Response to Referee #2

September 26, 2025

We thank the reviewer for their careful reading and constructive suggestions. We address each point below.

Major concerns

Since your ML prediction actually predicts the distribution (or quantiles), it is a pity that the discussion on how to use the quantiles to develop a IWC or Dme flagging algorithm, especially given the fact that your results contain so many small IWC/Dme values that are apparently unreal but just ML artifacts because the training focuses on learning the distribution. For example, the fake "near-empty" clouds near the freezing layer in Fig. 3 case, the much larger integrated IWC value compared to your retrieved IWP in the clear-sky regime in the Fig. 7 case, the spike in Fig. 6, and the "better-than-CloudSat" in the lower sensitivity threshold suggested in your Fig. 4 PDF comparisons. My guess is the PDF width from your prediction for these small IWC cases should be larger than your retrieved IWC value, but as the errorbar was never used for filtering, I don't know if that's the case or not. Ultimately if these become operational or research products for ICI, you'll be required to provide a quality flag or something similar to let the user know which retrievals are not trustworthy. My suggestion is to try playing with different thresholds (e.g., standard deviation, 75th quantile – 25th quantile, etc.) to develop a flagging algorithm and show the confusion matrix to demonstrate both clear-sky and cloudy-sky are accurately captured. Also, please use the flagging mechanism to update Fig. 3, 4, 5, 6, and 7.

- We thank the reviewer for this suggestion, which highlights a strength of our retrieval approach — the retrieval of quantiles of the PDF. We tested at thresholds based on quantile spreads, relative to the distribution mean. The 75th-25th spread was too narrow to flag problematic cases reliably, but the 99th-1st spread showed that the low-level clouds have the highest uncertainties, indicating that their predicted distributions are strongly skewed. In this scene, rain was present underneath the low-level clouds, potentially explaining their occurrence.

  We agree that such measures could form the basis of a flagging algorithm. However, our primary aim of this study is to demonstrate the capabilities and limitations of ICI profile retrievals, rather than develop an operational product. For this reason, we prefer not to apply filtering to the main results, as this would risk 'hiding' problematic cases from the reader. However, the findings resulting from this suggestion remain valuable, and should be included.

  We also note that excluding cases would affect the overall statistics of the retrievals. For this reason, we believe that the decision of whether and how to filter should be left to an end-user, depending on the application.

- **Changes to the manuscript:** Fig. 3 has been updated to include an additional subplot showing the spread between the 1st and 99th quantile, relative to the mean. The discussion around Fig. 3 has been extended to discuss the high uncertainties associated with the low-level clouds, and to describe how this metric could be used for a flagging algorithm in future applications.

With the same IWP value, the clouds could be top-heavy (i.e., developing), U-shape (i.e., mature), or bottom-heavy (i.e., decaying). The scientific value of profile retrieval mainly lies in being able to differentiate cloud vertical structure, and potentially understand better the system life stage. The three cases shown in Fig. 15 demonstrate that your algorithm could achieve this capability. However, the averaging kernel and DoF discussions all focusing on the mean vertical resolution for all training samples. I would strongly recommend updating the averaging kernel results for different types of cloud. Given the fact that your training samples are big, you can use some clustering method (e.g., PCA, k-clustering) to separate them into a few representative types, and then compute the results for each cloud type. There is no way that the 2.5 km vertical resolution can be achieved for all kinds of ice clouds between 5-15 km, so it would be much more appreciated if readers can be informed the real physical resolution that can be achieved for different cloud types. I'm especially interested to see how multi-layer clouds can be resolved in your profile retrievals.

- It is indeed true that different resolutions will be achieved with different cloud types. We explored the three suggested cloud types (multi-layer, top-heavy, and bottom-heavy) using simple physical rules (given in Table 1 of the manuscript). We considered using a clustering method, but mapping the clusters cleanly onto physically-interpretable cloud-types could be difficult. The averaging kernels were computed for each case and presented in the manuscript.

