Response to Reviewer 2

Thank you very much for your constructive and critical comments. Accordingly, this manuscript has been revised. Regarding the response, please note that the line numbers in your review do not exactly match with those in the latest manuscript and thus, responses are given based on comments.

Gil et al present a model based on Deep Neural Networks for estimation of HONO concentrations in urban environments using measurements of classical atmospheric pollutants and meteorological variables as input. Because HONO measurements are hardly available, the authors argue that using estimated HONO as input in photochemical models improves the calculation of the OH production rate and of O3 concentrations. This is an interesting and valuable piece of work, however, there are a few issues that should be resolved before this manuscript can be published in Geoscientific Model Development.

My main concern is the way the performance of the RND v0.1 model is evaluated and the conclusions and recommendations that are drawn from the model performance evaluation. For performance testing, both, the training set and the testing set are used. This is not correct, the training and validation data should only be used for model building, the performance assessment should only be done based on the test data. It is found that the model performance is much better for the period used for training and validation than for the test data. This is of course not surprising and indicates clear limitations of the model (e.g. over-fitting). The model performance assessment needs to be changed accordingly.

You are absolutely right that the data set used to train the model should not be re-used in the testing process. As illustrated in Figure 3, we divided the observation data into a training set (May-June 2016, May-June 2019) and a testing set (June 2018, April 2019, May-June 2021, and Oct-Nov 2021). This fact is more clearly expressed in the revised manuscript as follows.

> Line 220-222: RNDv1.0 and the RF model were tested using data obtained in June 2018, April 2019, and May–June 2021 and October–November in 2021, which were not used for RNDv1.0 training (Figure 3).

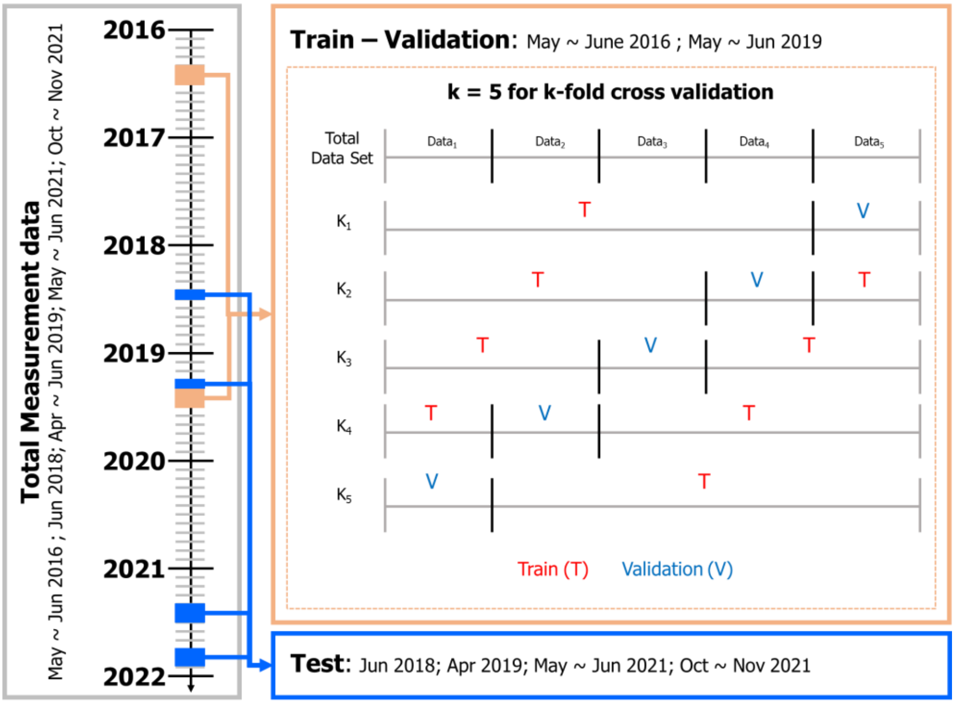

Figure 3. Training, validation, and test design to build RNDv1.0 using the measurement data. The k-fold cross validation was performed using randomly divided five subsets of the training data set.

