Review #1 Response

We thank the anonymous reviewer #1 for their thoughtful, detailed review of the manuscript, as it will improve the quality of the manuscript. Our response to each comment is provided below.

> *Is it possible that the NN is training on other factors and creating a water vapor simulation, because water vapor in the stratosphere is highly influenced by dynamics which is part of the NN training data. To truly demonstrate that the NN is retrieving water I suggest an experiment: fix the water vapor concentration to climatology. Then run the NN using the radiance variations in the other bands also inputting temperature and pressure. I suspect you will get the results shown in Fig 6 up to Hunga even though water vapor is not varying.*
>
> *I would also like to see a regression plot where the variation in the water vapor at the different levels is regressed against the various bands. This is a kind of standard step in feature engineering for machine learning. I suspect you will find the highest correlation between temperature and pressure and the other bands are contributing little. This should tell us if the NN is actually using water band variations.*
>
> *1. Channel Selection and Feature Engineering*
>
> *Figure 1 appears to demonstrate this, but it lacks a legend explaining the color code for the weighting functions. It looks like the weighting functions (dln(I)/Dln(H2O) are near zero before Hunga so it isn't surprising that the factor increase will be large.*
>
> *Given the potential for varying sensitivity, why not perform feature selection or regression analysis to identify the most informative channels? Why is it necessary to use 12 OMPS channels as inputs? What happens if you fix the temperature and pressure?*

We have updated Figure 1 to include the legend and adjusted the colors of each line to hopefully allow this to be better differentiated.

While a regression analysis is typical for feature engineering, it assumes that there is a linear relationship between the (possibly transformed) input-output pairs, but this is often insufficient for more complex problems, specifically problems that have complex non-linear relationships or highly correlated features, which is the case for this problem. Past work to use regression to determine a relationship between LP radiances and co-located H_2O profiles was unsuccessful, indicating the relationship between LP radiances and H_2O is more complex than can be captured by a simple regression analysis.

It is not strictly necessary to use 12 OMPS channels as inputs. In earlier stages, we also used ~50 channels, and while the results were similar, they were slightly worse than when we limited the number of channels. With straylight affecting LP's longer wavelengths, it's possible that the additional wavelengths complicated the relationship and inhibited the NN from learning to properly account for that, but this is a minor effect considering the similarity in results. The important aspect is that a wide range of wavelengths is used to capture the spectral behaviors of different aerosol and scene reflectivity conditions, which enables the NN to differentiate these effects from H_2O.

However, we have performed several tests that seek to answer the feature importance question through other means.

Our initial model setup allowed for a simple test of perturbations in the temperature/pressure profiles. In March 2025, there was a switch in the LP ancillary product from using the GEOS FP-IT data to the new GEOS-IT product, which exhibited a discontinuity in the temperature data on the order of a few degrees Kelvin. If the temperature/pressure data were primarily driving the H_2O predictions, then it would be expected to see differences between the model trained on the GEOS FP-IT temperature data but applied to the GEOS-IT temperature data, vs. a model trained exclusively on GEOS-IT data. We reapplied our methodology using the new GEOS-IT product throughout training and find our results for water vapor predictions in 2025 unchanged, indicating that the NNs are robust to small perturbations in temperature data.

We additionally investigated this question by training NNs without LP radiances or solar zenith angles, training NNs using climatological temperature/pressure profiles, and training on only LP radiances. When omitting LP radiance and solar zenith angles from training, we find that the model performs significantly worse, with larger root mean square errors and smaller R^2 values when applied to the test set (see Figure R1.1 below). The resulting tape recorder plot has worse agreement with the MLS tape recorder than what is presented in the manuscript, especially in the first weeks after the Hunga eruption as well as in 2025 (see Figure R1.2 below). Conversely, when using climatological temperature/pressure profiles, we find that the RMSE and R^2 values over the test set agree with those presented in the manuscript. Additionally, training on only the LP radiances also achieves similar RMSE and R^2 metrics as those presented in the manuscript. These results indicate that while the temperature/pressure data are useful, they are less important than the radiances when solving this problem. This is detailed in a new Appendix B.

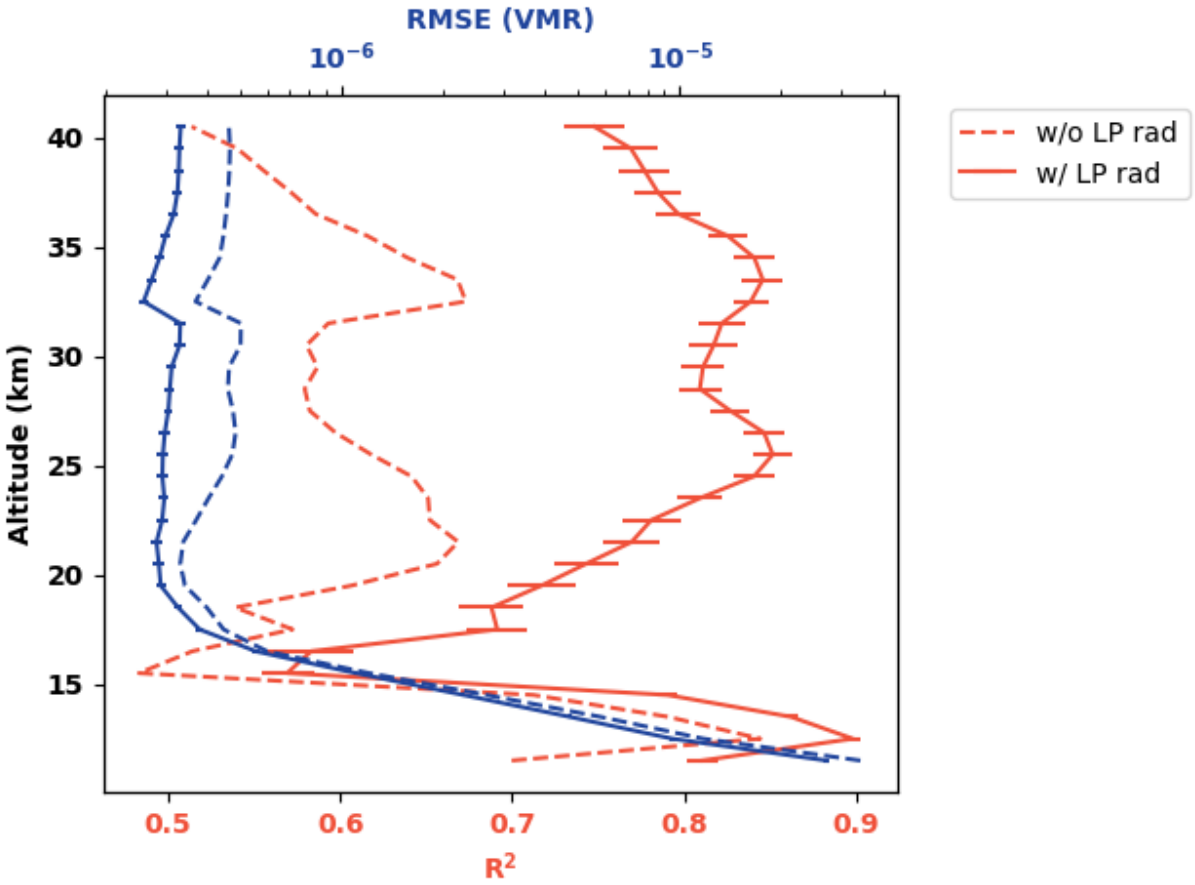

Figure R1.1. Like Figure 2 in the manuscript, except additionally showing the performance metrics for a NN trained on only temperature and pressure data (dashed lines). The degraded performance in the stratosphere above 15-17 km suggests that the LP radiances provide important information that enables the determination of a better solution for retrieving stratospheric water vapor.

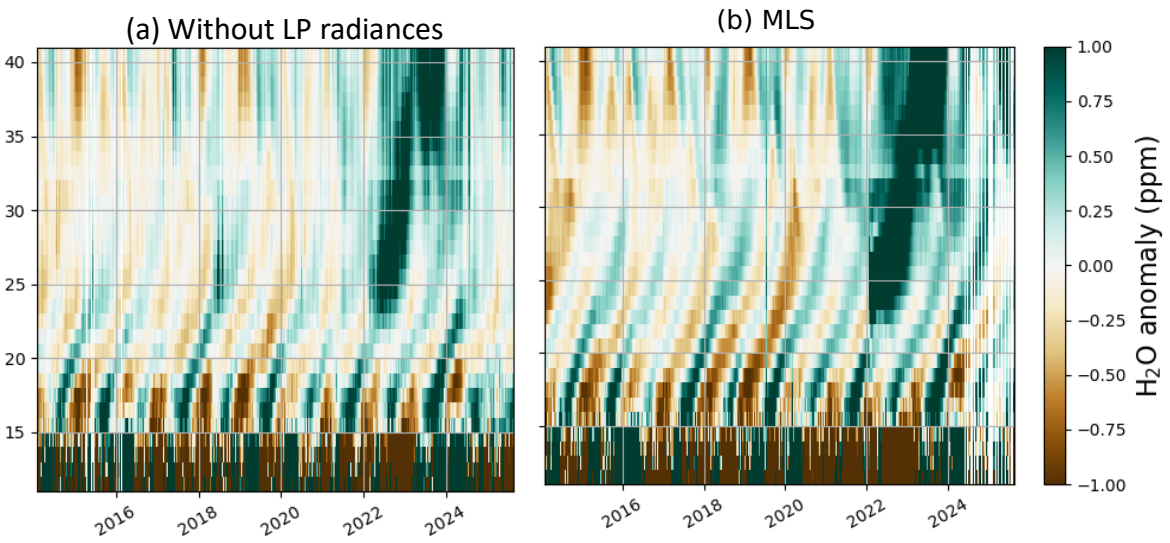

Figure R1.2. Like Figure 6 in the manuscript, except panel (a) shows results for the model trained only on temperature and pressure data.

