Reply to Referee 2 (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

Data Processing

The data processing methodology is unclear and lacks sufficient detail. The authors should enhance this section by:

- Including a flowchart in the Data section to visually illustrate the entire data processing workflow
  Thanks for the suggestion, we have added a flowchart to Section 2 that visualizes all data processing steps (see Fig. 2).

- Providing a table detailing each data source, including the spatial and temporal resolution, and any preprocessing steps applied to the input datasets.
  Information about preprocessing, as well as the spatial and temporal resolution of the data, is provided in Sections 2.1 and 2.2. However, we agree that a compact overview in the form of a table enhances clarity. We have included a table into the revised manuscript (see Table 2).

Satellite and Ground Station Pairing

- The statement "we associated the location of an in situ station with the VCD pixel or meteorological pixel whose center is nearest to the station's location" needs clarification. Does this refer to the center of the satellite pixel?
  Yes, it refers to the center of the satellite pixel since we do not apply a regridding. Thank you for this remark. We have clarified this in line 192 of the revised manuscript.

- If multiple ground stations fall within the same satellite pixel, how are these handled? Are they averaged, or is one selected?
  If multiple stations fall within the same satellite pixel, data points are generated separately for each station and included in the dataset. We are aware that if these stations measure different values, this adds noise to the dataset. However, this has the advantage of reducing the risk that the model overfits to the training data.

- GEMS pixel locations vary slightly with each scan due to orbital and observation geometry. Did the authors regrid the satellite data before co-location to ensure consistency?
  No, we did not regrid the satellite/VCD data onto a universal grid. This decision was made to minimize data manipulation and preserve the integrity of the original measurements.

Characteristics of Ground Stations

- What type of instruments are used at the ground stations? For example, are they chemiluminescent analyzers?
  Thank you for your question. We have been informed that the instruments utilize the chemiluminescence method, as described by Kley and McFarland (1980, Chemiluminescence detector for NO and NO_2). 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.

- Ground stations are often categorized as urban, background, or roadside. Did the authors use all station types, or restrict their analysis to specific types? The representativeness of the training data depends on this choice.
  No stations were excluded based on their location. As shown in Figure 1, stations are distributed across both urban and rural areas. As the spatial resolution of the satellite data is limited and vertical profiles vary between location types, it is to be expected that the relationship between NO_2 column and surface concentration is different for different location types. However, the approach taken here was to use all stations in the same way, and not to limit the training dataset to a certain type of station or to provide the station type as additional parameter.

Temporal Input and Data Loss

- The model will not produce predictions for the first few hours of each day, creating data gaps.

  Yes, if the model receives time-contiguous inputs, it is unable to make predictions for the initial hours of each day. This is a limitation which we included in the revised manuscript, see line 272 in Section 2.3.

- Cloud cover and other issues affecting satellite measurements in prior hours can propagate errors into the current hour's input, resulting in significant data losses during training and prediction. This cascading loss reduces the dataset from over 1.3 million data points for a 1-hour input window to approximately 350,000 for a 5-hour window. The authors should provide a clear justification for accepting this trade-off between increased data gaps and potential gains in model accuracy.

  To address this trade-off between data gaps and improved model accuracy, we have designed Experiment 2 (see lines 330–339 in the revised manuscript). Our findings indicate that models trained on the largest dataset ($\geq$ 1 million points) without time-contiguity perform worse than models trained on smaller datasets with time-contiguity. These results are discussed in detail in Section 5.2.

- Additionally, the authors should evaluate and discuss how these data gaps impact not only the training and validation phases but also the model's predictions and its applicability to real-world scenarios. This includes addressing potential limitations in the model's ability to generalize when encountering similar conditions in operational or extended applications.

  If a model is trained with time-contiguity, one cannot apply it to data points for which the time-contiguous features do not exist. Our study should be understood as follows: If time-contiguous features are not available for some data point, then use a model that has been trained without time-contiguity on the large dataset. However, as soon as we want to make a good prediction for a data point for which time-contiguous features are available, we have shown in Experiment 2 that better predictions are made by time-contiguous models. In real-world scenarios, this means that time-contiguous models support the non-time-contiguous model whenever this is possible. In the revised manuscript, we stressed this point in an additional remark after the rule of thumb, see lines 713-715, or already in line 594-596.

Justification for Input Variables and Preprocessing

- The paper lacks justification for the selection of input variables. A sensitivity analysis or variance inflation factor (VIF) analysis should be conducted to ensure the chosen variables are non-redundant and significant

  In Section 3.1, we describe the feature selection process used in our study. We used a simple criterion: We computed the correlation between VCDs and potential input features on the training datasets. Then we selected those features whose correlation to VCDs is at least 0.1, so that the features have a better chance of being significant. We cross-validated the correlation on 200 different potential training sets and observed that this criterion is very stable, meaning that (almost) always the same features are selected according to this criterion. However, we do not put a focus on redundancies among the features because of the following:

  Using time-contiguous inputs will presumably introduce some redundancy to the inputs if some of the features do not change much over time. A complete sensitivity analysis would not only require examining the choice of features, but also assessing the optimal time-contiguity for each feature. A complete sensitivity analysis would be out of scope for our study, and we chose not to deviate from our main focus, which is evaluating whether time-contiguous inputs are beneficial.

  As a compromise, we did a sensitivity analysis in Experiment 3 regarding three input features: VCDs, latitude and surface height. We explain the rationale for inspecting these features in Section 3.2 (lines 340–348 in the revised manuscript), and the results are discussed in Section 5.3.

