Response to review by anonymous referee #4

Opening remarks

Warm thanks to all four anonymous referees for taking the time to review our manuscript and for providing valuable and actionable feedback. Their comments and suggestions have resulted in revisions that present and discuss our results and limitations with greater precision, including important qualifications and specifications. We highlight some larger changes motivated by comments from multiple referees.

First, to answer how the provided uncertainty estimates can be applied. We previously illustrated the provided quantiles through uncertainty in an example scene. We now also demonstrate how to apply these quantiles to obtain probabilities for cloud FWP classifications.

More than one comment was concerning too strong claims on achieved horizontal resolution and effective resolution in the manuscript. We have reworded these claims with appropriate qualifications. To motivate that the method provides increased ability to resolve structure, we add a confusion matrix for FWP ranges. This matrix shows that low FWP estimates are never confused with high FWP estimates, and vice versa.

There were comments asking about the general value of achieving retrieval performance down to 10 or 40 g/m2. Now introduced a figure showing the cumulative distribution and cumulative contribution to the total FWP, derived from our simulation database. This helps clarify how sensitivity levels affect individual retrievals and the average global FWP.

Lastly, we have added clarifications, further details, and additional references when discussing other datasets, instrument information sensitivity, and the methods used to create our simulation database.

The remainder of this document directly addresses the referee's comments.

Replies on referee's comments

> In this paper, the authors present their product for frozen water path (FWP) from the AWS satellite. This is an inherently valuable contribution, as it represents the first dedicated FWP product derived from a satellite with historic wavelengths. Importantly, the algorithm uses contemporary ML methods and cites recent algorithm literature, which is a requirement for any new retrieval method given today's rapidly advancing computing capabilities.
>
> The authors present a strong evaluation of their product. They are transparent about the difficulties of performing an ideal evaluation, given the lack of proper overlap with an operational W-Band cloud radar. In the discussion, they make a well-supported argument that their simulated brightness temperatures are valid for this evaluation.
>
> Most of my comments concern the results section. Some reorganization of the text is needed: several lines currently in figure captions belong in the body

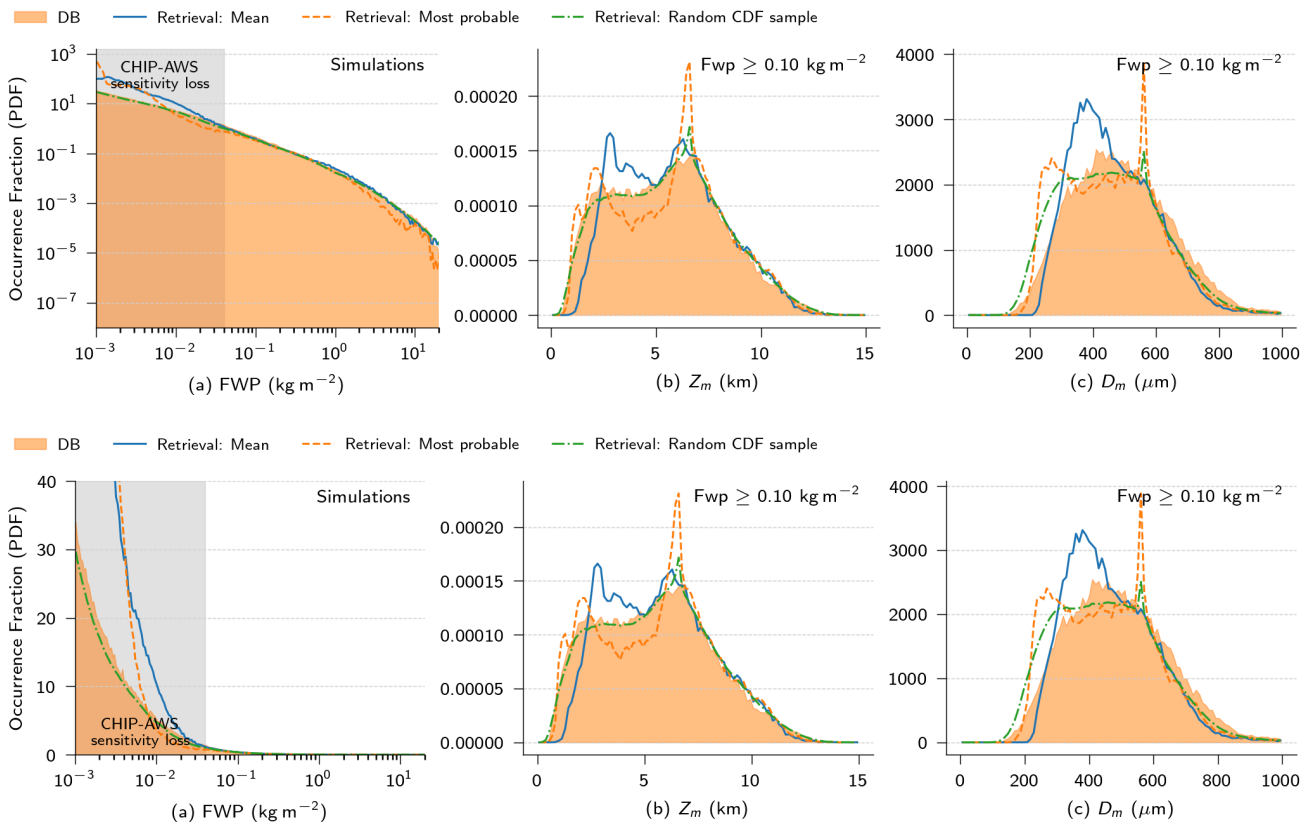
text, and some lines in the results section belong in the discussion. Specific examples are noted below, but in general, the results section would benefit from more objective and more descriptive discussion of the figures. Readers should be able to broadly visualize the figures and identify the key results from the text alone, and this is not currently possible.

