We'd like to thank the editor for handling our manuscript, as well as reviewer #1 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 #1 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.

The authors present a near real-time processor of Aura/MLS observations using a supervised neural network. The manuscript is easy to follow and shows that the processor has very good performance, very close to the operational processor. The new method presents a significant improvement compared to the previous near real-time processor based on a simplified optimal estimation method. I recommend the manuscript for publication, but I have minor comments that could be clarified by the authors.

General comments

1) I am impressed by the results overall, and more particularly with the ability of the model to capture the increase in H2O induced by the volcanic eruption, though the statistical weight of such events in the training dataset should be low. This illustrates the high potential of the model to capture special disturbances that occur over a restricted spatio-temporal range. However, I found that such abnormal conditions are not sufficiently discussed in the manuscript. Indeed, these are scientifically the most interesting cases but have a low impact on the overall statistical evaluation. For example, in Figure 4b, the increase in H2O at 100 hPa over India and part of Southeast Asia is clearly underestimated with ANN-NRT. This should be discussed in the manuscript and the authors should mention if they have found other cases where significant discrepancies were seen.

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.

Regarding the H_2O at 100 hPa example (Fig. 4b), the underestimations over India and Southeast Asia on that day are on the order of 0.5 ppmv and the OE-NRT algorithm seems to perform a little bit better. However, if we look at maps from two other days in that same week, shown in Figure 1 of this reply, we can see that ANN-NRT clearly outperforms OE-NRT in this region (as well as over Central America). While better than OE-NRT, the ANN again seems to underestimate the L2 results. Note that this is also indicated in Fig. 2h of the manuscript, where $H_2O > 5$ ppmv seem to be underestimated during that month. These apparent systematic departures of ANN H_2O from the L2 training set in the presence of strong convection warrant further investigation (although it will be hard to improve the respective ANN model, due to the statistical nature of machine learning approaches).

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.

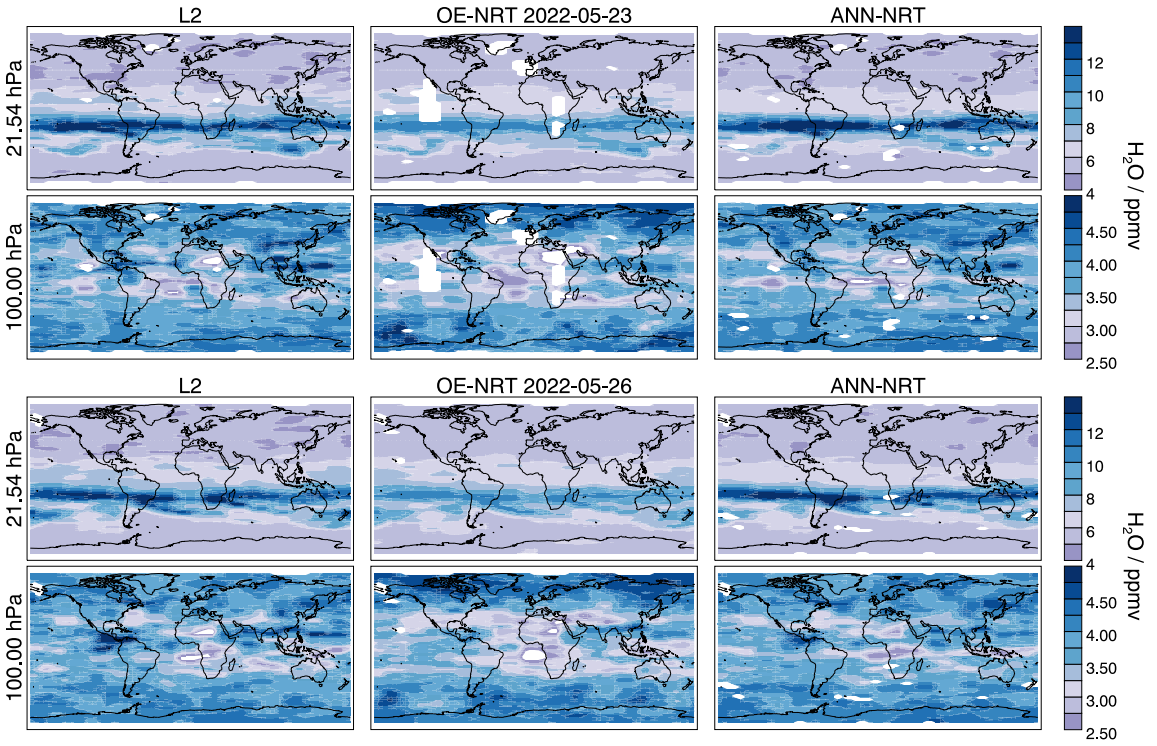

Fig. 2: Comparisons of L2 O_3 retrievals and ANN-NRT predictions at 100.00 hPa and 21.54 hPa.

