Responses to Reviewer2:

We would like to thank Fei Liu for her review of our manuscript. We have addressed her comments and suggestions and updated the manuscript accordingly. Our responses are indicated in red below the comments.

The authors develop a machine learning algorithm to infer surface NO2. The manuscript is well-written, and the results appear robust. I recommend the paper for publication after minor revisions.

General comments:

1. The authors state, "Currently, to our knowledge, there is little machine learning done to derive surface concentrations in less populated areas such as Canada." While this points out the motivation of the work, it would be helpful to expand on why this is significant. What differences between more populated regions (e.g., China, Germany) and less populated regions (e.g., Canada) would justify the need for this study? This additional context could better motivate the work.

   We expanded the text in the manuscript further to justify the work and the importance of NO2 in background areas:

   "However, to our knowledge, few studies have focused on applying such methods in less populated regions such as Canada. This gap is significant because the conditions in Canada differ significantly from those in densely populated regions: surface monitoring networks are sparser, emission sources are different compared to urban areas, and the limited spatial extent of measurement networks in northern Canada contribute to the challenge. These differences highlight the need to adapt and evaluate ML approaches under these conditions."

2. Section 2.4.2. I found it surprising that, in addition to NO_x emissions, SO_2 and NH_3 emissions were included as predictors. Have you tested the sensitivity of your model to these additional species? It would be worthwhile to elaborate on why these variables are expected to play a role in predicting surface NO_2. In Figure 3, I observe that NO_2 emissions are far more significant compared to other parameters, including TROPOMI NO_2. This observation seems inconsistent with the conclusion that "This is a feature that we explicitly wanted to see in the random forest prediction, as this means the random forest function is primarily driven by the satellite measurements of NO_2." Clarifying this potential discrepancy would strengthen the argument.

   Figure 3 includes the feature importance of all parameters used, including NH3 and SO2. We included the following sentence into the manuscript to clarify this more:

   "While the NO2 and NO emissions are directly related to the NO2 surface concentration, other pollutants can be helpful in the machine learning model as well. SO2 emissions are typically an indicator of industrial activity, such as refineries potentially affecting the surface NO2 concentrations differently compared to urban emissions. While NH3 emissions are typical indicators of agricultural and some industrial activities nearby, NH3 can also impact the

formation NOx (Pai et al., 2021). The importance of the various parameters is discussed in Sect. 3 (Fig. 3)"

With regards to the discrepancy in the text, we changed the sentence to:

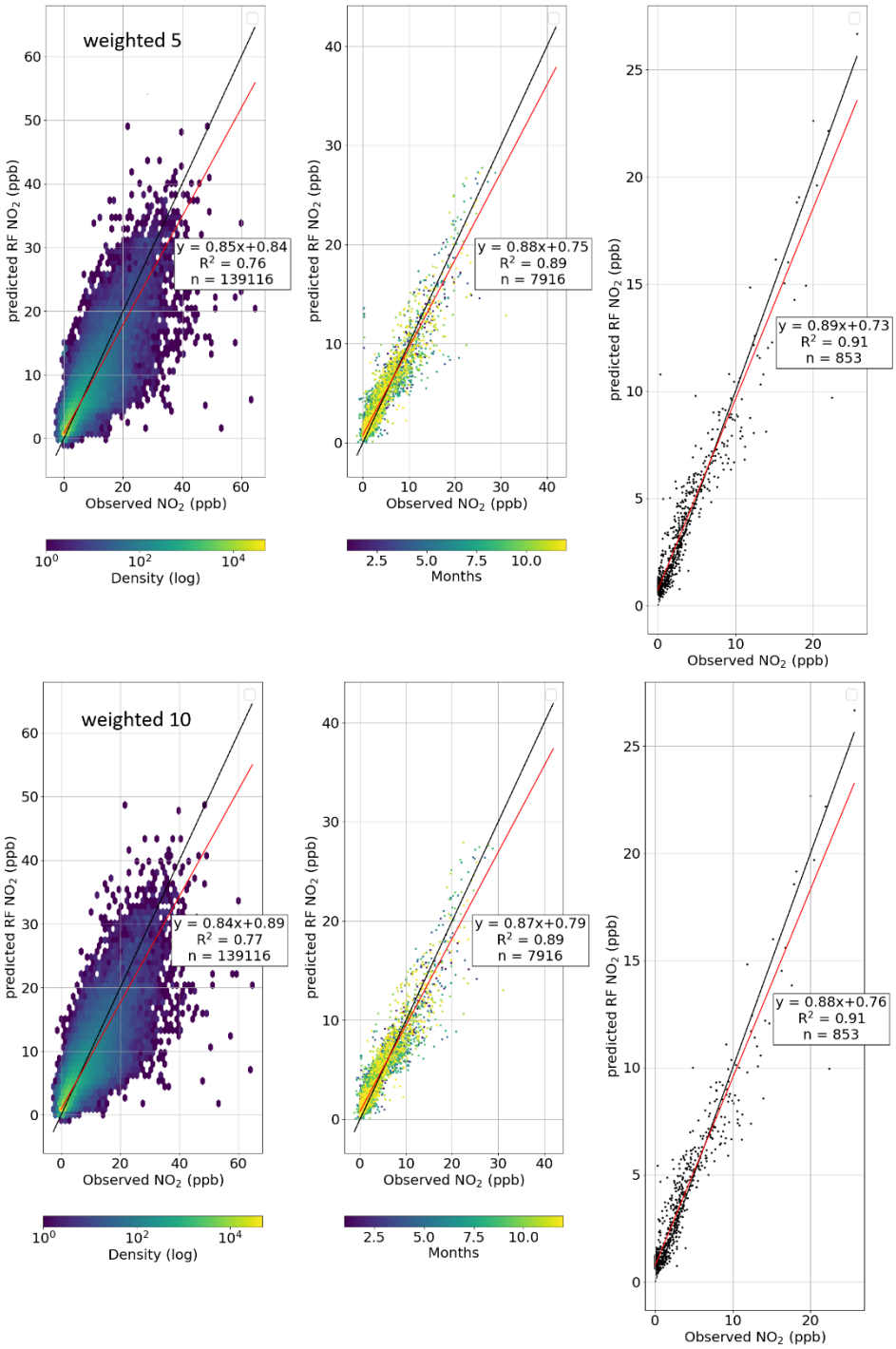
"…that the TROPOMI tropospheric column measurements are among the most important parameters for the prediction of the surface NO2 concentrations."

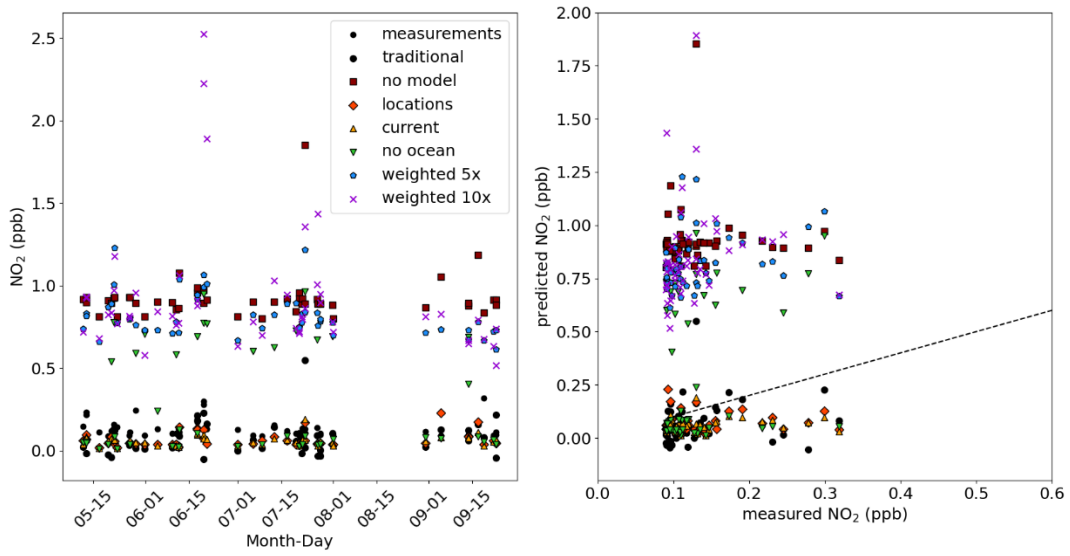
3. The authors mention that NO2 surface concentrations from the GEM-MACH model were used to augment the training dataset in remote areas. Would assigning greater weight to rural monitoring stations achieve a similar effect? If this has not been tested, it would be worthwhile to evaluate whether such an approach could improve the model's predictions for less populated regions.