The model performance was particularly poor for the test data from April 2029. The authors explain this by the fact that the conditions during April 2019 were different from the conditions covered by the training data. This points to another important aspect that is entirely neglected in the current manuscript: What are the conditions the RND v0.1 model can be applied with a performance as determined? What happens when the model is applied to conditions that are not covered by the training data (model applied to meteorological conditions and/or atmospheric pollutant concentrations outside the range covered in the training data)? It is very likely that applications of the proposed DNN model at other locations and during other times of the year will face this situation. It is necessary that this issue is addressed.

The authors say in the abstract and in the introduction section that the RND v0.1 model is proposed for calculation of HONO mixing ratios in highly polluted urban environments. In the results section, the model is described as being fit for application in any urban area (page 6, line 172). The conditions (in terms of air pollutant concentrations) where RND v0.1 can be applied should be made more clear.

This manuscript has been revised through several rounds of review, during which the recent HONO measurements performed in May-June and Oct-Nov 2021 were added to the test data set. In the new test set, extreme events that occurred in early winter are included. As a result, significant revisions were made to the manuscript, including sections “2.5. Model test” and “2.6. Bootstrap test and feature importance”.

In this study, testing data was obtained in a wider range of ambient air conditions than training data. It provides a good opportunity to evaluate the applicability of RNDv1.0. First, the performance of RNDv1.0 was compared with those of other models in Section 2.5. Then, the influence of input variables on the output results was analyzed in Section 2.6. In addition, detailed information is provided in Supplementary Information.

Actually, the extreme conditions are hardly constrained by model, causing the discrepancy between the measurements and model results. It is especially the inherent limitation of data-driven model. Given this fact, the RNDv1.0 is able to trace the variation of HONO when applied in the range of training data set, which is still difficult in other conventional models. We therefore suggested the conditions for the application of RNDv1.0 clearly as our training data coverage, and warning about its use in the different conditions. Consequently, it is reasonable to argue that this study demonstrates the applicability of RNDv1.0 to urban atmospheres enriched with NO_x.

The relevant parts are revised as follows.

> Line 228-244: The performance of RNDv1.0 was slightly lower than that of the RF model, but it well traced the HONO mixing ratio. Among the test dataset, the early winter (October–November) data are particularly valuable for demonstrating the applicability of RNDv1.0 because they stem from different weather conditions than the training dataset. For example, HONO mixing ratios reached over 4 ppbv when the daily average $PM_{2.5}$ concentration increased to 120 μg m^{-3} during severe haze pollution events. Therefore, in the next step, the performance of RNDv1.0 was compared for the two cases by dividing the test dataset into a group in which all input variables fall within the range of the train dataset and a group which does not meet this criterion. In RNDv1.0, there was no significant difference in performance between the two groups (Figure S5 and Table S2). When the data in which at least one input variable does not fall within the range of the training dataset were excluded from the test dataset, no significant difference was observed in the performance of RNDv1.0 between the two that meet same atmospheric conditions or do not meet the criteria (Figure S5 and Table S2). These extreme atmospheric conditions can make the model performance be worsened. Except for these extremes, RNDv1.0 well traced the

variation of the HONO mixing ratio. These results demonstrate the applicability of RNDv1.0, which is not strictly constrained by atmospheric conditions. The influence of input variable are further analyzed in the next section.

Line 314-324: RNDv1.0 was constructed using the measurements made in a high NO_x environment where the maximum NO_2 reached about 80 ppbv. During the measurement period, the HONO mixing ratio was increased up to about 7 ppb under the influence of air masses originating from China. When applying RNDv1.0 to regions or times heavily affected by transport, the model could possibly underestimate the HONO level without more detailed information, such as nanoparticles. Indeed, a previous study showed that HONO formation is closely related to the surface areas of submicron particles (Gil et al., 2021). Nevertheless, RNDv1.0 is advantageously a relatively inexpensive test for measurement quality control and location selection, and it supports the data used for traditional chemistry models based on the current knowledge of the urban photochemical cycle. Therefore, RNDv1.0 can serve as a supplementary tool for conventional forecasting models.

The paper is generally well written, however, there are rather many small linguistic errors such as missing articles (e.g. page 3, line 84; pg. 5, lines 139 and 140; page 6 line 159) and wrong grammar (e.g. should consequently be "training and validation" instead of "train and validation", and also often "testing" instead of "test". The manuscript should again be carefully checked and corrected.