2. Model Evaluation by Latitude

I recommend including performance metrics (e.g., RMSE, bias) as a function of latitude, which may also capture dependence on solar zenith angle, given its inclusion in the input dataset.

The percent bias between LP water vapor predictions and MLS per latitude is provided in the original manuscript; see Figure 4a.

We have added an additional panel to Figure 2 to show the relative RMSE and R^2 as a function of latitude as recommended. We find that the RMSE throughout the vast majority of the stratosphere is ~1/10 of the mean VMR. Below the tropopause, the RMSE is on the order of or larger than the mean VMR. For convenience, we provide that new panel below as Figure R1.3:

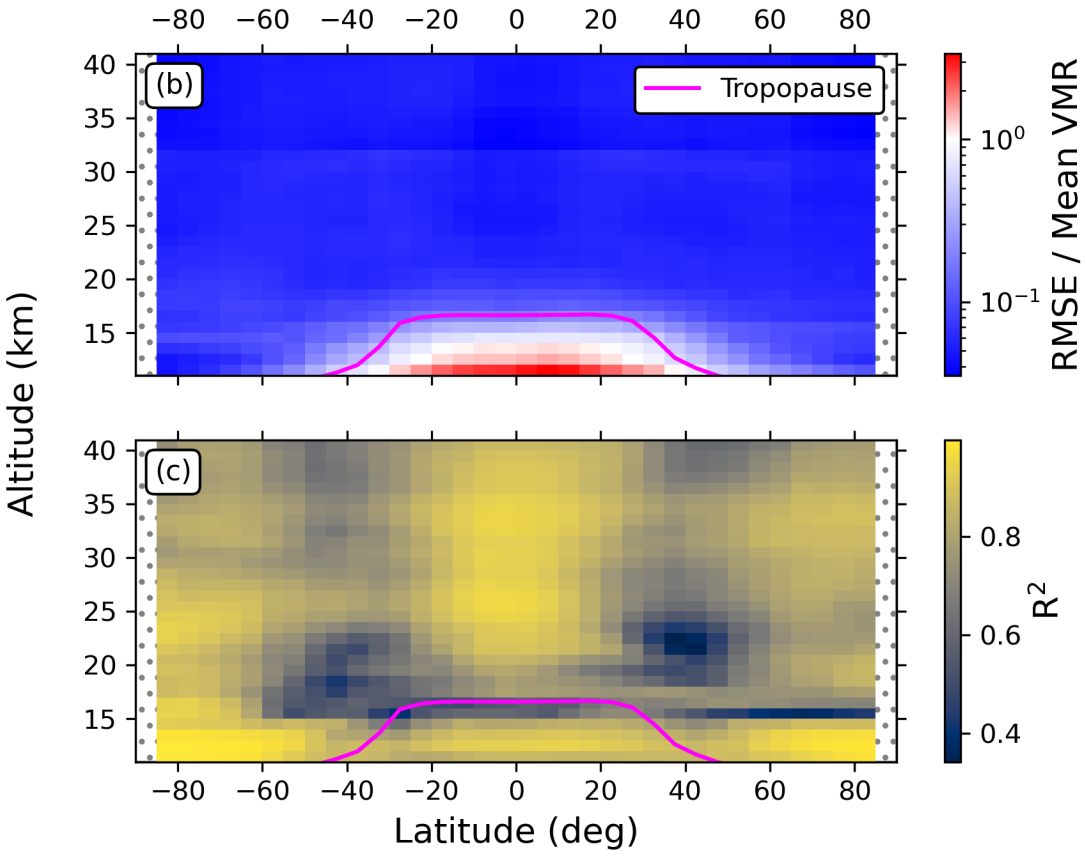

Figure R1.3. Plots of **(a)** relative RMSE and **(b)** R^2 as a function of latitude. The relative RMSE is shown on a logarithmic scale to better differentiate the transition in performance near the tropopause as well as minor variations in the stratosphere.

3. Ensemble Model Clarification

You mention determining ensemble size based on prediction stability. Is the ensemble size consistent across all profiles, or determined dynamically?

What differentiates each ensemble member, like architecture, initialization, or hyperparameters?

The ensemble size is constant across all profiles. As mentioned on lines 119-120, the architectures are identical among all ensemble members. Members are only differentiated by their random initialization. We have added additional text to better clarify this:

"Using the chosen architecture, we train an ensemble of 10 neural networks using a mean-squared-error loss function**; members are only differentiated by their random initialization**. **The size of the ensemble is held constant for all retrievals.**"

4. Normalization and RMSE Interpretation

What are the units in Figure 2? Does Figure 2 use absolute RMSE? Variables are normalized in each altitude, an absolute value may misrepresent performance. Consider plotting relative RMSE (e.g., RMSE divided by median water vapor at each altitude) to better contextualize errors, especially at lower altitudes (<15 km) where water vapor concentrations are naturally higher. This would also help clarify if the elevated RMSE near the surface is a true error or a reflection of larger absolute values.

Yes, Figure 2 shows the absolute RMSE, as indicated by the "VMR" units provided in the figure. However, it is a good point that this obfuscates how these RMSEs compare to the typical H_2O VMRs at each altitude, and using a relative metric would better contextualize these errors. We have updated Fig. 2 to show the absolute RMSE divided by the training data set's average H_2O VMR at each altitude, as suggested. We use the average rather than the median as the statistic had been previously calculated by the NN code.

5. The statement "errors increase below 18.5 km..." needs clarification. Do you mean that measurement density is higher in the troposphere, or that variability increases? Does the sample size vary significantly with altitude?

Yes, yes, and no, respectively.

The H_2O VMR is significantly higher in the troposphere, and absolute errors are also larger in this region.

When considering percent differences between the LP predictions and co-located MLS profiles, the variability of these differences is larger in the troposphere; the differences are typically within 10% in the stratosphere with extreme differences of ~20%, while in the troposphere they can exceed 50%. We believe this may be due to a saturation effect, as the increased scattering in the upper troposphere likely limits the accuracy and precision of our measurements in this regime.

In Figure 2, the sample size is identical at all altitudes.

6. Concerns About Temporal Coverage and Generalization

For the year dependence, Section 4.1 lacks clarity. You mention omitting 2024–present (Line 179), but training data is stated to cover 2014–2024 (Line 85). Did you use 2025 data? What is the exact time period excluded, and how does this affect inference quality?

Your explanation for 2024 being "special" is unconvincing – also see comment about Ruang above. The Hunga Tonga eruption occurred in early 2022, and the water vapor peaked shortly after. This does not justify 2024 as a critical component for training unless further supported by data.

Lines 178-180, where we discuss omitting data from 2024-present, describes a separate experiment conducted. The text has been updated to clarify this point:

> "**We carried out additional experiments where certain years were omitted from training and found that this** can be important for certain situations. When omitting 2015-2016, …"

For the model we presented in the manuscript (the model that is currently producing the LP H_2O products), training data covers 2014-2024 as described on line 85. Thus, the only difference between our presented model and the model from the separate experiment is the inclusion of 2024 data during training. 2025 data are not used in training at any point.

As discussed in the manuscript, the exclusion of 2024 data impacts inference quality when applied to data from March 2024 and onward. Note that this performance degradation is unrelated to Ruang, as Ruang did not erupt until mid-April 2024. The explanation for this poor performance is shown in Figure 6: in the first half of 2024, the MLS tape recorder shows significantly elevated H_2O above 30 km compared to the pre-Hunga period. These conditions are not well represented in a 2014-2023 training data set (where 30+ km H_2O enhancements are accompanied by different conditions than in 2024 and beyond), which leads to poor model generalization. By including some of these data in training, model performance significantly improves in this regime, and it generalizes into 2025 where MLS also shows elevated H_2O above 30 km.

7. Feature Design and Model Limitations

In your study, the year is not treated as an input feature. If year-to-year variation affects model performance, this could point to missing explanatory variables or insufficient feature engineering. You may consider a data imbalance or out-of-distribution (OOD) problem in your training.

Year is not treated as an input feature because year-to-year variability is implicitly contained within the LP radiances and, to a lesser extent, temperature/pressure data. Data imbalance was handled by subsampling the co-located data as discussed in lines 108-110, and the same lines also discuss a step taken to minimize the chances of an OOD problem. Despite that, it's possible that there is an OOD problem as the dimensionality of the problem (421 unique inputs) makes it difficult to truly determine this.

