We'd like to thank the editor for handling our manuscript, as well as reviewer #2 for reading our manuscript and providing numerous helpful suggestions for improvement.

We have carefully read through all the comments and questions and revised the manuscript accordingly. Please find our point-by-point response to reviewer #2 below. Here, the reviewer's general and specific questions/comments are formatted to be left-aligned text in bold font. Our responses are indented and formatted in regular font.

Here is a summary of the major changes in the revised manuscript:

1) Table 2 reports ANN performance metrics for both the validation and independent test data set.
2) We added additional information on the ranges for each hyperparameter and computational costs to section 3.
3) We added explanations on why the temperature ANN model appears to be more complex than other models.
4) 2) Tables 2 and 3 report the respective ANN performance metrics (RMSD, bias, and percentile differences) for each species in both their natural units (K, ppmv, ppbv), as well as percentages.
5) We added a subsection on data quality assessment to section 3.
6) We discuss areas in the global maps, where the ANN-NRT algorithm exhibits clear underestimations.

General comments

This paper presents new near real-time products of the Aura Microwave Limb Sounder (MLS) using artificial neural networks (ANN-NRT). The ANN-NRT show good performance and demonstrates the potential of applying machine learning to generate NRT products. The paper is clearly written and the study is well explained. I recommend the manuscript for publication, but I have some minor comments.

(1) Global maps show ANN-NRT is better than OE-NRT, but more discussion should be given to the special area of that ANN overestimates or underestimates.

These maps were originally thought of as simple examples. However, the reviewer is correct that we mainly focused on regions were the ANN-NRT performed well compared to the OE-NRT algorithm. We agree that it is only fair to point out areas where the ANN underperforms. However, we need to emphasize that these maps are generated from MLS observations sampled on a single day, which requires an area-weighted interpolation of the MLS orbit track. Also note that the discrete color bar can exaggerate discrepancies.

We added some additional discussion to the revised manuscript. Here, we emphasize regions where the ANN-NRT shows some larger discrepancies to the L2 results and mention possible reasons. First, we added this to the H_2O discussion:

"A notable exception is the area of increased H_2O over India and parts of Southeast Asia, where the ANN-NRT underestimates the L2-retrieved concentrations. This region is characterized by strong and deep convection during the monsoon months that affects the sampled radiance profiles and may introduce uncertainties into the ANN model predictions. Maps of 100.00 hPa-H_2O concentrations on other days during that week indicate that slight underestimations persist in this area; however, the ANN-NRT predictions generally are much closer to the L2 results than are the OE-NRT retrievals."

We also highlight an area with pronounced O_3 underestimations:
"The only obvious difference is the area of low concentrations over Antarctica, which is completely missed by the OE-NRT algorithm and is overestimated (in area) by ANN-NRT. Note that profiles sampled in this region are affected by radiances that are reflected by the surface (see Fig. 7d in Werner et al., 2021 and the relevant discussion), which might impact the reliability of the ANN predictions."

Finally, we added this part to the conclusions:
"Global maps of predicted H_2O and O_3 concentrations indicate that model performance may be affected by the presence of strong, deep convection, as well as by strong surface reflections over Antarctica. While the respective predictions agree better with the L2 retrievals compared to the OE-NRT results, more analysis is needed to explore potential improvements to the ANN setups."

Such improvements might be achieved by increasing the sample importance for cloudy profiles (i.e., telling the model to emphasize these profiles during training) or by adding additional features that indicate cloudiness.

(2) For performance evaluation of T model, I think it is more intuitive to use unit K rather than relative values. At least it should be described in the paper.

Tables 2 and 3 in the revised manuscript now summarize both Kelvin/ppmv/ppbv, as well as percentages. This not only provides more intuitive numbers for the temperature model, but also puts some of the large percentages for SO_2, HNO_3, and N_2O into perspective (i.e., there are very low concentrations at certain levels).

Specific comments

Line 96: I know brightness temperatures sampled over 2005–2022 are very large. However, it is better to describe the exact amount of input features for training, validation, and test data.

At that point in the manuscript, we wanted to give a very general overview of the theory and necessary steps to setup and train ANN models. Moreover, the exact number of samples varies from species to species due to the (i) differently sized data records, and (ii) number of successful MLS level 2 profile retrievals.

However, we agree that this is an important fact to cover in the revised manuscript. Therefore, we added the respective number of samples to table 1 and changed the relevant sentence in the manuscript text to: "It also provides details on the features that make up the input matrix for each ANN-NRT model, namely the start and end dates that define the training data record for each model, the number of total samples in that data record (determined by the number of successful profile retrievals), and the respective MLS bands, channels, and MIFs."