This manuscript was carefully checked, and errors were corrected.

Other comments:

Page 2, line 54-56: The authors write about "the" model and "this underestimation". It is unclear what model is meant, it seems that it is referred to photochemical models in general. Please make this clear and revise accordingly.

This sentence is rewritten as follows.

Line 74-76: In comparison, the WRF-Chem and RACM2 models captured approximately 67 %–90 % of the observed HONO in megacities such as Beijing

Page 3, line 70, should be "including data collection" instead of "including collecting data".

It is changed as you suggested.

Line 110-112: The RNDv1.0 development follows systematic steps that are similar to a general ML model construction workflow, including data collection, preprocessing data, building the DNN, training, and validating the model, and testing the model performance.

Page 4, line 95-97. The 10th and 90th percentile mixing ratios for the input variables are given. It is not mentioned what the time basis of these values are, are these hourly or daily values? The temporal resolution should be provided.

Statistics are provided for hourly measurements, which is clearly stated in the revised manscuript.

Line 135-136: The measurement statistics for the entire experimental periods are presented in Table 2 and Table S1.

Table 2. Input variables and their concentrations (10^{th}–90^{th} percentile of the hourly measurements), coverage, and scale factors for the RNDv1.0 model. Measurements were conducted in Seoul during May–June in 2016 and 2019.

Table S1. The range of input variables (hourly measurements) used in this study.

Page 4, line 102. Terminology "chemical and meteorological parameters" is not correct here. In the usual convention, the input variables are denoted as "variables" and not as "parameters". The parameters are their weights in a statistical model. Please change.

It is changed through the manuscript.

Line 101-103: This study aims to develop a user-friendly "reactive nitrogen species simulation using DNN' model (RNDv1.0) that estimates the HONO mixing ratios from the real-time measurements of criteria pollutants and meteorological variables.

Line 118-119: To construct RNDv1.0, measurement data were obtained, including HONO, reactive gases, and meteorological variables

Line 131-134: In addition to HONO, trace gases including O_3, NO_2, CO, and SO_2 as well as meteorological variables including temperature (T), relative humidity (RH), wind speed (WS), and wind direction (WD) were measured.

Line 141-144: As input variables, hourly measurements of chemical and meteorological variables were used, including the mixing ratios of O_3, NO_2, CO, and SO_2, along with T, RH, WS, WD, and solar zenith angle (SZA) to estimate the target species, HONO, as the output.

Line 194-196: The high IOA value signifies that the performance of RNDv1.0 is adequate, and it is capable of simulating the ambient HONO mixing ratio using the routinely measured criteria pollutants and meteorological variables.

Line 199-200: The RF model was constructed using the KFCV method and the same input variables as RNDv1.0

Line 233-237: Therefore, in the next step, the performance of RNDv1.0 was compared for the two cases by dividing the testing dataset into a group in which all input variables fall within the range of the training dataset and a group which does not meet this criterion. In RNDv1.0, there was no significant difference in performance between the two groups

Line 263-265: Thus, it is reasonable to state that RNDv1.0 constructed using routinely measured criteria pollutants and meteorological variables can sufficiently capture the HONO variability in the urban atmosphere.

Page 4, lines 105-107. The authors write that wind direction "should" be converted and there "should" be no missing values. From the text it seems clear that the authors have converted the measured wind direction and they have removed observations with missing values. I think the authors should rephrase the text so that it is clear what data conversion and selection steps have been done.

The relevant part is rephrased as follows.

Line 144-146: The WD in degrees was converted to a cosine value for continuity. In the last step of data processing, hourly measurement sets were removed from the input data set if any of the nine variables were missing.

Page 4, equation 1. I stumbled over the notation F1 and F2. It seems that these are simply the observed min and max of variable x. Why not denoting F1 and F2 as x_min and x_max? Would probably be more clear.

Yes, you are right. For clarity, the relevant part is reworded as follows.

Line 155-157: where x_{raw} is the raw data, x_{sca} is the scaled value, and the scale factors of F_1 and F_2 correspond to the maximum-minimum and minimum values of the input variable (X), respectively, which are listed in Table 2