> *8. In addition, given the relatively small number of input features except the 12 channels and model may be overfitting. I would like to see your support materials to make sure your model is not overfit.*
>
> *Please consider revisiting the input space, especially if training struggles to generalize beyond 2024.*

There are 421 unique input features (1440 input features when including redundant inputs for the image-based processing used) mapped to 30 output features. This is not typically considered a small number of input features in the ML literature. The model is not overfit as evidenced by various performance metrics being similar on both the validation (occasionally seen during training) and test (never seen during training) data. See Figure R1.4 below, which is analogous to the manuscript's Figure 2 except showing the metrics for the validation set in addition to the test set. The test and validation curves are nearly identical, indicating that the model generalized to unseen data and did not overfit the training/validation data.

Additionally, we show below in Figure R1.5 the median differences between LP and MLS for various years. Focusing in on 2025, we can see that the differences above 25 km are consistent with the years considered during training. Below 25 km, the errors for 2025 are around -6-7%, whereas the years considered during training are typically within 2%. However, it is important to note that there are significantly fewer LP-MLS co-locations in 2025 due to the MLS duty cycling that reduced MLS observations to ~6 days per month, which may be affecting these comparisons. Nevertheless, the errors in this regime are generally less than the LP-ACE comparisons shown in the manuscript's Figure 3b and they are comparable to the LP-SAGE comparisons shown in the manuscript's Figure 3c. If the model were overfitted, it would be expected that the 2025 errors would be significantly different than the years seen during training.

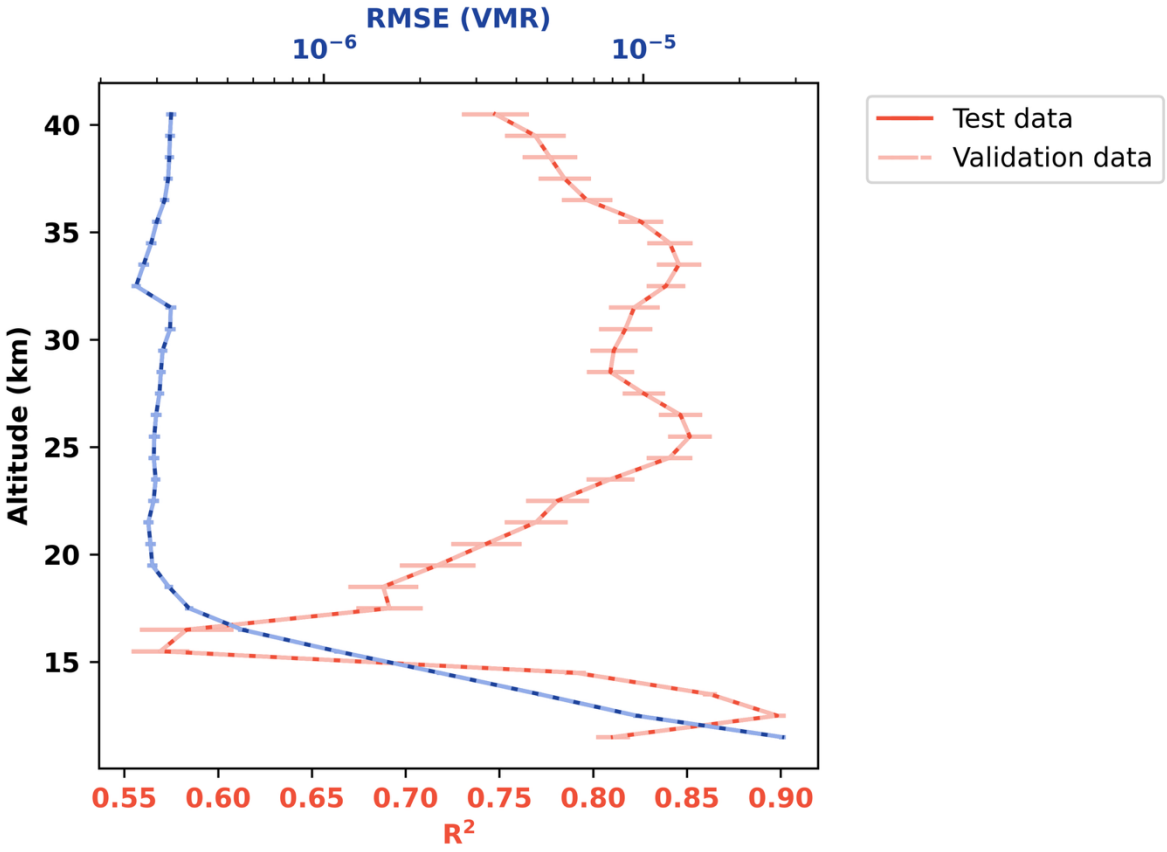

Figure R1.4. Like Figure 2 in the original manuscript, except also including the performance metrics for the validation data. The lack of a performance gap between the validation and test set metrics indicates that the model has generalized well to unseen data.

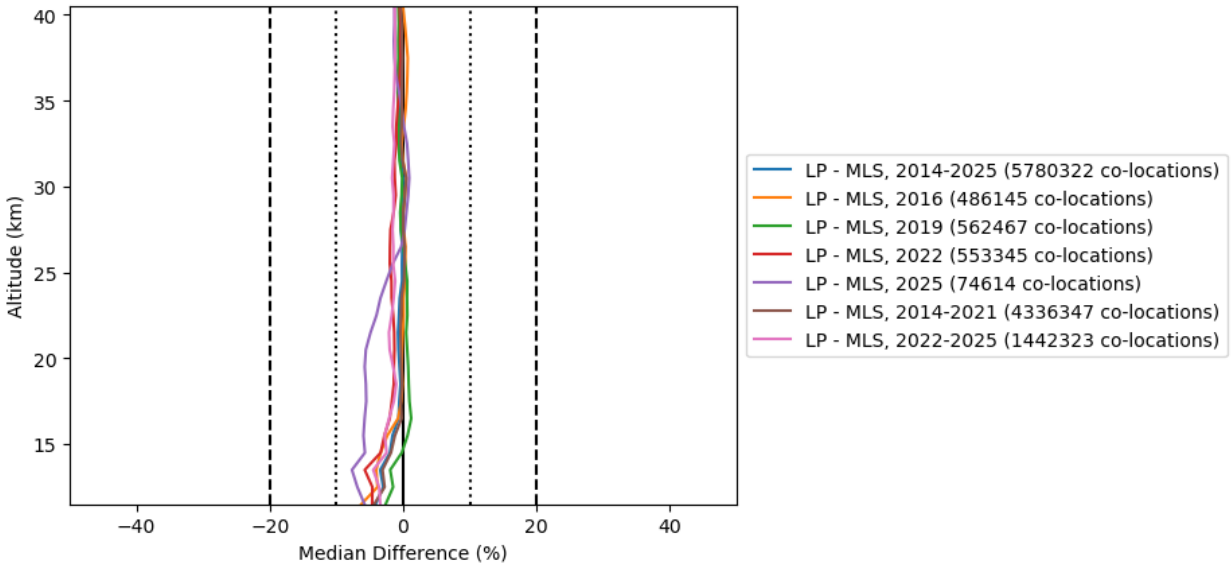

Figure R1.5. Like the manuscript's Figure 3a, but showing individual years as well as the pre- and post-Hunga periods.

9. You state that model errors may not related to aerosol loading in Line 187. I am just curious like a time series of model errors alongside aerosol concentrations (e.g., before and after the 2022 eruption), do error patterns increase during high aerosol periods?

Figure R1.5 above shows that the post-Hunga period has a 1-2% difference compared to the pre-Hunga period. Figure R1.6 below shows the requested time series plot. Each vertical line shows the average percentage difference between LP-MLS co-locations for each LP orbit (LP has 14-15 orbits per day). Near the top of the plot, three large eruptions are marked (Calbuco, Hunga, and Ruang). While there are occasional orbits with larger errors than the average, they are not correlated with major eruptions, consistent with what was reported in the manuscript. These additional details are in a new Appendix C.

