Reviewer: 2

Using the newly-developed space-time extremely randomized trees (STET) model, this study is aimed at estimating the 1-km-resolution PM2.5 surface concentrations across China. Besides meteorology, land surface conditions and population, a space term and a time term representing the spatial autocorrelation and temporal variation of PM2.5, respectively are also included to derive the PM2.5-AOD relationship. Overall this manuscript is well written, and potentially improves our understanding regarding how to retrieve the PM2.5 concentrations from AOD products and other auxiliary data. However, before I recommend this manuscript to be published, the authors should carefully address and clarify my several comments.

Response: We appreciate the time and effort the reviewer spent on this manuscript and the insightful comments and constructive suggestions. In light of your opinion, we have carefully revised our manuscript. The responses to the questions raised in your report are as follows.

General comments:

1. The relationship between (surface layer) PM2.5 and AOD might largely depend on the compositions (including aerosol water, as Reddington et al. (2019) indicated that aerosol water uptake and hygroscopic growth would also impact the AOD), vertical profile and size distribution of PM2.5. Thus, I find that some results in Figure 2 are confusing, and needs further analysis and clarification: 1) In Section 3.2, it is unclear that how the importance scores of all selected independent variables and spatiotemporal information to PM2.5 estimates for the STET model are calculated.

Response: We agree with you and we have mentioned this in the manuscript and cited the references. In addition, the importance score is described in more detail in the revised manuscript. The importance score of each independent variable used to estimate $PM_{2.5}$ is calculated based on the Gini index (GI). We have added a more detailed description in Section 3.3 of the revised manuscript as follows:

" … the GI index is selected to calculate the importance score of each independent variable on $PM_{2.5}$ estimates because of its higher accuracy and stability as a variable importance measure, especially for continuous variables with low signal-to-noise ratios (Jiang et al., 2009; Calle and Urrea, 2011), expressed as

$$GI(\omega) = \sum_{n=1}^{N} \omega_n (1 - \omega_n) = 1 - \sum_{n=1}^{N} \omega_n^2 \ , \qquad (2)$$

where *n* represents the number of the categories (N = 1, …, *n*), and ω_n represents the sample weight of each category. The importance of one feature (X_j) on node *m* is that the GI changes before and after node *m* branching:

$$\Delta GI_{jm} = GI_m - GI_l - GI_r \ , (3)$$

where GI_l and GI_r represent the GI of two new nodes after branching. The importance score for one feature (IS_j) in then the extra-trees with *k* trees (*i* = 1, …, *k*), calculated as

$$IS_j = \sum_{i=1}^{k} \Delta GI_{ij} = \sum_{i=1}^{k} \sum\nolimits_{m \in M} \Delta GI_{jm} \quad ,(4)$$

where ΔGI_{ij} represents the importance of X_i in the i^{th} tree when the node of feature X_i in decision tree j belongs to set M. Finally, an additional normalization approach is performed to all obtained importance scores for each feature."

2) Why RH turns out to be a much less important parameter, and it has an importance score that is only slightly higher than those negligible parameters do. RH is an important factor determining the aerosol compositions and water uptake, and recent air quality studies (e.g., Sun et al., 2014; Zheng et al., 2015) showed that high RH conditions facilitate rapid production of secondary PM.

Response: We agree with you that RH should have a large influence on the production of $PM_{2.5}$. However, a potential reason why RH turns out to be less important is that high RH conditions are potentially highly related to cloudy/rainy days, especially in summer, when there is a high probability of missing AOD retrievals. In addition, this importance score only represents the importance of features in splitting during the extra-tree construction, not the contribution of features to $PM_{2.5}$ in physical mechanisms. We have clarified these in in Section 3.3 of the revised manuscript as follows:

"The $PM_{2.5}$-AOD relationship might largely depend on the compositions (e.g., aerosol water, Reddington et al., 2019; Jin et al., 2020). High RH conditions and precipitation should have large influences on the production and removal of $PM_{2.5}$ (Sun et al., 2014; Zheng et al., 2015). However, RH and PRE turn to be less important with overall low importance scores in the STET model, which may be attributed to the fact that aerosol retrieval algorithms only work under cloud-free conditions when RH is relatively low. More importantly, the calculated importance score only represents the importance of features in splitting during the extra-tree construction, not the contribution of features to $PM_{2.5}$ in physical mechanisms."

