5)-(7) (Page 8 Line

193,194,198). When estimating the $PM_{2.5}$ concentration, the Phy-DL FMF and MODIS AOD are used as other stations. In other words, VE_f training data are not used to calculate $PM_{2.5}$, so these two sites use the same category of data as other stations in North China, and the period is not contradictory.

11. Page 14, Fig. 4: What is the correlation between STA and PMRS (RF-PMRS) and the RMSE or bias of the time series?

Response: Thank the reviewer for his comment to improve the quality of our manuscript. Fig. 4 displays the time series of $PM_{2.5}$ values for STA and PMRS (RF-PMRS) and it can show the difference of each day. Fig. 6 (Page 16 Line 374) plots the $PM_{2.5}$ bias distribution patterns between STA and PMRS (RF-PMRS) and it shows the relationship between them, including kinds of statistical indicators (RMSE, MAE and mean bias). As for the correlation coefficient (R), we have explained it in the text (see Page 16 Lines 368-371: **"RF-PMRS $PM_{2.5}$ values have a strong linear relationship with the ground truth at both sites, with R around 0.8 (0.82 at BJ and 0.78 at BC)."**). Besides, as Fig. R1 below shows, we have added bias time series plots in the revised manuscript (Fig. B1, Page 24 Lines 540-544) and related descriptions (Page 14 Lines 339-342). We hope it is clear now.

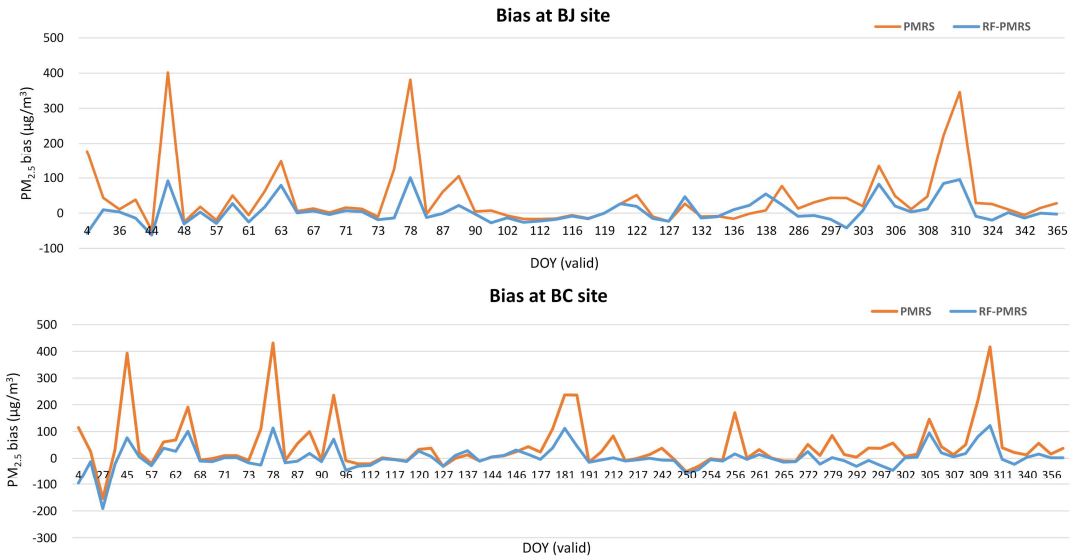

Fig. R1. The time series of PMRS/RF-PMRS $PM_{2.5}$ bias at the Beijing and Beijing-CAMS sites under their respective DOYs in 2017. The orange line represents the bias between the $PM_{2.5}$ values

of PMRS and stations, while the blue one indicates the $PM_{2.5}$ difference between RF-PMRS and stations.

12. Page 16, line 361: In the RF model estimating VEf, the authors include longitude and latitude as predictors, while the longitude and latitude of the sites in North China are out of training samples (Table 1). How can we trust the extrapolation of the RF model (and technically, RF is bad at extrapolation)?

Response: Thank the reviewer for his comment and the statement may be somewhat inaccurate. Machine learning relies on powerful data to build a known model, and then estimates the model value of an unknown range or time. This experiment requires multiple VE_f data pairs (training samples) to build a universal RF model to derive unknown VE_f. Its spatiotemporal continuity can be increased, so the "extrapolation" statement violates the original intention of machine learning. In this experiment, the VE_f model is trained based on RF to estimate the spatiotemporal range of VE_f continuously. The global sites were selected to consider the spatiotemporal differences of four main aerosol types, and North China did not go beyond the scope of training samples in this constructed global model. By comparing estimates with ground values through statistical indicators, it is also the most common and recognized method for remote sensing product validation today.

13. Page 16, line 361: This section mainly discusses the general performance comparison between PMRS and RF-PMRS. It would be helpful if the authors could elaborate more about the spatial or temporal distribution of biases for the two methods (e.g., which area or period shows larger improvement and why; what are the associated factors influencing VEf).

Response: Thank the reviewer for the constructive suggestions. As the review has mentioned, additional experiments can enable us to have a more comprehensive evaluation of the performance of our optimized method. For this, we have added relevant discussions. **1) RMSE spatial distribution at NC stations (Page 18-19 Lines 424-431, Fig. 9: Page 19 Lines 438-441).** In addition to the general performance

comparison, Fig. R2 (**Fig. 9 in the revised manuscript**) presents the annual average RMSE spatial distribution of PMRS and RF-PMRS $PM_{2.5}$ at NC stations. In the areas where the stations are clustered, the deviation also reduces significantly. Please see more details in the manuscript. **2) RF feature importance to evaluate associated factors influencing VE_f (Page 21 Lines 485-492, Fig. B2: Page 24, Lines 546-548).** The feature importance of RF is calculated to evaluate the contribution of model predictors to VE_f simulation. Fig. R3 below (**Fig. B2 in the revised manuscript**) shows the results by normalization (taking 100 as the total). It not only demonstrates the importance of predictors that influence VE_f, but also provides a basis for future model optimization.

This experiment does not show the surface distribution map, the main reasons are: 1) The experiment is based on the resolution of Phy-DL FMF (1°*1°), its resolution is coarse, and the number of meshes presented in North China is small (as Fig. R4 below shows). Therefore, the law of surface $PM_{2.5}$ is difficult to summarize. 2) The validation experiment is regional (North China) because the ρ empirical values and $PM_{2.5}$ truth values in other regions are insufficient, and more other research results are needed. 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.

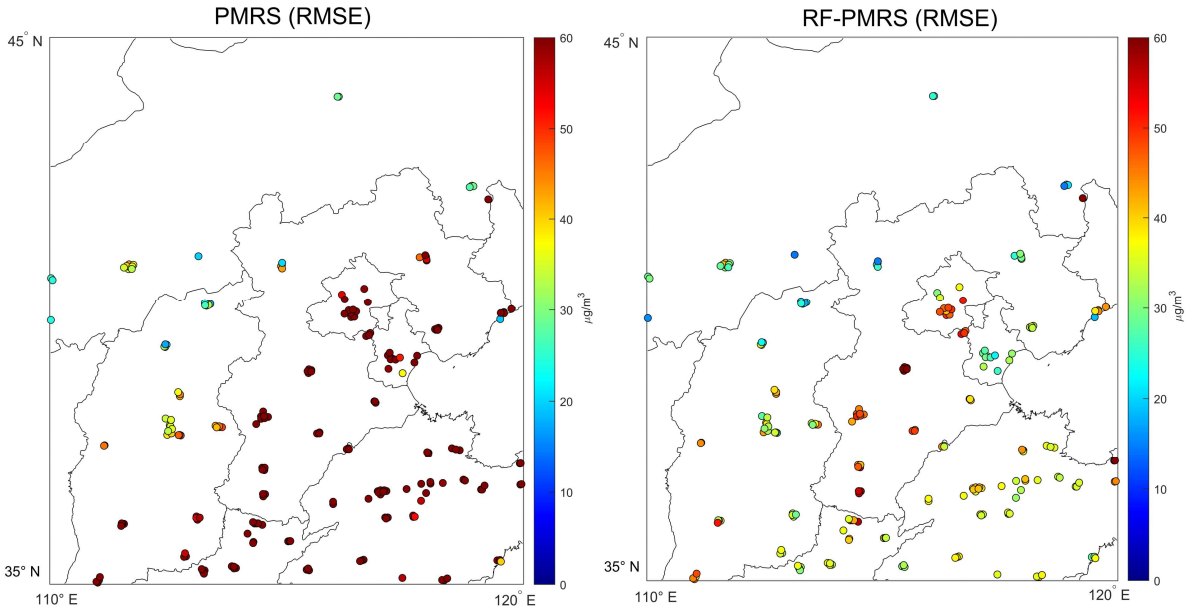

Fig. R2. RMSE of the year-average $PM_{2.5}$ concentration values between different models and ground stations (left: PMRS $PM_{2.5}$, right: RF-PMRS $PM_{2.5}$). Note that the top red of the RMSE legend indicates RMSE values equal to or greater than 60 μg/m³.

