Correspondence to editor's comments

I have gone through your response and it does yet answer the concerns of the reviewers. Both reviewers have raised the issue with generalization: how well would the model perform in other conditions. Your validation and testing datasets are very small covering quite narrow range of meteorological conditions. I understand that the amount of data you have is limited but still the concerns raised by the reviewers should be answered before the manuscript can be published in GMD. Now from the manuscript it appears that you have developed deep-learning model to your data but it does not show that you would have developed a model which is applicable elsewhere.

In 2021, we obtained additional measurement data sets during May-Jun and Oct-Nov, which were used to test the RNDv1.0. Therefore, the RND model was tested on measurements acquired in weather conditions different from those of the train dataset (Figure 3). The test results are presented in Figure 7(a): IOA = 0.68, MAE = 0.74, r = 0.55, and RMSE = 0.95. When the data in which at least one input parameters do not fall within the range of the train dataset is excluded from the test dataset, there is no significant difference 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).

It is particularly noteworthy that severe haze pollution events occurred in November 2021, when the daily average $PM_{2.5}$ concentration was raised up to 120 $\mu g\ m^{-3}$ and the HONO mixing ratio also increased to 4 ppbv or more in Seoul. Except for these extremes, RNDv1.0 traces well the variation of HONO mixing ratio.

It is good that you have tested the 1-layer ANN model. But in order to answer the reviewer concern you need to test also some simpler ML model(s) and add those to the manuscript. In you response you show comparison of you model and ANN, but this is for the training data. In general showing good correspondence with the training data (Figures 5-7 in the manuscript) does not tell how the model performs for "independent" datasets. Thus, after you have conducted additional simulations you should improve the model performance analysis with non-training data with proper scatter plots (similar to Fig A in the referee response) and statistical information (not just MAE and IOP but also RMSE, R). In addition, you need to answer the reviewer comment "Since the idea is to develop a model for others to use (of the shelf), it should be made very precise what are the capabilities and restrictions of using the developed model" by clearly stating the limitations and benefits of the model.

Pre-constructed 1-layer ANN model needs additional input parameters (boundary layer height and aerosol surface area), and unfortunately these data are not exist on test periods. Therefore as recommended, a random forest (RF) model was constructed using the same data and process of the RNDv1.0 construction and its results were compared with those of RNDv1.0, CMAQv5.3.1, and 1-layer ANN for the measurement data from 2016 KORUS-AQ campaign (Figure 5 & 6 and Table 3), and also for the test data (Figure 7). We are agreeing about your concern that “train” data should not be used to evaluate model performance in generally, so the comparison using 2016 measurement data was become a part of train-validation process (Figure 3).

Recently, we acquired HONO measurement data during May~Jun and Oct~Nov in 2021, so these data set are added in the test data set (Figure 3). By using this test data set which 2021 observation data added, the performance of RND1v1.0 and RF model were evaluated (Figure 7). The performance evaluation results using the test dataset, and the bootstrap (BS) test results of RNDv1.0 and RF clearly demonstrated that the ability of the deep learning model to simulate the HONO mixing ratio is more adequately in the urban atmosphere compared to the general machine learning model (Table 4). Statistical information including RMSE and r is provided for model evaluation (Figure 7 and Table 3).

In addition to these, the manuscript should be checked by language services as there are several issues with the language. In addition, at the end of page 7, the sentence on line 204 is unfinished.

The manuscript was thoroughly checked, and errors were corrected.