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) More generally, the authors do not show results for the whole test dataset (5% of 17 years corresponds to almost 1 year), in particular winter time which is strongly disturbed in the northern hemisphere. Is there a seasonal pattern in the results? Authors should clarify why the test data are well suited for describing the capability of the model and the limitations of such a choice (that could further be investigated in future studies). For instance, I would personally have used 2 entire years with very different conditions (e.g., SSW strength or QBO phase) to test the models.

We should have been clearer about the purpose of the independent test data set. The examples shown in the results section were not drawn from the test data set, but instead are new predictions made after each model was finalized. Note that the temperature model was trained on data sampled between 01/01/2005 and 05/31/2021. Meanwhile, Figs. 2–3 show comparisons between L2 retrievals and ANN predictions for July 2021.

The purpose of the independent test data set, and the validation data set to a certain extent, is indeed to evaluate model performance and test the ability of the model to generalize. For the temperature model, we have about 5 years and 1 year worth of profiles in the validation and test data set, i.e., about 4.8 and 1 million profiles. However, they do not comprise a continuous 5-year period or a single year of observations. Instead, these profiles are picked randomly from the full distribution and therefore cover all years, seasons, and geographical regions. If there is a close agreement between the performance metrics for the validation and test data set the model is able to generalize well for previously unseen data. Large discrepancies indicate poor model performance and unreliable predictions.

The example data in Figs. 2–4, as well as the figures in the appendix, are simply examples to illustrate performance going forward, i.e., after the models have been trained and evaluated. For example, it would not be possible to create maps like those in Fig. 4 of the manuscript from the test data set alone, because (statistically) only 5% of profiles of each individual day (~175) are part of the test data set (i.e., >5200 profiles from each month since 01/01/2005). Similarly, the analysis in Fig. 5 is based on a fixed independent data set. They are not used for model evaluation, although all predictions from now on could technically be considered an extension of the original test data set (i.e., an ever-growing amount of previously unseen profiles).

Note that we are constantly monitoring ANN-NRT performance. Fig. 2 of this reply shows an example of L2 O_3 retrievals vs ANN-NRT predictions from 04/09/2023 for two pressure levels. Similar to the metrics in the test data set, correlation coefficients are high with R>0.99 and very low biases <0.01 ppmv.

