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 comments by the referees.

Response to comments by referee #1

> This paper presents CHIP-AWS, an ice cloud product retrieved from the AWS radiometer. The retrieval algorithm is based on a Quantile Regression Neural Network (QRNN) trained using a database simulated by the ARTS forward model. In addition to retrieving ice cloud properties, the algorithm provides uncertainty estimates for each retrieval. Its performance is evaluated using both the simulated database and other ice cloud products, such as CCIC. Overall, the methodology and results are clearly presented. Several aspects require further clarification or additional discussion, as outlined in the comments below.
>
> (1.) In the Incidence Angle row of Table 1, it is stated that the satellite zenith angle is reported in the Level-1B (L1B) file. However, lines 114–117 indicate that the L1B file provides only the Earth-surface incidence angle, rather than the satellite zenith angle. These descriptions appear to be inconsistent and should be clarified. In addition, it would be helpful to include the equation used to convert the Earth-surface incidence angle to the satellite zenith angle in the discussion around lines 114–117. This equation is also needed for the footprint size calculations in Eqs. (5) and (6).

- Clarified this by updating the phasing in what were originally lines 114-117 from "Instead, zenith angles from Earth-surface footprint locations towards the satellite are available" to "Instead, satellite zenith angles are available, i.e. the zenith angles from Earth-surface footprint locations towards the satellite.".
- Updated Table 1: Changed label from Incidence Angle to Satellite Zenith Angle.
- What is referred to the incidence angle theta, is the same satellite zenith angle. Updated text to now refer to theta as the satellite zenith angle when describing Eqs. (5) and (6) to avoid confusion. Thus, satellite zenith angles reported from L1b files, and the averaged satellite zenith angles in our L2 files can be directly used as theta in these Eqs.

> (2.) Please check Eq. (4) again. The equation here calculates the particle diameter for each layer rather than the column-averaged particle diameter. Dm should be the FWC- weighted average of the layer-resolved particle diameters, following a similar formulation as Eq. (3).

- In Delanoë et al. 2014 Eq. (3), Dm is the volume-weighted diameter (melted-equivalent) at a specific vertical level. Our Eq. (3) is the same, except integrated over the vertical dimension z. This should be the mass-weighted mean particle size for the entire column.
- In the text we've added a reference to a derivation that shows this is the column weight averaged particle diameter: Amell et al. 2022, appendix A (eq. A9)

> (3.) For the simulation database discussion in Lines 259-264, it would be helpful to discuss why the CloudSAT CPR reflectivity observations are used to first retrieve ice cloud properties instead of directly using the 2C-ICE or DARDAR products. One reason I can think of is that using CPR reflectivity can keep the particle habit and PSD selection in the radar retrieval consistent with the selections in the following TB simulations. If there are more considerations, please clarify them.

- Indeed, it is to ensure we model our uncertainty from particle model assumptions, both when populating FWC for our simulation scenes, and in our forward simulations of AWS observations.
- We have added a paragraph motivating this in Sec. 4.1.

> Also, it would be helpful to provide more details about the radar retrieval method. For example, is it based on the optimal estimation method (OEM) or another approach? Do you retrieve only ice water content, or both ice and liquid water content? Since W-band radar signals are strongly attenuated by liquid clouds, how is the liquid water treated in the retrieval? Do you also consider the melting layer, and how is it handled?

- In Section "Simulation database" the following changes have been made:
  - A reference to a description of the radar inversion method (ARTS onion peeling) is now given in the text.
  - We have also specified that Liquid water content (including super-cooled) is populated from ERA5, while rain water content (for liquid hydrometeors) is retrieved from the radar backscatter.
  - Both LWC and RWC is taken into account when retrieving from radar and for the forward simulations producing the passive MW observations.
  - It is now also acknowledged that the melting layer is not included, as we don't have any scattering data for melting particles.

> Finally, have you compared the radar retrieval results with existing products such as DARDAR or 2C-ICE? The manuscript mentions that the mean values from DARDAR and 2C-ICE differ by about 20% (lines 459-460). It would be helpful to know whether your results are closer to DARDAR or 2C-ICE.

- As part of the introduction of the radar inversion method (Sect. Simulation database), it is now mentioned that FWP estimates are closer to DARDAR, in terms of zonal-means. References are given for further comparisons and details.

> (4.) For the surface emissivity model, apart from the TELSEM/TESSEM used for land/ocean surfaces listed in Table 2, lines 263-264 also indicate that the surface emissivity is changed based on the land-cover dataset. Do you use additional surface emissivity datasets for other surface type? If so, please describe this dataset around lines 263-264 and include it in Table 2 as well.

- Land and ocean are covered by TELSEM and TESSEM. For ice and snow, the emissivity is modelled by a flat surface informed by skin temperature and a scalar reflectivity. This is now specified in the paragraph and added to the table.

> (5.) Line 271-273. How many pencil beams are used in each 3D Gaussian footprint averaging?

- The number of pencil beam values and the along-track sampling distance is now specified.

> (6.) Figure 2: In the top panel of the right plot, the PDF shows an almost uniform distribution over approximately 5–20 kg m-2, which is quite broad. The uncertainty shown in the middle panel seems more reasonable. For cases with large FWP values, however, the uncertainty estimate would generally be expected to decrease, which is not consistent with the results shown here. Please provide further discussion on the possible reasons and the reliability of the uncertainty estimates.

- Possible causes for larger spread for this more intense case is now explicitly mentioned in the discussion of Figure 2. In short, the spread can be due to the particle models responding differently, contributing to a wider uncertainty. Another reason could be that some channels become saturated, preventing the FWP to be constrained better. Still, the estimate is certain that we have FWP greater than 5 kg/m2. It can be argued that this still is a fairly narrow estimate.

> (7.) Line 298: μ should be calculated as the integral of x over the PDF, not the CDF.

- Yes, the expected value can be calculated as suggested. However, in practice, we do in fact integrate quantile values x over the CDF (probability levels). This is formally expressed as a Lebesgue-Stieltjes integral.
- This is now better motivated in the text.