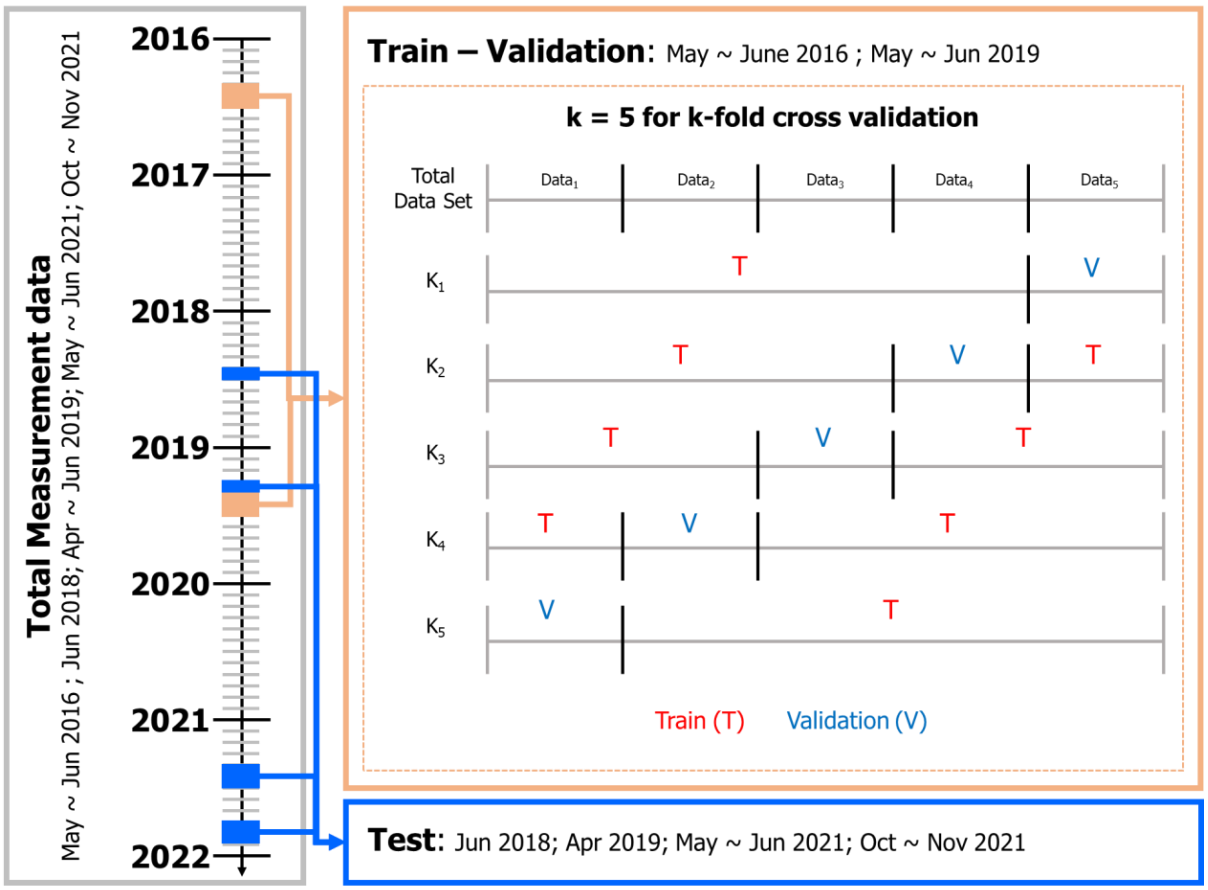

Figure 3. Design of training, validation, and test to build RNDv1.0 using measurement data. The k-fold cross validation was performed using randomly divided five subsets of training data set.

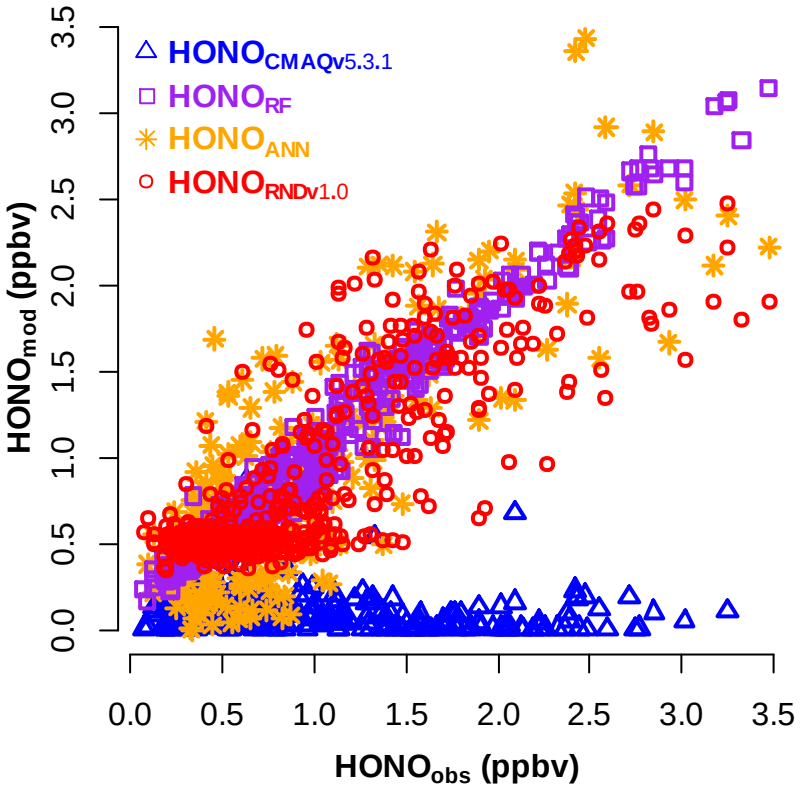

Figure 5. Comparison between measured HONO ($HONO_{obs}$) and calculated HONO ($HONO_{mod}$) using CMAQv5.3.1 (blue triangle), RF (purple square), ANN (orange star), and RNDv1.0 (red circle) during the KORUS-AQ campaign (may-June 2016)

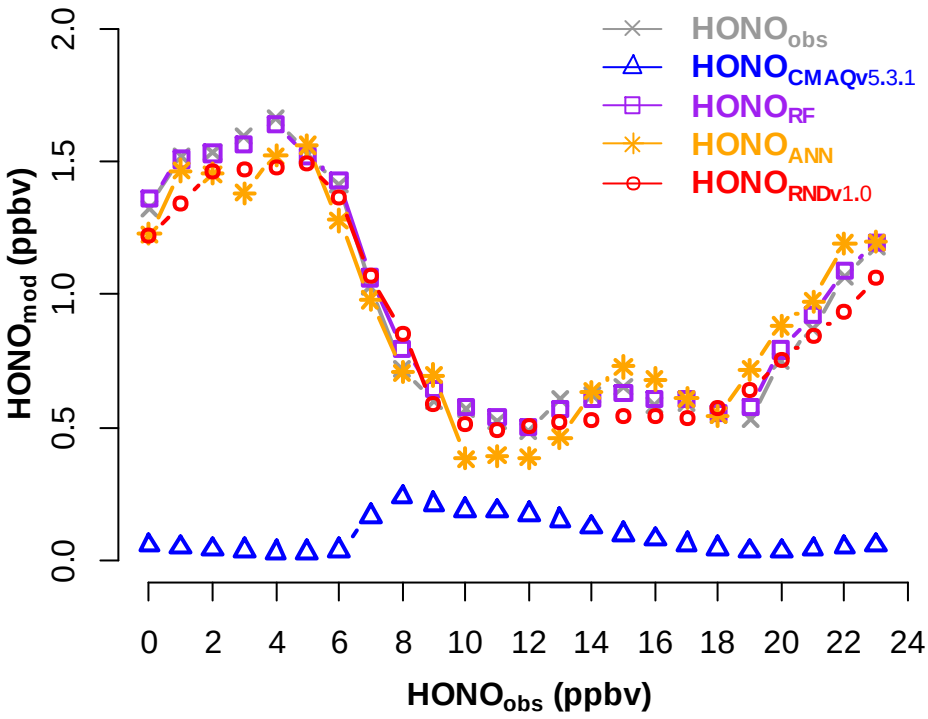

Figure 6. Average diurnal variation of measured HONO ($HONO_{obs}$) and calculated HONO ($HONO_{mod}$) using CMAQv5.3.1 (blue triangle), RF (purple square), ANN (orange star), and RNDv1.0 (red circle) during the KORUS-AQ campaign (may-June 2016)