Table 1: The number of neurons of T and O3 are much larger than other products, is it necessary? Why choose so many neurons instead of adding hidden layers? The MBS of T (i.e. 8192) is much larger than the others (i.e. 32), it should be discussed.

These discrepancies can be explained by the following reasons:

1) Development on the ANN-NRT models started because we were unsatisfied with the performance of the previous OE-NRT temperature results. Therefore, we initially only intended to replace the temperature product and to continue using OE-NRT for all other species. As a result, we almost overengineered that specific model and did not mind the immense computational costs associated with almost >5,000 neurons per layer. We also were content with increasing the mini-batch size to 8192, even though this required a significant amount of memory. We only cared about developing the very best model possible.
2) We made a mistake in Table 1; the O_3 model only has 400 neurons.
3) Regarding the number of neurons: we varied those between 100 and a predefined maximum, in increments of 100. We set that maximum to $\frac{2}{3} \cdot$ (the number of features

+ the number of labels), which is a widely-used (somewhat empirical) threshold. Increasing the number of neurons after that point usually makes very little sense; our experience confirms these findings.

Frankly, neither the large number of neurons or the large mini-batch size for the temperature model are necessary. In fact, as long as the number of neurons is ≥400 per layer, the overall performance metrics change very little (e.g., $\Delta R < 0.01$). Once we decided to also train models for the other NRT species, we decided to keep the mini-batch size lower to ease the computational costs regarding the amount of memory, as we found little to no improvement for the performance metrics. However, we decided to keep the already trained temperature model the way it was.

We added the considered ranges of each hyperparameter to section 3.1:

"We considered the following ranges and settings: J_{HL} = [1, 2], J_N = [100, 200, · · ·, 2/3·(n+k)] per hidden layer, AF=["relu", "tanh"], LRP=[n/a, 1e−6, 5e−6, 1e−5, · · · , 1e−1], GNS=[n/a, 1e−3, 5e−3, 1e−2, · · · , 1], and MBS=[32, 64, · · · , 8192].

We also added information on the computational costs of the training procedure:
"The computational costs associated with the training procedure of each ANN-NRT model, while dependent on the respective hyperparameters and size of the m x n input matrix, are generally as follows: it takes about one month to develop and train each ANN, using 12 CPUs and requiring ~ 100 GB of memory."

Finally, we added an explanation on why the temperature model is so much more complex:
"Note that the model setups for T, CO, and SO_2 differ from those of the other species. The T model is considerably more complex with comparatively high values of J_{HL}=5,078 and MBS=8,192. The ANN-based estimator for temperature was developed before those for the other products, with less regard for computational cost than was present in the subsequent development. The computationally more expensive temperature model is "overbuilt", but had already been trained so was used in this version of the NRT products."

Line 189: The SO2 statistics in Table 3 are based on the observations which were also included in training data set. So, the comparison of OE and ANN doesn't make much sense. Is there no other data for comparison?

We wanted to present statistics for the data covered in Figs. 2-3 and to present model performance for enhanced concentrations due to volcanic activity. However, we agree that the evaluation of the SO_2 model performance is problematic due to the inclusion of trained data. We acknowledge that fact in the manuscript when we say:
"Of special note is the ANN-NRT setup for sulphur dioxide (SO2). Volcanic eruptions are the primary source of stratospheric SO_2. As a result, we decided to train the SO_2 ANN model on MLS observations around major volcanic eruptions, namely those of Kasatochi, Calbuco, Sarychev, Nabro, Raikoke, and Hunga Tonga-Hunga Ha'apai (e.g., Pumphrey

et al., 2015; Millán et al., 2022). While ANN-NRT performs well in reproducing elevated SO_2 concentrations associated with the Hunga Tonga-Hunga Ha'apai eruption, the training data is limited and the model may suffer from overfitting (i.e., learning specific characteristics of known eruptions well to the detriment of generalization)."

We agree that adding more information about model performance for actually unseen data are necessary. Therefore, we added performance metrics for predictions in May 2022 to table 3, as well as the following sentence in the manuscript text:
"Note that two sets of SO_2 statistics are shown: one set based on MLS observations in January 2022, which are affected by the Hunga Tonga-Hunga Ha'apai volcanic eruption and were included in training data set, and a second set based on samples in May 2022 with no volcanic influence."