> (8.) Line 316-319. What about cases with values larger than 1 kg m-2? There should be many cases in this range.

- These are included, but in a linear scale instead. I.e values below 1kg m-2 are log scale, values above are linear. This is now said explicitly in the text.

> (9.) Figure 5. the '> =' sign is missing.

- Thank you. This appears to have happened during Copernicus processing, since submitted file does not have them missing. We will be double check for final submission.

> (9.) Figure 6 and 8. The figures here show the density distribution of the retrieved parameters versus the true values. The conditional PDF (p(Y|X)) in Bayesian theory, however, represents the probability of the simulations given the observations, which is a diˆerent concept. I do not think these two types of density distributions are equivalent. Can you give comments about why these plots are referred to as conditional density p(Y|X)?

- Figures 6 and 8 are not showing the joint density distribution of retrieved from database (Y) vs. true database value (X), but are showing the conditional PDF, i.e. the retrieved from database given true database value.
- In other words, these figures show the joint distribution P(X, Y) divided by the marginalised probability function for X (P(X)): $P(Y|X) = P(X,Y) / P(X)$.
- We don't think labels or the text is directly misleading in this regard. We have made sure the text surrounding these Figures are explicit that its the probability of an estimated value given the database value or in the case of Figure 8 P(CCIC|CHIP-AWS).

> (10.) Line 402-404. For the comparison of these two plots, does it indicate that the AWS- CHIP results have higher horizontal resolution, or are they noisier? The CCIC results appear to be smoother and more continuous.

- This is a valid concern that we have now tried to address in the text thanks to this comment and a similar comment by referee 3.
- We have extended the description in the text to more carefully talk about the suspected increase in ability to resolve smaller systems. We now differentiate between three levels of structure; the finest (pixel ~10km) based variation, slightly larger (at least 40km2) (comprises of multiple pixels/footprints 4-10) we think are genuine systems being resolved, and the larger scale, where CCIC and CHIP-AWS agree.
- That high FWP values are predicted in areas with low FWP through noise is unlikely. This is demonstrated by an added confusion matrix showing that high FWP values are never confused with low FWP values/clearsky.

> (11.) Line 416-417. In line 414, it is indicated that CCIC is trained with 2C-ICE. Intuitively, the CCIC results should best reflect the statistical characteristics of 2C-ICE. Since CHIP- AWS does not use 2C-ICE data, why do the CHIP-AWS results match 2C-ICE better than the CCIC results? Could you provide more discussion on this point?

- With CCIC trained on 2C-ICE, we can expect the retrievals to be calibrated against 2C-ICE. That the mean CCIC values should match the 2C-ICE mean. Other statistical properties depend on physical or practical limitations.
- We agree with the reviewer that "CHIP-AWS is closer to 2C-ICE than CCIC" is too strong of a statement. We do think though that for individual values that CHIP-AWS can be closer to 2C-ICE. We have now qualified and explained this in the text.

Response to comments by referee #2

> Work is performed carefully; however, I think some more information on the underlying physical sensitivities and (different) information content of the data would make the paper even more valuable.

> General comments:
> [1.] Physical meaning of retrieved products. The abstract states: “The accuracy in FWP should be unprecedented among estimates based on passive satellite data”. This is a strong statement that requires more argumentation. The manuscript needs to stress the complexity of ice particle size distributions (see Bartolomé García et al. 2024), transitioning from tiny particles (sub-visual cirrus) to large snow aggregates, and the information content of different retrieval products. For example, the selective sensitivity of the used data records (MODIS, CICM) that are affected by saturation in the VIS/IR signals.

- Regarding the complexity of particle size distributions (PSDs), we agree, there is considerable complexity and uncertainty in assumptions for PSD in relation to ice particle sizes. In our view, the method in this work presents a significant contribution toward tackling this and making these uncertainties explicit: by using an ensemble of particle models (habits + PSDs), and representing this uncertainty through distributions in the retrievals.
- The abstract and the introduction have been adjusted to make this contribution more clear.
- Introduction also includes a discussion on sub-mm and the gap it fills between MW and IR/Vis. There is a reference to Wu et. al 2024 that directly tackles the comparison between different data records and the role for sub-mm.
- The abstract has been updated to acknowledge that a proper validation can't be made (no co-located data). However, the analysis performed shows better accuracy over previous passive retrievals.

> As FWP is not a widely used variable, the difference to other product names (CIWP, see Stubenrauch et al., 2024) needs to be made clear.

- At first occurrence frozen water path is introduced, and it is stated that the new nomenclature is used to qualify that all frozen hydrometeors are taken into account. Product names like Cloud Ice Water Path, leave the ambiguity of what is meant by cloud.
- For a further motivation for FWP, refer to Eriksson et al. 2026. This is refereed in the text.

> Anticipating certain user questions: Could you make a rough estimate of how much of the global average FWP is missed by conventional products and how much more your product (CHIP-AWS) can sense?

- Thank you for the heads up on this anticipated question. In the introduction, we now added a paragraph that states conventional passive datasets have been shown to be roughly 50% compared to CloudSat CPR datasets in zonal mean comparisons. Two references are included. There it is also stated that our product gives values comparable to the CPR based datasets.

> and give some recommendations (I don’t expect detailed answers) on...
> How can one identify a “clear-sky” pixel and with which uncertainty?

- We thank the reviewer for this clarifying question. This is a point we would like to make more clear in our manuscript. The answer on what is clear-sky depends on what is meant by clear-sky. Clear-sky can mean different depending on the sensor's sensitivity. From an observational perspective taking sensor sensitivities into account, it makes sense to define clear-sky below the threshold where such an instrument would

sense ice. Since quantiles are provided, this can be answered using any threshold one likes to produce a probability that a case is below the threshold.
- In order to demonstrate this, and make this application more clear, we have extended the sub section "Example Scene", with an example providing classification probabilities for clear sky and other classes.
- In addition, a new confusion matrix figure also provides recall and precision for FWP < 1g/m2.