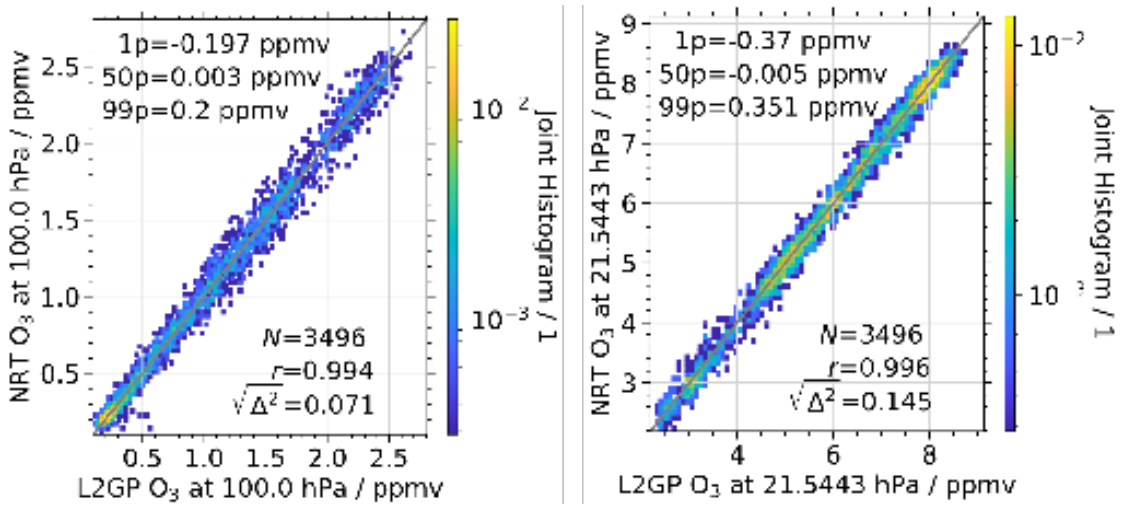

Fig. 2: Comparisons of L2 O_3 retrievals and ANN-NRT predictions at 100.00 hPa and 21.54 hPa.

We made a several changes to the revised manuscript to make these points clearer:

1) We include the metrics for the validation data set in Table 2 and directly contrast them with the metrics for the test data set. They are very similar.
2) We added the following description to the manuscript text:
   "These scores were derived by comparing the ANN-NRT predictions with the respective MLS L2 results for all MAFs in both the validation and an independent test data set. The distinction between the two is important. Following the discussion in Ripley (1996) and Russel and Norvig (2009), the validation data is used for hyperparameter tuning and to prevent overfitting during model training. To truly evaluate the performance of a trained model, a completely independent test data set is necessary. However, the performance scores for the validation and test data set should be similar and large discrepancies are an indication that the trained model does not generalize well (i.e., the model performs worse for previously unseen data). Note that of the ~ 3500 daily profiles MLS observed since 01/01/2005, ~ 875 and ~175 randomly selected samples are included in the validation and test data set, respectively."
3) We've added the following statement at the beginning of section 4:
   "This section presents comparisons between MLS L2 profile retrievals and the respective OE-NRT and ANN-NRT predictions. These observations were made after the respective ANN-models were developed, trained, and evaluated and serve as examples of model performance going forward."

3) Regarding the vertical resolution of profiles predicted with ANN-NRT. This issue is not addressed in the manuscript and could be clarified. If I understand the NN setting correctly, the vertical resolution of the predicted profile is the same as that of the level 2 operational product (here I am referring to the resolution derived from the operational averaging kernels and not the retrieval levels spacing). Am I right? This could be clarified.

ANN-NRT retrievals are trained to duplicate the L2 operational OE retrievals, and thus have vertical and horizontal resolution no better than that inherent in the OE retrievals' averaging kernels. However, the production OE retrieval uses multiple, overlapping scans of the atmosphere to "tomographically" retrieve a set of adjacent profiles, while the ANN-NRT relies upon radiances only from the nearest radiometric scan of the atmosphere to retrieve a given profile. This difference between 2D and 1D radiance inputs would be expected to have significant impact on horizontal (along-track) resolution and more subtle impacts on vertical resolution, but as ANN-NRT retrievals do not produce averaging kernels, it is difficult to make quantitative comparisons. This is a topic for further research beyond the scope of this paper.

The ANN-NRT models perform a mapping between (1) The MLS L1B brightness temperatures (sampled at 125 minor frames, or scan levels) and the operational L2 data products at their respective 37 or 55 retrieval levels (depending on the species). The models do not approximate any parts of the forward model or retrieval algorithm.

In other words, in the training phase, the temperature ANN learns the relationship between MLS brightness temperatures sampled in different bands/channels/minor frames and the operationally retrieved T at 100 hPa, 82.5 hPa, and 68.1 hPa, as well as the respective retrieval levels below and above. In the prediction phase it subsequently provides an estimate of T at these exact levels, thus providing an estimate of the eventual operational L2 profile retrieval.

We've added a short summary at the beginning of section 3 of the revised manuscript:
"This section described the theory, training process, settings, performance evaluation, and data quality assessment of the updated, ANN-based NRT algorithm. The goal is to train ANN models on all valid MLS L2 standard product retrievals over 01/01/2005–04/30/2022 and their associated, nearest brightness temperature profiles. Since the MLS L2 standard products are used as labels (i.e., ``truth'') during training, the best-case output of each ANN is a computationally-inexpensive, high-fidelity preview of the L2 profiles."

We've also added the following clarification to section 3.1:
"Here, the labels are values from individual profiles of a specific MLS retrieved L2 atmospheric constituent. Therefore, the size of k is determined by the number of retrieval levels of the respective MLS L2 product."

