Point-by-point response, Reviewer 1

I appreciate the effort that the authors have put in assessing spatial uncertainties associated with their predictions. Unfortunately, this is often not a priority in global mapping papers, so I welcome this focus on uncertainty very much.

Answer: We thank reviewer 1 for the positive view of our work and appreciate the suggestions to improve the manuscript. The color 'light blue' denotes changes we made according to the suggestions of reviewer 1. Please note that all section and line numbers refer to the mark-up version of the manuscript. Figure numberings are shifted as we moved one figure from the end to the beginning of the paper as recommended by Reviewer 2.

The manuscript is structured well, but overall fairly lengthy and is written in a relaxed, almost conversational writing style - which I like, but from sometimes it's perhaps a bit too much. Another suggestion would be to move most of the tables and figures to the supplemental to improve readability. With 12 figures and 5 tables in the main text, I found it sometimes hard to navigate and find the most relevant results.

Answer: We agree that the original manuscript is quite long.

Correction: We improved the writing style and removed redundant information. This resulted in many small changes that are infeasible to list here. All changes are noted in the mark-up version of the manuscript. We moved Figs. 5 and 7 to annexes C and E, respectively. As a result, the manuscript was shortened by three pages.

I must admit that I know close to nothing about ozone and what variables structure its spatial patterns in the troposphere, but essentially the model is predicting O3 using latitude, altitude and human development (nearby nightlight); collectively explaining the majority of variation in the model. However, latitude and altitude are both merely a proxy for temperature and radiation; which are probably the actual main drivers of O3 levels (L153). Why not use them directly as predictors in the model? There are various high quality global layers available.

Answer: The purpose of this study is to map ozone using only static geospatial features and to evaluate the reliability of maps produced with these features. We recommend using meteorological data for future ozone mapping studies.

Correction: We now mention the use of static input features up-front in the introduction, line 72f: "This study builds upon Betancourt et al. (2021b) who proved that ozone metrics can be predicted using static geospatial data." We discuss the aspect of meteorological data use for future studies in Sect. 4.4, line 569f: "The use of meteorological data as static or non-static inputs can be beneficial to further increase model performance and allow time-resolved mapping."

With the very strong effect of absolute latitude in the model (~25% global importance), the predicted pattern will strongly reflect absolute latitude – which is clearly visible in the final map (fig 11).

Answer: We agree and found the latitude dependency consistent with our expectations. This property of the map was mentioned in the Visual analysis section (3.3.2).

Correction: We rephrased the sentence to be more explicit: “The global importance of ‘absolute latitude' shows through a latitudinal stratification and a clear north-south gradient in Europe, the US and East Asia.”, Sect. 3.3.2, line 431f. We also discussed this property of our map in more detail in Sect. 4.2, line 487ff: “Ozone is affected by meteorology (temperature, radiation) and precursor emissions (Sect. 1). The fact that there is no continuous increase of ozone towards tropical latitudes shows that the mapping model at least qualitatively captures the influence of low precursor emissions in the tropics. The importance of ‘absolute latitude’ also indicates that the model can be improved by including temperature and radiation features from meteorological data.”.

Next, I have a couple of doubts with the current “area of applicability” (AOA) approach, where features are weighted by their respective importances as in the model. By weighting the features, you’re essentially using absolute latitude and altitude (twice, including relative altitude) as to define feature space in which you apply the model. But as above; it’s not latitude or altitude that define this space, but temperature and radiation (L153). A major drawback, I think, from scaling the features used in the AOA analysis using their respective importances, is that you’re now basically assuming you can predict everywhere on the same latitude, save for some places (some of) the fewer important features fall outside the sampled space.

Answer: Scaling features according to their global importance is necessary as this global importance determines the features on which the model is basing its predictions. Basing the applicability of a random forest on features it seldomly uses to make a decision would be difficult to justify. Furthermore, not scaling the features at all would make correlated features multiple times important for applicability through the course of dimensionality. For example, the ‘nightlight’ in different radii around a station would have triple the importance of ‘absolute latitude’. We agree with Meyer et Pebesma (2021) who state that the scaling ‘lacks a formal statistical argument’. Nevertheless, our analysis shows that the unmodified method remains the best solution for this study.

Even though ‘absolute latitude’, ‘altitude’, ‘relative altitude’, and ‘nightlight 5km’ are most important, the other features are not ignored when defining the applicability. For example, large parts of the southern hemisphere are not predictable, even if they have exactly the same 'absolute latitude' as the northern hemisphere. Of course, the area of applicability might look different if other features would be used to train the model.