  Since there is no unique definition for each cloud-type, adjusting, for example, the mass fractions led to differences in the results. Also, IWP, IWC, and Z_{m} are already restricted to ensure near-linearity, so further filtering leaves few cases, causing instability in the results. Therefore, exact values were unstable, but overall qualitative trends can be described. Results are presented in the manuscript.

  Ideally, we would apply further restrictions, e.g. looking at only top-heavy clouds with IWP $>$ 1 kg m^{-2} to target anvil clouds. However, not enough cases remain to produce stable results.

- **Changes to the manuscript:** A new figure has been added to the manuscript (Fig. 12), showing averaging kernels derived for the three cloud classes suggested by the reviewer. The discussion has been extended to present these new results. A new table has also been added (Table 1) to clearly present the five subsets of IWC now used in the averaging kernel calculations. The conclusion now includes the new results. Regarding averaging kernels, we also made very minor changes to the original IWC averaging kernels, including labels and a small change in some resolutions due to a code error. The conclusions remain the same.

Minor points

As mentioned in this work, the operational ICI products include mean mass height Zm and mass-weighted column averaged Dme. Could you check if your retrieved IWC can give you the mean mass height that's consistent with Zm, and your Dme and IWC profile retrievals can yield agreement with mass-weighted column averaged Dme? I'm especially curious about the former.

- To check this, we produced distributions of Z_{m} derived in three ways: from database IWC, from retrieved IWC, and a direct retrieval of Z_{m}. Similar distributions were made for D_{m}, deriving the values from both IWC and $D_{\mathrm{m,\,IWC}}$ profiles. The results show better agreement for Z_{m} than for D_{m}. However, this is somewhat expected since D_{m} must be calculated using both retrieved IWC and retrieved D_{m} profiles, which can amplify existing inaccuracies.

- **Changes to the manuscript:** Fig. 6 has been extended to include two extra panels displaying distributions of Z_{m} and D_{m}. The text has also been updated to discuss the results seen in the updated plots.

Your ML retrieval results suggest degradation happens above 12 km, but later on your averaging kernel experiments find the vertical resolution can be achieved stably below 15 km. Why this discrepancy?

- We believe that the reviewer's impression of degradation above 12 km stems from Fig. 3, where low-IWC are missing at altitudes above 12 km. Our wording in the discussion may have been misleading; this behaviour reflects poorer performance for low-IWC cases rather than at the specific altitude. These cases are difficult to detect due to ICI's low sensitivity to thin cloud, and this result is independent of altitude, as seen in the lower left corners of all panels of Fig. 2.

  In contrast, our conclusion of reliable performance up to 14 km was based on analyses over the entire IWC range, with particular focus on the mid-IWC regime where ICI is most sensitive. The averaging kernel analysis was performed on a subset of data with IWP > 0.1 kg m^{-2} and IWC > 0.01 g m^{-3}, thus excluding thin cloud cases. The kernels therefore remain stable not despite, but partly because of, the exclusion of low-IWC cases.

- **Changes to the manuscript:** The discussion around Fig. 3 has been updated to better clarify the potential reason for missing clouds in Fig. 3.

In the averaging kernel experiment, Dme response function is bi-model, but IWC response function is not. Do you know why they are inconsistent? Does this suggest ICI is mostly useful for sensing ice particles in anvils and cloud bottom? But the vertical resolution of 5 km at 10 km strikes me... Please elaborate your thoughts.

- Not too much importance should be placed on the exact value of the measurement response. Values near 1 indicate retrievals largely based on data, while values near 0 indicate strong a priori influence. For this reason, we would hesitate to conclude that the measurement response implies usefulness for only anvils and cloud bottoms. Even at mid-altitudes (e.g. 7.5 km), the response remains high (0.85-0.9), showing that the retrieval is still useful.