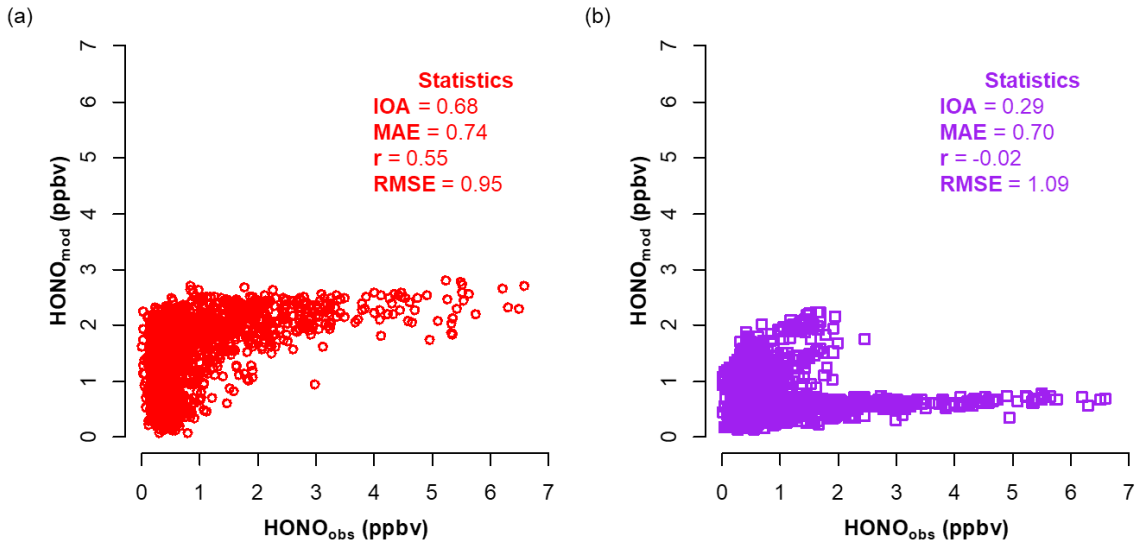

Figure 7. Relationship between measured HONO ($HONO_{obs}$) and modeled HONO ($HONO_{mod}$) using (a) RNDv1.0 and (b) a Random Forest model for the test dataset.

