The manuscript by Eriksson et al. addresses a fundamental challenge in quantifying atmospheric ice mass in satellite retrievals, GCMs, and GSRMs. It includes large a amount of results and summarises the main problems in both comparison between different observational products and their comparison with models. The paper is well written and organized but needs some changes before publishing.

Most of the following comments are rather questions to the authors to further explain their findings. However, several tables and figures are missing a full description and I recommend minor changes to the figures in order that the legends do not overlap with the plotted data. In some parts of the manuscript, I miss discussion/comparison of the results with the findings of others (see general and specific comments). Maybe the biggest issue I find is, to me, unclear satellite uncertainty. Do you find it as 10% as assumed in "satellite uncertainty" product (DARDAR +-10% is chosen), or 15% which you bring in the conclusions as you have showed through your results, or 40% based on the other paper? This needs to be clear and I find it very strange that the result from another paper (Austin et al. (2009)) is given in the abstract as an estimate for uncertainty. There are also some minor comments regarding parts which need to be clearer in order for the reader to fully understand the content. Finally, I have given some suggestions which do not require a lot of additional work but that I feel would improve our understanding on retrieval uncertainty and their comparison with simulations.

General comments

Abstract + Conclusions: While the abstract overall gives a beautiful picture of what to expect from the article, I cannot but notice how authors' estimation of retrieval uncertainty of 40% is misleading and not a result of the analysis performed. As mentioned later in the text, that uncertainty is reported in Austin et al. (2009) for FWC. In the conclusions, a 15% spread coming from the analysis is mentioned. I believe that this has to be clarified and reflected in the abstract and conclusions.

In several figures, the legends overlap with portions of plotted data. Please reposition the legends outside of the plotting area or where they do not obscure lines.

The captions under several tables (e.g. Table 3, Table 7) and figures do not clearly specify the geographic region and/or time period to which the results refer. Please revise the captions to provide this essential contextual information.

Given the length of the manuscript and the number of the results presented, it would also be helpful to restate the relevant time period in the main text when referring to tables or findings introduced in earlier sections. This would improve clarity for readers and reduce the need to look back through the manuscript for contextual information.

I would find it useful when describing own results to include information about others' findings. In some parts of the text, there is a nice comparison but on several occasions the text is written along the lines of "this has also been visualized/compared/analysed in…"

Satellite uncertainty in section 3.6, Figure 5 but also later leaves me with not being sure I understand what the grey area refers to (e.g. description in Figure 5: "The grey area represents the expected range according to older satellite observations (Sec. 3.6), based the zonal mean of DARDAR for July and August during 2007 to 2010)". Could you describe more clearly what you consider as satellite uncertainty? Is it the mean values during those 4 years and adding +-10% of the shaded area as described in Line 386? Also, does the DARDAR data here include only daytime or nighttime too? I

would also appreciate if the authors could add the mean value line. I was wondering if the assumed uncertainty is larger than the standard deviation of the DARDAR data for the years given? Moreover, since you are comparing DARDAR and 2C-ICE extensively, I would recommend to include and compare that DARDAR data (2007-2010) with 2C-ICE (2007-2010). It is also available on open source: TIWP and CIWP data used for accepted JGR paper 2020JD032848RR

Table 4 does not include information about the period (neither specific year or season – I gather it refers to the annual 2015 mean? Since you use DARDAR 2007-2010 as reference data (+- uncertainty), I think it would be useful to add the mean FWP for those years into the table, or later in the text, to understand overall multi-year variability of the data.

Generally, I am not sure what to expect from CCIC data but am a bit surprised with such large values since machine learning product uses passive remote sensing, especially in the tropics as I would expect that passive remote sensing will saturate fast (therefore, it is probably the result of ML statistics rather than retrieval?). I would appreciate if the authors could expand on it.

Section 4.2: The section is a bit unclear to me and maybe some conclusions are not very easy to reach. The manuscript is already long, but maybe adding some text to the appendix would help readers.

I find Figure 9 very interesting. I would be very much interested, if possible, to see what annual cycles other satellite data has (even for shorten time period).

General comment on GSRMs: Have you excluded the first 10 days of spin-up following Stevens et al. (2019), Lang et al., (2021), Corko et al. (2025 …?

In my opinion, it would be preferable if the authors compared the model results with observations at least with respect to the season during which the DYAMOND model simulations were conducted. Despite not being nudged, models were initiated with ECMWF SSTs and therefore the convection is highly influenced with the period of simulations.