We also developed a second SO_2 ANN-NRT model, where the training data set is based on all MLS observations over 01/01/2005–04/30/2022. We included performance metrics for the test data set in table 2, as well as the following discussion in the revised manuscript:
"As mentioned in section 2, stratospheric L2 retrievals in the absence of elevated levels of SO_2 can be considered noise, and comparisons between L2 and ANN-NRT results are difficult (R<0.26 and bias >11%). If the training data set is increased to include all MLS retrievals between 01/01/2005 and 04/30/2022, instead of just focusing on periods of volcanic activity, the associated correlation coefficients and biases slightly improve to 0.37 and <7%, indicating a better ability to predict noise. However, further analysis indicates that this model performs slightly worse for profiles containing elevated SO_2 concentrations; correlation coefficients for such profiles in the test data set are decreased by about 0.05 (R=0.52 compared to R=0.57), while the RMSD increases by about 0.31 ppbv (5.72 ppbv compared to 5.41 ppbv). Since the main objective of the SO_2 NRT is to detect volcanic activity, we decided to employ the model trained on the reduced (volcanic only) data set."

Line 237: All metrics get better with the increasing data except the absolute bias in Fig. 5(c), it should be discussed.

Unfortunately, after closely analyzing the different predictions and metrics, we frankly do not have a good explanation on why the bias does not decrease with increasing training data set size.

One possible explanation is that the observed biases between predictions are very small, especially compared to other species. Looking at the new table 3, the mean bias for the full data set is 0.16 K (<0.1%). The bias range for the smaller training data sets is 0.11-0.13 K, which is very similar. Keep in mind that these are absolute biases, so the true average is even smaller at 0.05–0.06 K. At this point, the biases for all ANN-NRT models are close to being negligible and increasing the size of the training data set does not further reduce the bias.

Note that the average of the 99th percentile of the difference between the L2 retrievals and ANN predictions decreases with an increase in training data size; illustrated in Fig. 1 of this reply. We find similar results for the 1st percentile of the difference.

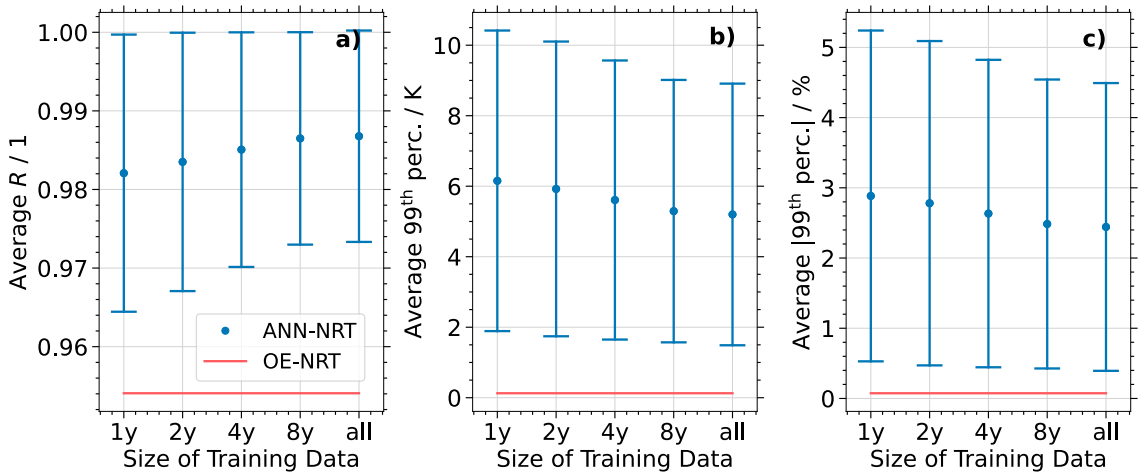

Fig. 1: (a) Correlation coefficient as a function of size of the training data set. (b) Similar to (a), but for the average of the 99th percentile of the difference between L2 and predicted temperatures. (c) Similar to (b), but shown as a percentage.

We unfortunately do not have another explanation; all predictions look very similar. Even though this lack of an explanation is rather unsatisfactory, we added some extra discussion about the small biases and the 99th percentile differences to the revised manuscript:

"A very small increase in the averaged absolute biases for the *T* models is observed. However, these absolute biases are in the range of 0.11-0.16 K (0.05-0.06 K if both positive and negative biases are averaged) and can be considered negligible. Note that similar analysis for the 1st and 99th percentile of the difference between MLS L2 retrievals and each ANN-NRT model prediction shows a monotonically decreasing behavior with increasing training data size."