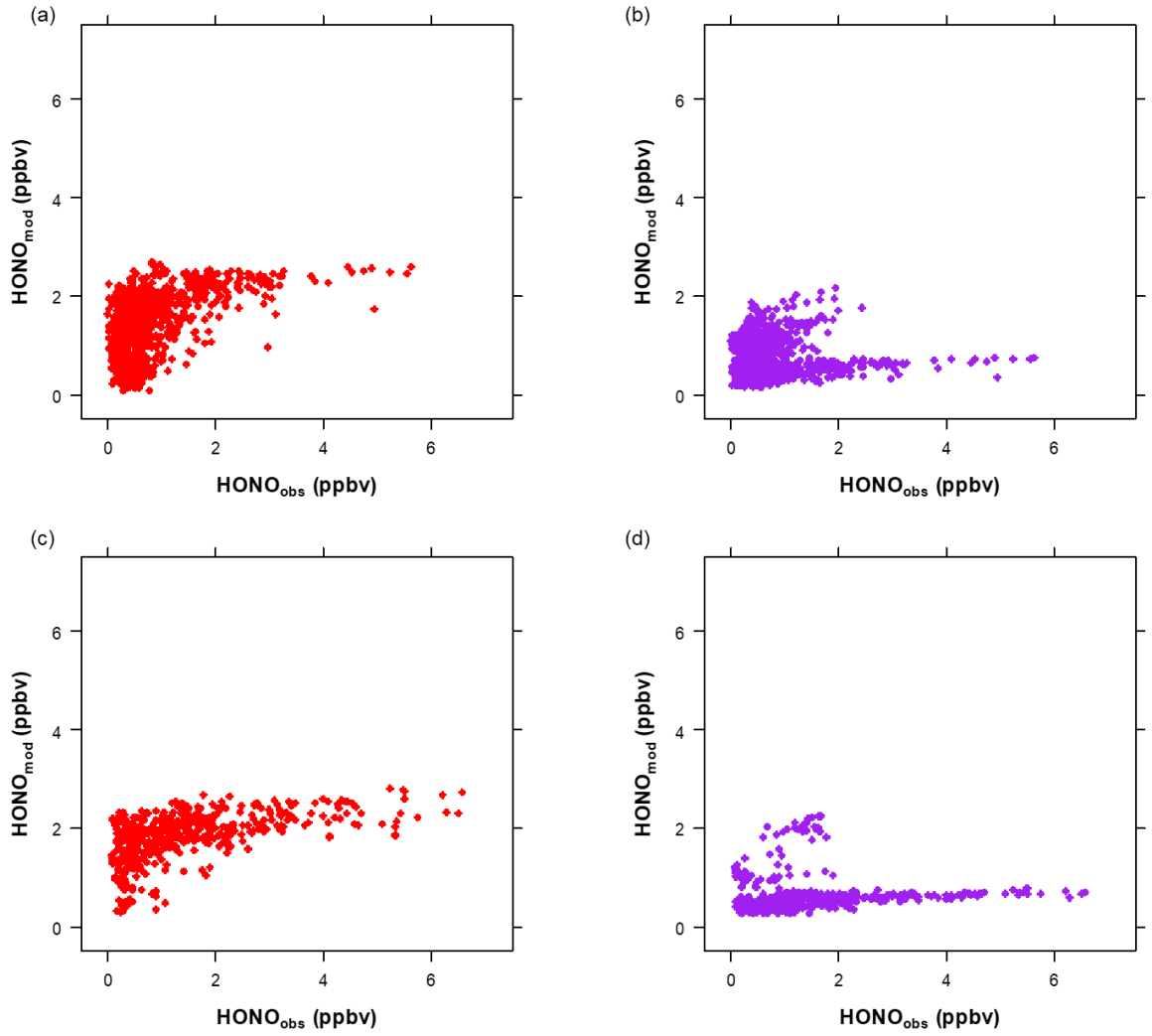

Figure S5. Relationship between measured HONO (HONOobs) and modeled HONO (HONOmod) using RNDv1.0 (red) and a Random Forest (purple) for the test dataset. (a) and (b) present data in which all input variables are within the range of the train dataset, and (c) and (d) are the others that do not meet the criteria.

Table 1. The performance of chemical transport model (CMAQv5.3.1) and machine learning (ML) models including Random Forest (RF), Artificial Neural Network (ANN), and RNDv1.0 on measurement data from 2016 KORUS-AQ campaign that were used for training.

	CMAQv5.3.1	RF	ANN	RNDv1.0
IOA	0.44	0.99	0.86	0.9
r	-0.07	0.99	0.81	0.84
MAE	0.82	0.1	0.38	0.27
RMSE	1.06	0.12	0.41	0.37

Table 4. The result of bootstrap test of measurement data used to train the RF and RNDv1.0 model. The greater the MAE, the greater the influence of variable.

Variable	RF		RNDv1.0	
	MAE	Feature Importance	MAE	Feature Importance
-	0.10	-	0.28	-
O_3	0.57	1	0.29	8
NO_2	0.24	4	0.59	1
CO	0.19	7	0.37	5
SO_2	0.17	8	0.34	6
Solar zenith Angle (SZA)	0.25	2	0.41	4
Temperature (T)	0.21	5	0.52	2
Relative humidity (RH)	0.25	3	0.52	2
Wind speed (WS)	0.20	6	0.34	6
Wind direction (WD)	0.13	9	0.29	8

Table S2. The performance of RNDv1.0 and a Random Forest (RF) model on the test dataset that is divided into 'in' where all input parameters fall within the range of the train dataset and 'out' that do not meet the criteria.

	RNDv1.0_in	**RF_in**	**RNDv1.0_out**	**RF_out**
IOA	0.71	0.28	0.82	0.73
r	0.55	-0.02	0.52	0.10
MAE	0.64	0.55	0.63	0.66
RMSE	0.86	0.87	0.96	1.24

The revised parts are as follows.

Line 24:

~ the several months from 2016 to 2021.

Line 25-27:

RNDv1.0 was constructed utilizing k-fold cross validation and evaluated with an Index Of Agreement (IOA), correlation coefficient (r), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

Line 76-77:

Monitor for AeRosols and Gases in ambient Ari (MARGA) (Xu et al., 2019),

Line 129-131:

The HONO mixing ratio was measured in Seoul using a QC-TILDAS system during May–June 2016, June 2018, and April–June 2019 (Lee et al., 2011;Gil et al., 2021) and a MARGA system during May–June and October–November 2021 (Gil, 2022).

Line 160-161:

Finally, 54.2 % of all available measurement data (2847) were used to construct and evaluate the RNDv1.0 in this study.

Line 187:

2.4. Model training and k-fold cross validation

Line 189-192:

The RNDv1.0 model was trained-and-validated and tested with HONO measurements obtained during May ~ June in 2016 and June in 2018, April ~ June 2019, and May ~ June and October ~ November in 2021, respectively (Figure 3). The number of data used for train-validation and test were 1122 and 1725, respectively.

Line 210-250: Re-write the train-validation-test chapter

The performance of RNDv1.0 was compared with that of other models, including Community Multi-scale Air Quality Model (CMAQv5.3.1, Appel et al., 2021), Random Forest (RF), and 1-layer Artificial Neural Network (ANN, Gil et al., 2021) using 2016 measurement data. A RF model was constructed using KFCV method and the same input parameters as RNDv1.0 (Figure S4). Their performance was evaluated by Mean Absolute Error (MAE), Root mean Square Deviation (RMSE), and Pearson correlation coefficient (r) (Eq. 4 - 6).

$$MAE = \frac{\sum_{i=1}^{n}|O_i - P_i|}{n}, \quad \text{(Eq. 4)}$$

$$RMSE = \sqrt[2]{\frac{\sum_{i=1}^{n}(O_i - P_i)^2}{n}}, \quad \text{(Eq. 5)}$$

$$r = \frac{cov(O,P)}{\sigma_O \sigma_P}, \quad \text{(Eq. 6)}$$

where σ and cov implies standard deviation and covariance, respectively.