Section 5.2 (The tropical region): out of curiosity, do you know if convection is more active/stronger during northern summer than northern winter? I would think that this is the case also due to more land over the northern hemisphere. Would that impact the comparison of AC (%) and in particular CC (%) from the winter DYAMOND model and observations from Table 3 (please add to the captions the area and time period where missing)?

Could you explain the 99.99 percentile? Are those outliers (1 or 2 cases or more)? I am surprised that, in 2C-ICE, it goes to nearly 18kg/m² while with DARDAR to only 9.6 kg/m²? In the convective case you showed in Figure 1, both DARDAR and 2C-ICE go up to nearly 10 kg/m² in the convective core, as well as in figure 2. Can you show the scene of that 99.99 percentile case? Do you know what is so different in DARDAR for extreme cases compared not only to 2C-ICE but also to AOP nominal that constrains DARDAR FWP to cca 10 kg/m² maximum?

Line 93: “Comparisons of 2C-ICE and DARDAR are surprisingly few; exceptions include Deng et al. (2013); Winker et al. (2024); Atlas et al. (2024).” It would be great if you cited those papers and their results when you compare DARDAR and 2C-ICE (I feel there is no unique answer regarding the differences between the two products). In my experience, I have always felt they were “less different” than what you show here. Do you have an idea what could be a reason for such a

difference? Do you think it would be different if you analysed a different time period? For example, in Atlas et al. (2024), Figure3 shows that in winter period (February data from the years 2007–2012) FWP from DARDAR is substantially larger than in 2C-ICE (opposite to what you found in general (table 4/figure 4 etc…). I feel that generally, this kind of discussions is missing in some parts of the manuscript.

Specific comments:

1. Line 26-28: "It is noteworthy that the data request document for the latest Coupled Model Intercomparison Project (CMIP6) in effect defines cloud ice as the ice categories considered by the model's radiation scheme." I feel citation would be good.

2. Line 32-33: "Later studies comparing cloud ice from GCMs with various satellite observation include Eliasson et al. (2011); Li et al. (2012); Jiang et al. (2012); Komurcu et al. (2014); Li et al. (2020)." This is an example of only providing citations without any information. Could you provide their results or reason for their mention?

3.Line 77-78: "In any case, total ice is still the only mass quantity that has a clear definition." Could be expressed better.

4. Line 78-80: "DARDAR and 2C-ICE have been used as the reference in many studies, but are generally used separately." I think a few citations should be inserted after "many studies".

5. Line 81-82: "DARDAR and 2C-ICE do not offer sufficient coverage for addressing these questions and a dataset representing the state of the art for passive retrievals is also applied (denoted as CCIC)." CCIC has not been defined yet, so please define it. I understand from the text that it is a machine learning (ML) product based on passive observations. Therefore, I would not call it retrieval but how you mentioned ML product.

6. Line 102-103: "As older passive retrievals have been shown to have a strong bias with respect to CloudSat-based ones, they are excluded from this study." What do you consider as older passive remote sensing? I feel citations are missing.

7. Section 2.1.3. Passive dataset. Could you explain it a bit more? I am not sure what to expect from this kind of product. IF 2C-ICE has certain amount of ice connected to certain region and season, e.g. over the tropics where passive sensors saturate fast (and will retrieve much lower FWP), machine learning will indicate that there is too low FWP (compared to 2C-ICE) and artificially prescribe statistics from 2C-ICE despite non-physical retrieval relative to passive remote sensing? In that sense, which part is coming from retrieved values from passive remote sensing and which from "copying" the 2C-ICE retrieval? I think it would be helpful to add a few sentences about the product.

8. Line 190-191: I would replace "resolution" into a grid spacing. "Resolution of 5 km" means that the model grid length is a fraction of that (say, around 1 km or less). Therefore, when one refers to the actual model grid length, then "resolution" should be replaced with "grid length". "Resolution" can be used in a general sense, as in "low resolution models". One can also say "5 km grid" in place of "grid length of 5 km". I think there are a few places, also in the tables, where I feel it needs to be changed to grid spacing.