  Although we cannot fully explain the measurement response shape, similar oscillations appear in classical OEM retrievals (Forkman et al., 2016), typically when the effective vertical resolution is high relative to the altitude spanned by the measurement response. However, we realise that referring to the measurement response as a 'response function' may lead to over-interpretation. Rather, it is defined per level, and should not be labelled bimodal as if it were a continuous function.

  Regarding the resolution of 5 km at 10 km, the poorer resolution compared to IWC is an expected result, since more information is needed to constrain particle size than the location of ice. Also, the resolution estimate summarises *average* behaviour across a range of cloud types, as the reviewer rightly noted in an earlier comment, and does improve for only lower-IWP cases.

- **Changes to the manuscript:** References to the 'response function' have been updated to the 'measurement response'.

Line 170: why using the mean instead of the peak of the predicted PDF? The PDF could be very skewed for many cases.

- We agree that the predicted PDFs are often skewed, which motivates our use of a method that retrieves non-Gaussian PDFs. We chose to use the mean as our final retrieved value since it is a measure of the full PDF and thus captures any skewness. As a result of the reviewer's earlier point on quality flagging, there exist cases where the 99th-1st quantile spread is large, indicating a strong skew and therefore a difference between the mean, median, and peak. We have also checked retrievals using the median of the distribution and found systematic underestimation of IWC. Skew in the higher-IWC end of the PDF therefore pulls the mean correctly upward, showing the importance of capturing the full PDF.

  That said, we recognise the merit of the reviewer's point. In our ongoing work on retrievals based on real satellite observations, we are evaluating the use of the PDF peak. However, small numerical issues arise when constructing the PDF from the retrieved CDF. This instability necessitates smoothing the PDF, which complicates taking the peak as the estimate.

We appreciate the reviewer's interest in this choice, which motivates our ongoing exploration of the two approaches.

5b vs. 5a: I can understand why your training database has low bias near the tropopause because your database doesn't include the CALIPSO measurements while DARDAR has. But why your PBL cloud ice are also low-biased compared to DARDAR?

- We assume the reviewer refers to the discrepancy at 0-1 km in Fig. 5a and 5b, where our mean IWC appears slightly high-biased (perhaps a mistype in the reviewer's comment). This bias arises from the way we treat the radar clutter zone during the radar reflectivity inversions. Specifically, directly above the surface in the radar clutter zone, the reflectivity values are rejected, and the value directly above the clutter zone is copied to the altitudes below. This leads to slightly higher near-surface cloud ice compared to DARDAR.

8a: There is a consistent and significant low-bias in the 1:1 correlation, but then the MFE is very small, indicating the retrieval results are good. Why? Is MFE a good measure for such a situation? Maybe you should use RMSE or MAE?

- We thank the reviewer for this helpful comment, which flagged potentially misleading aspects of our error summaries. We have revisited our use of MFE for $D_{\mathrm{m,IWC}}$ and agree that it may misrepresent retrieval accuracy at higher values. Following the reviewer's suggestion, we calculated the RMSE (expressed as a percentage of the truth). The RMSE shows an increase at higher $D_{\mathrm{m,IWC}}$, which is more inline with expectations. Relative errors remain largest at low $D_{\mathrm{m,IWC}}$, since low values inflate relative RMSE. The trough near $\sim$50 μm seen in MFE disappears with RMSE, which is also a potentially misleading feature, as it only reflects the median crossing the 1:1 line in a region of low sensitivity. We therefore agree with the reviewer that RMSE is preferable for $D_{\mathrm{m,IWC}}$, but maintain that MFE is a more appropriate metric for IWC due to the several orders of magnitude spanned by this variable.
- **Changes to the manuscript:** Fig. 8 has been updated to use the RMSE instead of MFE. The discussion has been updated accordingly.

References

Forkman, P., Christensen, O. M., Eriksson, P., Billade, B., Vassilev, V., and Shulga, V. M.: A compact receiver system for simultaneous measurements of mesospheric CO and O_3, Geoscientific Instrumentation, Methods and Data Systems, 5, 27–44, https://doi.org/10.5194/gi-5-27-2016, 2016.