It would be useful if the authors would include a comparison with previously published tropospheric ozone predictions, based on mechanistic models. Of course, the authors did do a substantial test of their model’s validity and stability – but these really only hold true for this particular model and dataset; and don’t provide insight in how the results compare to other tropospheric ozone predictions.

Answer: The focus of this paper is the explainability and the exploration of mapping with static geospatial input features. We agree that a comparison with model data products is beneficial. We found the maps generated by DeLang et al. (2021) to be the most suitable, as they cover the same spatial domain and also aggregate in time. The Visual analysis Section (3.3.2) describes several spatial ozone patterns that can be found in both maps. We decided against a thorough quantitative analysis of the differences between the two maps, as this would go beyond the scope of this already rather long paper.

Correction: We now mention the similar patterns more explicitly in the scope of the visual analysis (Sect. 3.3.2, line 442f): "The spatial ozone patterns described here can also be found in ozone maps generated by traditional chemical models such as the fusion products by DeLang et al. (2021).".

Further points

1. The statement on L440 (*"The effect of 'absolute latitude' on predictions is consistent with what is known about ozone formation – ozone generally increases toward the equator"*) seems to be the opposite of the patterns that are predicted (Fig 11), with lower values near the equator, and higher values in temperate regions?

Answer: We agree this is misleading. The statement is only correct for the US, Europe, and Asia, not for the equator region itself. Ozone values are low in the equator regions due to a lack of precursor emissions.

Correction: We reformulated two sentences: Sect. 3.1.3, line 358f now reads "A lower 'absolute latitude' value leads to an increased ozone value prediction.". Sect. 4.2, line 483ff now reads "The effect of 'absolute latitude' on predictions is consistent with known ozone formation processes, i.e. ozone production generally increases where more sunlight is available. [...] Ozone is affected by meteorology (temperature, radiation) and precursor emissions (Sect. 1). The fact that there is no continuous increase of ozone towards tropical latitudes shows that the mapping model at least qualitatively captures the influence of low precursor emissions in the tropics.".

2. The selection of the threshold distance for 'non-predictable' data (ie., the upper whisker of all the cv distances), is seemingly arbitrary. It is in line with the AOA paper by Meyer and Pebesma, but neither that paper provides a statistical reasoning for picking this particular threshold.

Answer: The cross validation distances might contain outliers that would make the area of applicability unrealistically large. It is common to take the upper whisker as a robust threshold for outliers. We have tried the 95% percentile as an alternative to flag outliers with similar results.

3. The manuscript includes various subjective interpretations of the results. I believe the manuscript would benefit from a more objective wording. Some examples:

- L310: *"East Asia is a special case because the ozone value distribution is rather narrow there"*
- L325: *"Very high 'nightlight in 5km area' values"*

- L406: *“are considered significant”*
- Figure C1: *“really high values*

Correction: Thank you for pointing this out. We eliminated subjective language throughout the manuscript. Regarding the examples given by reviewer 1,

- We removed the statement in line 310.
- Line 359 now reads “High ‘nightlight in 5km area' values lead to lower predicted ozone concentrations.” Here we refer to ‘high’ as used in the color scale of Fig. 6: SHAP does not use absolute feature values but differentiates between high and low values.
- Line 450 now reads “We define a 5 ppb change in RMSE score or predicted ozone values as significant (Schultz et al. 2017).”
- The caption of Fig. D1 now reads “[…] The maximum standard deviation exceeds 3.5 ppb for less than 20 realizations. […]”

4. The authors do use “not shown” rather often, a total of 8 times in the entire manuscript. I guess this is ok, but if the authors feel that the data/results aren’t necessary to show, arguable the entire section can be removed. If not, I would suggest placing the evidence for the statement in the supplemental materials.

Correction: We agree. We went through all eight occurrences of ‘not shown’. We concluded that the information given was often irrelevant to the study. One instance of ‘not shown’ remains in line 348 because the similarity of the feature combinations of the cross validation sets is relevant for the study, but the exact values of the dissimilarity indices are not.

5. Figure 3: are these ‘example data points’ points in the AQ-bench dataset or in the raster data you’re predicting?

Answer: These are data points we found in the gridded data.

Correction: We clarified this in the legend of Fig. 4.