4) For low SNR cases, the authors mentioned that the NN tends to smooth the noise compared to the operational product. Is this effect could be related to a degradation of the vertical resolution similar to the regularization effect in the OE method?

"Smoothing the noise" can be thought as more of a symptom than anything the ANNs actively do. What actually happens is that the models fail to establish a successful mapping between the features and labels, i.e., the model cannot determine any meaningful relationship between the input and output.

One can test this easily with a few simple examples. In a first test, we trained a model to predict a sine curve pattern, ie., the features are angles between 0 to 360 degrees and the labels are the sine of the features. Note that, since this is just a demonstration, we did not tune any of the hyperparameters and instead used some default settings of one hidden layer, 1 neuron, a "relu" activation function, and L2 regularization with a parameter of 5e-4; the split between training, validation, and test data is 90/10/10%.

Figure 1a of this reply show the results of this test. The test data (orange dots) nicely follow the original sine curve (blue dots) and the correlation coefficient is 1.00. If we set all input features to 0, the model will fail to establish a successful mapping between the features and labels. This is illustrated in Fig. 1b of this reply. Here, the model basically predicts 0 for all angles, i.e., the average value. Some slight deviations from 0 can be observed occasionally, which can likely be attributed to (i) insufficient regularization in the model, and/or (ii) imbalanced training and test data (i.e., they draw from slightly different distributions).

A second experiment simulates conditions closer to noisy MLS L2 retrievals, shown in Fig. 1c of this reply. Here, the features are again values between 0 and 360 degrees, but we set the labels to random values between 0 to 1, with an average of 0.5. Again, the model will fail to establish a successful mapping between features and labels and will simply predict the average value at all times. This is shown in Fig. 1d of this reply.

Note that this behavior is similar to other machine learning architectures, like GBDT, where the model attempts to predict residuals from an average value.

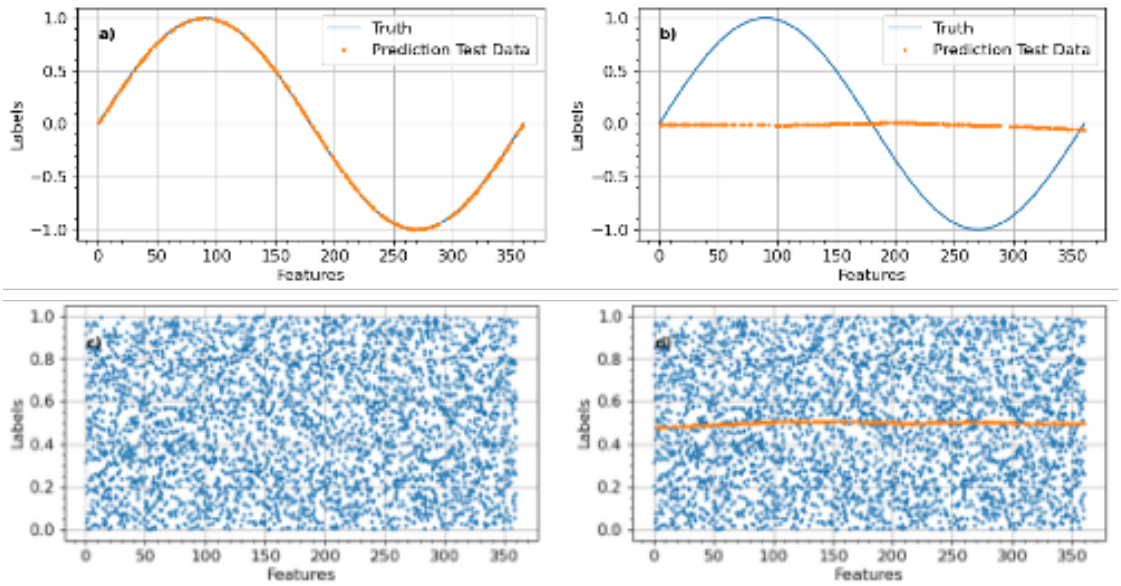

Fig. 3: Demonstration of ANN predictions for ill-defined problems.

We slightly tweaked a statement in the abstract:
"…, where the ANN models fail to establish a functional relationship and tend to predict zero."

Specific comments

Line 87: "n" is already used to define the number of input features. It would be clearer if another letter is used for the number of neurons per hidden layer.