How many cirrus clouds will be missed?

- Again, the answer here depends on the definition for cirrus clouds in terms of FWP. But taking (1g/m2 < FWP <= 10g/m2) to mean thin cirrus, our method struggles. There is not enough mass to get a strong signals for these channels.
- We hope that the added classification probability example and the confusion matrix help answer this more directly in the answer. The lower sensitivity for these cases is now acknowledged in the text.

Do I also need to apply a threshold of 0.1 kg/m2?

- The retrievals give a distribution of values that expresses the uncertainty. When computing probabilities from these estimates no threshold should be applied.
- However, the retrievals for Z_m and D_m are generally more certain when FWP is higher, so the point estimate (the mean or most-probable) will likely be off when FWP is low. So, a threshold on FWP of 0.1 kg/m2 has been found to work to filter mean point estimates for Zm and Dm with more certainty.
- In our presented results and analysis, we have opted to filter with FWP 0.1 kg/m2 for Zm and Dm for certain comparisons (occurrence fractions and conditional PDFs), in order to not have means from uncertain retrievals skew the result.
- Other comparisons, such as the zonal mean, local means, and Hovmöller diagram do not use this filtering, instead they are weighted by the FWP.
- The Section "Conditional densities", its conditional probability densities Figure, and related text has been updated to address this. The figure now includes conditional means for two filtering levels: FWP >= 0.1 kg/m2 and FWP >= 0.01 kg/m2, to demonstrate how the filtering affects precision for different values.

How much is a retrieval of 40 g/m2 worth?

- Added Figure showing the CDF for reference FWP values in the simulation database and the cumulative contribution to the total FWP.
- From this, values below 10 g/m2 and 40 g/m2 stand for 46% and 64%, respectively of all reference FWP values in the simulation database. However, values below 40 g/m2 contribute very little to the total FWP sum/average.
- Note, however, that even though we have a stated sensitivity down to 10 g/m2 or 40 g/m2, depending on how we compare, we still get statistically reasonable means when accounting for values below this. Since we are following a Bayesian approach, our distribution does account for values all the way down from the prior (simulation database).

Is Dm basically only a measure of snowfall size?

- Dm is a measure of all ice hydrometeors in the atmosphere, a mass weighted size estimate. Not specifically snow, and not snow reaching the ground.

[2.] Clearness and completeness of the provided information: At several

instances (see my specific comments), more accurate formulations are needed. One example is the term bias, which is frequently used in the paper, but its actual meaning is not always clear, as truth is not really available.

- Agreed. We specified conditional bias for the conditional mean results and discussions. Other uses of bias where datasets are compared and there is no clear "true"-value have been replaced with "difference".
- Called it across-track dependence instead of across-track bias.

Information on intercomparison data, e.g. MODIS, CLARA, is missing. The Discussion section is a bit confusing – here suddenly SPARe-ICE is appearing. More information on accuracy and limitation these data is needed. For the outlook the, authors should also consider intercomparisons with ground-based reference sites, e.g., ACTRIS/Cloudnet.

- MODIS, CLARA and CCIC are now introduced in the "Data sources" section under the new subsection "Comparison datasets".
- SPARe-ICE is now also explicitly introduced. However, we opt to keep the introduction of this dataset in the discussion, since it is only there to discuss and place our MFE results into context.
- Mentioned EarthCARE and ACTRIS Cloudnet in the outlook for validation work.

3. Length of the record: It is never really stated how long the data record for the presented analysis of CHIP-AWS is. I find it really important to have exactly a (full) one-year period (March to March) to avoid seasonal imbalances. MODIS or ERA5 could be used to assess the interannual variability.

- Text has been updated to explicitly state that we provide data for the full year starting from January 2025.
- In the zonal mean, MODIS and DARDAR show the interannual variability.

Minor point, but in many instances: One sentence does not make a paragraph.

- Thanks. Manuscript has been revised to avoid this.

Specific Comments:

Abstract: (you should mentrion the product name (CHIP-AWS for branding)

- Thanks. Added.

L4: Detailed simulations- here information on the blend of Cloudsat retrievals and ERA5 should be mentioned. For later: Any retrieval can only be as good as the underlying database. The authors are a bit short on information (and motivation) how they construct the database. Explain why not only ERA5 information is taken.

- Clarified that simulations use hydrometeor information from CloudSat radar reflectivities and ERA5 for background-information.
- The "Simulation database" section now further explains the CloudSat CPR inversion approach and justifies its use.

L11:” The accuracy in FWP should be unprecedented among estimates based on passive satellite data”. This is a strong message see major comment #1.

- See response to major comment 1.

  L12: “significantly” could be nearly anything – quantify

- Thanks. Quantified swath size.

  Introduction:

  - L21 – first dataset, duration missing -> first full year data set

- Adopted as suggested.

  - L34 225-1064 nanometer not micrometer; add (about 3mm) in parentheses after 94 GHz

- Adopted as suggested.

  - L33- maybe refer to the figure in the Buehler paper for the different wavelength respnse

- Added reference to Fig. 3a of Buehler 2007.

  - L35 – “spatial” is irritating – either vertical or radial

- Should have been vertical. Now fixed.

  - L39 – I would add climate and the changing water cycle as further motivatin

- We've added a short first paragraph to the introduction to motivate why FWP retrievals are important.

  - L43 – “applying machine learning” somehow sounds too fancy. ML is basically uses as regession

- Specified that Machine-Learning is used to train a non-linear predictor.

  - L47 – submillimetre vs far infrared?
  - L47 – for clarity mention the two conventional ones by name

- Specified that sub-mm fills the sensitivity gap to ice mass between MW and infrared.

  - L73 – “all hydrometeor classes” – hydrometeors haven’t been introduces yet. I think it is important to mention the arbitrary (model-dependent) distinction between ice and snow water content and the fact that kilometer-scale models introduce denser hydrometeor classes (graupel, sometimes hail) to enable the representation of strong convective events, i.e., thunderstorms.

