Reply to Referee 1 (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)

February 27, 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:

- A flowchart for the data processing workflow was added (see Fig. 2).
- A table with detailed information on the temporal and spatial data resolution and preprocessing steps was added (see Table 2).
- Some of the models in Section 5.3 were trained again and tested on seasonal data (see the new Section 5.4).
- These models were tested for different times of the day (see Section 5.4).
- We increased the size of symbols within plots for better readability.

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

Janek Gödeke

General Comments

- (...) The figures and results clearly indicate the use of earlier data improves the performance of these machine learning models overall. I would have liked to see a small discussion about how these improvements might change over the course of a day. The GEMS observations are probably much less accurate in the morning and late evening (high angles, less sensitivity to the surface). The morning is furthermore limited in earlier time-contiguous observations but the evening is not. How does the performance of the final result change as a function of time? (...)

  Thank you for the suggestion! We have analyzed the test performance of some of our models from Section 5.3 (trained on whole-day data) for different times of the day and incorporated a corresponding discussion into the new Section 5.4. Additionally, we have mentioned in the outlook (Section 6) that a potential avenue for future research could involve training models specifically on subsets of data, such as morning-only data. It would be interesting to investigate whether this specialization might lead to further improvements in performance. However, implementing such an approach would necessitate tailored, time-specific hyperparameter tuning, which is beyond the scope of this study. Therefore, we only mention this as an outlook.

- Two rather basic models are used (linear regression and Random Forest), which the authors chose to more easily isolate the performance changes. It's not clear how the performance with time-contiguous data would change in other model setups. Do you expect those to have the same gains in performance?

  We agree that Random Forests are basic models in the sense of a limited number of tunable hyperparameters, but for regression tasks, they are in general powerful and competitive.

  At the outset of this study, we also experimented with Neural Networks (NNs) for estimating surface NO_2. While we observed similar results to those obtained with Random Forests, the training time for NNs was considerably longer. Therefore, and due to the large number of hyperparameters and architectural design choices for NNs, conducting as many experiments with NNs as we did with Random Forests would have been outside the scope of our study. This is why we chose to focus on Random Forests, but we expect similar performance gains also for Neural Networks.

  In the revised manuscript, we added a remark to the introduction of Section 4, in which we mention that we also did a few experiments with NNs and observed similar performances. But since they were much more time-consuming, we could not do the same number of experiments as we did for Random Forests.

Specific Comments

- Line 44: Change to “the measurement of lower tropospheric gases is not accurate”

  Thank you for the suggestion. We have changed that in the revised manuscript.

- Line 44: “This is why most studies estimated daily” doesn't follow from your previous statement. The estimate they give is still at a specific time, not a daily average which is implied here. Clarify this sentence.

  In fact, there exist both types of studies: Kim et al. (2017) predict surface NO_2 at the specific satellite observation time. Di et al. (2020) predict daily averages of surface NO_2. Therefore, we changed the sentence as follows: ”Since satellites in low-earth orbits provide observations at most once a day, most studies either predicted surface NO_2 at this specific satellite observation time (e.g., Kim et al. (2017)), or they estimated daily (e.g., Di et al. (2020)), monthly or annual averages of surface NO_2.”

- Line 105: I'm confused... where did j come from? The above equation uses t-k+1 (no t-j mentioned.)

  Thank you for your comment. In the input vector (line 106 in the revised manuscipt), $t-k+1$ refers to the earliest time since k is the time-contiguity. On the other hand, j is a variable which takes values in the set $\{0, 1, ..., k-1\}$. So $t-j$ stands for all times between $t-k+1$ and t. We have specified this in line 107 of the revised manuscript.

- Line 148: Would be useful for context to summarize accuracy of the NO2 product you use, both for troposphere and stratosphere. And how does this change over a day?

  We agree that an investigation of the impact of uncertainties in the satellite columns on the predicted surface concentrations would be interesting, but unfortunately, the GEMS IUP-UB product does not yet have full error propagation. The tropospheric NO2 VCD error is therefore estimated with 25%. The main uncertainty results from the assumptions used in the calculation of airmass factors, in particular for surface reflectivity, NO2 vertical profile and aerosol loading. As the reviewer points out, uncertainties are expected to be larger in the morning when the boundary layer is shallow and smaller around noon and in the evening. Uncertainties introduced by the stratospheric correction can be important over clean regions but can be neglected over pollution hotspots. We have added this information to Section 2.1.1 of the revised manuscript.

- Line 153: The TM5 model may leave residual structure in the results... maybe mention resolution here.

  The TM5 model has an hourly temporal resolution with a spatial resolution of 1° × 1°. As the model a priori is interpolated in space and time, no obvious structures from the coarse model resolution are visible in the data, but the lack of detail still may impact the results. We included this information to Section 2.1.1 of the revised manuscript.

- Line 175: What kind of sensors are used? What is accuracy of the sensors?

  Thank you for this question. We have been informed that the instruments utilize the chemiluminescence method, as described by Kley and McFarland (1980, Chemiluminescence detector for NO and NO2). We have included this information to Section 2.1.3 of the revised manuscript. However, we were also advised that the specific types of instruments may vary, along with their accuracy. We do unfortunately not have detailed information regarding these variations.