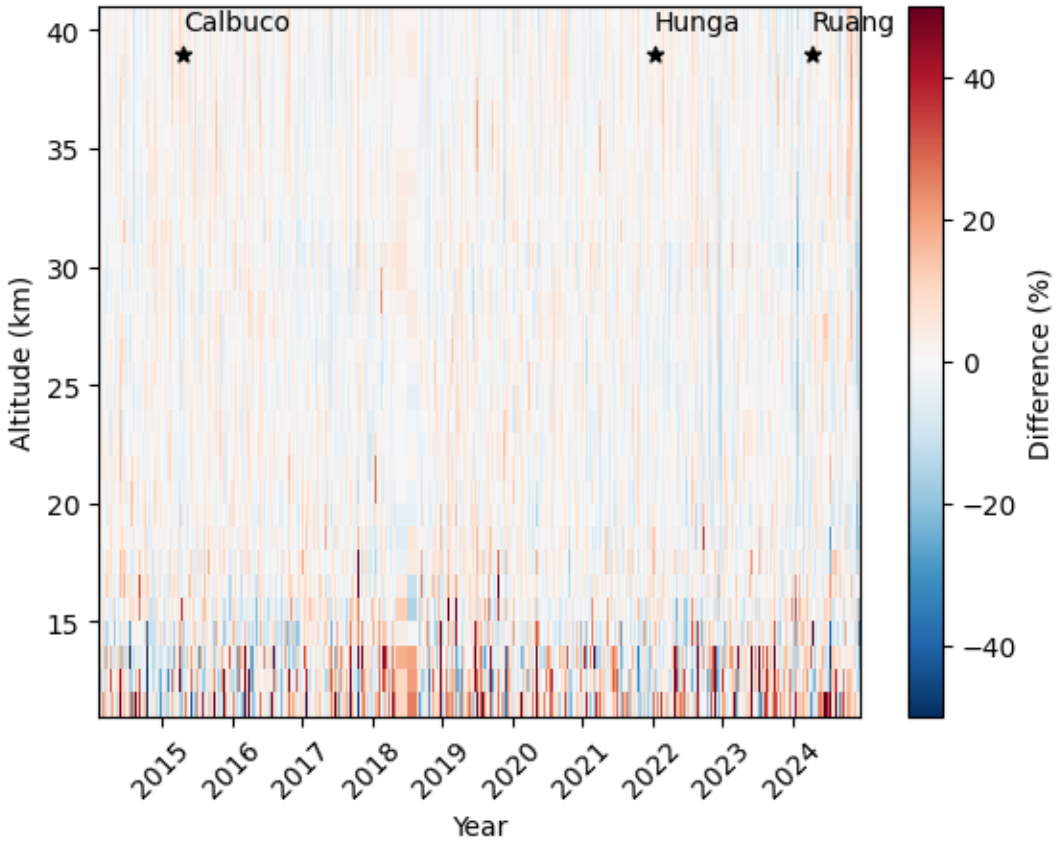

Figure R1.6. Time series plot of the average percentage difference between LP-MLS co-locations per LP orbit. Three major eruptions are denoted on the plot for context.

10. Comparisons and Justification of External Datasets

While comparisons with SAGE, ACE, and MLS are common, their measurement techniques differ significantly from OMPS-LP as you stated in the manuscript. This limits the interpretability of these comparisons. Since your model is trained on MLS water vapor, it makes most sense to validate primarily against MLS. In other words, the result shows differences, but these may stem from discrepancies between MLS and other datasets – see the MLS data quality and description document (Livesey et al., 2022), not from your model. The same remark can be applied to comparisons with M2-SCREAM.

Yes, we agree on this point. The primary validations are with MLS, as that is what we trained on. The purpose of including comparisons with additional instruments is to show that the model is more generally applicable, that is, it doesn't *only* work where LP is co-located with MLS. The differences between LP and the other instruments are indeed a product of the discrepancies between MLS and those other datasets, as our NN approach mimics the MLS product (e.g., line 235-236, "… or MLS is biased low in this regime and our LP product has inherited this bias."), but it is an important element to show the generalization of the approach.

(Noting that there is no comment #11)

12. Figure 8

The claim that the NN methodology reduces drifts may be overstated. If the MLS data exhibits a decadal trend and your model was trained with shuffled input, it would be expected to replicate that trend. It does not make sense to me the model can do drift correction automatically. Please investigate and explain the reason for the difference before attributing it to NN drift correction.

It is a good point raised here and by the other reviewer that the presented results do not support a conclusion of the NN model reduces MLS drifts, as we did not compare trends in water vapor derived from LP and MLS with more accurate frost point measurements. We have revised the text to remove mentions of drift reduction and instead focus on the presented trends:

"Regarding water vapor trends, the LP product generally shows ~~greater trends in the troposphere and~~ weaker trends in the stratosphere when compared with MLS. **In some locations (particularly south and southeast Asia, central Africa,**

and central America), the LP product shows greater trends in the upper troposphere. Where both products show …"

The NN is attempting to minimize the mean squared error across all training and validation cases, so the differences in trends are a product of that process. Presumably, the difference in trends is related to differences in instrument performance over time between LP and MLS, that is, the sensors will not degrade in the same way, but it is beyond the scope of this work to definitively determine this.

13. NOAA-21 Application (Section 4.7)

While it's reasonable to apply the trained model to NOAA-21, the manuscript doesn't clearly justify the value of this step.

You acknowledge a bias/shift between SNPP and NOAA-21 radiances, which already limits comparability. The bias between two OMPS radiances obviously reflects in the inference. The statement in Line 290, suggesting the model may implicitly account for radiance bias, is likely overstated given the model's simplicity and data.

The justification for this step is given at the end of Section 3.3:

"Finally, given that OMPS LP is onboard the NOAA-21 satellite and planned to launch onboard two additional satellites in the coming years, we apply our SNPP-trained model to NOAA-21 OMPS LP measurements to determine whether our model can generalize to future iterations of the same instrument …"

We have revised the text to more clearly explain why this is important:

"NOAA-21 OMPS LP has an insufficient period overlapping with MLS measurements (2023 – present, with MLS only taking measurements for ~6 days per month since May 2024), inhibiting the use of NOAA-21 OMPS LP data for our NN training methodology. Given the imminent termination of Aura MLS and that the SNPP satellite will presumably cease operations before the end of NOAA-21 or the subsequent JPSS satellites, we test the application of our SNPP-trained model to NOAA-21 OMPS LP measurements to determine whether our model can generalize to future iterations of the same instrument …"

14. Figures and Presentation

Figure 1. Missing legend. Please indicate what each color represents.

Done

Figure 6. Consider adding a third panel showing the difference between Figures 6a and 6b to better highlight anomalies or patterns not captured by direct comparison.

Done

Figure 7. Since Figures 7a and 7b are expected to show similar results due to the consistent retrieval, they may be redundant. Consider removing 7a and 7b, and retain 7c, which provides more useful spatial comparison.

We appreciate the suggestion, but we think that it is important to include both panels (a) and (b) to highlight the close agreement between LP and MLS.

Line 180. The Ruang aerosols may have created problems in the April 2024 period

As discussed above, Ruang erupted in mid-April 2024, but the issues emerged in March 2024, indicating they are unrelated to the Ruang eruption.

Line 207. Water vapor in the stratosphere doesn't have a diurnal cycle so why would time co-location make any difference unless the NN is using other gases such as O3 or temperature?

The NN does indeed rely on temperature in part, as shown above in comment #1. In the lower stratosphere, dynamics drive changes in trace gas concentrations; differences of several hours between measurements can lead to LP and MLS measuring different air masses with different concentrations of H_2O, which could limit the accuracy of the trained model. However, even when isolating this variable, the number of co-locations is a main limiting factor, as found by our experiment where we restrict the data set size for MLS co-locations (lines 208-210).

Review #2 Response

We thank the anonymous reviewer #2 for their thoughtful, detailed review of the manuscript, as it will improve the quality of the manuscript. Our response to each comment is provided below.

> *My main concern is that "when omitting 2024 during training, the model begins producing severely inaccurate predictions by March 2024." The authors claim that by including a small fraction of 2024 data during training, the model continues performing accurately up to the present time. However, Figure 6 will suggest otherwise: the MLS tape recorder in 2025 differs considerably from the one estimated by OMPS LP, indicating that the model is already producing inaccurate predictions, presumably because the state of the stratosphere continues to evolve as the Hunga plume moves.*

To better assess this, we have updated Figure 6 to include the difference between the LP and MLS tape recorder plots to more clearly show where the LP product exhibits deviations from MLS. While there is a brief period in 2025 where LP is overpredicting H_2O by more than the average difference over the 2014-2024 training period, by the end of 2025 it has returned to differences comparable to the training period. If necessary, future work could be to re-train the NNs with the inclusion of a small amount of data from this period that shows larger than normal errors to better inform the model of these conditions.

Additionally, we show Figure R2.1 below, which is like the manuscript's Figure 3a except that it includes a few individual years, including 2025, as well as the pre- and post-Hunga periods. Focusing in on 2025, we can see that the differences above 25 km are consistent with the years considered during training. Below 25 km, the errors for 2025 are around -6-7%, whereas the years considered during training are typically within 2%. However, it is important to note that there are significantly fewer LP-MLS co-locations in 2025 due to the MLS duty cycling (that lead to reduced number of MLS measurements, ~6 days per month), which may be affecting these comparisons. Nevertheless, the errors in this regime are generally less than the LP-ACE comparisons shown in the manuscript's Figure 3b and they are comparable to the LP-SAGE comparisons shown in the manuscript's Figure 3c.