- The text is rewritten to avoid using "hydrometeor classes". We agree that atmospheric models do not make a clear distinction between cloud ice and snow, but as we are here not comparing to any such data, a discussion of that topic would be a distraction.

  - L75 – “report” is not the right name – mostly they are even prognostic. Suggest to simply say “consider”

- Changed to GSRMs provide values for all hydrometeor classes.

    - L82 – retrieval PRODUCT

- Thanks. Changed to retrieval variable, since we've referred to the L2 files with all variables as the retrieval product.

    - L89 – I find the sentence confusing: “operational machine learning models for cloud retrieval have exclusively been trained on existing retrievals” . Maybe I misunderstood what is meant with “operational ML” (then it should be better explained) but tons of satellite retrievals are run on simulated measurements (mainly from atmospheric model data).

- We have removed operational to avoid confusion. We meant ML methods that have been trained through co-location with existing retrievals.
- We ask the reviewer to point out if we've misunderstood the comment.

    - Similarly L90 “For reasons discussed below, this more complex approach could not be avoided” sounds weird.

- The sentence is rewritten.

Sources and reference data:

- It would be helpful for the reader to get a rough idea of the retrieval and validation strategy before presenting the subsections to understand for which purpose they are needed for.
- For example, I was first expecting that ERA5 output (including hydrometeors) would be used for the simulations.

- The data source subsection "Chalmers Cloud Ice Climatology (CCIC)" has been renamed to "Comparison datasets", which introduces CCIC, MODIS, and CLARA and the way they are used for comparisons: Zonal means, occurrence fractions, co-located samples.
- This also makes it more clear that ERA5 is not used for any comparisons. It is used for background simulation input (not hydrometeors).

    - L101: 145 over a +/ - ??degree range

- Added. +/- 55 deg

    - L128: Give number of pressure levels. There is significantly more detail in the model layer output though I don’t believe it is important. More importantly, I am missing a discussing how physical consistency between thermodynamic state and ice hydrometeors is guaranteed? As an example could you have 20% relative humidity and high FWP?

Specified that all 37 pressure levels are used, but are interpolated into our simulation grid.

A good remark, indeed 20% relative humidity in a region with high ice mass could in principle be found in the input to our radiative transfer simulations. We spent considerable time trying to devise a correction procedure, starting with Rydberg et al (2007), but never found a satisfactory solution. We likely improved the realism for parts of the data, but with the cost of adding artefacts. For example, we applied corrections in Ekelund et al. (2020), but those resulted in decreased agreement with observations for "clear-sky" conditions (a

disagreement around the "peaks" in their Fig 10), but we kept the corrections as they should hopefully increase the fidelity for the data of concern for that study (the lower end of Tb-s). However, for actual retrievals we cannot neglect the artefacts for the clear-sky part, which in fact constitutes a dominating part of the data. Therefore, we have for the time being put the corrections to the side. This is not ideal, but not critical. At the time of Rydberg et al (2007), output from ECMWF was known to have a low bias in upper tropospheric humidity, the quality of ERA5 is much improved. We have manually inspected a significant number of scenes, and found the agreement between CloudSat and ERA5 satisfactory. That said, we should be clearer about physical inconsistencies, and changes to the text have been done in Sec 4.1 and Sec 6 "Discussions".

Rydberg, B., Eriksson, P., & Buehler, S. A. (2007). Prediction of cloud ice signatures in submillimetre emission spectra by means of ground-based radar and in situ microphysical data. Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, 133(S2), 151-162.

Ekelund, R., Eriksson, P., & Pfreundschuh, S. (2020). Using passive and active observations at microwave and sub-millimetre wavelengths to constrain ice particle models. Atmospheric Measurement Techniques, 13(2), 501-520.

> - L131: why is only cloud water content used? I would also recommend the inclusion of rain as light rain amounts are frequent and even present in cases of no surface precipitation.

- Text updated to state that rain is populated in our simulation scenes through inversion from the CloudSat CPR reflectivities, with a reference for more details.

> - L141: Why 2015? Maybe one could see from the longer-term MODIS, CCICs data set that this is an “average” year?

- The choice for 2015 was arbitrary, the data was available and ready to use at the start of this project.
- We have checked that this year was not an special with respect to ENSO. This is now stated in the text.

> - Subsections 2.3 and 2.4 don’t provide information on the uncertainty of the described products. Specifically what is the limitation of CCICs wrt high FWP?
> - CLARA and MODIS Data (and their uncertainties) are not described here and not mentioned in the data availability section.

- Sections 2.3 and 2.4 have been combined into one section "Comparison datasets".
- Here, brief comments on underlying methods, limitations, and if datasets provide uncertainty estimates are now added. References for more details are also included.

> **Data product**:
>
> - I would be curious how strongly the three main retrieval products are correlated?

- A good suggestion. We have now added a new table (Table 4) to show the correlations between the retrieval variables in the simulated training data, in the test retrievals, and in the CHIP-AWS retrievals.

> - L159 – see my main comment on duration of the data set for analysis –

full year

- Specified that we are providing data a full year of data from January 2025.

    - L167 – provide also angle and swath width – then you don't need it later

- Adopted.

    - L188 – also Eriksson et al., 2020 don't use the term spatial sensitivity but refer to vertical one..

- Thanks. Changed to radial. Those equations consider the radial footprint of the observations.

    - L196 – "These restrictions are put in place to reduced the scope" -> These restrictions are put in place to reduce the complexity..

- Adopted.

    - L197 - Variability can mean many different things, please explain your definition used here

- Thanks. Removed variability and state that the 75th percentile increases closer to the equator.

    - L206 – at least mention that ERA5 is coarser

- Yes, it is a good point that ERA5 is much coarser than our discussion on footprint sizes of ~10km.
- However, ERA5 was used to populate background information in our simulations that are then used to train the ML model. CloudSat is still the source used for populating the simulations with hydrometeors. No ERA5 data is used to perform the actual retrievals once trained.
- We argue that ERA5 coarseness should not impact our footprint-by-footprint resolution and leave this discussion on ERA5 out, to keep the section focused.