3) Furthermore, the parameter of precipitation could significantly impact the removal of PM, but is negligible in the STET model. Both RH and precipitation are associated with cloud, and what is the uncertainty for the predicted $PM_{2.5}$-AOD relationship caused by the treatment of AOD data on cloudy dates?

Response: We agree with you that the precipitation should have a large influence on the removal of $PM_{2.5}$. However, it shows the lowest important score and is negligible because remote sensing aerosol retrieval algorithms cannot work when clouds are present, so there are no AOD retrievals on rainy days. Similarly, the importance score only refers to the importance of features in splitting during the extra-tree construction and not the contribution of features to $PM_{2.5}$ in physical mechanisms. We have added this description to Section 3.3 of the revised manuscript (See above comment):

2. The authors declared that STET model exhibited a strong predictive power and could be used to predict the historical PM2.5 records in the Abstract Section (in Line

39). This conclusion could be inappropriate as the authors only tested the year of 2017. Emissions were not expected to change greatly between 2017 and 2018. Actually, I doubt the applicability for the STET model. The space and time terms seem confusing to me, and the former term is represented by the geographical difference between two pixels, while the latter term is represented by the difference for a given pixel on different days in a year. I think they might be "residual terms" to implicitly resolve the "unknown parts" unexplained by other independent parameters. I mean, the authors need more independent parameters that could explicitly explain the PM2.5 compositions, vertical profile and size distribution. Why not emissions for different precursors (e.g., SO2, NOx and VOCs) as well as fine size dust are included as independent parameters?

Response: $PM_{2.5}$ changes dramatically in space, and varies over time, showing significant spatiotemporal heterogeneities and patterns. Thus, introducing the spatial and temporal terms account for the spatiotemporal autocorrelations of $PM_{2.5}$ between different points for each day and between consecutive time series at the same place. In addition, per your suggestion, we have included emissions for main precursors and fine-sized dust as independent parameters to enhance the STET model and improve the estimation of $PM_{2.5}$ in Section 3 of the revised manuscript as follows:

"Different with our previous study (Wei et al., 2019b), pollutant emissions for different precursors (including SO_2, NO_x, CO, and volatile organic compounds) and fine-sized dust are also employed to help explicitly explain the $PM_{2.5}$ composition, collected from a multi-resolution emission inventory for China (Zhang et al., 2007)."

In addition, we have updated and re-described in detail all the results in Sections 3 and 4. Results show that the model performance is overall improved.

3. Equation 1 is confusing. What is the R2 for each linear regression? Are these two linear regressions consistent with each other? Why not to average the Terra and Aqua data directly?

Response: We have replaced the regression method with the average approach per your suggestion and clarified this in the revised manuscript as follows:

"Terra and Aqua MAIAC AOD retrievals are thus averaged for each pixel on each day to form a new dataset and enlarge the spatial coverage."

4. The description for the STET method in Section 3 is not readily to understand. Please add clarification (better to include a schematic) so that ACP readers with less experiences in machine learning could generally understand the fundamentals of the STET method.

Response: We have added clarification and a schematic of the STET model in Section 3.4 of the revised manuscript as follows:

"For the enhanced STET model, all the selected independent variables are first input into the ERT model, and the random splits (S, a_i) are established according to the whole of training data samples; then totally different *K* attributes are selected randomly from all attributes according to spatial and temporal differences; then *K* random splits are generated (s_1, …, s_k), and a split (s*) is selected by calculating the score measure function, i.e., Score(s*, S); then split node (S) is completely randomly generated to establish an extra tree; last the extra tree ensemble is built using the similarity method. Detailed information on ERT algorithm can be found in Geurts et al. (2006). Figure 4 illustrates the schematic of the enhanced STET model."

