We'd like to thank the editor for handling our manuscript, as well as reviewer #2 for reading our manuscript and providing numerous helpful suggestions for improvement. We have carefully read through all the comments and questions and revised the manuscript accordingly. Please find our point-to-point response to reviewer #2 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 that the reader understands 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>Major comments:</u>

1. Generalizability and data-splitting strategy

The use of a 98%/1%/1% split for training, validation, and testing raises concerns regarding the generalizability of the ML model. While the absolute number of samples in the validation and test sets is large, the strong spatial and temporal correlations inherent in satellite observations mean that a random split does not guarantee independence. For example, the 10 June 2023 wildfire case is drawn from the 04/2023-01/2025 period used for training. Given that 98% of the data are included in the training process, the test set likely contains many samples that are spatially and temporally adjacent to training samples. Under this split strategy, the reported test-set performance may largely reflect re-

prediction of patterns already seen during training rather than true out-of-sample generalization.

A more robust evaluation would involve temporally or regionally independent splits (e.g., holding out entire months, seasons, or geographic regions), or comparison with fully independent third-party observations such as in situ or ground-based measurements. As currently implemented, the 98%/1%/1% split limits the interpretability of the reported test results.

We thank the reviewer for this important comment regarding the use of a random 98/1/1 split and the implications of spatiotemporal correlations in the data. We addressed this comment in several ways:

(1) Clarification of the model objective
We would like to clarify that the goal of the proposed framework is not to replace optimal estimation (OE) with a fully independent, stand-alone machine learning model that was trained on a limited historical dataset and subsequently used for long-term standalone predictions. Instead, the model is designed as part of a hybrid OE–ML system that emulates and extends OE retrievals for the same observing system. In this framework, the ML component operates alongside OE, fills gaps where OE retrievals are unavailable or fail, and is periodically retrained as new OE retrievals become available.

(2) Interpretation of the random data split
We agree that a purely random split does not enforce strict independence, as atmospheric states and observations are correlated in space and time. However, in the context of the intended application, this is not a limitation, but a reflection of the operational setting in which the model is intended to be used. The objective of the split is to evaluate how well the model reproduces OE retrieval behavior across the distribution of atmospheric states and viewing geometries sampled by the instrument, rather than to test fully independent spatiotemporal generalization.

To ensure that the split still provides a meaningful evaluation, we verified that the training, validation, and test sets have statistically consistent distributions (via Kolmogorov–Smirnov tests) and exhibit highly consistent performance metrics. This indicates that the model does not overfit and generalizes well within the sampled observational distribution.

(3) Generalization and independence considerations
We acknowledge that this setup does not constitute a stringent test of independent spatiotemporal generalization. However, such independence is not the primary requirement for the intended hybrid application, where the model is expected to remain closely tied to the evolving OE solution through periodic retraining. In this sense, the model is designed to “interpolate” within the relevant observational manifold rather than to extrapolate far beyond it.

(4) Additional temporal extrapolation test (Fig. 4)
To address the reviewer's concern more directly, we have added a new experiment (Fig. 4) in which the model is evaluated on a day outside the training period (summer of 2025). This provides a complementary test of temporal extrapolation and demonstrates that the model retains strong predictive skill under these conditions. We emphasize, however, that such standalone predictive capability is not the primary objective of our hybrid framework, but rather a secondary property.

These points have been clarified in the revised manuscript to better distinguish between interpolation within the sampled distribution and extrapolation beyond it.

Changes to the manuscript regarding (1)

We added clarifying statements throughout the manuscript. In the abstract we added: "The framework is designed to emulate and extend the OE retrieval, rather than replace it, by providing full spatial coverage and enhanced resolution consistent with the underlying physical solution."

In the introduction we say: "Importantly, the ML component is designed to emulate the OE retrieval and its associated diagnostics, rather than to replace or surpass the underlying physical solution, thereby extending OE--derived information to full spatial coverage. The TROPESS-HYREF framework is therefore intended as a hybrid OE-ML system, in which the ML model operates alongside OE, for example by filling gaps in retrieval coverage, and can be periodically retrained as new OE results become available."

