Reply to both Referees (Preprint egusphere-2024-3145 - Hourly surface nitrogen dioxide retrieval from GEMS tropospheric vertical column densities: Benefit of using time-contiguous input features for machine learning models)

April 9, 2025

We would like to thank the Editor and the Referees for their comments and suggestions which helped us to improve the quality of our manuscript.

The following major changes were implemented in the revised version of the manuscript:

- In Figures 10 and 11, we have changed the coastlines to white and removed the black frame around the colorbars. This ensures that there is no ambiguity regarding the black mask for missing data.
- We clarified the choice of latitude as an input feature in lines 344-353.
- We remarked in lines 223-226 that experiments on a small dataset indicated that filtering negative VCDs has presumably a very small impact on the performance and would be neglected for future studies.
- In lines 603-607 of Section 5.2, we mentioned an interesting task for future work: Inspecting further potential differences when switching between models of different time-contiguity.

Below, we give a point-by-point response to your reviews.

Janek Gödeke

Reply to Referee 1

- I am not convinced with the reply about the exclusion of negative values. I note that the other reviewer also had a similar concern in their original review. In their response, the authors note that "Negative VCDs, so negative concentrations, have no physical meaning. This is why we excluded them from both the training and the test data to increase the quality of the dataset." It is not true that negative values have no physical meaning in these kind of satellite data, which are in effect differential measurements. There are a number of reasons that negative values many occur. The first is fitting a spectrum with random noise. Consider a measurement of a completely "clean" atmosphere with no tropospheric NO2 and a column of zero molecules/cm2. If this is measured by a detector with any noise, we would expect the observations to be distributed about zero (i.e., half will be positive and half will be negative.). Excluding all negative values in an analysis with this data will bias results high, where in truth each small negative value is more or less as significant as each equally small positive value.

  In addition, systematic uncertainties in spectral fitting inputs, like cross sections, could cause negative biases in the data in clean regions. Furthermore, the fact that a tropospheric column is being derived from a total column measurement using an estimated stratospheric column is another source of a potential negative bias in a tropospheric column. Even if the model cannot deal with negative values (which I guess is the case), I think there needs to be more of a discussion of how excluding these values could affect the results, rather than brushing over this point. I would think as well as affecting model performance, it needs to be explained what will happen to the negative VCDs when the final model is applied to a map of VCDs to derive concentrations. Can these be used, or are you at risk of losing data over clean regions that are actually useful (again, potentially biasing results)?

  We agree that negative columns must not be removed when computing averages of satellite data as that would create a high bias. Here, the situation is slightly different as not the target quantity (the in-situ observations) but one of the inputs (the satellite data) is filtered. In our opinion, there is no reason to expect a high bias of the predicted surface concentrations just because one of the input variables to the random forest is limited to positive values.

  However, we agree with the reviewers that negative satellite columns are usually found over regions with low tropospheric columns, and therefore, the filter we applied leads to a loss of predictions for such situations. In this sense, the ensemble of predicted values may be biased high.

  We tested the effect of the filter on a small data set and found very little changes. This is probably due to the fact that applying this filter only leads to a reduction of the dataset by less than 0.5%. In hindsight, it probably was not useful to implement this filter. Unfortunately, repeating the study without this filter would be a large effort at this point, but this lesson will be taken into account for future work. In the revised manuscript, we mentioned this in lines 223-226.

  When Random Forests are trained on non-negative VCDs only, they are still able to make reasonable, but potentially biased, predictions over clean regions with negative VCDs as inputs. In this case the Random Forests would treat negative VCDs as being zero.

- I am also still confused by the use of latitude as a predictor variable in such a small region as Korea (even if the justification is that it has been used in a previous paper over Switzerland). I am still not convinced it is a useful variable on which to focus, but I suppose at this point would it would require a lot of work to redo the model and paper. The authors responded to me and the other reviewer (also with this concern) but I am not sure they have changed the paper in any way to address this comment. It might be helpful to add a line or two to the paper to clarify the choice of this predictor.