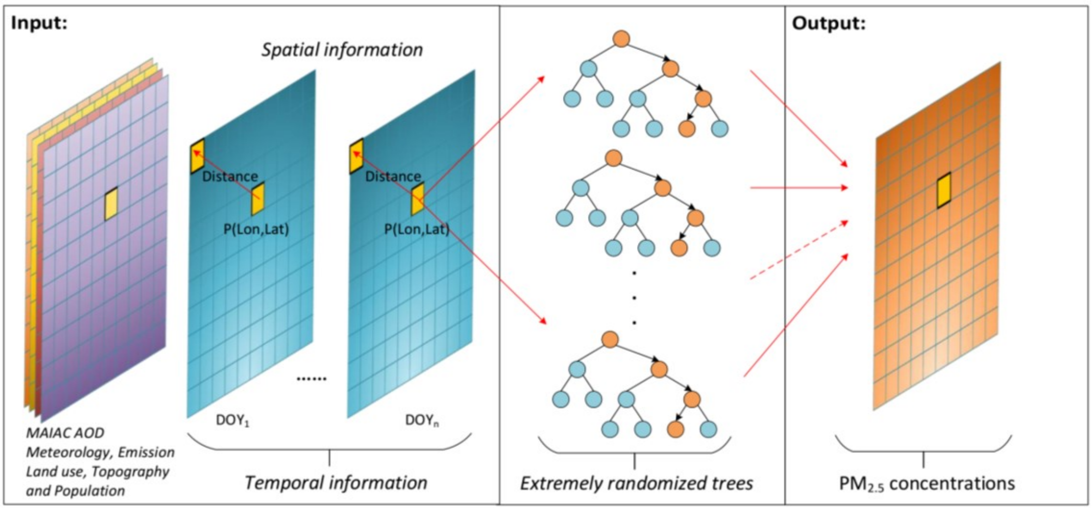

Figure 4. Schematic of the enhanced STET model developed in our study.

5. In Figure 7, what is surprising is that I see a good positive correlation pattern between R and RMSE. Generally, a good model performance is associated with a high R and a low RMSE against observations. Please check and clarify.

Response: We have verified the numbers, which are correct. Mathematically speaking, R^2 and RMSE are two independent measures of a correlation between two variables whose correlation depends on the slope of the regression between the two, higher for a regression slope closer to unity. Since the slope varies from site to site, they may not show the same spatial patterns. We have taken a closer look at the spatial patterns of these quantities and added the following text attempting to give a physical explanation (section 4.2.1 of the revised manuscript):

"In general, high R^2 with overall large RMSE but small MRE values are observed at the beginning and end of the year (in winter). This is because $PM_{2.5}$ concentrations vary more and are always high due to the greater amount of pollutant emissions caused by heating or frequent dust storms. By contrast, lower R^2 with overall small RMSE and large MRE values are observed in the middle of the year (in summer) because air pollution levels are lower."

Specific comments:

1. Line 48, the "evenly dispersed" is confusing, and is conflict with the "PM2.5 shows great spatial and temporal heterogeneities" in Line 80.

Response: Corrected.

2. Line 175, better replace "differences" by variation.

Response: Corrected.

3. Line 227, typos: Figure 2 or Figure 3?
Response: Corrected.

4. Line 247, what is definition for MAE and MRE?
Response: We have provided definitions of these evaluation indicators in the revised manuscript.

5. Figure 9, typos: the year is 2018 or 2017? Also please add the season labels for each plot.
Response: Corrected.

Reddington, C. L., Morgan, W. T., Darbyshire, E., Brito, J., Coe, H., Artaxo, P., Scott, C. E., Marsham, J., and Spracklen, D. V.: Biomass burning aerosol over the Amazon: analysis of aircraft, surface and satellite observations using a global aerosol model, Atmos. Chem. Phys., 19, 9125-9152, 10.5194/acp-19-9125-2019, 2019.

Sun, Y., Jiang, Q., Wang, Z., Fu, P., Li, J., Yang, T., and Yin, Y.: Investigation of the sources and evolution processes of severe haze pollution in Beijing in January 2013, Journal of Geophysical Research: Atmospheres, 119, 4380-4398, 2014.

Zheng, G., Duan, F., Su, H., Ma, Y., Cheng, Y., Zheng, B., Zhang, Q., Huang, T., Kimoto, T., and Chang, D.: Exploring the severe winter haze in Beijing: the impact of synoptic weather, regional transport and heterogeneous reactions, Atmos. Chem. Phys., 15, 2969-2983, 2015.