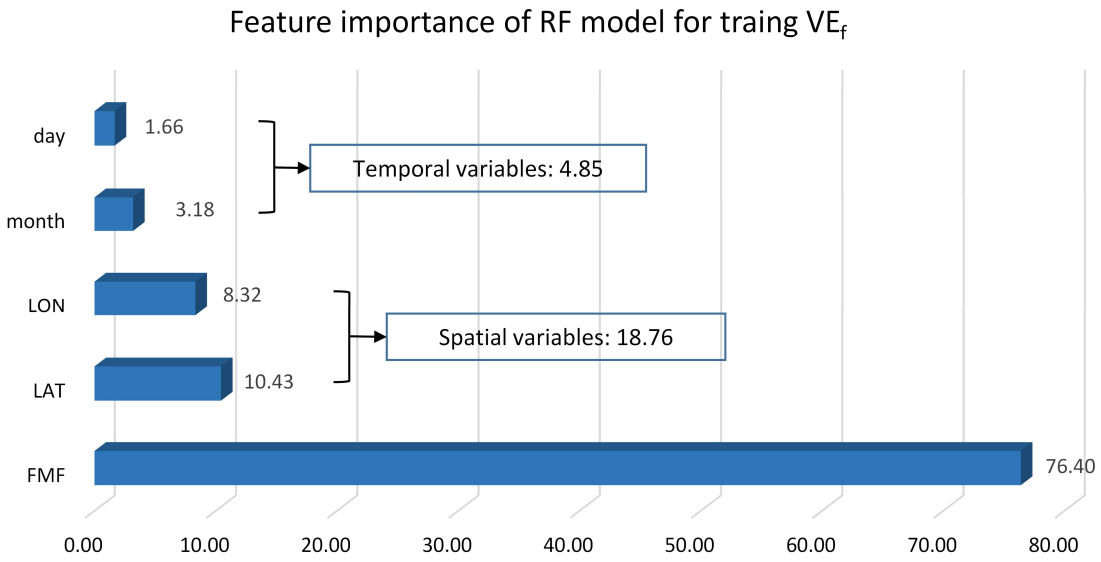

Fig. R3. The predictor importance results (normalized) of the RF model for training VE_f.

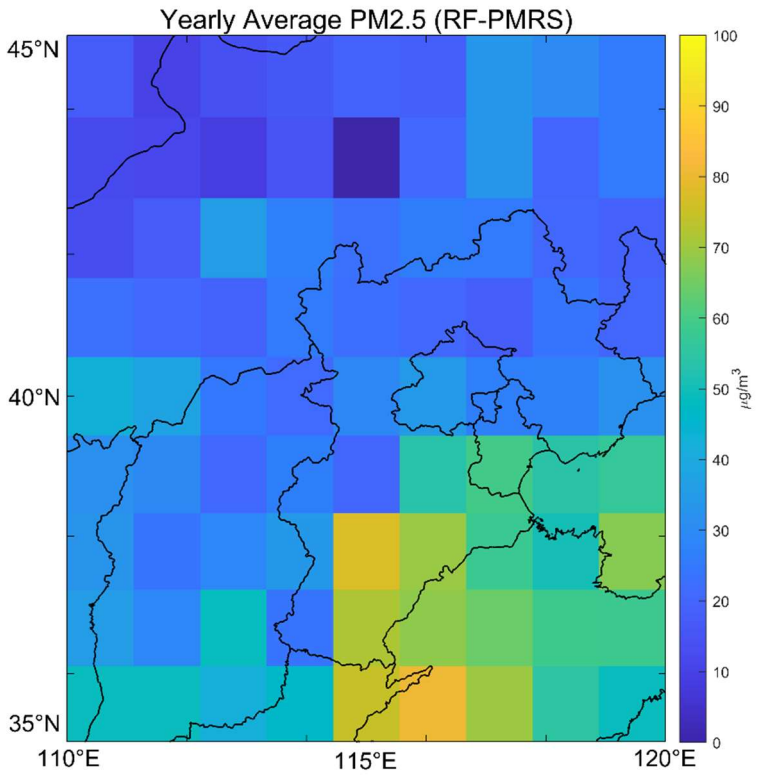

Fig. R4. Map of annual average $PM_{2.5}$ values of RF-PMRS in North China.

14. Page 17, lines 382-385: What do the high-value points mean? The high values of PM2.5 concentration or VEf? I guess the underestimation of VEf would lead to the underestimation of PM2.5 in RF-RMRS.

Response: We gratefully appreciate the reviewer for this comment. Firstly, we feel sorry for making an ambiguous expression here. The statement "high-value points" here means high $PM_{2.5}$ concentration (especially greater than 150 μg/m³), which has been explained in the original manuscript (Page 17 Lines 400-401). The specific meaning of the word can be known in relation to the context and Fig. 7 (validation scatterplots of $PM_{2.5}$ results). To clarify the expression, we have changed it from "high-value points" to "high-value $PM_{2.5}$ points". Secondly, in many previous studies (Ma et al., 2014; Li et al., 2017), there is an underestimation of high $PM_{2.5}$ values. In our experiment, it may be caused by the insufficient quantity of high-value $PM_{2.5}$ points (only 1319 out of 28305), which is difficult to adequately reflect the fitting effect of the method. As the scope of future experiments expands, this discussion can be refined. In a word, the proposed method solves the overall overestimation problem of PMRS and the underestimation phenomenon is not obvious, which is trustworthy.

References:

Ma, Z., Hu, X., Huang, L., Bi, J., and Liu, Y.: Estimating ground-Level PM2.5 in China using satellite remote sensing, *Environ. Sci. Technol.*, 48, 7436-7444, https://doi.org/10.1021/es5009399, 2014.

Li, T., Shen, H., Zeng, C., Yuan, Q., and Zhang, L.: Point-surface fusion of station measurements and satellite observations for mapping PM2.5 distribution in China: Methods and assessment, *Atmospheric Environ.*, 152, 477-489, https://doi.org/10.1016/j.atmosenv.2017.01.004, 2017.

15. Page 18, line 414: Is this experiment also based on 2017 and North China? Please specify.

Response: Thank the reviewer for this comment and we feel sorry for making an ambiguous expression here. The accuracy comparison of PMRS using MODIS/Phy-DL FMF is conducted at BJ and BC sites in 2017. And we have added statements about the experiment time (Page 12 Lines 295-296, Page 20 Line 451), and we hope it is clearer now.

16. Page 19, line 430: Is this experiment also based on 2017 and North China? Please specify.

Response: Thank the reviewer for this comment and we feel sorry for making an ambiguous expression here. Section 5.2 compares the performance of different decision tree (DT) models. This experiment takes the same region and time as the RF model to train VE_f and validate VE_f, that is, the period in section 4.1 (Page 13 Line 310). The experimental information has listed in Table 2 (Page 13 Lines 308-309), and you can refer to Table 1 (Page 4 Lines 119-121) for the specific time and site locations. We hope it is clearer now.