Thanks for noticing. We switched the index to "j" (a common letter to describe an index) and the total number to a capital "J", both in the manuscript text and in Table 1.

Line 93: is the levels of the predicted profile the same as the number of levels of the operational product?

This is correct. The labels of each ANN model are the respective operational retrieval products. Therefore, the predicted profiles exist at the exact same vertical levels as the MLS L2 products.

We've added a short summary at the beginning of section 3 of the revised manuscript:
"This section described the theory, training process, settings, performance evaluation, and data quality assessment of the updated, ANN-based NRT algorithm. The goal is to train ANN models on all valid MLS L2 standard product retrievals over 01/01/2005–04/30/2022 and their associated, nearest brightness temperature profiles. Since the MLS L2 standard products are used as labels (i.e., ``truth'') during training, the best-case output of each ANN is a computationally-inexpensive, high-fidelity preview of the L2 profiles."

We've also added the following clarification to section 3.1:
"Here, the labels are values from individual profiles of a specific MLS retrieved L2 atmospheric constituent. Therefore, the size of k is determined by the number of retrieval levels of the respective MLS L2 product."

Table 1: I understand that the hyperparameters are defined by a set of tests but the differences between the models could be discussed. Why the number of hidden neurons is much smaller for the H2O model than for T and O3? Why is the tanh activation preferred over Relu for some species? (It is considered that Relu make the training more efficient)

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 (as well as a "tanh" activation function).
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.

Regarding the use of the "tanh" vs "relu" activation functions: We found that for almost all models the performance was determined by the combination of activation function and normalization. Apart from the O_3 model, the use of "relu" only produced higher performance scores when combined with Gaussian noise layers. Whenever L2 regularization was associated with higher performance (or no regularization was preferrable, like for the CO models) it was in combination with "tanh" layers. Note that we are not saying this is a universal characteristic, but something unique to the MLS NRT setup. Similar to our findings for the temperature model: differences in the performance metrics between the different model setups were very small as long as we had a sufficient number of neurons per hidden layer. However, the reviewer is correct: the models with "relu" activation functions converged a lot faster than the "tanh" models. This is one of the large benefits of the "relu" activation function. Of course, there are also some disadvantages, like the fact that neurons with negative values get eliminated.

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."

Line188/Table 3: Are the scores calculated for the same periods as Figure 3?

Yes, these metrics were calculated for the same time period as in Fig. 2, 3, A1, and A2. We clarified this in the revised manuscript:
"A summary of average performance metrics is given in Table 3, derived for the same time period as is used in Figs. 2–3 and Figs. A1–A2. Specifically, the presented metrics are: *R*, the average absolute RMSD, and average absolute bias for each species and the two NRT algorithms, as well as the averages of the 1st and 99th percentile of the differences to L2 (as a proxy for the minimum and maximum deviations)."

Line207: "Here the ANN … , and the results are close to L2 data": there is a clear underestimation of the H2O vmr over india and East-Asia. This issue could be mentioned and what could be the reason?

- Yes, these are random.

Line 219/Line 241: Would it be possible to complete a small training dataset with simulated data?

Yes, this could be done. Testing performance from a model that was trained on simulated data would be an interesting analysis. It would prove the feasibility of providing NRT products for a completely new instrument, for which no previous data record exists. Since machine learning approaches are statistical in nature, using a wide array of synthetic composition profiles and radiance data should in theory provide reliable predictions. Such an approach is not dissimilar to calculating look-up tables of synthetic observations for a wide range of viewing geometries and cloud variables in MODIS-like cloud property retrievals. As long as the radiances accurately describe the actual (noisy) observations and the set of composition profiles cover a wide array of possible atmospheric states, that approach should yield reliable results. Again, a retrieval approach based on look-up tables is very similar.

We think that such an analysis goes far beyond the scope of our paper and would be best suited for a separate study. For the MLS NRT retrieval we fortunately did not require synthetic data sets due to the large MLS data record.

