Overview

This study presents a machine learning approach using OMPS-LP radiances to retrieve water vapor, with MLS water vapor profiles serving as the training target. The method employs a neural network (NN) model trained on 12-channel OMPS-LP inputs, and the manuscript outlines the application and evaluation of the resulting product.

Overall, this may be a valuable contribution, particularly in leveraging machine learning for satellite-based water vapor retrieval. However, I am concerned that the NN is not actually retrieving water vapor since the OMPS-LP instrument spectral resolution is too broad to capture the NIR water vapor lines in the 0.95μm region (Figure A, below). The authors need to prove that the system retrieves water vapor. In addition, several aspects of the study require clarification or improvement.

Specific comment on retrieval

SAGE III/ISS uses the 930 to 950 nm channels to retrieve water vapor. The SAGE spectrometer has a spectral resolution of ~3 nm (Davis et al. 2021). The SAGE water vapor retrieval is very noisy even with the high spectral resolution and solar radiation source. OMPS-LP uses scattered radiation, and spectral resolution is ~40 nm in this wavelength region. Looking at Figure A below, I have a hard time seeing how OMPS-LP can detect water vapor at all unless it is highly elevated (e.g. manuscript Figure 1).

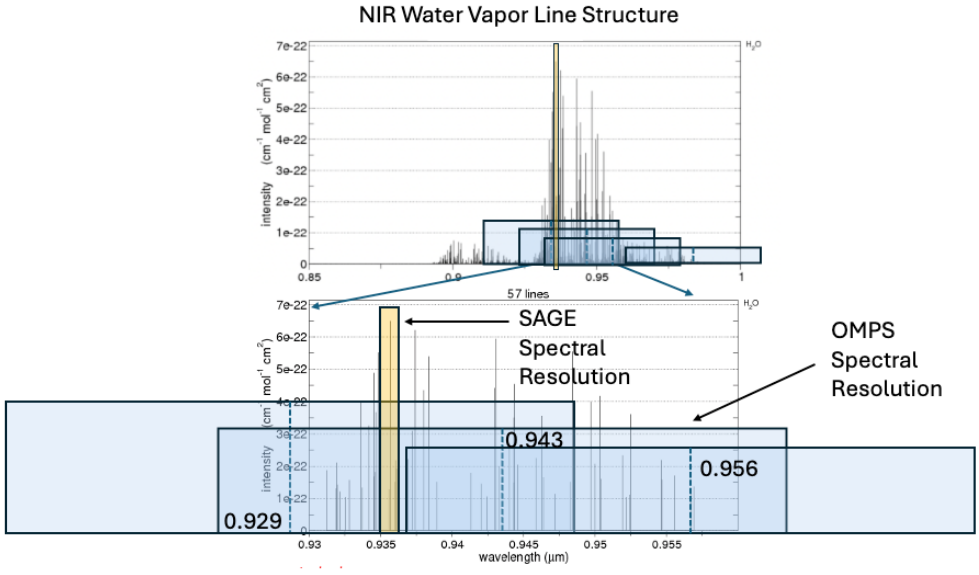

Figure A. NIR water vapor line intensity (lower figure is a blow up of the upper figure). Blue boxes show the OMPS spectral resolution and yellow box shows the SAGE spectral resolution for comparison. Given the narrowness of the lines and spectral width, the changes in water vapor will be difficult to detect.

At background levels (~ 4.5 ppmv), the authors are getting ~0.25 ppmv variation (Fig. 6b). Given the wide spectral resolution of OMPS the channels 929, 943, 956, 970, and 983 reported on line 73. 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.

Detailed comments and questions for revision:

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?

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.

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?

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.

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?

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.

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.

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.

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?

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.

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.

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.

14. Figures and Presentation

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

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.

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.

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

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?

Reference:

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* 126.2 (2021): e2020JD033803.