    - L214 – what does unreasonable mean – how can this be reproduced? I don't expect a full description but maybe a link to a reference

- Text now specifies how unreasonable satellite altitudes and incidence angles are filtered out.

    - L225 – How strongly are the variables correlated in the training, test and application? Do you see any changes? It is also very interesting to investigate the correlation with temperature, we also suggest including the zero degree line in the figures. One could also do a 2D histogram with zm and the height of the zero degree isotherm for consistency checks.

The correlation between variables, and whether the correlations are maintained in the retrievals, is an interesting point. We have now added a new table (Table 4) to show the correlations in the simulated training data, in the test retrievals, and in the CHIP-AWS retrievals.

We do in fact see a small change in the correlations, where the correlation between FWP and the other two variables increases for the test set retrievals. Encouragingly, the increase is

much smaller in the CHIP-AWS retrievals. An observed increase could be explained by the fact that we may not always have enough information available in the observations to infer each variable entirely independently. The correlations are not so high to suggest that we can never infer independently, and we suspect that this behaviour varies across atmospheric conditions and/or the ranges of the variables. This is supported by the low degrees of freedom at low FWP, observed in Fig. 1 of Duncan et al. (2026). This is certainty interesting, and we plan to do a deeper analysis into the correlation behaviour moving forward.

- A 0degree isotherm for transect along the example scene has now been added.
- A nice suggestion regarding the 2D histogram with Zm and 0deg isotherm as an extra result. However, due to practical reasons and time limitations, we wont act on this suggestion. We hope the database comparisons, Hovmöller diagrams, and zonal means results sufficiently demonstrate the retrieval performance regarding this variable.

    - L230: There needs to be more discussion on the choice of the threshold and implications for analysis – when do we have a (precipitating) cloud?.

- At first mention the reason behind the threshold of FWP >= 0.1 kg/m2 for Zm and Dm is explained, with a reference to the conditional distribution section for more motivation of the threshold level.

Retrieval Method:

- L295 – for me it would have been helpful to mention early that a binning of 0.01 for the CDF would be used
- What is the choice for the chosen output quantiles – why not 0.05 and 0.95?

- The binning of 0.01 is now stated in Section "Inversion with QRNN".
- Choice for quantiles 0.02, 0.16, 0.5, 0.84, and 0.98, is now provided: chosen to line up with with the median, +/- 1sigma, and +/- 2sigma.

    - L 313: I guess randomly chosen?

- Correct. This is now stated in the text.

    - L320 This is a good opportunity to advise the user if a “trustworthy” FWP is retrieved..or what the likelihood is that this pixel has no FWP?

- This is now introduced and demonstrated in the Example scene, where classification probabilities for a scene are shown.

    - L330 – specifiy substantial

- This is now specified.

    - L333: THE mean estimate

- Fixed.

Results:

- L336 – agrees is a too strong word -> represents the observed atmospheric state

- Adopted.

    - L340 – what is meant with accurate? Just the RT? Then I would say first a representative atmospheric state and then an accurate RT.

- Thanks. Adopted.

    - L343: best-performing – best understood surface interaction

- Thanks. Adopted.

    - L353: As a rule of thumb values below….

- Thanks. Adopted.

    - L358: There is some slight discrepancy at the coldest temperatures already seen here over ocean – so one could start here already start the discussion on intense convection

- Thanks for the suggestion.
- We opt to keep the discussion of intense convection focused on land. Over ocean, we see that we have some simulated scenes with lower Ta than in observations.
- For intense convection over ocean being missed by our simulations, we would expect to see lower observed temperatures than our simulations can provide. Currently the opposite is the case, which is not as much of a problem since it gives us a greater range of cases to base retrievals on.
- However, when discussing land, the complexities of modelling deep convection are now brought up.

    - L360: You could also make a correlation matrix between different channels

- Our interpretation is that the reviewer is suggesting to show ill-posed-ness of the observations. To our understanding, we can have ill-posed cases that happen with uncorrelated channels. We opt to leave such a correlation matrix out of this manuscript, since that analysis is not comprehensive.

    - L362 – specify where one can see this? Furthermore, intense convective cores – you didn't describe beforehand why these are tricky, e.g. graupel, hail..

- Text has been updated, to specify that we attribute low Ta values for AWS21 with tropical deep convection.
- Added a short motivation as to why deep convection can cause issues.

    - L374: I think the figures (going down to 1 gm-3) overemphasize the importance of the really low values. Looking at the helpful Fig 5.

- Thanks for pointing this out.
- The new figure with the cumulative distribution of reference FWP from the simulation database and the cumulative contribution to the total FWP is used to help put importance of very low FWP values into context.

    - L409: bias-free – how can you judge as truth is there; away -> higher?

- From addressing the main comment on use of the word bias, we've decided to remove

the paragraph discussing CCIC's bias.

    - L417: I am missing a discussion on the physical interpretation of Fig.8 Basically what can be seen what not?

- Figure 8 is now more thoroughly described and interpreted in the text. Now describing that for given CHIP-AWS estimate values, the co-located CCIC estimates tend to overestimate FWP below 1kg/m2 and underestimate above.

    - Section 5.34. The MODIS data set was not introduced before. To my knowledge it was never constructed to include precipitating ice so there is no surprise about the difference

- MODIS is now introduced in the "Data sources" section. The caveat that MODIS retrieves the ambiguous cloud ice water path instead of all frozen hydrometeors (FWP), is now pointed out in the text, both in the introduction and in the results "Spatial distribution" section now called "Local means".

    - L439 “The CHIP-AWS estimates have no coastline artefacts that cannot be attributed to orographic features. “ That is confusing. Orography makes an artifact?

- Artefact was an incorrect word choice. This word is removed.

    - L442: cases : do you mean ocean and land? Maybe regimes is better: there is no clear indication where one can see the effect of the dry cases ?