- Line 183: "We assume" – this seems like something that should be clear in a user guide or the information could come from the data producers upon request. Is this a fact or are you really making an assumption? Without more information it could also be assumed that 1:00UTC is describing the monthly average from 00:30-1:30 UTC. I generally find the time stamp discussion confusing. Wouldn't it make sense to label this example as 2021/01/23/02 since two datasets at least are occurring around 2:00 UTC?

  Unfortunately, there is no user guide available for the data. Within the dataset there is only one time label given per data point, e.g. 01:00 UTC. Therefore, we inspected the correlation between in situ surface NO_2 and VCDs at different hours, which suggested that the label 01:00 UTC in the in situ dataset refers to measurements between 01:00 and 02:00 UTC. However, in response to the reviewer's comment, we enquired again about the correct interpretation of the time label in the in-situ data and were informed that they indeed should be read as indicating measurements for the previous hour. This misinterpretation of the data is very unfortunate and could only be fixed by repeating all steps of the analysis, which is not possible at this point. However, the above mentioned tests showed only a small dependency towards changing the interpretation of the data by one hour, which gives us confidence that the conclusions of the manuscript are not affected by this mistake.

  Regarding the time stamps, we chose the same stamp that is used also within the VCD dataset. By that, we wanted to avoid confusion when using the VCD dataset.

- Line 209: Maybe I don't know enough about how these models work, but I don't understand how these negative values can be excluded, or why they have to be. Can you give some more justification? If the model is trained on a dataset that is biased at low column values of NO2, how does this affect results? If you don't care about the bias but can't handle negatives, why not add a background amount to make all the negatives positive to maximize use of all data? If you want to use the column values later to estimate surface NO2 in a given location but have negative values and haven't considered them in the model, how can these be used?

  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. By doing this consistently on both training and test data, the model does not suffer from any bias, because the model is trained and tested on the same type of data. One should not test these models on data points with negative VCDs (then we would agree that there could be some bias).

  For future tasks, one could also train and test new models on the larger datasets in which negative VCDs are included. However, we doubt that increasing the dataset is always beneficial, as it could also be disadvantageous, because it could make the data less interpretable by the model.

- Table 1: I think it would be useful to re-define N and give its unit here in caption.

We have added this information to the revised manuscript.

- Line 314: I'm not really clear about why latitude should get included at all as a feature in the first place. It's good to see later that its inclusion doesn't matter much, as the tropospheric VCD should have very little dependence on latitude in a physical way. Presumably the correlation in Table B1 is moderately high because in Korea the NO2 sources are dominated by a few cities including Seoul in the North, but the latitude is not the cause of enhanced tropospheric NO2. It could be important for other gases and larger domains, but not trop NO2 in a tiny area like South Korea.

  We agree that the inclusion of the coordinates might be problematic and shouldn't matter much. However, other studies have used spatial coordinates for predicting surface NO_2. Mainly over large regions, such as the USA (e.g., Gharemanloo et al. (2021)) or China (e.g., Li et al. (2022), Qin et al. (2020)). But also over smaller regions, such as over Switzerland (e.g., deHoogh et al. (2019)).

  We took spatial coordinates (longitude/latitude) into consideration during feature selection (Section 3.1) because we wanted to check:

  - Although spatial coordinates only slightly differ within Korea, couldn't there be a small helpful information for the model from spatial information?
  - Is there an additional risk for spatial overfitting when taking spatial coordinates as an input? This is why in Experiment 3 (Section 5.3) we made the same analysis without using latitude as an input.

- Figures 2 and 3: Not a big deal but I'm not sure why left column has to be included... seems redundant with middle column which provides a more complete result.

  We agree that the left column is a bit redundant. Our motivation for including it, though, is to stress that although the standard deviations in the middle column are quite large, the curves of individual station splits are more or less parallel. We think this will make it easier for the reader to understand that there is always a benefit from time-contiguous inputs, for every individual split and not only on average.

- Line 653: Here and earlier, I'm not clear why you would want to use this model outside of Korea with no VCD input (also, the focus of the paper seems to be GEMS – i.e., satellite observations). Can you elaborate under what circumstances this would be useful? I would expect it to be pretty inaccurate without the VCD, especially in regions with no monitors, and not as useful as a physical model output from something like CAMS or GEOS-CF.

  Excluding VCDs as an input for the models did lead to worse predictions over Korea (see Section 5.3). However, we would actually have expected an even larger decay in performance. Nevertheless, we believe that VCDs as an input will prevent the model from spatial overfitting, which may be a less significant effect for a small area like Korea. This motivates the investigation within future work whether the models without VCD-input will indeed perform much worse at locations outside of Korea. Further, a model that also works outside the region at which it has been trained is desirable in practice, because by that one could get a broad picture about surface NO_2 pollution, even if no data from ground monitors are available.

Technical Comments

- English issues

  We have corrected the English issues that you have pointed out.

- Figure 2 and 3: In these and other figures, the linewidth, symbol size and sometimes font size are very small and hard to read on my screen. There are not many points, so there is a lot of room to improve the figures by making lines and symbols larger in future plots.

  Thank you for this suggestion. We have used larger symbols and lines in the updated version of the manuscript.