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-to-point response to reviewer #1 below. Here, the reviewer's general remarks, as well as the 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) We emphasized the fact that the new ML-based algorithm is not replacing the OE retrievals but is employed to fill the retrieval gaps due to computational limitations. New statements in the abstract, introduction, and conclusions section ensure the reader that the ML model is not a stand-alone algorithm that is trained once and is designed to provide TROPESS products from now on but instead is run alongside the OE code to enhance the operational products rather than replace them.
2) We present results for more of the predicted variables. Whereas the initial manuscript version only showed results for total column concentrations, total column retrieval errors, and degrees-of-freedom, the revised manuscript adds panels for the column averaging kernel at 162 hPa in Fig. 2. There is also a new Fig. 4, which presents maps for the column averaging kernel at 383 hPa, as well as global statistics for the full column averaging kernel, total and tropospheric column concentrations, and the degrees of freedom for global predictions over a full day. We also discuss evaluation statistics in more detail.
3) Related to the second major change: We add a bit more discussion on generalization of the model, both in the evaluation section and in the discussion on the new Fig. 4. That Figure shows global predictions for a day that is not included in the training period (instead about 6 months past the last training day) to demonstrate the ability to predict on unseen data. It is important to note that this is not the intended application of the hybrid ML algorithm (see the first major change).
4) We reorganized section 4.2 to improve readability.

<u>Overall Feedback:</u>

I think that this paper is great and worthy of publication with only minor revisions. Most of the feedback below consists of recommendations about different analytical techniques that may help improve the paper.

One specific point of feedback is a rather general critique of the value of the simple linear regression analysis for statistically comparing similar datasets. In the limit that correlations approach 1 linear regression plots start to provide limited visual evaluation ability – or to put it more bluntly everyone has seen a good looking regression and they often look less than usefully similar. There is a more robust approach to construct comparisons like this known as the "Bland-Altman Plot", that also helps to incorporate additional statistical information about the data and better poses the fundamental

question: "could variable B statistically replace variable A". Figure 2 in your paper currently does a reasonable job of displaying the other dimensions that can matter for such a regression; given that they display retrievals, errors, and degrees of freedom.

I would recommend at least looking at Bland-Altman plots in response to this review and potentially including such analysis in the paper itself. In particular, Bland-Altman offers a better visual framework for handling data comparison tasks if the distribution of data is not linearly distributed or has uniquely variable uncertainty in either of the compared datsets. This is particularly true for variables with non-gaussian variability (e.g., logarithmic distributed variables such as optical thickness) or heteroscedastic uncertainty/variability. It looks to me as though these considerations might matter for the datasets in panels a,b, and c of Figure 2 – whereas panel d appears clearly normally distributed at all scales. One further relevant concern here is that neural network architectures such as yours are largely tuned toward gaussian process prediction and can struggle (without adequate consideration) to handle heteroscedastic variability in datasets because of the common isotropic noise assumption [Stirn et al., 2022].

An example of demonstrating a situation where analysis with Bland-Altman can significantly improve your analytical toolkit can be found in Knobelspiesse, et al. (2019). This paper explores an instrument intercomparison for radiometric polarimeters – which exhibit non-gaussian distributions in observed radiances as well as heteroscedastic variability in in the degree of linear polarization (DoLP) uncertainty. The example therein is discussed in section 3.C and summarized visually in Figure 8 and Figure 9. The links below summarize the methodology and has a python notebook demonstrating examples.

https://github.com/knobelsp/BlandAltman?tab=readme-ov-file
https://colab.research.google.com/github/knobelsp/BlandAltman/blob/main/BlandAltman.ipynb

Furthermore, as a ML retrieval example of how heteroscedasticity can cause issues with application of machine learning methods - the cloud microphysics retrievals in Miller et al., 2020 struggle to handle retrievals across the whole range of variability of the retrieval datasets. This is because of the statistical distributions of radiances and DoLP have rather heteroscedastic dependencies on the geophysical variables attempting to be retrieved.