9. Figure 2: The information about the area of analysis is missing.

10. Section 3.3 and text related to Figure 3: To me, it appears that PDFs of FWP (Figure 3a) are very similar and do not vary a lot. In my understanding, everything under 5 g/m² can be detected only by LIDAR, which also loses its sensitivity for <1 or 2 g/m². Therefore, I am not sure if I would focus a lot on the FWP < 10^{-3} kg/m². Also, I believe PDF (Figure3a y axis) should not have units. Regarding Figure 3b: could you please explain in more detail how you calculated it? I find values on y axis very low and very different than in the papers you have cited (e.g. Sokol and Hartmann (2020); Atlas et al. (2024).). I believe it is influenced by the number of bins?

11. Figure5: Even though I appreciate the effort to use newly available data, I would still be very careful about using and interpreting EarthCARE data, as it is still experimental and in the validation process. As authors may be familiar with, there will be a few products available regarding the FWP and the product they use is just one of them. Therefore, at this stage, I would rather exclude this experimental data. However, in case the authors want to keep and show it, it has to be clear in the text, caption related to the figure and the legend, that this is still experimental data, yet to be verified, and instead of using "Earthcare", in the legend and elsewhere, the name of the product "CPR_CLD_2A" should be written because it could make an incorrect impression of EarthCARE data.

12. Line 442-443: "When those outliers were excluded, the remaining models showed differences of about a factor of 6 – comparable to our findings." This result is surprising as it would turn out that there is no overall improvement in the model spread between old CMIP3 and new CMIP6 generations. This is the opposite of what you mention in the abstract. Could you please comment on this and include citations in the results' part that support the progress between the model generations?

13. Line 448: "Most models participating in the CMIP6 and CMIP6-HighRes underestimate the FWP when compared to satellite retrievals." In my opinion, this is clearly an understatement. Nearly all models in Figure 7 underestimate FWP (except one in each group and some small parts of 3-4 models are barely within the "huge" uncertainty of satellite observations). This might change once it is clear what you mean by satellite uncertainty (see my general comment)

14. Line 469-471: "Despite the large variability of FWP in between the models, we can still conclude that none of the models overestimate FWP compared to the satellite retrieval, and therefore none of the models can be falsified. However, it must be considered unlikely that non-reported ice masses can explain the gap to the satellite range for all models." I do not understand what the authors wanted to say here, e.g. "none of the models can be falsified"? Is it always "non-reported" ice as some GCMs include precipitating (snow) ice due to radiation scheme?

15. Figure 10(11) and related text: Since you are showing CCIC for February in Figure 10, why not include (calculate) its mean value and compare it with that value in DYAMOND models, instead of comparing mean FWP from models with annual CCIC and commenting it is about 3% lower than in February? That way, you would not need to speculate in the 1st paragraph in 5.1 section. In Figure 11, I appreciate the comparison with DARDAR for January to March in the years 2007-2010. Would you also consider adding CCIC zonal mean line from Figure 10? You are working with amazing data and interesting results which could, with minimum additional work, contribute to a very rich analysis (you could also calculate and comment the mean FWP for DARDAR 2007-2010 for January to March which is shown in Figure 11. As expected, it can already be seen in Figure 11 that, even though models mostly underestimate FWP, they are all in the range of DARDAR data when you compare it with the same winter season.

16. Line 544-546: ”Model and observational distributions agree most closely at high occurrence fractions, where they exhibit similar peak values, while most models show higher fractions than the observations around FWP ≈ 10^{-4} kg m-2.” As I already commented before, you are describing FWP of 0.1 g/m², values which observations will most probably not retrieve, not even lidar? Figure 12: Again, I do not understand such small values on y axis. See also my comment 11.

17. Conclusion (Line 650-654): Here you mention that you showed around 15% spread, in the text for comparison with models, you assume DARDAR data as a reference with 10% spread, but in the abstract “estimate“ uncertainty up to 40% (because what Austin et al 2009 found in their paper). And this 40% in the abstract, in my opinion, is misleading of what to expect from the paper since the spread you show is “much” lower. Also, mentioning that EarthCARE data could indicate the 40% uncertainty, based on preliminary, yet to be verified, one product would further mislead the reader and I find it not to be appropriate.

Technical corrections:

A minor point: in several places the manuscript does not clearly distinguish between satellites and the instruments they carry (e.g., CALIOP lidar on CALIPSO, the radar on CloudSat/EarthCARE). This appears to be a writing oversight. Please review the text to ensure that satellites and their respective instruments are correctly identified throughout.

Line 39: GSMRs should be GSRMs

Line 432: “However, we still compare the observations following Waliser et al. (2009), Jiang et al. (2012), Li et al. (2020), and others.” – compare models/them with the observations