17. Page 20, line 448: The authors should consider adding more discussions, including 1) uncertainties of the embedded RF approach (e.g., out-of-sample issue mentioned in comment #12 and the uncertainties of PM2.5 estimation associated with different data sources), 2) spatial or temporal distribution of biases for the two methods (see comment #13).

Response: We gratefully thank the reviewer for his constructive suggestions. For the out-of-sample problem mentioned by the reviewer, we used the IV method to experiment on the external accuracy during the verification of the VE_f model, and during the $PM_{2.5}$ numerical validation, its AOD and FMF data were replaced, which did not involve the VE_f training data, so it was also an out-of-sample validation to a certain extent. At the same time, the purpose of the experiment is to optimize the complex parameters in the physical model (i.e., to improve VE_f), without modeling other variables involved in $PM_{2.5}$ calculation, and the variable selection is based on the current optimal or most commonly used remote sensing products. Their uncertainties only exist in instrument measurement or statistical release. Please refer to the website and the official instructions for each data. Interpretations of experimental procedures and validation methods can be found in specific comments: Response #7 and #9.

As the review has mentioned, additional experiments can enable us to have a more comprehensive evaluation of our optimized method. For this, we have added relevant

discussions. 1) **RMSE spatial distribution** at NC stations (Page 18-19 Lines 424-431, Fig. 9: Page 19 Lines 438-441). 2) **RF feature importance** to evaluate associated factors influencing VE_f (Page 21 Lines 485-492, Fig. B2: Page 24, Lines 546-548). A detailed explanation can be found in specific comments: Response #13. Thank the reviewer again for his wise comments which largely help us improve the paper quality.

Technical comments:

1. Page 2, line 37 & Page 13, line 314: The word "trends" is misused. Fig. 4 displays the "time series" of PM2.5 values in 2017. In my opinion, "trends" is often used to describe a long-term increase or decrease in the data, which is not the case in Fig. 4.

Response: Thank the reviewer for pointing out this problem very much. We apologize for misusing the word 'trends' due to misunderstanding. Based on the comment, we have replaced this term in the full text, with the following modifications: 1) Page 2 Line 37, 'time series change'; 2) Page 14 Line 331, 'the time series of $PM_{2.5}$ values'; 3) Page 14 Line334, 'the variation'; 4) Page 15 Line 352, 'Three $PM_{2.5}$ time series'. The above changes are highlighted in the revised manuscript, and we hope that the expression is now clear.

2. Page 19, line 419: Please specify DOY in the main text.

Response: Thank the reviewer for this comment. In fact, we explained 'DOY' in the heading of Table 5 (Page 20 Lines 463-465). For clarity, we have added detailed explanations where 'DOY' first appears (Page 15 Lines 353-354), and we have also elaborated the meaning accordingly in the main text (Page 20 Lines 453-455). Hope it is clear now.

Response to Comments of Reviewer #3:

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.

General Comments:

This manuscript used the Random Forest machine learning method to improve the calculation of the parameter VEf, which is the columnar volume-to-extinction ratio of fine particles, in order to improve the PM2.5 simulation. The results present the new method outperformed the traditional method. In addition, this study combined the model-based and observation-based data to further improve the accuracy of PM2.5 simulation. However, this manuscript should have better organized the structure. The data, method and result sections should have the clear clue. Second, eq. 3 is used to calculate the ground truth of VEf, but why VEf should be estimated by the PMRS (eq. 8) and the RF method, which is very confused for the logic of this study. The RF model is trained by the spatial and temporal variables, leading to that the relationship will depend on the different locations. As a consequence, the ML model may not represent the intrinsic physical relationship. Third, the result section only selects two similar sites to estimate the performance. It could not be enough and could be better using more sites with different aerosol types. Finally, there are several obvious typos in the manuscript, and the English language is poor. I think the authors should be asked to have the manuscript proofread by a native English speaker before the article can be considered for publication in a scientific journal. Therefore, I would recommend for a major revision.

Response: We gratefully thank the reviewer for his comments and constructive suggestions. As the reviewer described, we do have deficiencies in terms of article structure, English grammar, and experimental explanations, and we apologize for that. Therefore, we have made improvements in the following areas: 1) adjust the structure

of the article; 2) supplement the experimental analysis; 3) revise the English grammar.

Firstly, to show the logic of the paper more clearly, we have separated data and experiment results into two sections and reorganized the manuscript. Sections 2 to 4 are arranged in the order of "Data - Methods - Experiment results". The specific modifications are:

1) The general introduction of the article structure: Page 4 Lines 101-104.
2) Section 2. Data: from Page 4 Line 106 to Page 7 Line 173.
3) Section 3. Methods: from Page 7 Line 174 to Page 12 Line 304.
4) Section 4. Experiment results: from Page 12 Line 305 to Page 20 Line 447.
5) The Discussion (from Page 20 Line 448) and Conclusions (from Page 22 Line 494) parts respectively become the fifth and sixth sections of the article.

At the same time, the serial numbers of the corresponding equations and figures have also been changed, which are all highlighted in the text.

The second question has been answered in detail in **Response #10** and the third question has been replied to specifically in **Response #7 and Response #19**. The grammatical errors have been corrected. In fact, we have done several rounds of grammar corrections for this paper. According to the reviewer's comment, we have re-edited the paper and revised the full-text grammar.

Other minor problems have also been responded to and revised, and all changes have been highlighted in cyan color in the manuscript. An item-by-item response to the comments raised by the reviewer follows.

Minor issues:

1. Line 62-64: The authors say the machine learning is the powerful tool. But in the next sentence, you write "the regression is affected by the distribution and density of ground stations". It is confused that what the challenge is for the ML method. Is it correct the ML methods cannot achieve better performance for the second category methods? Do you try to compare your method with these methods?

Response: We gratefully appreciate the reviewer for his comment. The principle of the regression method is to establish the relationship between a variety of variables and

ground $PM_{2.5}$ site values, including generalized linear model, two-stage hierarchical models, geographically and temporally weighted regression (Zhang et al., 2021). To better characterize the complex nonlinear relationship between the $PM_{2.5}$, AOD, and other possible influencing factors, machine learning (ML) methods have been widely applied to establish data-driven models for $PM_{2.5}$ mass concentrations. ML has strong nonlinear relationship capture ability and data fitting ability. Most experiments show that ML can greatly improve $PM_{2.5}$ accuracy and mapping results (Li et al., 2017; Wei et al., 2020). Wei's experiments comprehensively compare the performance of different regression methods, as shown in the following table (refer to Table 2 in Wei et al., 2020).

Based on the advantages of ML, our experiments focus on the integration of physical methods and ML methods and improve the accuracy of physical parameters. So, we didn't add additional regression model comparisons. In terms of model selection, different ML models are suitable for diverse research data, and decision tree (DT) models can better fit experiments with fewer variables, such as this study. For comparison, except for RF, the Extremely Randomized Tree (ERT) and Gradient Boosting Decision Tree (GBDT) models have also been established. And we choose RF that best suits our experimental data as the optimization model. Details of the comparison can be found on Page 20 Section 5.2.

However, ML methods also face challenges and need to be improved. There are two main challenges: 1) Site dependency; 2) Lack of physical meaning (Irrgang et al., 2021). We have all pointed out its shortcomings in the article (Page 2 Lines 64-65, Page 3 Lines 69-70). As for the semi-empirical physical approach, it has strong physical significance and derives the $PM_{2.5}$ mass concentration independently of in situ observations. The above advantages of physical methods are also the shortcomings and challenges of ML. In order to complement each other, an optimization method is proposed in this experiment. We hope it is clear now.