The measured HONO mixing ratios correlated well with those calculated except for the CMAQ (Figure 5), which not only severely underestimated the measured HONO, but also

failed to represent the diurnal variation (Figure 6). The statistical information about the performance of the four models is given in Table 3. The mean HONO mixing ratio measured and calculated from CMAQ, RF, ANN, and RNDv1.0 was 0.94 ppbv, 0.09 ppbv, 0.95 ppbv, 0.88 ppbv, and 0.89 ppbv, respectively. Of the four models, RF showed the best performance, followed by RND. ANN has advantage of being able to calculate HONO more accurately than RND using more input variables, but it results in a lower data capture rate (41.5 %) compared to RND (97.7 %) or RF (85.3 %).

2.6. Model test

The RNDv1.0 and RF models were tested using June in 2018, April in 2019, and May ~ June and October ~ November in 2021 (Figure 3). Of the test dataset, the early winter (October ~ November) data is particularly valuable for demonstrating the applicability of the RNDv1.0 because they were produced in different weather conditions from those of the train dataset. Note that the RF performance was the best among the four models in train-validation process (Figure 5). Interestingly, the performance of RF was much worse than RNDv1.0 in test (Figure 7). The IOA and correlation coefficient of RF were extremely low (0.29 and -0.02, respectively), which are similar to or worse than those for CMAQv5.3.1 (Table 3).

The performance of RNDv1.0 was slightly lessened, but it well tracing the HONO mixing ratio. When the data in which at least one input parameters do not fall within the range of the train dataset is excluded from the test dataset, there is no significant difference 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). And this test dataset includes severe haze pollution events when the daily average $PM_{2.5}$ concentration was raised up to 120 μg m^{-3}, and the HONO mixing ratio also increased to 4 ppbv or more in Seoul. Except for these extremes, RNDv1.0 traces well the variation of HONO mixing ratio. These test results, therefore, are convincing evidence for the applicability of the RNDv1.0 to the estimation of HONO levels in the urban atmosphere.

Line 253:

2.6. Bootstrap test and feature importance

Line 265-272:

In contrast, O_3 was the most important variable for RF. This is likely due to the distinct inverse relationship between O_3 and HONO in the diurnal patterns and O_3 variations over a wide range. In conjunction with the evaluation of test presented in the previous section, the results of feature importance for the two models demonstrates the ability of the deep learning model to simulate the HONO mixing ratio more adequately in polluted urban areas compared to the general machine learning model. Thus, it is reasonable to argue that the RNDv1.0 constructed using routinely measured criteria pollutants and meteorological parameters can sufficiently capture the HONO variability in the urban atmosphere.

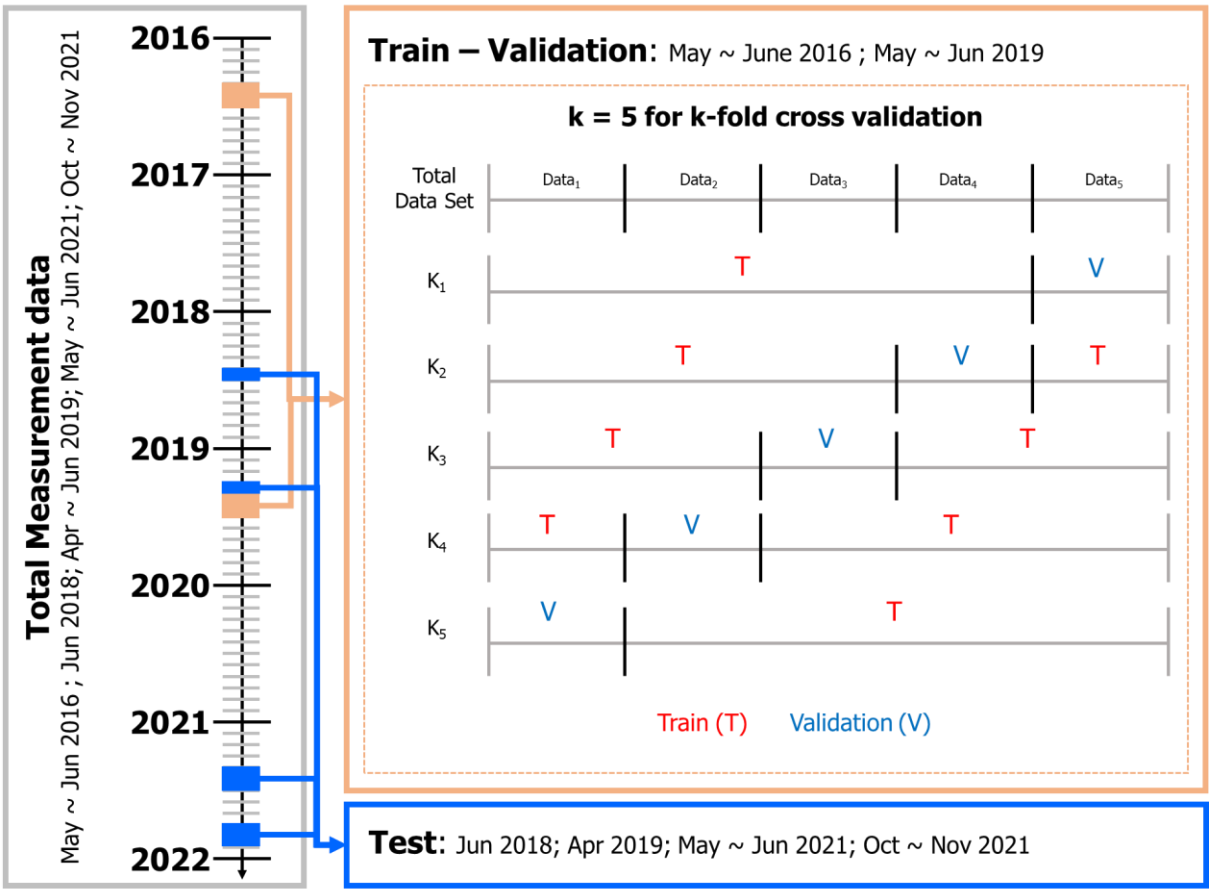

Figure 3. Design of training, validation, and test to build RNDv1.0 using measurement data. The k-fold cross validation was performed using randomly divided five subsets of training data set.