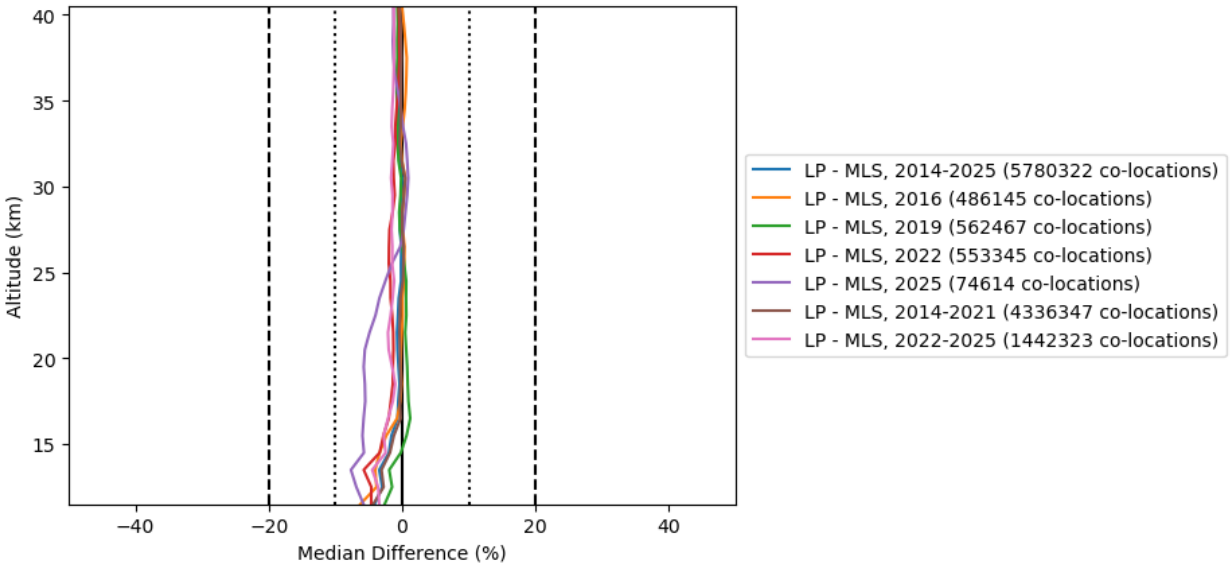

Figure R2.1. Like the manuscript's Figure 3a, but showing individual years as well as the pre- and post-Hunga periods.

The rationale that "In 2025, the NNs perform reasonably well below 30 km, indicating that there is sufficient sensitivity for the determined approximation to remain accurate when applied to unseen data" may also be faulty, Hunga excess water vapor has not reached those levels, when it does, it may affect the estimates as well, as the training has not seen those type of values.

We have adjusted the language to be more balanced:

"In 2025, the NNs perform reasonably well below 30 km, **suggesting** that there is sufficient sensitivity for the determined approximation to remain accurate when applied to unseen data, though there is a slight overestimation ~0.25 ppm) in mid-2025 between 25--30 km. We therefore advise that users exercise caution when using the OMPS LP H_2O product above 30 km in the tropics. **Additionally, as the excess SWV from Hunga continues to evolve, it is possible that those conditions will be different enough from the training data that the NNs' predictions become inaccurate; we will continue to monitor their performance to ensure the LP NN-based algorithm continues to provide reasonable H_2O profiles.**"

Overall, while the OMPS LP may accurately predict water vapor under "stable" conditions (i.e., prior to Hunga), it appears it cannot predict post-Hunga unless the

training dataset includes some representation of those behaviors. Given that training with ACE-FTS and SAGE III/ISS failed, the absence of MLS data for training means the model will likely fail to capture the true state of the stratosphere as the plume evolves.

As discussed in the comment above, there is a short period in 2025 within the tropics where the model overpredicts H_2O, but this bias becomes insignificant by the middle of 2025. When considering results across all latitudes, the model is performing satisfactorily, with biases vs. MLS typically less than the differences between MLS, ACE-FTS, and SAGE III/ISS. Based on these results, we might expect the model to have small periods of increased bias in certain locations as the stratospheric water vapor continues to evolve, but we do not expect the model to fail in general. Once stratospheric water vapor returns to pre-Hunga levels in the coming years, we expect the model to perform well, given that it accurately handles the pre-Hunga period. However, if there is another Hunga-like event, we expect that the model may not perform satisfactorily given that the conditions will likely differ from Hunga and therefore not be represented by the training data set. If such an event occurs, we will assess the accuracy of the LP H_2O profiles using data available at that time and take appropriate actions as necessary.

Furthermore, it is not clear whether the authors applied the MLS quality screening criteria to properly filter the MLS data. The MLS quality document is available at https://mls.jpl.nasa.gov/data/v5-0_data_quality_document.pdf

If the authors did apply the MLS quality screening, this should be clearly mentioned in the text, perhaps in the Data Curation section. If they did not, the analysis should be repeated using the quality screening to avoid retrieval artifacts, a priori influences, etc.

This is a good point. Yes, we did apply the MLS quality screening as recommended in the MLS data quality document, with one caveat: we did not apply the screening criteria between January 15, 2022 through February 12, 2022 as they have been reported to filter out valid profiles shortly after the Hunga eruption. We have added the following text in Section 3.1 accordingly:

> "We apply the recommended MLS data screening criteria except for the period of January 15 -- February 12, 2022, as it has been shown that the recommended screening criteria filter out many valid profiles affected by Hunga (Millán et al. 2024)."

The same goes for ACE-FTS, the ACE-FTS screening criteria can be found at

https://doi.org/10.5194/amt-8-741-2015

Yes, we are applying the recommended ACE-FTS screening criteria. We have clarified this at the end of Section 3.1:

> "We also consider a similar methodology but for ACE-FTS and SAGE III/ISS data **with recommended screening criteria applied to both datasets**. **We utilize** co-location criteria …"

> *Also, how does the ensemble standard deviation (used as an uncertainty estimate as discussed in line 121) compares with the MLS precision error. Are there comparable? What is the vertical resolution of the OMPS-LP water vapor dataset. Is it the same as MLS even though the OMPS-LP dataset will be reported in 1km spacing?*

From the MLS data quality document, MLS's typical retrieval precision is ~0.3 ppmv in the stratosphere, growing to ~1-5 ppmv at the greatest pressure levels considered in our study. By comparison, on an average day the NN ensemble's standard deviation profiles over the LP record yields uncertainties of ~0.12 ppmv in the stratosphere (with some individual days averaging as much as ~0.35 ppmv) and ~2-5 ppmv in the UTLS (with some days averaging up to 10 ppmv). The NN predictions cannot be more precise than the training data; therefore, it suggests that the standard deviation among NNs underestimates the uncertainty, particularly in the stratosphere.

This underestimation is at least in part due to neglecting a source of uncertainty. The NNs are predicting an MLS-like profile; the standard deviation represents their uncertainty in that process. Yet, MLS profiles also have some uncertainty, which is not considered by that calculation. We have therefore revised the uncertainty estimation to be the ensemble's standard deviation combined in quadrature with the reported precision of MLS. This ensures that the LP product's uncertainty is greater than the corresponding MLS uncertainty.

Logically, the vertical resolution of the LP product cannot be less than the LP radiances (~1.8 km). Since the NNs are trained on the MLS H_2O product which has a vertical resolution of 2.8 – 3.8 km over the pressures considered in this study, it's possible that the resulting LP resolution is between 1.8 – 3.8 km. However, we have not yet performed a detailed analysis of the vertical resolution, and so at present we conservatively estimate the vertical resolution at >4 km.

> *The manuscript needs to include a table discussing the OMPS LP water vapor characteristics (e.g., precision, horizontal resolution, vertical resolution)*

Thank you for the suggestion, we have added a table with this information in Section 4.1.

Lastly, have the authors performed a feature importance analysis? In other words, how much of the water vapor information is coming from the LP radiances, how much from the FP-IT fields, and how much from the solar zenith angle? It is possible that the neural network primarily relies on temperature and pressure information to estimate water vapor, with the LP radiances contributing very little. Is it possible that some channels are contributing and other are not? How robust would the water vapor estimates be to jumps or changes in the reanalysis fields? The authors could modify these fields to assess their impact on the estimated water vapor.

Our initial manuscript allowed for a simple test of perturbations in the temperature/pressure profiles. In March 2025, the LP ancillary product switched from using the GEOS FP-IT product (it was discontinued soon after this time) to the new GEOS-IT product, which exhibited a discontinuity in the temperature data on the order of a few degrees Kelvin. If the temperature/pressure data were primarily driving the predictions, then it would be expected to see differences between the model trained on the GEOS FP-IT temperature data but applied to the GEOS-IT temperature data, vs. a model trained exclusively on GEOS-IT data. We reapplied our methodology using the new GEOS-IT product during training and find our results in 2025 unchanged, indicating that the NNs are robust to small discontinuities in temperature data.

We additionally investigated this question by training NNs without LP radiances or solar zenith angles, training NNs using climatological temperature/pressure profiles, and training on only LP radiances. When omitting LP radiance and solar zenith angles from training, we find that the model performs significantly worse, with larger root mean square errors and smaller R^2 values when applied to the test set (see Figure R2.2 below). Conversely, when using climatological temperature/pressure profiles, we find that the RMSE and R^2 values over the test set agree with those presented in the manuscript. Additionally, training on only the LP radiances also achieves similar RMSE and R^2 metrics as those presented in the manuscript. These results indicate that while the temperature/pressure data are useful, they are less important than the radiances when solving this problem. This is detailed in a new Appendix B.