Table R2. Comparison of different regression methods in previous studies (refer to Table 2 in Wei et al., 2020)

Model	Spatial resolution	Model validation					Predictive power		
		R^2	RMSE	MAE	Slope	Intercept	Daily R^2	Monthly R^2	Literature
GWR	10 km	0.64	32.98	21.25	0.67	21.22	–	–	Ma et al. (2014)
TSAM	10 km	0.80	22.75	15.99	0.79	15.31	–	–	Fang et al. (2016)
Gaussian	10 km	0.81	21.87	–	0.73	17.97	–	–	Yu et al. (2017)
RF	10 km	0.83	18.08	–	–	–	–	–	Chen et al. (2018)
GAM		0.55	29.13	–	–	–	–	–	
DBN	10 km	0.54	25.86	18.10	0.55	24.56	–	–	T. Li et al. (2017b)
Geo-DBN		0.88	13.03	08.54	0.86	6.39	–	–	
Two-stage	10 km	0.77	17.10	11.51	0.76	11.64	0.41	0.73	Ma et al. (2019)
Two-stage	6 km	0.60	21.76	14.41	0.85	8.63	–	–	Yao et al. (2019)
GRNN	3 km	0.67	20.93	13.90	0.62	22.90	–	–	T. Li et al. (2017a)
GWR	3 km	0.81	21.87	–	0.83	9.44	–	–	You et al. (2016)
D-GWR	3 km	0.72	21.01	14.59	0.79	12.92	–	–	He and Huang (2018)
Two-stage		0.71	21.21	13.50	0.73	16.67	–	–	
GTWR		0.80	18.00	12.03	0.81	11.69	0.41	–	
XGBoost	3 km	0.86	14.98	–	–	–	–	–	Chen et al. (2019)
ML	3 km	0.53	30.40	19.60	0.53	25.3	–	–	Xue et al. (2019)
ML + GAM		0.61	27.80	17.70	0.61	21.2	0.57	0.74	
MLR	1 km	0.41	20.04	30.03	0.41	30.03	0.38	–	Wei et al. (2019a)
GWR		0.53	23.28	19.26	0.61	20.93	0.44	–	
Two-stage		0.71	18.59	14.54	0.71	15.10	0.35	–	
RF		0.81	17.91	11.50	0.77	12.56	0.53	–	
STRF		0.85	15.57	9.77	0.82	9.64	0.55	0.73	
STET	1 km	0.89	10.35	6.71	0.86	6.16	0.65	0.80	This study

References:

Zhang, Y., Li, Z., Bai, K., et al.: Satellite remote sensing of atmospheric particulate matter mass concentration: Advances, challenges, and perspectives, *Fundamental Research*, 1, 240-258, https://doi.org/10.1016/j.fmre.2021.04.007, 2021.

Li, T., Shen, H., Zeng, C., et al.: Point-surface fusion of station measurements and satellite observations for mapping PM2.5 distribution in China: Methods and assessment, *Atmospheric Environ.*, 152, 477-489, https://doi.org/10.1016/j.atmosenv.2017.01.004, 2017.

Wei, J., Li, Z., Cribb, M., et al.: Improved 1 km resolution PM2.5 estimates across China using enhanced space-time extremely randomized trees, *Atmos. Chem. Phys.*, 20, 3273–3289, https://doi.org/10.5194/acp-20-3273-2020, 2020.

Irrgang, C., Boers, N., Sonnewald, M., Barnes, E.A., et al.: Towards neural Earth system modelling by integrating artificial intelligence in Earth system science, *Nat. Mach. Intell.*, 3, 667-674, https://doi.org/10.1038/s42256-021-00374-3, 2021.

2. Line 75: It is confused what the relationship between the sentence "PM2.5 concentration was estimated … " and the previous half sentence is?

Response: Thank the reviewer for this comment and we feel sorry for causing this confusion. As stated in the text, the semi-empirical physical approach takes the physical theory as the basis and S is a complex physical parameter that needs to be optimized in the $PM_{2.5}$ calculation steps (Page 3 Lines 66-68, Lines 72-73). The first half sentence ("Raut and Chazette (2009) introduced a specific extinction cross-section to simplify

the expression of S") explains the method of optimizing S. $PM_{2.5}$ calculation is based on this optimization method, and the second half of the sentence continues to explain the data used in the process. For clarity, we have brought the description of the data to the beginning of the sentence (Page 3 Lines 73-75). We hope it is clear now.

3. Line 97: RF is not a deep learning method.

Response: Thank the reviewer for this comment. In fact, the deep learning here (Page 4 Line 97) is an introduction to the Phy-DL FMF dataset, not an explanation of RF. Phy-DL FMF dataset is generated by a hybrid retrieval algorithm of a deep learning method and physical mechanisms. As for RF, it is a machine learning method (Page 2 Lines 94-95). We carefully examined the full text and found no false claims that classify RF as a deep learning approach.

4. Line 133: what does the "PVSD" stand for?

Response: Thank the reviewer for this comment. We have explained **"PVSD"** in detail where it was first mentioned, please see page 8 Line 196 (Line 128 in the original manuscript). In the article, PVSD means "particle volume size distributions". In general, abbreviations in the paper only need to be explained once, so we did not write the full name of "PVSD" on Page 8 line 201 (formerly Line 133).

5. Line 148: please add the reference for the Eq 8.

Response: Thank the reviewer for this comment. Section 3.1.2 (Page 8 Line 203) is a detailed description of the PMRS method, and we introduced its provenance when it first appeared (Page 3 Lines 77-78). According to the comment of the reviewer, we have added the reference for Equation 10 (formerly Equation 8).

6. Line 152: how to define the "uncertainty"?

Response: Thank the reviewer for this comment. Actually, we have quoted only the exact words of the previous article (Zhang and Li, 2015), whose evaluation of the results uses the indicator "uncertainty". The original text is stated as follows: "The

$PM_{2.5}$ remote sensing formula suffers uncertainties not only from a parameterization scheme, e.g. FMF-VEf relationship and hygroscopic growth function, but also from measurement parameters. Following error propagation theory, errors on $PM_{2.5}$ can be written as:

$$\frac{\delta PM_{2.5}}{PM_{2.5}} = \sqrt{\left(\frac{\delta AOD}{AOD}\right)^2 + \left(\frac{\delta FMF}{FMF}\right)^2 + \left(\frac{\delta PBLH}{PBLH}\right)^2 + \left(\frac{\delta VE_f}{VE_f}\right)^2 + \left(\frac{\delta f_0(\mathrm{RH})}{f_0(\mathrm{RH})}\right)^2 + \left(\frac{\delta \rho_{f,dry}}{\rho_{f,dry}}\right)^2}$$."

And the 34% uncertainty result is derived from the above formula. Please see Zhang and Li (2015) for details. **In our study, the indicator "uncertainty" is not used, so the definition is not explained in the manuscript.**

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.

7. Line 155: You mentioned that aerosol type and spatiotemporal variables could affect the regression. It could be better to discuss their importance in the result section.

Response: We gratefully thank the reviewer for his constructive suggestions. According to the reviewer's comments, we have added relevant discussions. **1) RMSE spatial distribution at NC stations (Page 18-19 Lines 424-431, Fig. 9: Page 19 Lines 438-441).** In addition to the general performance comparison, Fig. R5 (Fig. 9 in the revised manuscript) presents the annual average RMSE spatial distribution of PMRS and RF-PMRS $PM_{2.5}$ at NC stations. In the areas where the stations are clustered, the deviation also reduces significantly. Please see more details in the manuscript. **2) RF feature importance to evaluate associated factors influencing VE_f (Page 21 Lines 485-492, Fig. B2: Page 24, Lines 546-548).** The feature importance of RF is calculated to evaluate the contribution of model predictors to VE_f simulation. Fig. R6 below (Fig. B2 in the revised manuscript) shows the results by normalization (taking 100 as the total). The contribution of spatiotemporal variables is about 1/3 of FMF, which indirectly affirms the credibility of RF feature learning.