- Thanks. Adopted

    - L446: Better surface emissivities..there should be also some reference to the literature on that

- Thanks. Adopted.
- We assume the reviewer refers to references to what current emissivities we are using for ice and snow. This has now been specified in Section 4.1 and Table 2.

    - L449 and following: I don’t understand the meaning of bias here because no truth is there.

- Following from main comment on use of bias, comparison is now made in terms if "difference" instead.

    - L460: what is the correlation between DARDAR and 2C-ICE?

- It would be interesting to see this correlation. However, we want to keep manuscript focused, and a comparison between other datasets is not our goal.
- There are previous works that perform such comparisons, see Eriksson et al. 2026, and Duncan and Eriksson 2018.

    - L469..and knowledge of the physics.

- Added that the results are also consistent with the expected limited information content deeper into clouds with the sensors used for MODIS and CLARA.

    - L470: In case of multi-layer clouds radar – lidar is not a good reference

- Thanks. This is now mentioned.

    - L484 . Refer to where the retrieval accuracy is mentioned.

- Thanks. This was referring to the sensitivity limits in Sec. 5.2.1 and 5.2.2. Revised to make this more clear and added reference the sections.

Discussion:

    - I got a bit confused with this chapter which seemed to be more like a summary than a discussion.

- Removed repetition of results that in the beginning that are not really discussed.

    - L487 and following: Which are the three orders of magnitude – be specific when one can “trust” the retrieval. As there are specific uncertainty values for each pixel, what would be the recipe for using them? What is meant with “entire relevant range”?

- Thanks. Specified the magnitudes, removed "entire relevant range".
- An example for how to use the uncertainty to obtain probabilities is now presented in Sec. "Example scene"

    - L489: “The comparison database values are radar retrievals” What is meant by that?..or is this the start of a new paragraph?

- This sentence on radar retrievals is misleading. Now only referring to the simulation database reference values. This was referring to that the simulations are driven by radar measurements: ARTS onion peeling, that is now properly introduced in the "Simulation Database" section.

    - L491: Now the discussion goes to Spare-Ice which has not been introduced before - why is it not included with the MODIS and CLARA comparison?

- Rephrased the first mention of SPARE-ICE in the discussion to explicitly introduce it and state that it is only included to provide context to our MFE results through their reported MFE results.
- SPARE-ICE is not included with the other datasets, since SPARE-ICE is currently not currently maintained operationally, whereas CLARA and MODIS are established datasets with operational processing. To the best of our knowledge, only SPARE-ICE provides a MFE estimate for FWP in literature.

    - L501: Why not ground-based estimates from Cloudnet? This could also provide information on zm.

- Mentioned expanded comparisons the outlook.

    - L504: “Recent machine-learning approaches trained on radar retrievals have achieved similar results” Sounds very generic – where does the information come from – new architecture, new training??

- Removed recent, and specified that we are talking in terms of zonal means. I.e. the ML-based approaches, thanks to their training loss-function, tend to minimize the average error.

Conclusions:

- L526: Simulations can mean anything – also atmospheric model simulations. In this sense the sentence in L534 should be reformulated" This is, to the best of our knowledge, the first simulation-based passive retrieval method"

- Reformulated to: "This is, to the best of our knowledge, the first passive retrieval method based on atmospheric radiative transfer simulations to do so (Duncan and Eriksson, 2018) (Eriksson et al. 2026)."

- L529: rather unspecific. Simulated Ta distributions agree with their climatological counterpart observed over one year globally.

- Thanks. Adopted.

- L539: the effect of convective cores was never explained,

- This is now explained, see response to comment on L358.

- L534: "all three variable??

- Thanks. Specified that we are referring to FWP, Zm and Dm.

- There should also be some link to passive MW snowfall and SWP retrieval (Camplani et al., 2024) which use partly similar principles.

- Yes, this is very relevant related work to highlight. We've added it to the outlook section on interesting related paths.

Figures:

- not always all lines are explained

- Thanks. All lines and areas are now either explained or have a clear legend entry.

- Figure 1: Explain CDF on first occurrence. Give date and time. How was the transect line chosen? It would be very helpful for interpretation to plot the zero-degree isotherm (from ERA5) into the Zm transect.

- When discussing Figure 1 (example scene) and Figure 2 (classification probabilities) choice for transect is explained: arbitrary to intersect with regions of varying FWP.
- Date and time for start and end of pass is given in the first panel.
- A zero-degree isotherm is now added.

- Figure 2: scaling of middle CDF/pdf is ns irritation – better go only to 1 or even lower FWP

- Adopted. Middle and bottom figure share axis range of 0 to 2 kg/m2.

- Figure 6: Micrometer are missing, red and dashed lines need to be explained. The vertical scale of the MFE should be better visible for values of 100%

- Axis limits have been changed as suggested.

Tables

- Table 1: main (top 3) variables. Why is the mirror angle not included?

- This is now included.

- Table 3: Why not polarisation and bandwidth?

- The table lists inputs to our ML model. We refer to Table 1 in Eriksson et al. (2025) for complete channel specifications.

- Table 4: Are these really the best metrics? From these I would say it can only retrieve snow? What about relative error and only calculated above the threshold of 0.1 kgm-2? In the abstract you mention down to 40 gm-2 – but how trustworthy is such an estimate?

We agree that, for a variable such as FWP which can vary over orders of magnitude, a relative error is more appropriate, and chose MFE for this reason. However, we admit that some confusion here may have arisen from how we have presented results in the table. MFE is in fact a unitless and relative measure. It is an error of 86%, not an error of 0.86 kg/m2, which perhaps is why the reviewer made the association with snow.

We thank the reviewer for bringing up this point, since it is now clear to us that Table 4 can be misinterpreted. Furthermore, we now realise that the metrics shown here simplify the situation too much, since RMSE, MFE, and bias can vary across the range of each variable. The variation of MFE and RMSE are already shown the lower panels of Fig. 8, and bias can be understood from the curves in the upper panels. We have therefore decided to remove MFE and RMSE from Table 4. We hope that the metrics, and their complexity, are now clearer to the reviewer, and help to answer the questions regarding FWP thresholds. Over 0.1 kgm-2, the MFE of FWP is between 80% and 50%. Around 40 gm-2, the MFE is around 100%, i.e. only 40 gm-2.