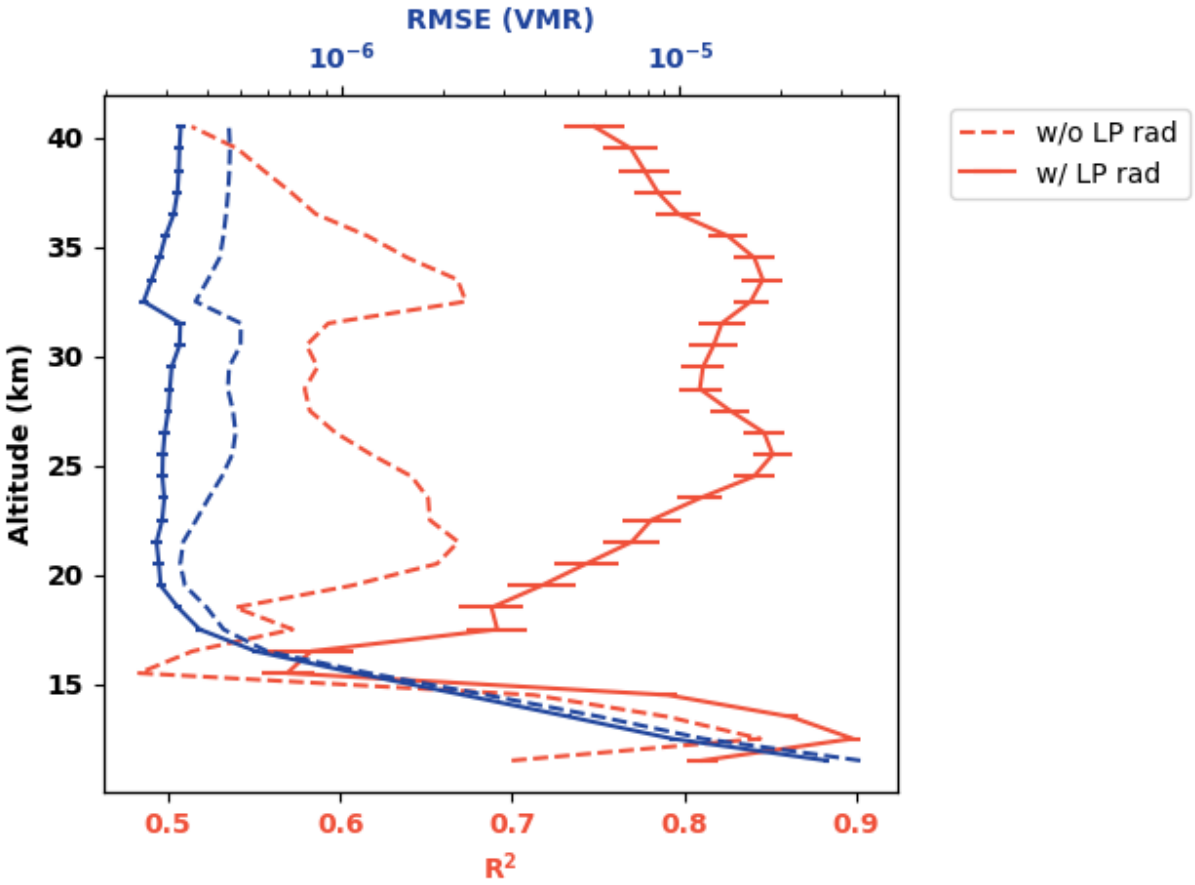

Figure R2.2. Like Figure 2 in the manuscript, except additionally showing the performance metrics for a NN trained on only on the temperature and pressure data (dashed lines). The degraded performance in the stratosphere above 15-17 km suggests that the LP radiances provide important information that enables the determination of a better solution for retrieving stratospheric water vapor.

Given that the OMPS-LP water vapor dataset exhibits a different trend from the MLS dataset (Figure 8) and has a limited vertical range (11.5–40.5 km, though effectively only up to 30 km, as the manuscript states: "We therefore advise that users exercise caution when using the OMPS LP H_2O product above 30 km."), the current title is somewhat misleading. A more appropriate title might be: "OMPS LP water vapor estimates based on a neural network trained on MLS water vapor."

The MLS-like nature of our product allows it to function as a continuation of the MLS water vapor record for a more limited altitude range, which is why we chose that name for the manuscript. We have updated the title to more explicitly identify that it uses NNs: "Continuing the MLS water vapor record with OMPS LP using neural networks".

L2 the name of the instrument is: Atmospheric Chemistry Experiment Fourier Transform Spectrometer (ACE-FTS) Please change accordingly here and elsewehere

Done

*L3 … Experiment III on the International Space Station *(SAGE III/ISS)**

Done

L10 retrieve -> please change to estimate or predict, retrieve is associated with the typical retrieval process (i.e., optimal estimation).

Retrieval algorithms perform an estimation of some properties based on remote sensing data. While our retrieval algorithm does not follow the traditional methodology, we have described it in detail to ensure readers are aware that it utilizes neural network predictions. We have updated the text in this line to better describe this:

"… a neural network-based **retrieval algorithm** to **estimate** SWV …"

L20 There is likely a better citation for UT water vapor, it has been known for decades. Please add other citations or at least add e.g., before Read et al 2022.

Done

L28 for MLS please cite Waters et al 2006
https://ieeexplore.ieee.org/document/1624589

Done

L29 For ACE-FTS please cite
https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2005GL022386

Done

L33 Please provide a brief description of the agreement between the instruments. These differences help set the stage for properly evaluating the OMPS LP water vapor measurements discussed here.

We have adjusted this sentence to provide the recommended context:

> "These instruments provide well validated H_2O products that **typically show median differences within ~10%** among each other as well as with ground-based and in-situ measurements."

*L35 now only operates *around* 6 days per month to *preserve measurement lifetime* and will continue …*

Done

*L37 please add */ISS* after SAGEIII (there was a SAGEIII meteop, so it is customary to call the SAGE III on the ISS, SAGE III/ISS). Please change SAGEIII to SAGE III/ISS elsewhere as well.*

Done

L51 retrieve -> estimate

This sentence is talking about the H_2O retrieval problem for OMPS LP in general, as the previous sentence discussed the challenges of radiative transfer-based retrievals of H_2O from OMPS LP.

L60 eruption. You need to explain why you are using before and after Hunga. You never mentioned how Hunga altered the stratospheric water vapor.

We have added additional context to this sentence as recommended:

> "… we investigate the sensitivity of LP to water vapor under conditions before and after the Hunga eruption, **which injected more than 50 Tg of water into the stratosphere (citations**)."

L70 the same -> identical. to ensure -> ensuring

Done

L71 differences -> variations

Done

L73 How was the sensitivity to H2O estimated? Degrees of freedom, smaller precision, please provide succinctly the details.

As mentioned at the beginning of this sentence, this choice was made based on the sensitivity kernels (partial derivatives or Jacobians) output by the RTM; wavelengths that showed larger Jacobians than adjacent channels were selected, as well as wavelengths that showed larger Jacobians than most other channels considered. We included a wide range of wavelengths to capture the spectral behavior of aerosol and scene reflectivity, which enables the NN to recognize those conditions and presumably differentiate them from H_2O.

L86 It is not clearly defined why the data needs to have aligned orbits, could the authors simply use all colocations that are within 6 hours and 100km. What is gained by using the 2 criteria listed in L82 and 83.

It's possible that using all colocations within 6 hours and 100 km would work. In the lower stratosphere, particularly in the mid and high latitudes, strong dynamical variability can lead to large changes in trace gas concentrations. By using strict co-location criteria, it minimizes the impact of geophysical noise. Additionally, most LP-MLS co-locations occur at high latitudes; this latitudinal sampling bias negatively impacts the trained model, which led to the resampling procedure described on L108-109.

The criteria listed on L82-83 begin to address that latitudinal sampling bias. Since each SNPP OMPS LP measurement differs by ~1° latitude, only considering days with orbits that have 60 consecutive colocations increases the chances that some of the events occur near the equator. A sampling bias still exists using this approach, but a side effect of this approach is that the training and validations sets only use data from certain days (roughly 1 every 2-3 days). All other days can therefore be used as tests on unseen data and ensure the NN is correctly solving the problem. If we mixed data from all days, some test cases might occur next to a training case, which could make the model appear more accurate than it is in reality.

L88: Due this means that the training was based on around 8 percent of the available OMPS-LP data, i.e, 200000 colocations / (250000meas per year x10years). If it is, perhaps the authors should mention this on the text, to improve the context.