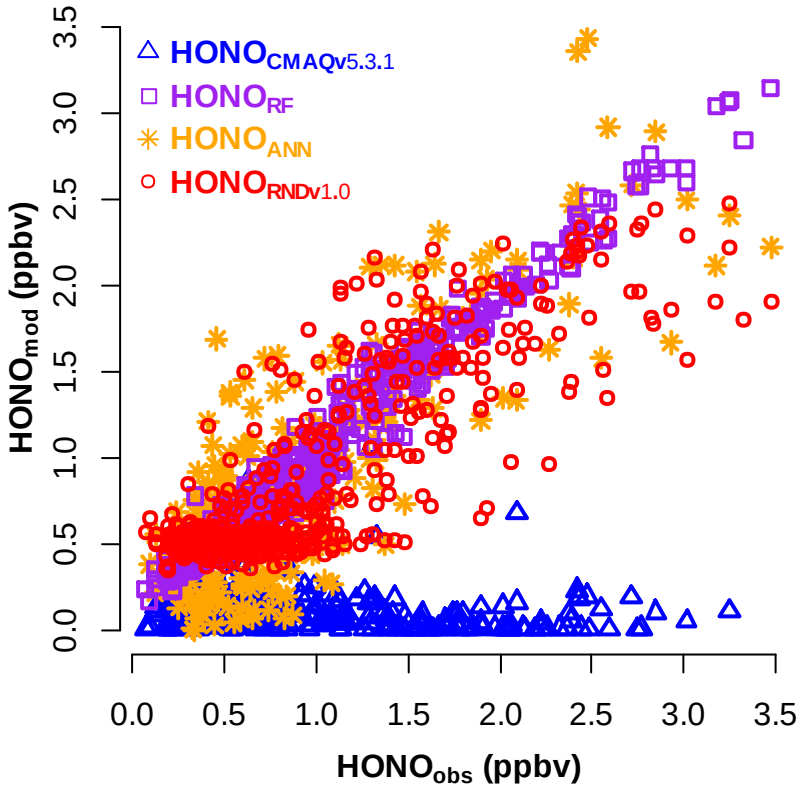

Figure 5. Comparison between measured HONO ($HONO_{obs}$) and calculated HONO ($HONO_{mod}$) using CMAQv5.3.1 (blue triangle), RF (purple square), ANN (orange star), and RNDv1.0 (red circle) during the KORUS-AQ campaign (may-June 2016)

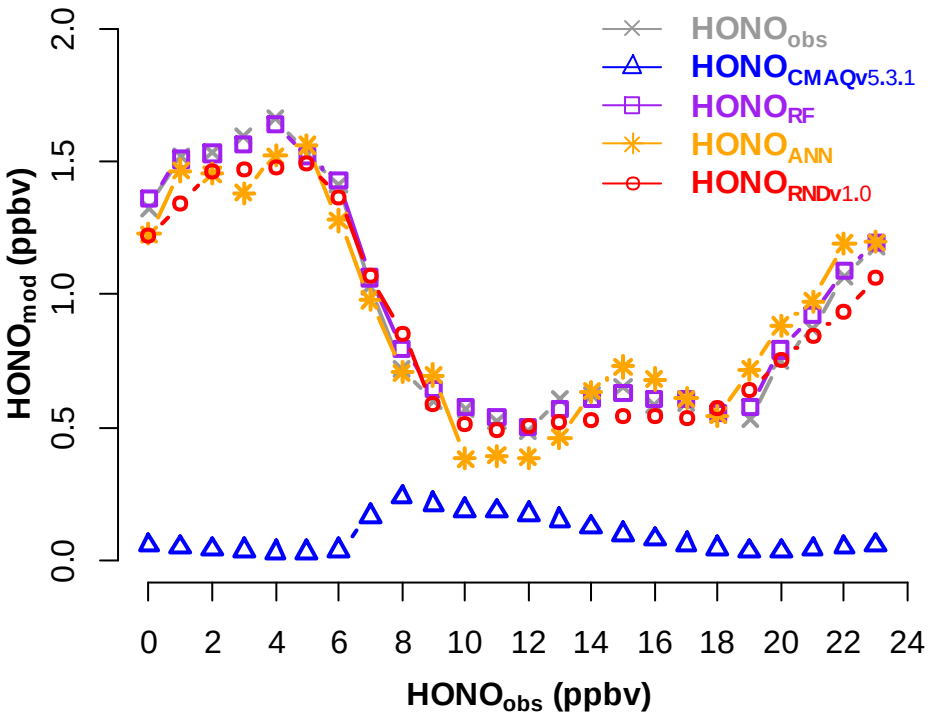

Figure 6. Average diurnal variation of measured HONO ($HONO_{obs}$) and calculated HONO ($HONO_{mod}$) using CMAQv5.3.1 (blue triangle), RF (purple square), ANN (orange star), and RNDv1.0 (red circle) during the KORUS-AQ campaign (may-June 2016)