Response to comments by referee #3

The manuscript "The Arctic Weather Satellite, introducing a new wavelength range for ice hydrometeor retrievals" by McEvoy et al. introduces a new frozen water path (FWP) product based on observations from the Arctic Weather Satellite (AWS), which provides the first operational Earth-viewing microwave observations at submillimeter wavelengths. The product is based on a pixel-wise retrieval algorithm using a database of reference hydrometeor properties derived from CloudSat Cloud Profiling Radar (CPR) observations. The dataset provides retrievals for the central portion of the AWS swath and is restricted to latitudes equatorward of $\pm 60^{\circ}$. The retrieval also exhibits increased uncertainties over high-latitude land surfaces.

The manuscript is generally well written and provides an adequate characterization of the dataset and retrieval methodology. However, several statements regarding the accuracy, spatial resolution, and physical information content of the retrieval go beyond what can be established from the analyses presented in the manuscript. In particular, only limited comparisons against independent observations are provided, making it difficult to determine whether some of the additional small-scale variability seen in the AWS retrieval represents genuine FWP structure or retrieval noise.

I therefore think that several claims regarding retrieval accuracy and resolved spatial structure should be removed, qualified, or reformulated.

Since these concerns primarily affect the interpretation and presentation of the results rather than the retrieval methodology itself, I recommend minor revisions.

Comments:

Title: The current title may overstate the novelty of the wavelength range. Submillimeter observations have previously been used for atmospheric and hydrometeor observations from airborne sensors and limb-sounding instruments. I suggest clarifying that the novelty concerns operational Earth-viewing observations, if that is what the authors intend.

- A good point. We have now specified that the sub-millimetre is the wavelength range and that it is novel in the context for operational ice hydrometeor retrievals.
- We opted not to qualify observations as earth-viewing, since limb-sounding can be argued to also belong in this category.

l. 11: Given the limited independent validation presented in the manuscript, I do not think this claim regarding retrieval accuracy is sufficiently supported. I suggest removing or qualifying it. The claim also seems difficult to reconcile with the stated limitations of the retrieval over land surfaces.

- We thank the reviewer for this point, which another referee also raised. To avoid overstating the claim, we have extended the abstract. We now directly concede that a proper validation was not made and why. We also state that the comparisons presented show that performance is better than other passive retrievals, but now qualify that there are issues over ice and intense convective cores over land.

l. 45: Please elaborate on the statement that passive retrievals do not provide new physical insights. It is not clear to me what distinction is being made here between the physical information provided by passive retrievals and that provided by the proposed approach.

- We thank the reviewer for pointing out the ambiguity in our wording. We intended to refer specifically to data-driven "black-box" machine-learning approaches, where assumptions are not necessarily transparent. This is in contrast to methods in which assumptions on, for example, microphysics are explicitly modelled, and therefore can provide physical insights. - We have now added a clarifying sentence to this paragraph, stating that data-driven "black-box" machine-learning retrieval approaches do not include independent and explicitly modelled assumptions about microphysics.
- We have also updated the introduction in a similar way, to more explicitly contrast with our approach.

l. 89: As I understand the methodology, the CHIPS-AWS retrieval is itself indirectly constrained by CloudSat-derived hydrometeor retrievals through the reference database used for its development. Please clarify this relationship when contrasting CHIPS-AWS with CloudSat-based products.

- Our simulations use CloudSat radar reflectivities as input. However, no CloudSat-derived retrieval product is used at any point. We are therefore not constrained by any of the CloudSat-based products to which we compare. We thank the reviewer for bringing this to our attention, since it is an important distinction, and have updated the

text to clarify that our simulations are driven by the radar reflectivities.

> l. 91: Please explain more concretely how the proposed approach enables improved physical understanding. What additional physical information can be inferred from the retrieval beyond the presence and magnitude of the retrieved ice hydrometeor signal?

- We thank the reviewer for the interesting question. Since our synthetic database is produced from simulations, all assumptions (e.g. microphysical) are explicit, and this allows us to gain physical insight. For example, if poor retrieval performance is observed for a set of observations, these cases can be compared to similar cases in the synthetic database. This allows us to assess the impact of our assumptions and the sensitivity of the retrievals to them. Consequently, we can evaluate the quality of our assumptions, which in turn informs our understanding of the real world.
- The text has been updated to clarify that a synthetic database produced from simulations allows us to explain the mapping from a physical state to simulated instrument observation, with all microphysics and other assumptions being explicit.

> l. 194: Is the primary reason for these restrictions the difficulty of accurately simulating observations at high incidence angles over high-latitude surfaces? If so, I think this should be stated more explicitly.

- The restrictions were primarily chosen to limit scope and complexity in the initial version of CHIP-AWS. Our strategy was to first restrict to conditions which we are confident we can simulate well, validate our results, and then relax these conditions in future versions.
- With that said, the restrictions are chosen with certain concerns in mind. The limited latitude range was primarily chosen due to suspected difficulties in simulations over snow- and ice-covered surfaces. The scan angles are limited due to a concern that too high incidence angles would result in a too slanted beam through the atmosphere, affecting the vertical column-retrievals.
- The concerns behind these restrictions are now mentioned in the text.

> l. 205: I do not think an effective spatial resolution can be inferred directly from the nominal footprint size or sampling characteristics. Please quantify the effective resolution using appropriate test data, or reformulate this statement in terms of the nominal footprint or sampling resolution.

- We adopt the suggestion of reformulating in terms of nominal footprint,

> l. 214: Please consider reformulating this sentence for clarity.

- It is now more clearly specified what values are excluded from averaging.

> l. 216 and following: Please specify whether (Z_m) is calculated along the vertical radar column or along the slanted AWS viewing geometry.