As the review has mentioned, aerosol type is an important factor affecting the value of VE_f. Therefore, when modeling, we use site data of different aerosol types, and the obtained model has a certain universality. However, due to limited experimental data, we only conducted experiments in North China (the main aerosol type is urban–industrial). The main reasons are: **1) insufficient $\rho_{f,dry}$ value** (Page 8 Line 210, Eq 8). As the empirical value in the semi-physical empirical model, the $\rho_{f,dry}$ value is often obtained by field measurement and induction. Table R3 below shows some of the $\rho_{f,dry}$ empirical values from previous studies. The insufficient $\rho_{f,dry}$ values hinder the derivation of $PM_{2.5}$ in other regions; **2) $PM_{2.5}$ public value of ground sites around the world is limited.** Accurate and sufficient in-situ $PM_{2.5}$ values are the basic guarantee for the verification of estimated $PM_{2.5}$ results. 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.

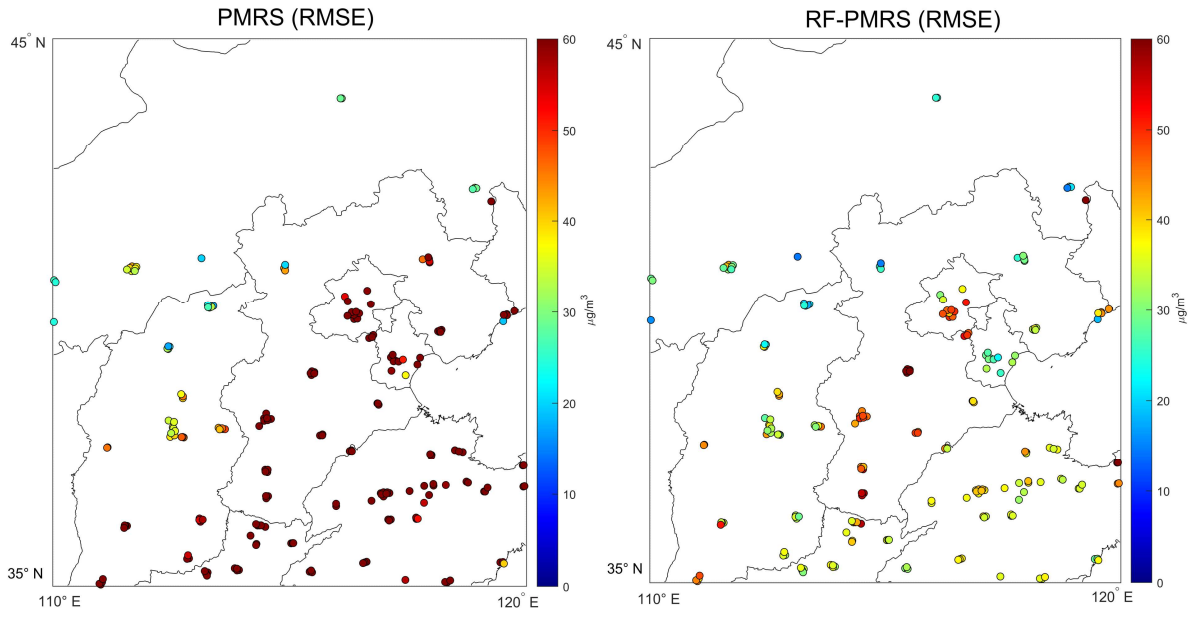

Fig. R5. RMSE of the year-average $PM_{2.5}$ concentration values between different models and ground stations (left: PMRS $PM_{2.5}$, right: RF-PMRS $PM_{2.5}$). Note that the top red of the RMSE legend indicates RMSE values equal to or greater than 60 μg/m³.

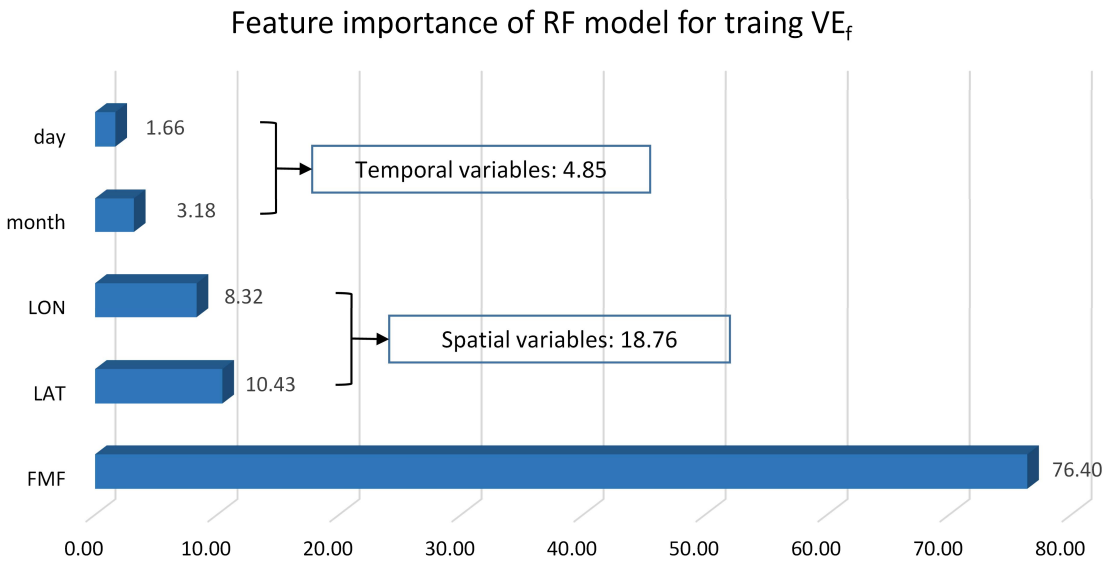

Fig. R6. The predictor importance results (normalized) of the RF model for training VE_f.

Table R3. The particulate matter density and volume equivalent diameter d_p from literature.

Location	Density (g/cm^3)	d_p (μm)	Sampling time	References
Texas, US	1.85 ± 0.14	2.04	Jul.–Oct., 1999	Hand & Kreidenweis (2002)
Atlanta, US	1.54–1.77	2.16–2.07	Aug., 1999	McMurry et al. (2002)
Mainz, Germany	1.83 ± 0.4–3.50 ± 1.0	2.04–1.65	1996–1974	Hanel & Thudium (1977)
Jungfaujoch, Switzerland	2.87 ± 0.59	1.76	4 Mar.–5 Jun., 1972	
Mace Head, Ireland	1.93 ± 0.04	2.01	13–30 Nov., 1971	
Deuselbach, Germany	1.82 ± 0.04–3.32 ± 0.27	2.05–1.68	Oct.–Nov., 1975	
Shanghai, China	1.7	2.09	Apr.–Aug., 2005	Gao et al. (2007)
Beijing, China	1.5	2.18		
Average	2.09 ± 0.53	1.99 ± 0.15	–	–

8. Line 173-178: this paragraph should be rephrased. It could be more clear if you firstly present the problem of the original FMF dataset and then introduce the benefit of the phy-DL FMF dataset, including the details of this dataset.

Response: We gratefully thank the reviewer for his constructive suggestions. Detailed descriptions of Phy-DL FMF dataset have been supplemented in the manuscript, including its generation process and performance (Page 5 Lines 138-146, Page 6 Lines 147-148, Page 10 Lines 243-245, Page 30 Lines 726-729). Also, for the convenience

of statements, we have adjusted the order of data introduction (Page 4 Line 107 to Page 6 Line 150, Section 2.1 to section 2.4). Other detailed modifications are: 1) abbreviation explanation, Page 5 Line 124; 2) definition of meteorological data types, Page 6 Lines 151-152. A description of the overall restructuring of the article can be seen in General Comments: Response.

9. Line 177: what does the "spatiotemporal continuity" stand for?

Response: Thank the reviewer for this comment. The phrase **"spatiotemporal continuity"** means wide space-time coverage with little missing data. To clearly express the meaning of this phrase, we have replaced it in the text. Please see Page 10 Lines 243-244. Hope it is clear now.