- The input variables differ in units and magnitudes, which could cause instability in model performance. Did the authors scale, normalize, or log-transform these variables before training? This critical preprocessing step is missing from the discussion.

  Thank you for this question. Yes, we applied an affine linear transformation to each feature, such that the mean of the training data is 0 and the standard deviation is 1. We have added this information to the end of Section 2.3 of the revised manuscript.

Choice of Models

- The authors used Random Forest and linear regression but did not justify these choices.

  The motivation for using Random Forests and linear regression is given in the introduction of Section 4: Random Forests have only few hyper parameters to tune. Therefore, one gets a clearer insight into whether time-contiguous inputs are beneficial. Additionally, Random Forests are well-suited and powerful for regression tasks. Although linear regression models are not competitive, they give a first insight into the experiments.

- More advanced machine learning methods, such as Artificial Neural Networks (ANN), Recurrent Neural Networks (RNN), or Convolutional Neural Networks (CNN), have been shown to better handle non-linear relationships and spatio-temporal dependencies in atmospheric data. The authors should explain why these advanced methods were not used or compare their results to them.

  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.

  We have avoided comparing our results, obtained with Random Forests, to methods from other studies due to the use of different datasets, primarily from different regions of the world. Such comparisons could risk being unfair in either direction.

Handling of Negative Values

- The authors ignored negative GEMS VCD values, which will bias the average toward positive values. Justification is needed for this choice.

  Thank you for the comment. However, we do not think the models suffer from any bias due to the decision of excluding negative VCDs, because the models are both trained and tested on this type of data. One should not test these models on data points with negative VCDs (then we would agree that there is some bias).

  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. Excluding negative input VCD values from a training set does not create a low bias in the target quantity unless you later feed negative VCD values to the model.

- Similarly, were there negative values in the in-situ measurements? If so, how were these handled? This needs to be explicitly discussed.

  Thanks for this question. No, there were no negative values in the in-situ measurements. We have added this information to the end of Section 2.2 (lines 224-226) and to the description of the data-preprocessing in Section 2.3.

QA Value Threshold and Bias

- The authors only used data with QA values equal to 1. This choice filters out cloudy conditions but potentially introduces a clear-sky bias since cloudy conditions can be associated with higher aerosol or NO2 levels. The authors should address this limitation and quantify its impact on results.

  The decision to only consider QA values equal to 1 implies that the model is not (reliably) applicable to situations in which there are cloudy conditions. We have added a note on that at the end of Section 2.2. Additionally, we have remarked that a potential future direction of this work could involve examining the effects of lowering the QA threshold. This would result in a larger, but more complex, dataset.

Inclusion of Latitude

- Including latitude as an input variable needs further justification, as the latitudinal variation over South Korea is minimal. The authors should explain the rationale behind this decision.

  We agree that the inclusion of the coordinates might be problematic. 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.

Section-Specific Comments

Section 5.2

The atmospheric lifetime of NO2 varies with season and time of day, and this variability likely influences model sensitivity. The authors should:

- Conduct and present seasonal and diurnal sensitivity analyses to account for these variations.

  Thank you for the suggestion. We have trained some of the models discussed in Section 5.3 and evaluated their performance on seasonal datasets. We included this to a new Section 5.4 in the revised version of the manuscript. Additionally, we examined their performance over the course of the day.

- Address potential biases from the limited temporal scope of training data (January 2021 to November 2022). For instance, why was data from December underrepresented, and why were only 23 months used instead of two full years?

  We received the in situ dataset at the beginning of December 2022. Therefore, data from December was not available at that time. We made a note on that in Section 2.1.3 of the revised manuscript. Further, a short remark of a potential bias due to the Covid-19 pandemic, has been added to the end of Section 6 of the revised manuscript.

- Discuss whether differences in valid data points across seasons (e.g., more data in summer due to fewer clouds) lead to seasonal biases in model training.

  In fact, after applying a filter for the qa-value, there are less data points in summer available, due to the monsoon. We added Table 3 to the revised document, which shows the contribution of each season to the total number of data points. Further, we have inspected the performance of some models from Section 5.3 at all different seasons. Details were added to the new Section 5.4.

Section 5.3

- The prediction maps show that the model has been applied beyond South Korea, including regions over the ocean, Japan, and North Korea. The authors should: Validate the model's performance in these regions by comparing predictions to in-situ measurements from other countries, such as Japan. This would demonstrate the model's transferability across different geographies.

  Thanks for the suggestion. Evaluating the model's performance outside of Korea would be an interesting extension of our work. However, accessing datasets from regions such as Japan and studying the model's performance there is left as a future task. This is mentioned in Section 6.

- The prediction maps also exhibit noticeable grid structures, likely originating from the meteorological ERA5 dataset. Did the authors interpolate the ERA5 data to reduce these artifacts? If not, why?

  We did not interpolate the ERA5 data to reduce these grid artifacts, as we did not want to modify the data solely for the purpose of improving the appearance of the plots. Our focus was on preserving the integrity of the original dataset, and we believe that these artifacts are a natural characteristic of the data.

- Clarify how gaps in GEMS data (e.g., due to cloud cover) were handled during prediction. The maps show no missing areas (Figure 9 and 10), suggesting the model was applied to cloudy data despite such data being excluded during training. Discuss the implications of using potentially contaminated data and its impact on model accuracy.

  The maps do show missing areas, because gaps in GEMS data were not considered during the prediction. Figures 9 and 10 (in the revised manuscript, these are Figs. 10 and 11, respectively) display these missing areas as black pixels. This is also noted in the description of Figure 9. In the caption of Figure 9, we state: "The black mask indicates missing data, e.g. due to clouds."