This manuscript by Eriksson et al presents a fascinating discussion of atmospheric ice mass in satellite retrievals, GCMs, and GSRMs. The paper is very thorough and reflects an impressive amount of work. It is well written and clearly organized, and the figures are effective.

The paper's greatest strength is the section focused on spaceborne remote sensing, which provided an intermediate-level overview of some of the many considerations that go into these retrievals. The authors' expertise was evident, and the level of detail is well calibrated for readers of ACP. I learned a lot from this section and, as the authors note, it fills a nice gap in the literature when it comes to comparison of these commonly used satellite products.

The GCM and GSRM sections were also strong, and the results in those sections will be of interest to the community. There were times when I was left hungry for more detail, but not every line of inquiry can be pursued in a single paper, and I think the authors triaged these questions effectively.

The majority of my comments below are suggestions or follow-up questions and do not necessarily require significant changes. Major comment #2 is probably the most significant and requires at least some additional detail. Otherwise, the paper seems publishable in its current form, and I defer to the authors to decide which comments are worth addressing in the paper.

-Adam Sokol

Major Comments

1. Section 4.2: The overview of regional trends in this section is relatively vague, and there are some conclusions presented that I don't think I would have reached myself. Nothing necessarily has to be changed, but my suggestions are:
    - "Similar spatial patterns are observed in the decadal FWP trends of FGOALS-f3-L and FGOALS-f3-H"…while I appreciate the optimism, I'm just not sure Figure 8 bears this out. Tre is good agreement in the Arabian Sea and just south of the equator in the Atlantic. But the eastern tropical Pacific is the region with by far the most significant changes in Pfreundschuh et al (2025) and Figs B2/B3, and here FGOALS-f3-L shows a clear positive trend, while FGOALS-f3-H is mixed with a regional average that is probably close to zero. There is also significant disagreement north of the equator in the Atlantic and west of Australia. I think this figure could give different impressions to different readers, all of which might be reasonable, which is why I think some more specificity in the text could improve this section.
    - Line 490-497: While I like the idea of distilling the GCMs into two simple groups, it just doesn't seem like the split is clear. I think it may be more accurate and interesting to comment on more specific features – such as the fact that most, but not all, models predict a negative trend in the eastern Pacific ITCZ, which is in line with CCIC and PATMOS-x, but not ISCCP and ERA5, in Pfreundschuh et al (2025). Interestingly, FGOALS-f3-L predicts a positive trend here. I'm guessing intermodel differences are closely related to differences in regional SST
2. With regard to the CC_n analysis and discussion of convective organization
    - I have some reservations about the use of globally uniform FWP threshold to identify deep convection. Several factors may lead to differences in FWP within deep convection – updraft intensity, depth of convection, surface temperature, etc. These things may systematically vary from region to region, meaning the use of a threshold based on the 97th percentile successfully identify convection in some regions but not in others. As a quick check, I looked at DARDAR v3 for the full 2008 year. In the eastern tropical Pacific (180W to 80W), the 97th FWP percentile

is ~0.4 kg/m2, while in the western Pacific (100-180E) it is 0.97 kg/m2. So, it seems likely that much of the convection in the E Pacific would go undetected when the global threshold is used. In Sokol & Hartmann (2020), we used a fixed FWP threshold to distinguish anvils and convective cores only over the very limited geographic regions from which the threshold was derived.

I realize that a similar argument could be made against any metric used to identify convection—none is perfect, but I think FWP introduces more complications than others. Using a different variable to identify convection would be a larger undertaking and probably does not make sense considering that this section does not have much to do with the rest of the paper. If the authors choose to retain this section, I think it is important to note the potential shortcomings of the FWP approach for identifying convection.

- It seems like the CC_n size/number is analysis is done for each day of output, and the daily results are used to generate the PDFs in Fig 14. Is the FWP threshold used to define convective cores the same for each day? Or does it reflect the 97th percentile of the FWP distribution just for that day? If the latter, I think this raises some additional complications, as even when the entire tropical belt is considered there may very different day-to-day statistics due to variability in the distribution and intensity of convection. Using a single metric for each model seems to make more sense, but does not fix the issues with regional differences described above.

- Putting aside the two points above, I am wondering about the results in Fig 14 – not the analysis itself, but rather how reasonable it is to compare object size in the GSRMs to CCIC. It was shown in earlier sections that the convolutions in the CCIC algorithm reduce the product's resolution relative to the input data, and it can be seen clearly in Fig 1d that the width of convective towers can be exaggerated relative to the other retrievals. Could the spatial smearing have an impact on the size/number of the convective cores being shown in Fig 14? I don't have the technical expertise to make this judgement, but hopefully the authors can provide some insight as to whether this may be an issue.

