**Review comments for "Development and validation of satellite-derived surface $NO_2$ estimates using machine learning versus traditional approaches in North America"**

The study focuses on a current important topic of estimating spatial surface $NO_2$ using high resolution satellite instruments such as TROPOMI. Authors use an RF based ML approach to derive surface $NO_2$ and compare these results with the ones derived using the traditional scaling approach, focusing on region of North America. An interesting approach of additionally including model data as training targets to include under-represented remote areas seems to overcome the typical under-prediction issues of ML models in the background regions. This is clearly depicted in this study and is an interesting find that can help with accurate spatial predictions, not only over urban regions where stations are usually located, but also over rural and remote regions, highlighting the importance of integrating ML, satellite observation and model data. The findings of the study can be of interest to the scientific community working on air pollution.

While the manuscript is generally well-structured, scientifically sound, and includes relevant results and figures, there are a few minor issues to be addressed before it can be accepted for publication. In particular, some sentences would benefit from improved clarity and framing. I encourage the authors to carefully revise the manuscript, especially the sections highlighted below, to better communicate their findings. Furthermore, the manuscript is missing a more focused and detailed comparison between the ML-based and traditional approaches, which forms the main objective of the study.

Specific comments

1. Line 55: Please mention the target spatial resolution of your model somewhere in the methodology.
2. Line 64: The changes in versions of TROPOMI through the years results in some differences between the TROPOMI NO2 values for high pollution levels. Could you please explain why you choose this version when a harmonized reprocessed version is available?
3. Line 88: Figure 1 indicates the location of the model parameters as well. Why do you consider the model parameter points over ocean, if you are not predicting there? Will that not impact or introduce some noise into the model as the meteorological conditions over ocean are quite different than over land and hence these datasets can act as noise. Care must be taken to analyze any negative impact of this.
4. Line 97: Citation required for Random Forest
5. Line 100: The logic behind the 0.1 threshold needs a one-liner explanation or a citation

6. Line 123: Please mention a line or two on the data preparation part that includes how and to what resolution the input parameters were gridded or regridded to, since they all are in different spatial resolution? Were there any transformations applied on the datasets to deal with their general skewness? If yes, please mention.
7. In Table 2 caption, please indicate shortly what does the combination daily/mid-day mean. This helps a reader to gauge all the necessary information from the table, without having to search in the main text. Also, of relevance here is the spatial resolution of input parameters
8. Line 130: Why not ERA5-Land for meteorology such as wind and surface pressure which is available at higher spatial resolution than ERA5 and is more relatable to surface variations.
9. Figure 2: The color scale for density is missing and needs to be added
10. Line 185: How is the feature importance derived here? I suppose, in your case it's from the inbuilt feature importance function from random forest? Please mention this in the methods.
11. Line 194: Do you mean "which results in remote areas predicting/computing much higher concentrations than there actually are"? Please check this sentence in terms of phrasing it right
12. Line 205 and 208: Figure 4a and Figure 4b seem to be missing. There is only one figure 4. The black dots mentioned as part of Figure 4a are missing. Please check this result.
13. Line 210: Please check the logical flow of this sentence and correct the same. The result is not communicated clearly. Also, it would be good to specify the direction of bias in certain discussions, i.e, positive or negative. Would recommend modifying the sentence into something like "This example highlights the high positive bias when only station measurements are used as random forest predictor. The RF model is unable to predict low concentrations due to lack of low-concentration scenarios in the training data"
14. Line 210: Does this mean that RF without model inputs, will generally not work well in background locations? How about during low $NO_2$ periods such as summer over urban regions? ML models usually overestimate low pollution levels and the inclusion of GEM-MACH model in your case, perhaps helps with the overall improvement in accuracy during low pollution levels (even over urban regions) and can also be generalized and stated as such in the study. This will be an important contribution and consider mentioning it after verification.
15. Line 220: More comparative discussion between Fig 5a and Fig 5b should be included, as this is also the focus as per the title of the paper.
16. Line 225-226. Please highlight in the map by overlaying correlation values or something similar to indicate the data filling capabilities of the model when latitude longitude is included as inputs. Yes, we can see the weird patterns at some locations, but which locations indicate the gap filling capacity is not clear or evident. Currently there is no statistical proof or indicator to show the same.

Please consider adding a figure (may be a zoomed in version) or a table with correlation values to indicate this.

17. Line 240: What is the reason for less accuracy in the traditional approach here?
18. Line 243: Looks like authors missed including something after “The observations and”. Please check this.
19. Line 269-270: The authors state that the RF-based current approach performs better than relying on CTM output alone. Where is this conclusion referred to in your result?
20. Line 272-273: Does the RF version without the model as input capture these exceedances? It would be interesting to verify if the inclusion of model data is lowering these peak predictions?

Looking forward to seeing the revised version of this manuscript with the above changes. I thank the authors for the manuscript and wish them a good luck.