Since my comments mostly concern the presentation of the science rather than the science itself, I recommend submission with minor revisions.

Major Comments

Occurrence fractions are not defined in the text and should be defined there rather than only in the figure caption. It is also unclear whether calculating bin occurrence is appropriate when FWP is measured on a logarithmic scale. Should these values be normalized by bin width?

- The definition of occurrence fractions is now included in the main text.
- Normalising by bin width results in a very large range of values, which makes it difficult to compare the distribution of FWP values for low and high FWP values. Therefore, we believe that, for the sake of comparing the distribution of values between datasets and database, that not normalizing is more informative. To illustrate this, we attach two examples of plots normalised with bin width, one with logscale and one with linear scale (for FWP only):



In both cases, it becomes difficult to compare differences between the datasets, compared to our current approach.

What is the purpose of presenting the mean value here? No analysis of the mean appears in either subsection referring to Figure 5. If its relevance cannot be addressed in the results or discussion, consider removing it.

- We do not entirely follow the reviewers comment, but we think that this may stem from ambiguous wording on our part. The sections that refer to Figure 5 (5.2.1 and 5.3.3) are directly describing and discussing the occurrence fraction of the mean FWP, Z_m, and D_m estimates. `Mean' does not refer to an average across cases, but to the mean of the retrieved distribution for each individual case, i.e. a single estimate of the retrieval variable. We have now updated the relevant text in these sections to explicitly state that we are comparing occurrence fractions of the mean estimates.

  Section 5.3.2 would benefit from a rewrite and reorganization. The results are currently described in only two sentences, which seems insufficient to merit a standalone subsection. The following two paragraphs appear better suited to the discussion section.

- We agree that the section was too brief to stand alone. We have now combined both of the CCIC comparisons sections into one: CCIC-case comparison and the conditional mean are now in Sec. "Co-located comparisons with CCIC".

  More quantitative results are requested in Section 5.3.3. Specifically, what range of occurrence fractions qualifies as results that "must be considered good for passive observation"?

- Qualified what is meant by good here: that occurrence fractions follow the expected distribution for a wide range of values, something other passive datasets have struggled to do. Added a reference to such a comparison.
- Added CDF and mass contribution figure. It motivates that we don't have to put to much weight on the bad occurrence fractions for lower values.

  More quantitative results are also requested in Section 5.3.4. What is meant by "slight positive bias," and how should "areas of deviation" be interpreted? Some of these requests may seem minor if the level of description elsewhere were more quantitative, but as currently written, this section relies on subjective language without supporting numerical results.

- We interpret this comment to be addressing Section 5.3.5 instead of 5.3.4, since both "slight positive bias" and "areas of deviation" appear in the former.
- We've computed latitude-weighted bias metrics and reported them in the zonal mean figure. In section 5.3.5, text has been changed to be mor quantitative, referencing these metrics.

  Minor Comments

  Section 3.1: Please clarify what is meant by "most probable" where it is defined.

- Changes made: A short description is given in the text and the section that defines "most probable" and "mean" is more clearly referenced.

  Line 250: Please clarify that this refers to an ice-scattering model and a PSD model, to avoid confusion with a static PSD.

- We have updated the text to use "PSD model" instead of "PSD", both here and elsewhere in the manuscript.

  Line 362: Additional detail on the difficulties associated with convective cores would provide better context for why they are currently missed and why the

authors plan to address them in future work.

- More detail as been added on difficulties with convective cores and our plans.

Section 5.1.2: This section may not warrant a standalone subsection, as it does not present any visible results. The "ongoing work" sentence may fit better in the introduction.

- We agree with the reviewers comment here, as the relevant results are placed in the appendix. We have opted to combined section 5.1.2 and 5.1.1 into a new section 5.1.1 "Simulation performance over surface types".

Section 5.2: Consider splitting Figure 5 into two separate figures. The bottom row is not discussed until four subsections later and includes two variables that are not used or introduced in the top row.

- The reviewer makes a very fair point. However, we have opted to keep these two figures together. We believe that readers will likely want to compare simulated and actual performance. Since the differences are subtle, keeping the figures together greatly aids in this comparison.

Line 430: This information should be presented while the figure is still being described, prior to the discussion of results.

- Agreed, this is now consolidated into the figure introduction.