SNPP OMPS LP measures ~2.5 million profiles per year, and we considered an 11-year period for the training data. As mentioned in Section 3.2, we subsampled the data set of 2,074,101 co-locations down to 1,137,100 to address a latitudinal sampling bias. Of those ~1.1 million co-locations, 75% are used for training and 15% are used for validation, or just over 1 million co-locations seen during training. This comes out to ~3.7% of the SNPP data over this 11-year period. Factoring in 2025's data drops that to ~3.4%. While SNPP OMPS LP also has data for 2012-2013, sample table differences mean our methodology cannot be applied without modification. That said, if those years are also considered, then our training and validation data set is ~3% of the total available SNPP OMPS LP data.

We have added text in Section 3.2 to provide this context to readers:

> "We then split these data into training (used to update NN weights), validation (monitors for overfitting during training), and testing (tests model generality on unseen data after training is complete) sets in a proportion of roughly 75%, 15%, and 10%, respectively. **For context, this results in the training and validation sets containing a total of ~3.7% of all available SNPP OMPS LP data over the 2014-2024 period considered.**"

Figure 1 caption: Hunga peak in the lower stratosphere. So upper troposphere -> lower stratosphere

While the Hunga plume and therefore change in Jacobians peak in the lower stratosphere (Figure 1d), the Jacobians (Figures 1a and 1b) peak in the upper troposphere, which is what is discussed in that sentence. However, upon reviewing this sentence, we think it is phrased a bit confusingly, so we have revised it to be clearer:

> "Panels (a) and (b) show the Jacobians for selected MLS H_2O profiles from before and after the Hunga eruption, respectively."

L89-90 where is this information coming from? All levels could be affected by apriori values, which is why the MLS data sets the precision to zero or negative. See MLS quality document.

The MLS quality document states for 316hpa "Occasionally erroneous low value < 1 ppmv and high value fliers are retrieved in the tropics, usually in clouds."

Based on inspecting a random assortment of MLS data files, the number of 316 hPa measurements that have a reported precision of -1 is around 5%, while the 261 hPa level is around an order of magnitude less frequent. We wanted to maximize the amount of data available for this study, so we chose to omit the 316 hPa level given that a non-negligible fraction were a priori-dominated. In future work, we will consider including this pressure level so that the LP profiles can have a lower minimum altitude, but it is unclear at present how this reduction in data set size will impact the accuracy of the NNs. Reviewing this sentence, it is phrased confusingly and does not clearly convey this point, so we have omitted that part of the sentence for conciseness:

"We limit the MLS water vapor profiles to ≤261 hPa~~, as the 316 hPa pressure level can be affected by the a priori profiles used in the MLS retrievals~~."

L91-93 Are you using the closest FPIT fields to the MLS measurement locations or are you interpolating in time and space to the MLS measurements times and locations. Please be specific.

We are using the FPIT fields interpolated to the LP measurement locations/times. We have updated this sentence to be more specific as recommended:

"We log-linearly interpolate the water vapor profiles from the MLS pressure grid to OMPS LP's geometric height (11.5-40.5 km in 1 km steps) using the NASA Global Earth Observing System Forward Processing for Instrument Teams pressures **interpolated in time and space to the LP measurements**."

We chose to use the FPIT fields at the LP measurements because (1) they were readily available, and (2) the strict co-location criteria used greatly minimizes differences in temperatures and/or pressures between the MLS and LP measurement locations.

L94 ACE -> ACE-FTS (here and elsewhere)

Done

L102 Are the FP-IT interpolated for the OMPS LP measurement times and locations, or are you using the closest fields in time and space?

We have updated the text to clarify this:

> "FP-IT pressures and temperatures **interpolated in time and space to the LP measurements**"

> *L131 the correct citation for MLS v5 water vapor dataset is*
>
> *Lambert, A., Read, W., & Livesey, N. (2020). MLS/Aura Level 2 water vapor (H2O) mixing ratio V005. [Dataset]. Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5067/Aura/MLS/DATA2508*

Thank you, the reference has been added.

> *L140 The drift was found in v4. In version v5 the drift was ameliorated. From Livesey et al (2021) "As a result of this correction, the MLS v5 H2O record shows no statistically significant drifts compared to ACE-FTS. However, statistically significant drifts remain between MLS v5 and frost point measurements, although they are reduced."*
>
> *I think the authors can simply delete the mentioned of the drift, that is, "We investigate whether our product shows similar properties as the MLS product by …"*
>
> *Or they can be more specific and say something like: Given the statistically significant remaining drifts between MLS v5 and frost point measurements, despite the applied drift correction (Livesey et al 2021), we investigate …*

Thank you for the recommendation, we have removed the text about the drift.

> *L194 please add *upper* before tropospheric*

Done

> *L194-197. How was the presence of tropospheric clouds determined? How was the presence of PSC determined? Through OMPS-LP measurements, MLS measurements, other? Please explain in the text. Could you show figures showing the lack of impact for tropospheric clouds as well as the impact due to PSCs?*

These were determined using the LP radiance measurements (Chen et al., 2016). The corresponding properties are reported in the LP aerosol and ozone product files. We have added this detail as requested:

> "In general, the error with respect to MLS is independent of the presence of upper tropospheric clouds **identified by the OMPS LP measurements (Chen et al., 2016)**. However, events affected by polar stratospheric clouds (PSCs) **identified by the OMPS LP** show a median bias around -2% at most altitudes."

Below are the requested figures. Figure R2.3 shows a 2D histogram of the LP – MLS percent differences vs. the aerosol extinction, both at 11.5 km. While there are a few outliers that show large errors > 200% (omitted for clarity), the vast majority of the histogram density is centered on ~0% difference at any given aerosol extinction value. At the largest aerosol values, the distribution is perhaps slightly skewed toward the negative, but it is not statistically significant. These additional details are in a new Appendix C.

Figure R2.4 shows the median differences between LP and MLS for measurements in the test set that are contaminated by PSCs. As reported in the manuscript, altitudes above the PSCs tend to feature a negative bias of a few percent, while altitudes below the PSCs feature an increasingly negative bias.

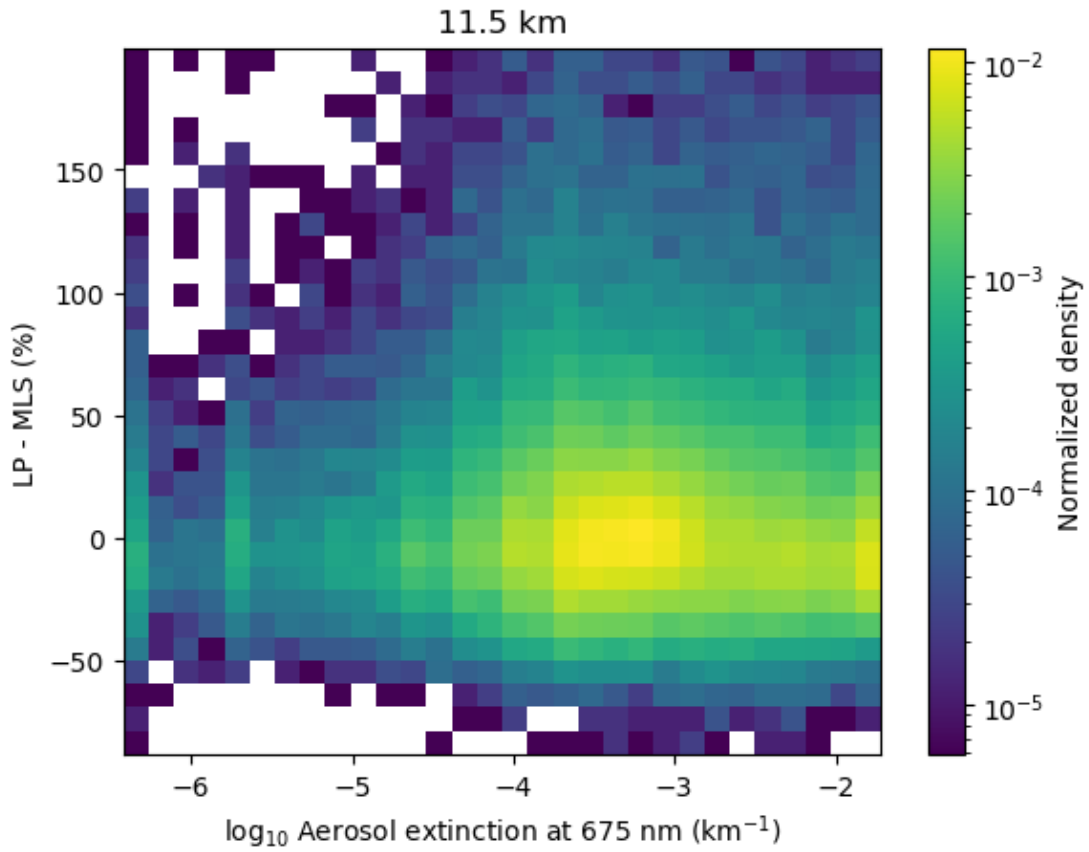

Figure R2.3. Histogram of the LP-MLS percent differences vs. OMPS LP aerosol extinction at 675 nm at 11.5 km.

