EGUsphere, referee comment RC2
https://doi.org/10.5194/egusphere-2022-946-RC2, 2022

Comment on egusphere-2022-946

Anonymous Referee #2

Referee comment on "An optimized semi-empirical physical approach for satellite-based $PM_{2.5}$ retrieval: embedding machine learning to simulate complex physical parameters" by Caiyi Jin et al., EGUsphere, https://doi.org/10.5194/egusphere-2022-946-RC2, 2022

General comments:

Jin et al. presented a semi-empirical physical approach for $PM_{2.5}$ retrieval from AOD and the validation results in China. The authors embedded the random forest (RF) model into the physical $PM_{2.5}$ remote sensing approach (PMRS) with a high-quality fine mode fraction product to estimate surface $PM_{2.5}$ concentration. Compared to the PMRS method, the machine learning embedded approach (PMRS-RF) showed better performance in $PM_{2.5}$ estimation, with lower biases and RMSE. The methods of the combination of machine learning and semi-empirical physical approach could be of high interest to the community, and they would be useful for $PM_{2.5}$ retrieval from satellite observations, particularly for regions with sparse ground measurements. The paper is well-written, and the ideas are presented clearly. However, the structure and the experiment designs are a bit challenging to follow. The uncertainties of the machine learning approach (e.g., out-of-sample performance) and spatial distributions of biases can be discussed more.

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.

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 VE_f validation, 2) a hold-out test of 2017 for $PM_{2.5}$ validation at Beijing and Beijing-CAMS sites, and 3) a generalization test for $PM_{2.5}$ validation within North China (not sure for which year). In addition, is it correct that AERONET AOD is used for calculating $PM_{2.5}$ 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.

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 VE_f 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 VE_f and surface $PM_{2.5}$ should be consistent.

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.

Specific comments:

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

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.

3. Page 9, line 216: How does the difference between station FMF and Phy-DL FMF influence surface $PM_{2.5}$ estimation?

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

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?

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.

7. Page 11, line 270: Is AERONET AOD used for calculating $PM_{2.5}$ 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 $PM_{2.5}$ 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.

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

(use different colors and shapes).

9. Page 12, line 288: The experiment period is a little bit confusing. The surface $PM_{2.5}$ validation is conducted for 2017, while the VE_f validation is based on the 10-fold CV and different hold-out periods. Also, please justify the test selection for 2017.

10. Page 13, lines 308-309: Was the VE_f based on RF derived from the hold-out experiment? Ideally, the VE_f 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).

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

12. Page 16, line 361: In the RF model estimating VE_f, 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)?

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 VE_f).

14. Page 17, lines 382-385: What do the high-value points mean? The high values of $PM_{2.5}$ concentration or VE_f? I guess the underestimation of VEf would lead to the underestimation of $PM_{2.5}$ in RF-RMRS.

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

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

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 $PM_{2.5}$ estimation associated with different data sources), 2) spatial or temporal distribution of biases for the two methods (see comment #13).

Technical comments:

1. Page 2, line 37 & Page 13, line 314: The word "trends" is misused. Fig. 4 displays the "time series" of $PM_{2.5}$ 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.

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