Comment on essd-2021-296

Anonymous Referee #2

Referee comment on "Full-coverage 1 km daily ambient PM$_{2.5}$ and O$_3$ concentrations of China in 2005–2017 based on a multi-variable random forest model" by Runmei Ma et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-296-RC2, 2021

General comments

This study aimed to estimate the 1-km resolution PM$_{2.5}$ and O$_3$ concentrations with a random forest model covering China. Spatiotemporal variations in PM$_{2.5}$ and O$_3$ distributions during 2005-2017 were further characterized. The modelling workflow is reasonable; however, the method section lacks some critical information and the source data can hardly support the spatiotemporal resolution of the predictions. The result presentation could be improved and some discussion on data uncertainties are needed.

Specific comments

- In the abstract section, the author presented the model fitting R2 from "sample-based division method" in line 31-33 but it is not clear if the R2 is from test data or from cross validation. Additionally, the model fitting R2 normally means the R2 during model fitting stage with the model fitting dataset. Please clarify the R2 here and throughout the manuscript.
- Line 55-56: Why did the high pollution events and unsatisfactory pollution control bring difficulties to capture pollution distribution? Are the pollution level higher than the monitor measurement range?
- Figure 1: The number of air quality monitors in China kept increasing during 2013-2015 and there are much more monitors in 2017 compared to 2013. This figure shows the 2017 monitors but describes it as average measurement concentrations during 2013-2017. How to deal with the monitors that are not available in 2013-2014?
- Line 81-82: the reference Wei et al. 2021, which was cited in the result section,

constructed the datasets with longer time series and 1-km resolution. What is the advantage of this work compared to previous works?

- Line 95-96: The resolution is 1-km but the author did not provide the projection information.
- Figure 3 and Figure 4: The PM2.5 ranges of the yearly plots are much smaller than those of the daily plots. Please shrink the x- and y-axis.
- Line 105-107: Why did the author use Aqua AOD but not Terra AOD for PM2.5 modeling? What percent of satellite AOD are missing?
- How did the author deal with the missing satellite retrievals to get full-coverage daily dataset? And what is the performance of the gap-filling method?
- The 2*2.5 degree GEOS-Chem simulations were used for 1-km resolution O3 modeling. Since the spatial resolution of GEOS-Chem simulation was too coarse compared to the O3 prediction and it ranked the second most important predictor, I doubt if the prediction could truly reflect O3 variations at local scale. Actually, none of the predictors for O3 modeling provides sufficient spatiotemporal information on variations at the 1-km daily resolution. The design of the O3 model is not solid.
- Line 113-114: The gridded GDP data needs a citation. The GDP data only cover year 2005 and 2010, how did the author assign GDP of other years? Similarly, the road data is of year 2016 but the road map of year 2005 could be considerably different from the road map of year 2016. How did the author consider this issue?
- Figure 4: the slopes of the daily and monthly plots are lower than 0.8, and the slopes of the yearly plots are lower than 0.7, indicating system bias.
- Section 3.2: How to calculate the feature importance and what does the "Value" in Table S4 mean? Why did the Value is in digital number in Table S4-1 but in percentage in Table S4-2. The values of some predictors, e.g. High speed road and Railway, are very low. Why did the author keep them in the model? The author used a whole section to discuss the importance of predictors, thus the Table S4-1 and Table S4-2 could be move to the main text.
- Figure 6: Figure S1 and Figure S2: This study produce 1-km PM2.5 and O3 data products, but only showed the national map and the quality of these figures could not reflect any local scale characteristics. Please zoom in at key regions to give the readers more details.
- Figure 6: Figure S1, and Figure S3: The spatial patterns over the west China are weird.
- Figure S2: Please explain the extremely low O3 concentrations over Tibet on the 2016 map and the weird spatial pattern in West China on the 2017 map.