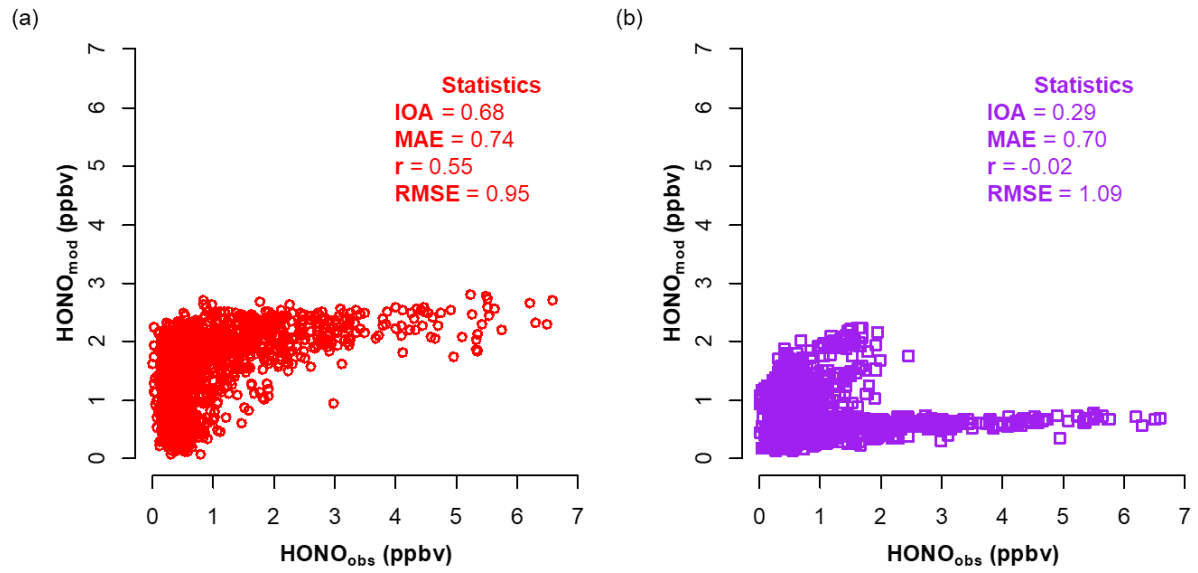

Figure 7. Relationship between measured HONO ($HONO_{obs}$) and modeled HONO ($HONO_{mod}$) using (a) RNDv1.0 and (b) a Random Forest model for the test dataset.

Table 2. The performance of chemical transport model (CMAQv5.3.1) and machine learning (ML) models including Random Forest (RF), Artificial Neural Network (ANN), and RNDv1.0 on measurement data from 2016 KORUS-AQ campaign that were used for training

	CMAQv5.3.1	RF	ANN	RND
IOA	0.44	0.99	0.86	0.9
r	-0.07	0.99	0.81	0.84
MAE	0.82	0.1	0.38	0.27
RMSE	1.06	0.12	0.41	0.37

Table 4. The result of bootstrap test of measurement data used to train the RF and the RNDv1.0 model. The greater the MAE, the greater the influence of variable.

Variable	RF		RNDv1.0	
	MAE	Feature Importance	MAE	Feature Importance
-	0.10	-	0.28	-
O_3	0.57	1	0.29	8
NO_2	0.24	4	0.59	1
CO	0.19	7	0.37	5
SO_2	0.17	8	0.34	6
Solar zenith Angle (SZA)	0.25	2	0.41	4
Temperature (T)	0.21	5	0.52	2
Relative humidity (RH)	0.25	3	0.52	2
Wind speed (WS)	0.20	6	0.34	6
Wind direction (WD)	0.13	9	0.29	8

Supplementary:

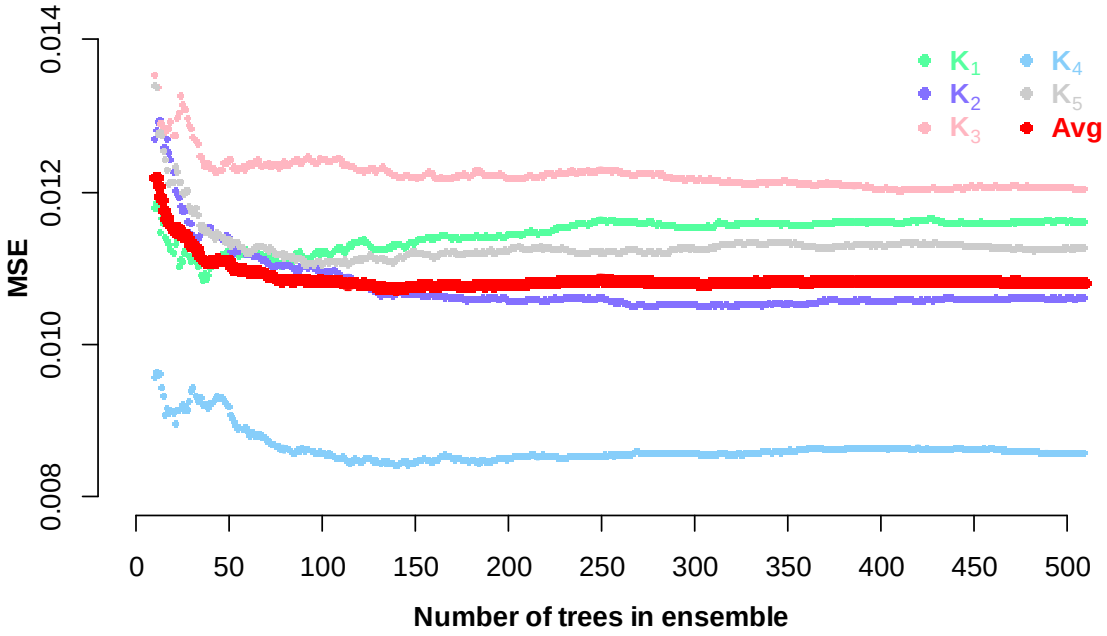

Figure S4. K-fold cross validation for Random Forest (RF) model by changing the number of tress in ensemble.