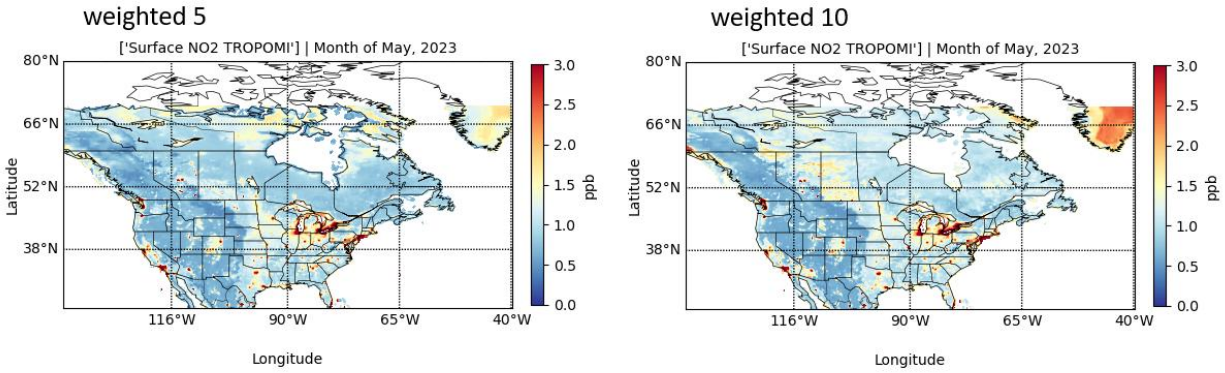
   We would like to thank Fei for the suggestion, we tested out weighting the background stations (stations with 0 population around) in two ways, weighting them 5 times more than all other stations and 10 times more, respectively. We created similar figures as Fig. 2,4, and 5. While Fig. 2 suggests very similar results as using the synthetic station data for the training, but looking at rural areas on the map or compared to the CAPMoN rural site it performs much worse. The results are shown below.

   This is an interesting analysis and worthwhile to elaborate on this in the manuscript. We included a brief discussion on this into the manuscript:

   "Additionally, we tested weighting stations that are in non populated areas five and ten times more than other stations. This method did not work as well as including synthetic stations, the NO2 surface concentrations in northern Canada were still too high even when rural stations are weighted more in the random forest training (see Fig. A4)."


weighted 5
y = 0.85x+0.84
R² = 0.76
n = 139116
predicted RF NO₂ (ppb)
Observed NO₂ (ppb)
Density (log)
y = 0.88x+0.75
R² = 0.89
n = 7916
Months
y = 0.89x+0.73
R² = 0.91
n = 853
weighted 10
y = 0.84x+0.89
R² = 0.77
n = 139116
y = 0.87x+0.79
R² = 0.89
n = 7916
y = 0.88x+0.76
R² = 0.91
n = 853





Specific comments:

1. line 170. 1e14 shall be superscript.

   We have changed it accordingly to: "…order of 10^{14} molec/cm2 (as a comparison, VCDs in polluted areas are on the order of 10^{16} molec/cm2)."