  Thank you for pointing out that we had not clarified the choice of latitude in the manuscript. In Section 3.2, we have now added in lines 344-353 some clarification for the selection of latitude as a predictor, in accordance with our response to the first review. Regarding the consideration of input features without latitude (but with surface height), as well as an alternative choice for feature studies in Experiment 3, implementing these changes would indeed require extensive model training and tuning, necessitating substantial revisions to the paper.

Reply to Referee 2

- Your response mentions the use of "models" and a "rule of thumb" while utilizing an ensemble of models trained on different time-contiguous features. However, critical questions regarding the model's practical application and potential biases remain. While you suggest using a Random Forest trained with k = j' + 1 when time-contiguous features are available, and switching to a model trained without time-contiguity when they are not. Did the authors analysis the bias or any potential differences when switching among these models trained with different time-contiguous features?

  In Experiment 2, we analyzed the difference between all different models in terms of their prediction accuracy with respect to different loss functions. For all five considered loss functions, we made the same observation that the usage of time-contiguous models are beneficial and similar rule of thumbs apply, see Table B2. However, we did not systematically assess other potential differences that may arise when switching between models trained with different time-contiguous features. For instance, it remains an interesting task for future studies to investigate whether the ensemble of such models yields consistent combined spatial patterns in predicted surface NO_2. We included this as a remark in lines 603-607.

- In your response, you state, 'Negative VCDs, so negative concentrations, have no physical meaning. This is why we excluded them from both the training and the test data to increase the quality of the dataset.' While it is true that negative concentrations are not physically realistic in the absolute sense, negative measured values can and do occur in real-world datasets due to measurement noise, particularly when the true values are close to zero. These negative values are not necessarily indicative of data quality issues but rather a reflection of the inherent uncertainty in the measurements. Furthermore, your claim that excluding these negative values does not introduce bias is actually incorrect. By systematically removing negative values, you are artificially shifting the mean of the dataset towards positive values. This will inevitably introduce a positive bias in any model trained on this altered dataset, regardless of whether the test data also lacks negative values. While your model might perform well on your artificially positive-shifted test data, it will not accurately reflect real-world scenarios where extremely low values are sometimes measured as negative. Skipping that part will definitely result a positive bias, especially for regions with low values. Also, I cannot see why the authors have to apply this additional filtering and cause addition missing data in their prediction. Therefore, the argument that the model will only be used on data without negative values, as you imply, is not well supported.

  We agree that negative columns must not be removed when computing averages of satellite data as that would create a high bias. Here, the situation is slightly different as not the target quantity (the in-situ observations) but one of the inputs (the satellite data) is filtered. In our opinion, there is no reason to expect a high bias of the predicted surface concentrations just because one of the input variables to the random forest is limited to positive values.

  However, we agree with the reviewers that negative satellite columns are usually found over regions with low tropospheric columns, and therefore, the filter we applied leads to a loss of predictions for such situations. In this sense, the ensemble of predicted values may be biased high.

  We tested the effect of the filter on a small data set and found very little changes. This is probably due to the fact that applying this filter only leads to a reduction of the dataset by less than 0.5%. In retrospect, it probably was not useful to implement this filter. Unfortunately, repeating the study without this filter would be a large effort at this point, but this lesson will be taken into account for future work. In the revised manuscript, we mentioned this in lines 223-226.

- Thank you for pointing out the black mask indicates missing data in your figures. But I would suggest using a different color for missing values, as your coast lines and lowest value of your colorbar are also black.

  Thank you for the suggestion! We experimented with different colors for masking the missing data but were not satisfied with the results. For example, a white mask drew too much attention away from the rest of the image. Therefore, we opted for the following alternative: We retained the black mask for missing data but changed the coastline color to white. Additionally, we removed the black frame around the colorbars to make it clearer that black is not the lower end of the color scale.