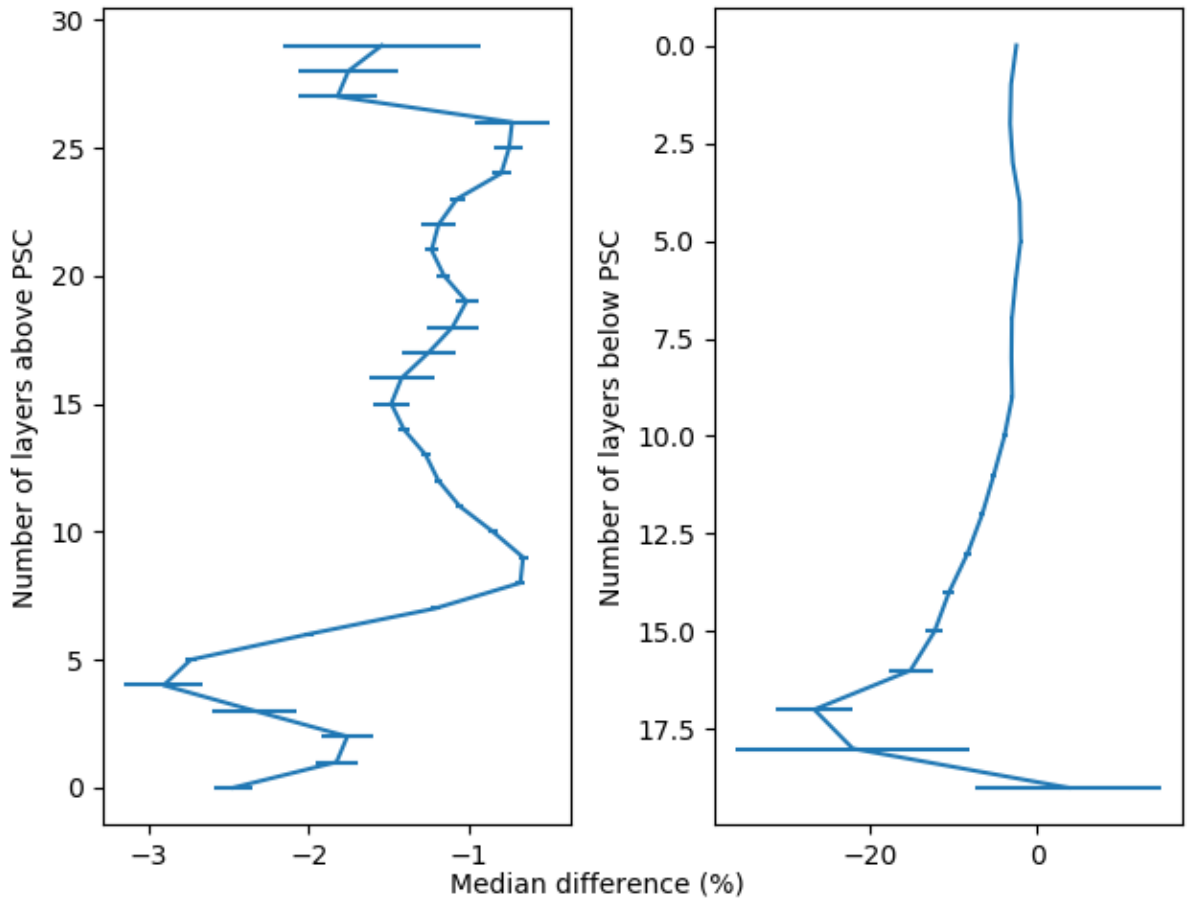

Figure R2.4. Median differences between LP and MLS for measurements in the test set that are contaminated by polar stratospheric clouds (PSCs). Error bars denote the standard error of the median.

L200. What was the coincident criteria for training ACE-FTS and SAGE III/ISS? How many coincidences were available. Why was SABER not considered for training?

The ACE-FTS and SAGE III/ISS coincident criteria were detailed at the end of Section 3.1:

> "We also consider a similar methodology but for ACE and SAGE III data using co-location criteria of within 1 day, within 2° latitude, and within 1113 km longitude (equal to 10° longitude at the equator), consistent with the criteria used in Davis et al. (2021)."

The number of coincidences was around 20,000 for SAGE III/ISS and less for ACE-FTS. This detail isn't important for the manuscript considering that the MLS data set subsampled to match the size of SAGE III/ISS + ACE-FTS co-locations did not yield an accurate model.

SABER was not considered for training because its yaw cycle leads to a bias in latitudinal coverage at different times of the year, which is expected to limit the usable latitudinal range of the NN. However, this would be an interesting topic to investigate in future work if we later find that the current methodology fails before the launch of HAWC.

L202 were negative -> were not satisfactory or were suboptimal.

We have revised the text to "not satisfactory" as recommended.

L209 on line 87 the authors mentioned 2 million colocations not 1 million, which one is the correct value?

On L87, the ~2 million colocations are all colocations that satisfied the criteria mentioned on L82-83. As discussed on L108-109, we subsample that ~2 million data set down to ~1 million for the NN training, validation, and testing. Here on L209, we are discussing that ~1 million data set used for the NNs.

L218 Why are you using the median and not the mean?

We chose to use the median for consistency with previous H_2O validation efforts in the literature. For well-behaved data, it provides the same (or nearly the same) value as the mean, while it is more robust than the mean for long-tailed distributions or extreme outliers.

L220 Why are you using the standard error of the median? You should use the standard deviation as a metric.

The standard error measures the accuracy of the calculated median for the dataset considered.

L230-232. Please delete the following sentence "This behavior is generally consistent with earlier studies, such as Davis et al. (2021) which shows a similar pattern of increasing differences between 15–40 km when comparing SAGE and ACE." The fact that SAGE and ACE also have differences increasing with altitude is pure coincidence and have no merit in this discussion.

Done

L235 our -> the

Done

L264-267: This analysis is not enough to conclude that NN reduce these drifts. The authors need to collocate the MLS and OMPS data with the balloon measurements and compute the drifts for both datasets. Figure 8 only shows that the long-term trends are different between MLS and OMPS.

This is a good point. We have removed mention of the drifts to instead focus on the trends. Future work should investigate the drifts between LP and frost point measurements to assess whether the LP product has inherited the same drift present in the MLS data.

L275 How do the authors determine that OMPS-LP is more accurate than the natural variability of H2O? This could also mean that OMPS-LP is capturing less variability?

In Figure 9, the variability of both the LP and M2-SCREAM products are shown alongside the variability of the differences between these products. LP's variability is slightly less than that of M2-SCREAM, but more importantly the variability of the differences between the products is smaller than the variability of either product. If LP were less accurate than the natural variability of H_2O, it would be expected that this plot would show the variability of the differences between the products to be similar to the variability of M2-SCREAM.

L278 Why is Figure 10 displaying SAGE II and SAGEIII (not ISS I presume). These datasets have not been discussed, I suggest deleting those lines. Also, Figure 10 looks pixelated. Why was SABER not considered for this figure?

We have removed the SAGE II and SAGE III lines from this plot, retaining only the instruments considered elsewhere in this manuscript (MSL, ACE-FTS, SAGE III/ISS, and OMPS LP) as suggested.

We're not sure why the image is pixelated as it looked normal before submission, but we have updated the image with a high-resolution PNG that hopefully addresses that problem.

SABER was omitted for convenience/time; this figure follows from Davis et al. (2021)'s Figure 10. MLS is the primary instrument we are validating against; ACE-FTS and SAGE III/ISS are considered to provide context when comparing this manuscript with past validation efforts as

well as to ensure that the NN models correctly predict H_2O profiles in locations besides where there are LP-MLS coincident measurements.

L307 should this be 30 km? after the tape recorder discussion.

Measurements outside of the tropics seem to be accurate even above 30 km. To better clarify, we revised the sentence at the end of the tape recorder discussion:

> "We therefore advise that users exercise caution when using the OMPS LP H_2O product above 30 km **in the tropics**."

And we have added this caveat in the conclusions:

> "Overall, we find that the LP product **generally** performs comparably to MLS over the 11.5-40.5 km altitude range considered, enabling the continuation of the MLS water vapor record for these altitudes. **The exception is in the tropics above 30 km, where a period in 2025 shows larger errors relative to MLS than those seen during the 2014-2024 training period; we advise that users exercise caution when using LP H_2O data in this regime**."

References

Chen, Z., DeLand, M., & Bhartia, P. K: A new algorithm for detecting cloud height using OMPS/LP measurements, Atmospheric Measurement Techniques, 9(3), 1239–1246, https://doi.org/10.5194/amt-9-1239-2016, 2016.

Davis, S. M., et al.: Validation of SAGE III/ISS Solar Water Vapor Data With Correlative Satellite and Balloon-Borne Measurements, Journal of Geophysical Research: Atmospheres, 126, e2020JD033 803, https://doi.org/https://doi.org/10.1029/2020JD033803, e2020JD033803 2020JD033803, 2021.

Millán, L., et al.: The Evolution of the Hunga Hydration in a Moistening Stratosphere, Geophysical Research Letters, 51, e2024GL110 841, https://doi.org/https://doi.org/10.1029/2024GL110841, e2024GL110841 2024GL110841, 2024.