10. Line 192: it is confused here why don't use eq 3-5 to directly calculate the PM2.5?

Response: We gratefully thank the reviewer for this comment. Equations 5 to 7 (originally Equations 3 to 5) are the specific process of calculating VE_f, which calculates the result at the site points based on the AERONET ground data. Despite the stability, there are still some limitations to the spatiotemporal analysis due to the sparse and uneven distribution of the ground sites. **Our study is aimed to reconstruct the $PM_{2.5}$ data with wider coverage, that is, from point scale to surface scale (S-FMF to Phy-DL FMF)**. Therefore, we use ML to establish a model and optimize the simulation results of VE_f. The specific reasons can be summarized in the following two aspects. **1) Space coverage**: ground FMF (S-FMF) is point data with a sparse distribution. Phy-DL FMF is areal data, which is applicable to a wider range of experiments. **2) Accuracy:** S-FMF is a true value with high accuracy. Experiments show that the precision of Phy-DL FMF is equivalent to that of S-FMF (Yan et al., 2022). In order to expand the applicability of VE_f model, this study uses the ground values as the truth to train the RF model and uses the planar FMF (Phy DL FMF) to estimate $PM_{2.5}$ in a wider range and more locations.

References:

Yan, X., Zang, Z., Li, Z., Luo, N., Zuo, C., Jiang, Y., Li, D., Guo, Y., Zhao, W., Shi, W., and Cribb, M.: A global land aerosol fine-mode fraction dataset (2001--2020) retrieved from MODIS using hybrid physical and deep learning approaches, *Earth Syst. Sci. Data*, 14, 1193-1213, https://doi.org/10.5194/essd-14-1193-2022, 2022.

11. The title of step 1-4 should be more clear.

Response: We gratefully thank the reviewer for this suggestion. In general, the steps for obtaining $PM_{2.5}$ are "VE_f truth value calculation - RF model construction - VE_f value estimation - $PM_{2.5}$ calculation by formula". Steps 1-4 here mainly show the process of RF modeling (Page 10 Line 252 to Page 12 Line 290).

Step 1: Calculate the VE_f truth value as the output of model training. We generalize it to "VE_f calculation".

Step 2: Select related variables as the input of the RF model. We generalize it to "VE_f-related variables selection".

Step 3: Train the RF model. We generalize it to "RF model establishment".

Step 4: Estimate results through the RF model and validate the accuracy. We generalize it to "Accuracy validation".

At the same time, Fig. 3 also clearly depicts the flowchart of the process (Page 11 Line 260, Fig. 3). In summary, we believe that the title clearly states the main purpose of each step, and the text also describes in detail how each step is completed. Hopefully, our explanation will make Steps 1-4 here clearer to the reviewer. If there is still something unclear to the reviewer, please feel free to let us know.

12. Line 212: use the correct reference for RF.

Response: Thank the reviewer for pointing out this problem very much. In the original manuscript, we cite Yang's paper, which used an optimized RF model to estimate $PM_{2.5}$ concentrations. It is an article on applying RF for remote sensing estimation. Based on the comment from the reviewer, we have added an illustrative article on the RF model as a reference (Page 12 Lines 279-280; Pages 29-30 Lines 706-709).

13. Line 214: The introduce of RF is not clear. RF doesn't learn in the random manner.

Response: Thank the reviewer for the comment. Random forest makes predictions by integrating the results of multiple decision trees, which makes up for the shortcomings of weak robustness of a single decision tree. **Its randomness is mainly reflected in the following two aspects.** 1) When generating training subsets. Let the original training set have N samples, and for each tree, n training samples are extracted randomly as the training set of the tree, repeated K times, and the training subset of K groups is generated. 2) When building an optimal model (decision tree). Let the total number of features be M, and randomly select m features in each sub-dataset for node splitting, so that the decision tree grows and the optimal learning model is constructed. Fig. R7 below shows the specific process of RF.

In the manuscript, we simplify the above statement. Based on the suggestion of the reviewer, we have made some changes to make it clearer how RF works (Page 12 Lines 277-283). Meanwhile, Fig. 2 (Page 9 Line 219) and step 3 in Fig. 3 (Page 11 Line 260) depict the flowchart of the RF algorithm. We believe that the explanations are sufficient to give readers an understanding of how RF works, and the specific mechanism can be found in the cited references (Svetnik et al., 2003; Belgiu and Drăguţ, 2016).

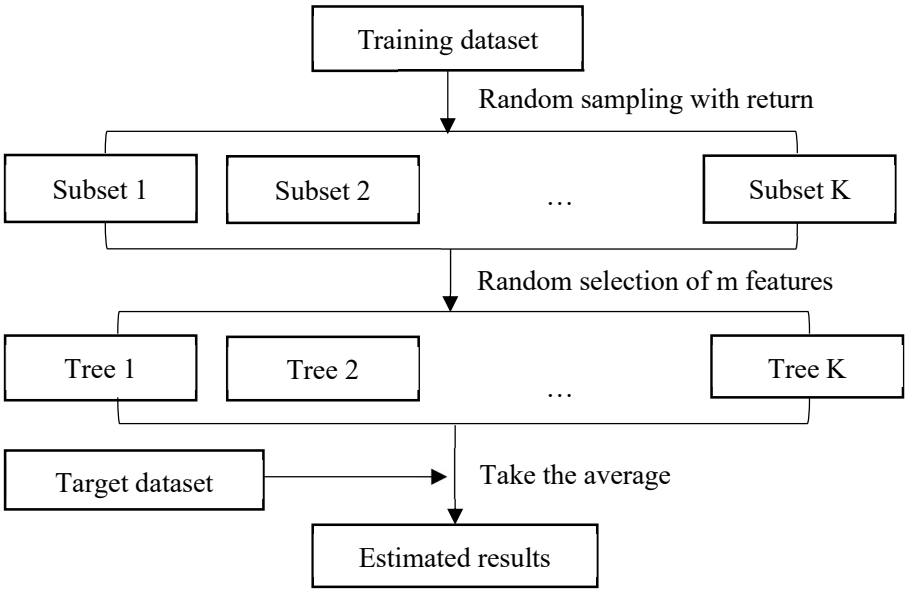

Fig. R7. RF algorithm flow chart. K is the number of training subsets, and m is the feature number for node splitting.

14. Line 221: briefly introduce the cross-validation and isolated-validation.

Response: Thank the reviewer for the comment. To make the structure of the article clear, an introduction to cross-validation and isolated-validation is placed in Appendix A1 (Page 22 Lines 511-518). And there is a prompt "see Appendix A1" in the paper (Page 12 Line 290). For clarity, we have added some leading words in the manuscript (Page 12 Lines 289-290). Hope it is clear now.

15. Line 235: add the reference for MCD19A2.

Response: Thank the reviewer for the comment. We have added the reference to MCD19A2 at the place mentioned by the reviewer (Page 5 Line 125). And the specific download website is listed in the Code and data availability section (Pages 24-25, Lines 551-553).

16. Line 287: how to align the location for different source dataset?

Response: Thank the reviewer for the comment. When calculating $PM_{2.5}$ results, data from different sources are spatially matched with reference to the latitude and longitude of the ground $PM_{2.5}$ values (as the verified true values). The purpose of this matching is to validate the $PM_{2.5}$ accuracy by comparing the estimation results with the ground values. It is important to note that the experimental data were not upsampled or downsampled (i.e., their original resolutions were maintained). Therefore, the row and column numbers corresponding to the site location (latitude and longitude) are found in different source datasets, and the value of the corresponding grid is extracted. Finally, the matching of the data pairs is complete. **Fig. R8 below shows a schematic diagram of location alignment**. Due to changes in the structure of the article, the introduction to data matching has been moved to the Methods section (Page 12 Lines 293-296). We hope it is clear now.