Some preliminary thoughts on such a new study are:
1) The synthetic profiles need to be representative of actually observed profiles, and should cover as much of the full dynamic range as possible. We would expect to need an order of magnitude of 100,000 profiles to develop a reliable model.
2) For each of these profiles we need to simulate the relevant MLS radiance observations, which is computationally expensive.
3) ANNs might not be the ideal machine learning architecture for such an application. They tend to learn a specific problem very well, such as measurement uncertainties

and idiosyncrasies in the applied forward model and inversion algorithms, which result in uncertainties in the retrieved composition profiles. Synthetic data might look just different enough compared to actual measurements/retrievals that the ANN predictions become unreliable. Other architectures, such as Gaussian Process Models or Gradient Boosted Decision Trees, are more robust with regard to noisy data.

We think this topic is well worth exploring in a separate study, but that it requires a lot of additional efforts and considerations.

However, we ran a small test to, at the very least, confirm the feasibility of such an approach. Instead of creating a large set of possible atmospheric states and running a forward model on each to create synthetic MLS radiances, we used simulated radiances for day 51 in 1996 as input for our ANN-NRT temperature model. That data set is part of our testing procedure for the MLS retrieval algorithm. Note that the ANN-NRT models were trained on the relationship between a set of noisy MLS radiances and noisy MLS L2 retrievals. Applying these models on noise-free radiances and climatological temperature profiles introduces considerable uncertainties.

The results are shown in Fig. 4 of this reply. Panels a and b show scatter plots of predicted vs modelled temperatures at 100.00 hPa and 21.54 hPa, respectively. While model performance is worse compared to our analysis for actually observed MLS radiances and retrieved temperature profiles, it still performs reasonably well. Correlation coefficients are 0.95 (100.00 hPa) and 0.93 (21.54 hPa). The RMSD>2 K is in the range of the results in table 3 of the manuscript. These metrics are also worse than the ones based upon a single year of MLS observations (see Fig. 5 of the manuscript).

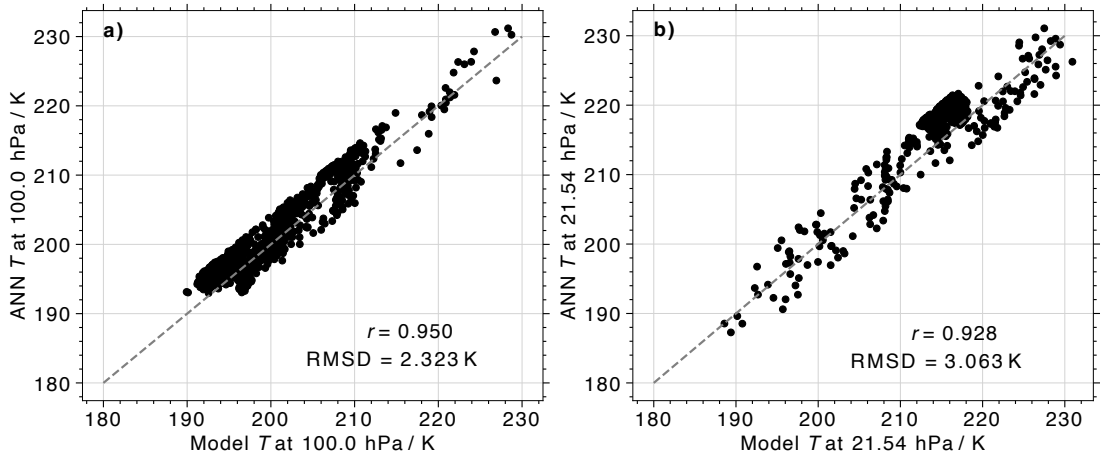

Fig. 4: Model performance for simulated temperature profiles and radiances.

Line 244: I don't understand the sentence "The previous version…". Do the authors mean: The previous version of MLS NRT data products (OE-NRT, Lambert et al., 2022) is replaced with predictions from an artificial neural network (ANN).

This is indeed confusing. The ANN approach was developed and implemented in phases, starting with the temperature ANN model and only later extended to cover all other NRT species as well. An ANN-based model has been used operationally for NRT temperature

since the end of 2021, as documented in the previous version of the MLS NRT user guide.

However, we don't think this distinction is necessary and will only confuse potential readers. We therefore simplified the first paragraph of the conclusions in the revised manuscript to:
"The previous version of MLS NRT data products (OE-NRT) is replaced with predictions from an artificial neural network (ANN). This manuscript describes the setup and evaluation of ANN models for all MLS NRT species. Starting in January 2023, all MLS NRT data products are based on this new approach (ANN-NRT)."