- The text is updated to specify that all retrievals of FWP, Zm and Dm are estimates of the vertical column and not along the slanted AWS viewing geometry.

> Section 4.1: Please specify whether any spatial averaging or aggregation is applied to the reference FWP values before they are used in the retrieval database.

- We have updated the text to clarify that radar inversions and simulation are performed

in a grid that follows CloudSat's native sensor grid in its along-track trajectory. In the altitude dimension, radar reflectivities are interpolated to a grid defined by us.
- Regarding spatial averaging, simulated brightness temperatures are integrated and weighted with a Gaussian to simulate AWS's footprint, which we now address in section 4.1.

> l. 321: Please specify the number of neurons in each layer of the network.

- The text has been updated to specify that each layer has 512 neurons.

> l. 330: Please provide a quantitative range for the differences described here as "substantial."

- Provided a range of expected remap-distances.

> Fig. 4: Please consider using frequency rather than channel names for the labels. This would make the figure easier to interpret without requiring familiarity with the AWS channel nomenclature.

- We thank the reviewer for pointing this out. We strived to do this in all figures, but must have missed this one. It has now been fixed.

> Fig. 6: The micron symbol appears to be missing.

- We thank the reviewer for pointing this out. Unfortunately, this looks like an issue arising from the processing by Copernicus, since the original file does not have this issue. We will keep an eye out for this.

> l. 404: The presence of greater spatial variability or finer-scale structure does not by itself demonstrate that the retrieval contains more information about actual cloud or FWP structure. Some of the additional variability in the CHIPS-AWS estimates also appears consistent with pixel-scale retrieval noise. I suggest avoiding an interpretation of increased spatial variability as increased physical information unless this can be demonstrated against an independent reference.

- We have extended the description in the text to more carefully talk about the suspected increase in ability to resolve smaller systems. We now differentiate between three levels of structure; the finest (pixel ~10km) based variation, slightly larger (at least 40km2) (comprises of multiple pixels/footprints 4-10) we think are genuine systems being resolved, and the larger scale, where CCIC and CHIP-AWS agree.
- That high FWP values are predicted in areas with low FWP through noise is unlikely. This is demonstrated by an added confusion matrix showing that high FWP values are never confused with low FWP values/clearsky.

Response to comments by referee #4

> 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.

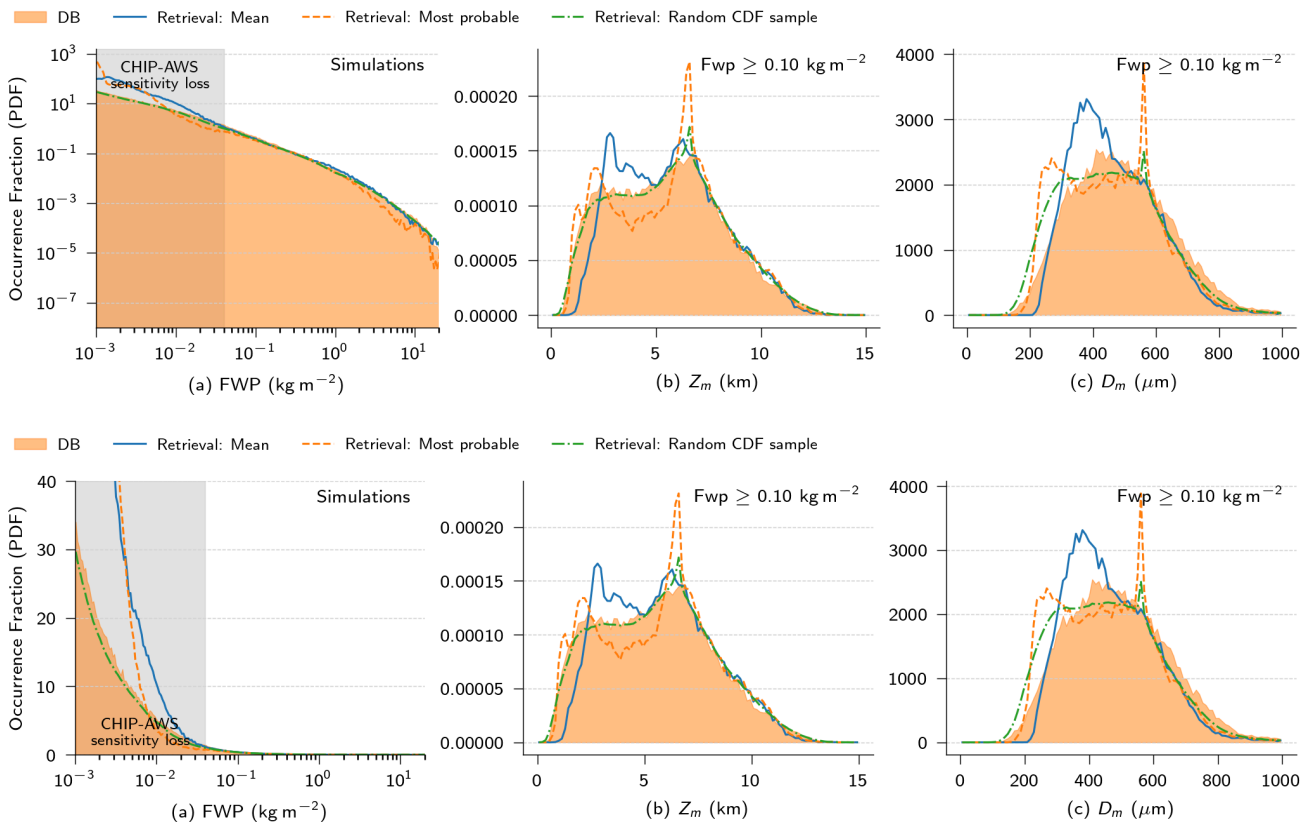
Most of my comments concern the results section. Some reorganization of the text is needed: several lines currently in figure captions belong in the body text, and some lines in the results section belong in the discussion. Specific examples are noted below, but in general, the results section would 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.