Other small additions in section 3 remind the reader of the complementary nature of the ML model.

Changes to the manuscript regarding (2)

We added the following paragraphs to the evaluation discussion:
"Performance metrics are consistent across training, validation, and test datasets, with nearly identical correlation coefficients ($|\Delta r| < 0.0006$), normalized RMSD values ($|\Delta\text{RMSD}| < 0.21\%$, apart from the last two column AK levels deep in the stratosphere where values of effectively 0), and biases ($|\Delta\text{bias}| < 0.12\%$) for all predicted variables. This indicates that there is no evidence of overfitting and that the model exhibits stable behavior across the available datasets.

We note that the use of a random split does not enforce strict independence between training, validation, and test datasets, as atmospheric states exhibit strong spatial and temporal correlations. In this study, the purpose of the split is therefore not to assess fully independent generalization, but to evaluate how well the model reproduces OE retrieval behavior across the distribution of atmospheric states and viewing geometries sampled by the instrument.

To ensure that the split remains representative, we verified that the distributions of key variables are statistically consistent across training, validation, and test datasets using Kolmogorov–Smirnov tests. Combined with the consistent performance metrics reported above, this indicates that the model generalizes well within the sampled observational distribution. This behavior is consistent with the intended hybrid OE–ML application, in which the model is designed to operate on the same observational distribution as OE."

Changes to the manuscript regarding (3)

We added this paragraph right before our new Fig. 4:
"The close agreement between the ML predictions and OE retrieval shown in Fig.3 primarily reflects interpolation within the sampled observational distribution, as this day is part of the training period, rather than fully independent spatiotemporal generalization. To further assess the performance of the ML model beyond the training period, Fig.4 shows a global scene for 8 June 2025, which lies outside the training range (04/2023–01/2025). This experiment represents a limited temporal extrapolation test rather than a comprehensive assessment of long-term model stability. The results demonstrate that the model retains strong predictive skill under these conditions, indicating that it can generalize to unseen temporal states to a certain extent. However, we emphasize that such standalone predictive capability is not the primary objective of the framework. The model is designed to operate as part of a hybrid OE–ML system, benefiting from periodic retraining and remaining closely tied to the evolving distribution of OE retrievals."

Changes to the manuscript regarding (4)

Figure 4 is similar to Figure 3 but shows global data for a day outside the training data set. Model performance is very similar to that shown in Fig. 3.

2. Evaluation of predicted diagnostics

A key claimed advantage of the TROPESS-HYREF framework is its ability to predict retrieval diagnostics such as averaging kernels, DoF, and retrieval errors. However, the evaluation presented in the paper focuses largely on CO column concentrations. Additional assessment of the predicted diagnostics would strengthen the paper. For example, how accurate and stable are the ML-predicted averaging kernels relative to OE? Are the predicted errors statistically consistent with OE-derived uncertainties? How suitable are these diagnostics for downstream applications such as data assimilation?

We thank the reviewer for highlighting the importance of evaluating predicted diagnostics such as averaging kernels, retrieval errors, and degrees of freedom.

In response, we have extended the analysis in several ways:

(1) **Averaging kernels (AKs)**
We now include additional evaluation of column averaging kernels, both in terms of spatial structure (Fig. 4b,d,f) and vertical characteristics (Fig. 4g), in addition to model evaluation (Fig. 2c,d). The results show excellent agreement between ML and OE, with nearly identical global mean profiles and differences well within the natural variability of the OE retrieval. Quantitatively, the median differences are below 1.5%, with the majority of predictions within 10% of OE values.

(2) **Retrieval errors and degrees of freedom (DoF)**
We have added probability density distributions of the differences between ML and OE total and tropospheric column retrieval errors (Fig. 4h), as well as DoF (Fig. 4i). These distributions are centered near zero and exhibit narrow spreads, indicating that the ML model accurately reproduces both the magnitude and variability of these diagnostics. For example, the majority of retrieval error predictions lie within ±6.11% of OE values with a median difference of 0.04%.