> We agree with the reviewer that, for correlations approaching unity, simple regression plots can mask potential issues, particularly for ML-based predictions with non-Gaussian characteristics. The model can give the impression of strong overall performance while still exhibiting deficiencies at the tails of the distribution.
>
> Following the advice provided in the referenced papers and Jupyter notebooks, we added panels showing Bland-Altman plots to the revised Fig. 2. These plots strengthen our initial analysis, as the distributions show no issues at the tails or magnitude-dependent slopes. Note that the Bland-Altman plots are in addition to the regression plots (similar to the Jupyter notebook examples), as most readers are familiar with this representation. We also added the following discussion to the manuscript:

"To complement the regression analysis, we employ Bland-Altman plots to further assess the agreement between the OE and predicted results. This approach highlights systematic differences and potential heteroscedasticity that may not be apparent in standard correlation-based evaluations, especially for non-Gaussian or magnitude-dependent variability, and has been shown to provide a more robust framework for intercomparison of geophysical datasets (e.g., Knobelspiesse et al., 2019). Panels b, d, f, and h show the Bland-Altman distributions, with the paired mean of predicted and OE values on the x-axis and the difference between the two datasets on the y-axis, along with three horizontal dashed lines. The central line denotes the mean difference (bias), while the outer lines show the 95% limits of agreement (mean $\pm$1.96 standard deviations), indicating the interval that contains approximately 95% of the differences under the assumption of normally distributed residuals.

The distributions show no evidence of magnitude-dependent bias or systematic slope. The proportion of data points within the 95% limits of agreement (95%, 95%, 96%, and 95%, respectively) aligns closely with expectations for normally distributed residuals, and 79%, 80%, 86%, and 78% of points fall within $\pm$1 standard deviation. This indicates well-behaved error distributions with no pronounced evidence of heavy tails or heteroscedastic spread, suggesting that the model adequately captures variability across the full dynamic range. This is particularly relevant for machine learning retrievals, which can otherwise struggle in the presence of heteroscedastic relationships between observables and geophysical variables (e.g., Miller et al., 2020), and indicates that such effects are not evident in our ML predictions. Minor increases in spread near the peak of the distributions reflect regions of highest data density and do not indicate systematic bias or magnitude-dependent variability.

The close agreement between predicted and OE-derived diagnostics indicates that the ML model not only reproduces the retrieved state, but also captures the associated sensitivity and uncertainty characteristics of the OE solution. This consistency is critical for downstream applications, such as data assimilation and model evaluation, where the proper interpretation of retrievals depends on the availability of reliable averaging kernels and error estimates."

Specific Feedback:

1. **Fig 3e,f – could you match the number of significant figures in each of these colorbar labels? That might also help you keep the formatting of the numbers from one bar from nearly overlapping with the other as they are now.**

   We matched the number of significant digits in the colorbar labels of the revised Fig. 3. We also made sure to follow the same advice for the new Fig. 4.

2. **I would recommend a more clear and early statement of *why* you want to look at the power spectrum for scale breaks in section 4.2. This will help keep the reader's attention on the importance of the relevant result when you finally introduce the**

figures. I understand that this is a good way to test and demonstrate its value beyond a simple sub-grid interpolator. But I think perhaps you could reorganize this section to address this. For example you mention scale-breaks on line 161 but it takes until line 180 to explain of why spatial variability and power spectrum is of interest.

We agree that the initial structure of section 4.2 was a bit awkward. We restructured that section with the following intent: (i) motivation and hypothesis framing, explicitly defining the conditions under which higher-resolution information is required, (ii) method 1 applying simple linear interpolation, (iii) method 2 performing scale analysis (including theory, implementation, and results), (iv) robustness check, and (v) conclusions.

More specifically, here is the revised introduction to this section that better motivates the two approaches:
"A key question is whether the ML product captures physically meaningful CO variability below the nominal 0.80° TROPESS retrieval resolution, or whether it primarily behaves as a spatial interpolation of the OE field. While interpolation can reconstruct smooth fields between observations, it cannot introduce new information at unresolved spatial scales. To address this, we analyze the spatial variability of the ML and OE products using spectral methods, which provide a scale-dependent view of their structure.

To investigate this, we employ two complementary approaches: (i) comparing the ML-predicted CO fields with linearly interpolated TROPESS CO retrievals, and (ii) analyzing the spatial power spectral densities $E_I(k)$ to identify scale-dependent variability, particularly in the sub-0.80° domain. Together, these approaches allow us to assess whether interpolation is sufficient to capture the underlying structure, or whether additional variability persists at smaller scales that requires the higher-resolution information provided by the ML model."

We also moved some technical detail of the power spectrum calculation further down (after the equation), and added some discussion on the power spectrum behavior for the linearly interpolated data:
"For comparison, the power spectral density of linearly interpolated OE fields exhibits a similar large-scale slope ($\beta \approx 1.70$) but a much steeper decay of variability toward smaller spatial scales ($\beta \approx 3.28$), reflecting the smoothing inherent in interpolation. While the apparent scale break shifts slightly, its exact location is sensitive to the fitting procedure and binning choices and is therefore not interpreted further."

Line 179: Typo – "cahnges"

Fixed