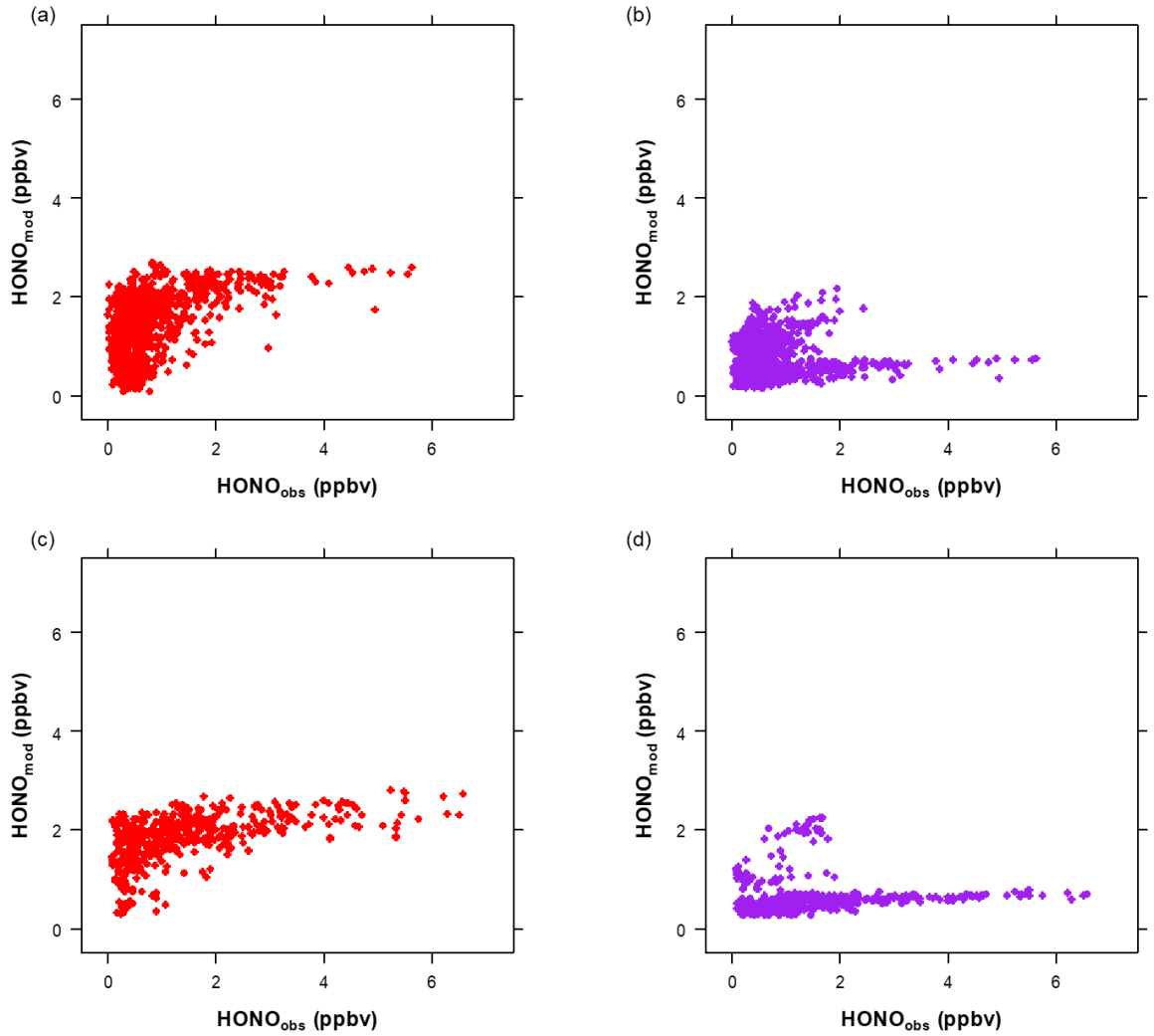

Figure S5. Correlation between measured HONO (HONOobs) and modeled HONO (HONOmod) using RNDv1.0 (red) and RF (purple) using test data set. (a) and (b) present data in which all input variables are within the range of the train dataset, and (c) and (d) are the others that do not meet the criteria.

Table S1. The range of measurement variables used in this study.

Maximum	**O_3** (ppbv)	**NO_2** (ppbv)	**CO** (ppbv)	**SO_2** (ppbv)	**SZA** (°)	**T** (°C)	**RH** (%)	**WS** (m s^{-1})	**HONO** (ppbv)
Train-Validation	205.6	82.3	1112.5	13.4	126.512	32.9	99.1	7.586	3.5
Test	132.0	88.5	1500.0	10.4	163.600	32.5	100.0	8.179	6.6
Minimum	O_3 (ppbv)	**NO_2 (ppbv)**	**CO (ppbv)**	**SO_2 (ppbv)**	**SZA** (°)	**T** (°C)	**RH** (%)	**WS** (m s^{-1})	**HONO** (ppbv)
Train-Validation	0.8	2.4	137.3	1.0	14.195	8.6	10.6	0.010	0.01
Test	1.0	1.7	165.6	0.1	14.900	-2.2	9.7	0.270	0.10

Table S2. The performance of RNDv1.0 and a Random Forest (RF) model on the test dataset that is divided into 'in' where all input parameters fall within the range of the train dataset and 'out' that do not meet the criteria.

	RNDv1.0_in	**RF_in**	**RNDv1.0_out**	**RF_out**
IOA	0.71	0.28	0.82	0.73
r	0.55	-0.02	0.52	0.10
MAE	0.64	0.55	0.63	0.66
RMSE	0.86	0.87	0.96	1.24