(3) **Statistical consistency with OE diagnostics**
While explicit uncertainty estimates for AKs and DoF are not available in the OE framework, the close agreement in both mean structure and distributional behavior (including near-zero-centered differences and narrow full-width-at-half-maximum values) indicates that the ML-predicted diagnostics are statistically consistent with those derived from OE, within the variability of the retrieval itself.

(4) **Suitability for downstream applications (e.g., data assimilation)**
A full assessment of suitability for data assimilation is beyond the scope of this study. However, the demonstrated agreement in AKs, retrieval errors, and DoF suggests that the ML-predicted diagnostics preserve the key characteristics required for such applications. A more comprehensive evaluation in an assimilation framework is the subject of ongoing work and will be addressed in future studies.

These additions and clarifications have been incorporated into the revised manuscript.

Changes to the manuscript regarding (1) and (2)

Here is the new discussion for Fig. 4 (focusing on the diagnostic variables):
“In addition to column concentrations, Fig. 4 demonstrates that the ML model accurately reproduces key retrieval diagnostics. The column AK exhibit a maximum at 383 hPa and is shown in panels b, d, and f, showing strong agreement in both spatial structure and magnitude. The median difference is 0.67%, and the majority of ML predictions lie within $\pm$8.50% of the OE results. The global mean AK profiles (Fig. 4g) are nearly identical between OE and ML, with differences well within the natural variability of the OE retrieval. Median differences over all vertical levels are within 0.001 and 90% of predictions are within 0.028 of the true AK value at any level. This translates to 1.5% and 10%, respectively, at levels where the AK is noticeably different from zero, i.e., in the troposphere above the surface.

Likewise, differences in retrieval errors (Fig. 4h) and degrees of freedom (Fig. 4i) are centered near zero and exhibit narrow distributions, with full-width-at-half-maximum values of $0.02 \cdot 10^{18}$ molec. cm^{-2} and 0.032, respectively. This behavior indicates that the ML–predicted diagnostics are statistically consistent with those derived from OE, within the intrinsic variability of the retrieval."

Changes to the manuscript regarding (4)

We added this sentence at the end of the subsection:
"Together, these results demonstrate that the ML framework not only reproduces the retrieved state, but also captures the associated sensitivity and uncertainty characteristics of the OE solution for unseen atmospheric conditions, while resolving finer spatial structures beyond the native OE sampling. These characteristics suggest that the predicted diagnostics retain the key properties required for downstream applications such as data assimilation, although a full assessment within an assimilation framework is beyond the scope of this study."

3. Claims regarding performance relative to OE

Some statements suggesting that the ML system may "outperform" the OE retrieval are concerning. Given that the ML model is trained to reproduce OE results, it is unclear how it could outperform the OE retrieval in a physical sense. Clarifying that the ML system improves coverage and computational efficiency, rather than retrieval accuracy relative to OE, would help avoid overinterpretation.

We thank the reviewer for this important clarification and fully agree that a machine learning model trained on OE retrievals cannot surpass the physical accuracy of the OE solution itself.

The intent of the TROPESS–HYREF framework is not to outperform or replace the OE retrieval, but to emulate and extend it by providing full spatial coverage and increased spatial detail consistent with the underlying physical solution. In this sense, the ML component is designed as part of a hybrid OE–ML system that reproduces OE retrievals while enabling additional applications such as gap-filling and rapid prediction.

The statements in the manuscript referring to improved representation of sub-grid variability (e.g., Fig. 5) are not meant to imply that the ML model exceeds the accuracy of the OE retrieval. Rather, they reflect that the ML predictions retain spatial variability that would otherwise be smoothed when OE results are interpolated onto a regular grid. In this context, the comparison is between ML predictions and interpolated OE fields, not between ML and the OE retrieval itself. The ML model therefore provides a spatial refinement of OE-derived information, rather than an independent improvement of the physical retrieval.

To avoid potential ambiguity, we have revised the Abstract, Introduction, and Conclusions to explicitly state that the ML model is designed to complement and extend

the OE retrieval, rather than replace or surpass it (see also our response to comment #1). We have also carefully reviewed the manuscript to ensure that no statements can be interpreted as suggesting that the ML model exceeds the physical accuracy of the OE retrieval.

We hope this clarification resolves the reviewer's concern.

