Reviewer: 3

I noticed that the same authors published a very similar paper in ES&T, https://pubs.acs.org/doi/10.1021/acs.est.9b03258. The only difference is between PM2.5 and PM1.0. However, the ACP paper needs originality.

Response: We would say that the two papers are similar but also differ in many regards that are grossly summarized as follows:

(1) They deal with different pollution quantities: PM_1 and $PM_{2.5}$, whose emission sources, formation and transport mechanisms, and health impact are all different. As such, both the figures and text of the manuscripts differ considerably. Their ratio varies greatly, ranging from less than 0.5 to greater than 0.9 at both spatial and temporal scales, especially in heavily polluted regions due to different influential factors (Wei et al., 2019b). The two papers may thus be regarded as a series of companion studies that do not undermine their respective scientific originality. The reviewer is invited to compare them to see how different they are.

(2) The estimation approaches used to derive PM_1 and $PM_{2.5}$ are similar but also differ in several aspects. While the same kind of machine learning method, namely, the space-time extra-trees (STET) model, is used for retrieving PM_1 and $PM_{2.5}$, there are numerous differences in their applications. For retrieving $PM_{2.5}$, we have 1) used different input parameters by adding the aerosol precursor gases (SO_2, CO, NO_x, VOC, fine-size dust) from pollutant emission inventories; 2) corrected the satellite retrievals of AOD with reference to ground-based measurements; 3) modified the feature selection approach using the Gini index; and 4) improved the determination of spatiotemporal information. We have clearly described these differences in Section 3 as well as in the introduction of the revised manuscript.

Moreover, the manuscript has some fatal defects, (1) It does not work well with high pollution events, which is paid more attention.

Response: Like similar studies, ours suffers from a limitation of having relatively large errors under severely polluted conditions whose causes are further explained, per the reviewer's suggestion. This is a common problem reported in many previous studies. We have added the following text to the revised manuscript (Section 5.1): "We find that all traditional statistical regression models, and machine and deep approaches reported in previous studies underestimated $PM_{2.5}$ concentrations under highly polluted conditions with poor regressions (i.e., slope < 0.9, and intercept > 6 $\mu g/m^3$) between measurements and retrievals of $PM_{2.5}$ in China, a common problem. Potential causes are: 1) There are large estimation errors in AOD retrievals under severe pollution conditions in China (Wei et al., 2019c). This is further rooted to the fundamental limitations of satellite-based AOD retrievals, i.e., the non-linear to reflectance and the high sensitivity of the single-scattering albedo (Z. Li et al., 2009); 2) High AOD does not correspond to high $PM_{2.5}$ concentrations because their ratio is highly variable over space and time, affected by both natural and human factors; 3)

The number of samples for high-pollution cases is small, hindering the ability to train the model."

It appears that all approaches suffer from this inherent limitation, which should thus not be regarded as a "fatal defect" of our study, more importantly, the comparison results suggest that our model can more accurately capture the high pollution events with a larger slope of 0.86 and a smaller intercept of 6.16 $\mu g/m^3$ with reference to other models reported from previous studies (Table 2).

(2) Such method seems falling into a dead cycle, the results were compared by the observations which were used to fit the parameters. I do not think it works with another independent database. Some similar comments were pointed by the other two reviewers.

Response: We do not think the method itself is a "dead cycle", but do make more efforts to enhance the validity and effectiveness of the validation approach. Three independent validation methods are applied, ensuring that the training and validation data are independent, as described in Section 3.5, copied below:

"Different from our previous study, three independent validation methods are performed to verify the model's ability to estimate $PM_{2.5}$ concentrations. The first independent validation method, i.e., the out-of-sample cross-validation (CV) approach, is performed by all data samples using the 10-fold CV procedure (Rodriguez et al., 2010). The data samples are divided into ten subsets randomly, and nine (one) of them are used as training (validation) data. This approach is repeated ten times, and error rates are averaged to obtain the final result. This is a common approach to evaluate the overall accuracy of a machine learning model, widely adopted in most satellite-derived PM studies (T. Li et al., 2017a, b; Ma et al., 2014, 2019; Xiao et al., 2017; He and Huang, 2018; Chen et al., 2019; Wei et al., 2019b; Xue et al., 2019; Yao et al., 2019).

The second independent validation method, i.e., out-of-station CV approach, is similar to the first one but performed using data from the monitoring stations to evaluate the spatial performance of the model. Data samples collected from different spatial points make up the training and testing data, and the relationship between spatial predictors and $PM_{2.5}$ built from the training dataset is then estimated for each testing. The third independent validation approach tests the predictive power of the model. It is performed by applying the model built for one year to predict the $PM_{2.5}$ concentrations for other years, then validating the results against the corresponding ground measurements. This approach ensures that the data samples for model training and validation are completely independent on both spatial and temporal scales."