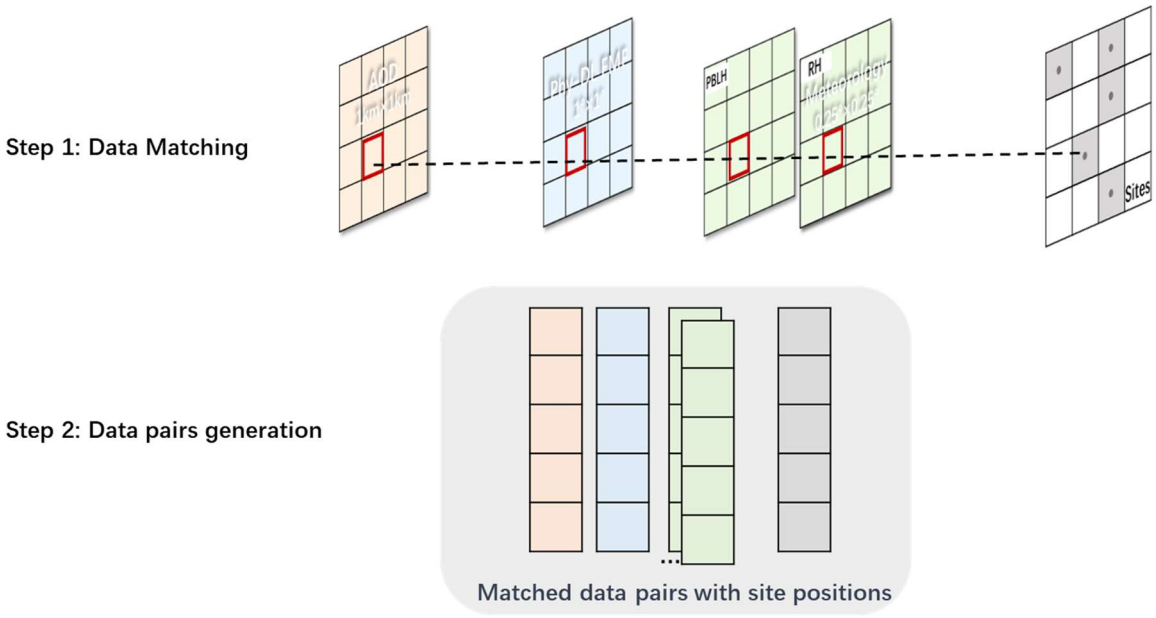

Fig. R8. Spatial matching diagram of data from different sources. The orange and blue grids represent MCD19A2 AOD and Phy-DL FMF datasets, respectively. The green grids indicate meteorological data while the $PM_{2.5}$ site observations are shown in white grids.

17. Line 289: does the empirical value introduce uncertainty?

Response: Thank the reviewer for the comment. As the empirical value in the semi-physical empirical model, the $\rho_{f,dry}$ value is often obtained by field measurement and induction. Table R3 above shows some of the $\rho_{f,dry}$ empirical values from previous studies. In our study, we select the 1.5 g/cm^3 as the $\rho_{f,dry}$ value for the NC region. As can be seen from Table R3 above, there are certain variations in the empirical values in different regions, and there will be errors (uncertainty) between the values in Beijing and other places in the NC region. However, our experimental area is not large, and we can use 1.5 g/cm^3 to represent the $\rho_{f,dry}$ of the whole region, which has been applied in previous articles (Zhang and Li, 2015; Li et al., 2016). Thanks to the reviewer for the reminder again. In future experiments in other regions, we will note the uncertainty of this empirical value.

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.

18. Table 2: which sites the performance statistics are for? Do you try to compare the performance with the polynomial regression?

Response: Thank the reviewer for the comment. Table 3 (originally Table 2) shows the performance statistics of the RF model for training VE_f. **Based on 9 AERONET sites around the world**, the RF model was trained and its performance was validated. Please refer to Table 1 (Page 4 Lines 119-121) for the specific time and site locations. Also, an introduction to cross-validation and isolated-validation is placed in Appendix A1 (Page 22 Lines 511-518).

After building the VE_F model, input the variables of the same year to get the VE_f of this year, and then deduce $PM_{2.5}$ through formula (8) (Page 8 Line 210). **Then, We compare the $PM_{2.5}$ results of the proposed method (RF-PMRS) with the PMRS method in a comprehensive way (Section 4.2-4.3, Page 13 Line 323 to Page 20 Line 424). The PMRS method is based on a simple polynomial to obtain VE_f, and then calculate the $PM_{2.5}$ value.** Because $PM_{2.5}$ sites are widely distributed and the $PM_{2.5}$ value is the final target value, the comparison with polynomial regression is based on the $PM_{2.5}$ value.

And we feel sorry for the unclear description of the main experiments. Therefore, we have added an experiment information table (Table. R4 below) in Section 4. It includes the validation object, study region, study period, and temporal scale of three main experiments. Some descriptive statements are also adjusted or added in appropriate places. The specific modifications to this comment include: (1) Page 12 Lines 306-307, Page 13 Line 308: added Table 2 and related introduction; (2) Page 12 Lines 288-289: abbreviation addition corresponding to Table 2; (3) Page 6 Lines 164-168: explicit statement to the experimental area and period; (4) Page 12 Lines 293-298: adjusted from the original data part to the method introduction part; (5) Page 17 Lines 381-382: added statement of the experiment information. And we hope the experiments are presented in a clear way now.

Table R4. A brief information summary of the experiments conducted in our study.

Experiment	Object	Region	Period	Time scale
Model performance for training VE_f	VE_f	Global scale (Nine AERONET sites)	CV: Training period in Table 1 IV: Isolated-validation period in Table 1 (See Appendix A1)	Daily
Accuracy evaluation of PMRS/RF-PMRS	$PM_{2.5}$	Two AERONET Sites: Beijing, Beijing-CAMS	2017	Daily
Generalization performance of RF-PMRS	$PM_{2.5}$	North China region	2017	Daily

19. Line 309: why do you choose these only two sites?

Response: Thank the reviewer for the comment. The validation experiment is regional (2 AERONET sites in North China) because the ρ empirical values and $PM_{2.5}$ truth values in other regions are insufficient, and more other research results are needed. The details are described below. **1) insufficient $\rho_{f,dry}$ value** (Page 8 Line 210, Eq 8). As the empirical value in the semi-physical empirical model, the $\rho_{f,dry}$ value is often obtained by field measurement and induction. Table R3 above shows some of the $\rho_{f,dry}$ empirical values from previous studies. The insufficient $\rho_{f,dry}$ values hinder the derivation of $PM_{2.5}$ in other regions; **2) $PM_{2.5}$ public value of ground sites around the world is limited.** Accurate and sufficient in-situ $PM_{2.5}$ values are the basic guarantee for the verification of estimated $PM_{2.5}$ results.

In the future, with the expansion and disclosure of relevant experimental data, we will verify our proposed method in a broader range and continuously optimize it from all aspects.

20. Line 367: could you explain why the difference is not significant in terms of R, but there is a big difference for RMSE and MAE?

Response: Thank the reviewer for the comment. In fact, we are also puzzled by the results of this experiment on model performance. R represents the correlation between the estimated data ($PM_{2.5}$ results of PMRS/RF-PMRS) and the truth value (ground data),

but it does not provide a comprehensive representation of the accuracy of the estimated data. **Based on the R and RMSE calculation formulas, R depends on the trend of the two values and RMSE depends on the absolute difference between the two.** As shown in the scatterplot (Fig. 7, Page 18 Line 406), the RF-PMRS $PM_{2.5}$ values are distributed around the 1:1 reference line evenly, with lower RMSE and a slightly higher R. The overall results of PMRS are overestimated, and R is similar to the RF-PMRS results, probably because its variation trend is similar to the ground values. Taking the y=sinx and y=2sinx functions as examples (Fig. R9 below), the correlation between the two is very strong (R=1), but the RMSE is as high as 0.77. The trend of the two corresponding datasets varies similarly (PMRS results and ground values), so R may be large. **However, from the perspective of scatterplot distribution and overall performance, the RF-PMRS method successfully solves the problem of partially estimating abnormal $PM_{2.5}$ concentrations (RMSE has dropped significantly) and still greatly improves the accuracy of $PM_{2.5}$.** As the cause about R difference cannot be determined, it is not detailed in the text. In future research, we will continue to explore the data pattern in more experiments and further explain this phenomenon.