Minor comments:

L96-97: The description of the forward and backward processes may be confusing for general readers, as it does not explicitly mention the backpropagation algorithm. A brief clarification would improve readability.

We agree that the original wording may have been unclear for readers less familiar with neural network training. We revised the sentence to explicitly reference backpropagation and briefly clarify the forward and backward passes.

Here is the updated text in the manuscript:
"For each training iteration, batches of samples are passed through the model in a forward pass to compute predictions, followed by a backward pass in which model weights are updated via backpropagation. Each mini-batch contains 8,192 samples. Further details on these parameters and their impact are provided in Reed et al. (1999); Goodfellow et al. (2016); Werner et al. (2021)."

L101: The time range (04/2023–01/2025) is critical information and should be mentioned in the Data section.

We agree that this information is useful for context and have added the training period (04/2023–01/2025) to the Data section. We also added a clarification in the Conclusions emphasizing that the model is intended to operate alongside OE with periodic retraining using newly available data:
"In the current implementation, the model is trained on approximately two years of TROPESS OE retrievals; however, ongoing work focuses on incorporating regular retraining using newly available OE data. This ensures that TROPESS-HYREF continuously adapts to evolving atmospheric conditions while maintaining consistency with the OE retrieval, reinforcing its role as a complementary extension rather than a stand-alone predictive model."

Section 3: Feature preprocessing is not described: were different input features (radiances, latitude/longitude, UTC time, a priori values) normalized or scaled prior to training?

We agree that this detail was missing. We added a brief description of the preprocessing and sanitation steps in the model setup section, clarifying that invalid samples were removed using quality and range filters, extreme outliers in the label distributions were masked using percentile-based tail cutoffs, and both features and labels were standardized prior to training. The exact paragraph is:

"Prior to training, these inputs and outputs were filtered to remove invalid samples using a set of basic quality filters, including non-finite values, fill values, extreme outliers, failed retrieval quality flags, and target values outside the valid retrieval range. In addition, extreme outliers in the label distributions were masked using percentile-based tail cutoffs, and both input features and output labels were standardized before training."

L106: Is it necessary to use the full spectrum, or would a reduced set of CO-sensitive channels suffice?

We thank the reviewer for this insightful question regarding the choice of input channels. We explored reduced channel sets focusing on CO-sensitive spectral regions and found that using the full spectrum provided slightly improved performance. This is likely due to additional information contained in other channels (e.g., temperature and humidity sensitivity), which is implicitly used by the OE retrieval but must be learned by the ML model. Given the available computational resources, the use of the full spectrum does not pose a limitation.

We added a brief clarification to the manuscript:
"We tested reduced channel sets focused on CO-sensitive regions but found that using the full spectrum provided slightly improved performance, likely due to additional information on atmospheric state variables (e.g., temperature and humidity)."

L188-189: The mesoscale processes associated with the identified spectral break should be discussed more explicitly.

We agree that a more detailed discussion of the underlying physical processes would provide additional physical insight. However, such an analysis is beyond the scope of this study, which focuses on the retrieval framework itself. We revised the sentence to clarify this point:
"This feature is consistent with the de-correlation length scale of CO (not shown) and indicates a transition to enhanced small-scale variability. A detailed attribution of the underlying processes is beyond the scope of this study."

Figure 4: A direct comparison of power spectral densities between ML-predicted CO and interpolated OE CO would be informative and could further clarify the added value of the ML approach.

We agree that such a comparison is informative. We performed this analysis and found that interpolated OE fields exhibit a similar large-scale slope ($\beta \approx$1.7), but a much steeper decline at smaller scales ($\beta \approx$ 3.3$), reflecting the smoothing effect inherent to interpolation. We added a corresponding statement to the manuscript summarizing this result: "For comparison, the power spectral density of linearly interpolated OE fields exhibits a similar large-scale slope ($\beta \approx$-1.70) but a much steeper decline at smaller scales ($\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."

To maintain clarity of the figure, we chose not to include an additional curve, as the key effect is already demonstrated by the interpolation comparison in panels (a)–(b) and the PSD behavior shown in panels (c)–(d).