Minor comments (all optional)

1. As discussed by the authors, the satellite retrievals assume that, below some temperature threshold, all condensate is ice as opposed to liquid. The authors mentioned differences in these thresholds between datasets and how suspected mixed-phase layers are treated. I'm not sure what to make of these different ways of dealing with mixed-phase clouds, namely, could these assumptions result in a significant overestimate of FWP, or in the authors' opinion are the simplifications all reasonable? We know that liquid can exist—even dominate—at temperatures well below the thresholds used by the retrievals (which from section 2.1.2 seem to vary from 0 to -7 C). Does this matter? Or is the mass of liquid that may mistakenly be classified as ice likely inconsequential, since ice particles tend to be larger and more massive? This was a question I was left with at the end of the paper and I would be interested to know the authors' thoughts given their expertise.
2. It may be interesting to mention in the introduction or conclusion efforts by the modeling community to avoid the arbitrary separation of ice into "cloud ice" and "precipitating ice" categories, namely the introduction of the P3 microphysics scheme which uses a single ice category to avoid these problematic distinctions (see Morrison and Milbrandt 2015; https://journals.ametsoc.org/view/journals/atsc/72/1/jas-d-14-0065.1.xml)

3. The text uses both g/m2 and kg/m2 for FWP. I suggest unifying all values to kg/m2 to match the figures, which makes the back-and-forth referencing to figures easier for the reader.
4. (optional) While there is significant attention devoted to the FWP occurrence fraction distributions, much of the paper is focused on *mean* FWP (be it global or zonal). This is reasonable of course, and I suspect most readers of this paper will agree that quantifying the total mass of atmospheric ice is intrinsically interesting. But it may be worthwhile noting in the conclusion that for many purposes—e.g., cloud radiative effect (CRE), in the case of my own interests—mean FWP is not the relevant metric. Berry & Mace 2014 (https://agupubs.onlinelibrary.wiley.com/doi/10.1002/2014JD021458) have a very nice paper demonstrating that mean FWP does not tell you much about CRE, since mean FWP is heavily influenced by the FWP of convective cores but the thinner clouds are what largely determine CRE. So, while models may be very biased when it comes to mean FWP, they may still do a good job on CRE for the right reasons.

Line Comments

Line 35: By "not represented in their output", do you mean that it is not on the standard list of output variables for projects such as CMIP, or that these models are not built to output precipitating ice at all? Assuming the former, a clarification might be useful.

Line 39: GSMRs -> GSRMs

Line 38-40: I'm wondering what the authors mean when they say that ice microphysical processes have a direct physical representation in GSRMs. Clouds themselves are explicitly resolved, which is certainly a big improvement but microphysical processes are still parameterized.

Line 93: see also Figure 4 in Gasparini et al (2025) and Figure 1 in Sokol et al (2024). It could be worth mentioning the quite substantial differences between DARDAR v2 and v3, as shown in both of those figures in addition to the cited Atlas et al (2024)

Section 2.1.3- I think it would be helpful to include 1-2 sentences of what measurements exactly the CCIC product is ingesting in and what it is putting out. As is, the background on CCIC is scattered in a few places, and I am not exactly sure what my expectations for the product should be – i.e., is it most sensitive to small particles because it relies on passive IR measurements, most sensitive to larger particles because it is learning from CloudSat retrievals, or is the idea that it is doing both? Based on other ML-based retrievals, I think it is using merged, passive IR with overlapping CloudSat measurements to produce a 2C-ICE-like retrieval for every column of IR measurements. A sentence or two concisely describing this might be helpful.

Line 237: Did the authors confirm whether DARDAR indeed detects *no* cloud ice for these thin cirrus, or if the amount of ice detected just falls below the colorbar cutoff of 10^-5 kg/m3? These seems to be some ice detected by DARDAR that does not appear in Fig 1a—for example, Fig 1c shows nonzero DARDAR FWP at latitude~3.25 N, but in panel a these columns appear cloud-free.

Line 370: sensitive -> sensitivity

Section 3.2: I wonder if there would be better agreement between these DARDAR and 2C-ICE at low IWP if nighttime observations were used, when the lidar backscatter signal is easier to distinguish from background noise. If the authors agree, this might be a worthwhile point to mention.

Section 3.4: the unexpected agreement between DARDAR and AOP is quite striking. If the authors have any more speculation as to why AOP is so much closer to DARDAR than 2C-ICE, it could be interesting to include.

Section 3.7: the authors may also wish to mention IceCloudNet as another advancement in the ML-based retrieval category, although the recently published version is not global in coverage: https://journals.ametsoc.org/view/journals/aies/4/4/AIES-D-24-0098.1.xml

Line 470: I'm not sure what is meant by "none of the models can be falsified"

Line 471: Fig 6 shows that there is not much of a systematic difference in global mean FWP between models that include falling ice in their FWP and those that do not. It would be interesting to know if this is also the case for the zonal-mean FWP picture. While my gut tells me systematic differences are unlikely, they certainly might be plausible considering differences in governing processes between tropical and mid-latitude ice. No further analysis necessary, but if this can be easily done it might be interesting.

Line 540-541: the tropics are references earlier in the paragraph, but it might be good to reiterate here that this sentence applies only between ~20S and ~40N

Table 2, Table 6, Fig 12 lines 520, 543 (probably others that I've missed) – while many will make the connection, I recommend changing "GFDL" to "FV3" to match the name of this model in DYAMOND project