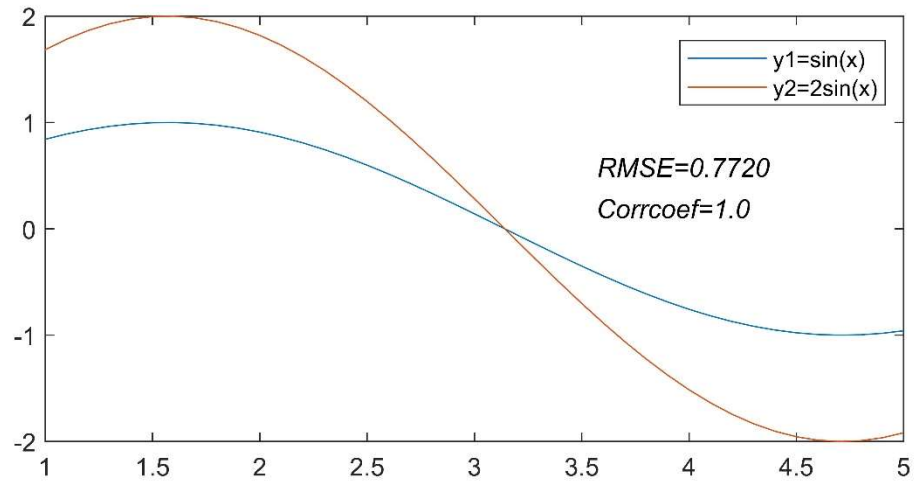

Fig. R9. Illustrative diagrams of y=sinx (the blue line) and y=2sinx (the red line).

21. Line 383: you mentioned the underestimation could be caused by the high-value points. It seems that the RF model is overfitting. Do you try to reduce the

overfitting?

Response: Thank the reviewer for the comment. The reviewer may have misunderstood the meaning we expressed in the manuscript and we feel sorry about it. We found some underestimation of $PM_{2.5}$ estimates in the high-value $PM_{2.5}$ region (Page 17 Lines 402-403 (originally Line 383)). And in many previous studies (Ma et al., 2014; Li et al., 2017), there is also an underestimation of high $PM_{2.5}$ values. In our study, this phenomenon is attributed to a small number of high-value fitted data (only 1319 out of 28305), rather than the high-value points as the reviewer claims.

In the experiments, we use two methods to verify the performance of the RF model. Among them, 10-fold cross validation (CV) is to validate the internal accuracy of the model (recorded during training), and **isolated-validation** (IV) is to validate the temporal generalization of the model, that is, the **external accuracy** of the model. **Table R5 below (Table 3, Page 13 Line 320) shows the statistical results in CV and IV experiments are similar, indicating that the RF model has no obvious overfitting phenomenon.**

Also, **when training the RF model, we have considered the overfitting phenomenon and selected the optimal combination of parameters.** The four parameters of RF are adjusted. Fig. R10 represents the parameter tuning process. It shows the correlation coefficient r changes with (a) the number of trees, (b) maximum depth, (c) maximum number of features when splitting, (d) minimum number of split samples. By observing the impact of parameter changes on model performance, the four parameters of RF are adjusted to 60, 10, 2, and 8 to prevent overfitting. It can ensure high accuracy while improving training efficiency. Please see Appendix A3. Parameter adjustments of the RF model, Page 23 Lines 531-537 for details.

Table R5. Performance statistics of the RF model for training VE_f. N represents the number of data, and VE_f has no unit.

	R	RMSE	RPE	MAE	N
Cross-validation（CV）	0.974	0.076	32.9%	0.034	6463
Isolated-validation（IV）	0.975	0.067	29.8%	0.037	814

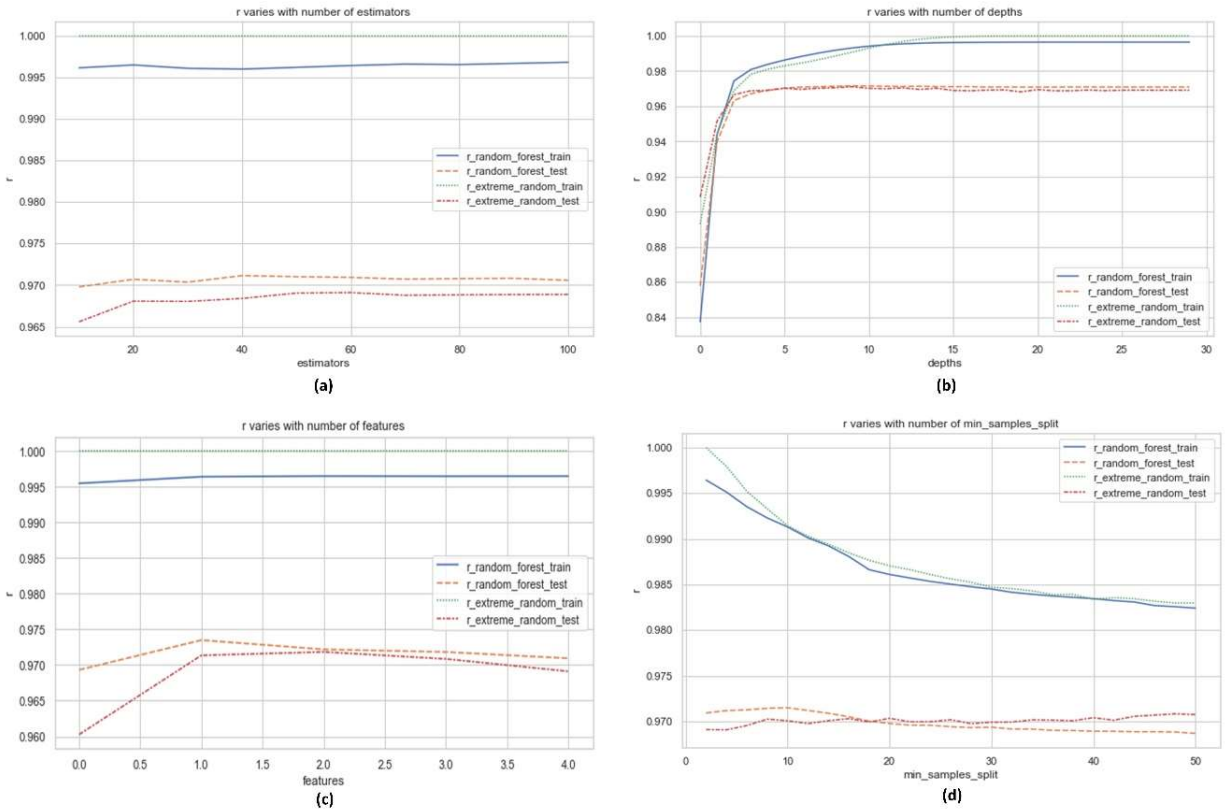

Fig. R10. The experiment results of four parameters adjustments of RF and ERT. (a)-(d) represent the correlation coefficient r changes with n_estimators, max_depth, max_features, min_sample_split, respectively. The blue and orange lines denote the results of the training dataset and test dataset of RF, while the green and red lines denote the results of two datasets of ERT.

References:

Ma, Z., Hu, X., Huang, L., Bi, J., and Liu, Y.: Estimating ground-Level PM2.5 in China using satellite remote sensing, *Environ. Sci. Technol.*, 48, 7436-7444, https://doi.org/10.1021/es5009399, 2014.

Li, T., Shen, H., Zeng, C., Yuan, Q., and Zhang, L.: Point-surface fusion of station measurements and satellite observations for mapping PM2.5 distribution in China: Methods and assessment, *Atmospheric Environ.*, 152, 477-489, https://doi.org/10.1016/j.atmosenv.2017.01.004, 2017.

Thanks again to the reviewers for their patience and constructive comments. We know that our experiments still have a lot to improve. With the expansion and disclosure of relevant experimental data, we will optimize it from all aspects in the future.

Meanwhile, thank the reviewers for the recognition of our work. We will further explore the combination of atmospheric mechanism and machine learning on the $PM_{2.5}$ retrieval methods.