Response to review by Adam Sokol

> *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.*

- We thank the reviewer for noticing the amount of work that is behind the study, finding the efforts worthwhile and understanding the aims and practical constraints of the manuscript. It is also appreciated that the review is clear about what are strong suggestions and what we can consider as optional.
- Please, find below replies to your comments. We gave emphasis to points where there are commonalities with the input from one or both of the other referees, still carefully considering all points raised.

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. There 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.*

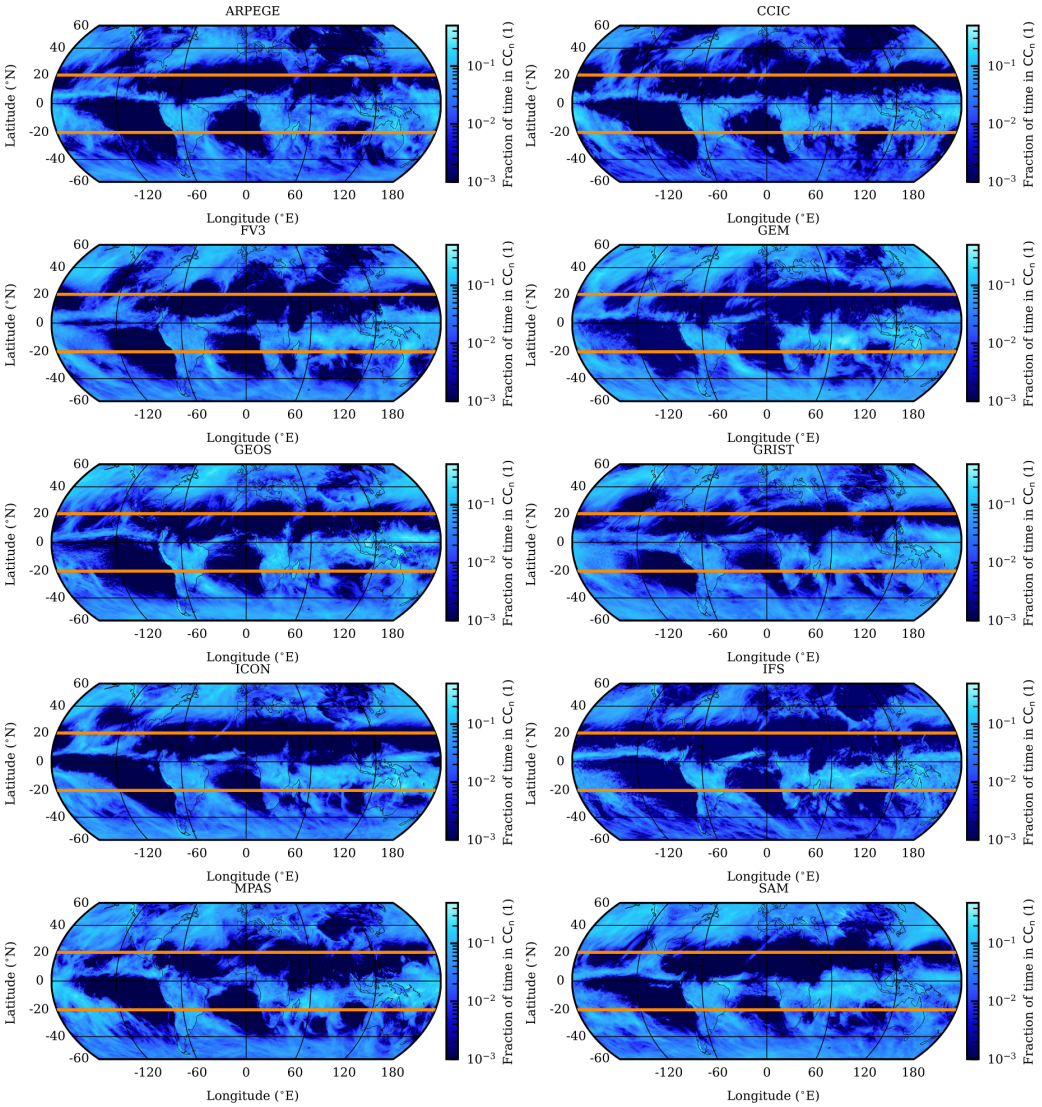
- We agree that Figure 8 can give different impressions to different readers. After re-evaluating this part of the manuscript we concluded that Figure 8 does not provide additional novel information and therefore we decided to remove it.

> - *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*

- Following the reviewer's recommendation, we have rewritten the text.

> *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.*

- We understand the concern. First of all, we have added comments to the text to remind the reader that we analyze high FWP regions, and not convection directly.
- It is correct that the highest FWP values tend to be localised to some regions, but it should still be reasonable to assume that FWP and the convective strength are related, and we are then hopefully looking at the areas of highest convection both with CCIC and inside the models. In fact, we relax the test of the models by ignoring exactly where inside the tropics the highest FWP values occur. That said, as shown by the figure below, the spatial patterns of CC_n occurrence fractions are similar between the models, and in agreement with CCIC. The figure also shows that several areas have significant portions of CC_n. The coloured lines indicate the region we use in Secs. 5.2 and 5.3, and for which the FWP threshold has been derived (individually for each dataset). The found threshold has, for this figure, also been applied outside of the tropics, for context.



> - *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.*

- We use the later, but see why it could be misunderstood. We now write: "with a common threshold for the complete time period".

> - *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.*

- In the discussion of Fig. 15, we wrote "while the limitations in CCIC's horizontal resolution can play a role
for areas below about 1000 km².", but it is correct that the limited horizontal resolution shall be considered already for Fig. 14. We have changed the text accordingly, and put significantly more emphasis on the 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.*

- We thank the reviewer for the trust in our expertise, but want to clarify that we are developing passive retrievals and our knowledge on radar observations is not as deep. There were some clues in the text. In Sec. 2.1.2 we wrote "Cloud liquid droplets are too small to generate significant back-scattering at 94 GHz, but the DARDAR approach still assumes that there exist no super-cooled liquid droplets with a size matching drizzle or stronger rain." As Table 5 shows, ignoring the liquid cloud water gives an underestimation (as the attenuation is underestimated). Assuming that all back-scattering below 0C should be incorrect, a fraction should be due to larger liquid droplets. That is, our assumption here is that DARDAR makes an overestimation of FWP, but we feel uncertain about the magnitude. The test in Table 5, setting the temperature threshold at -5C, should be a worst case estimate. In any case, we have here two effects impacting the retrievals with different signs, and the net effect could be close to zero (for large scale averages, not single retrievals).
  We think it is would be distracting to discuss these issues in detail in the text. However, in our rewriting of Sec. 3.6, we found it suitable to make some comments in this direction.

> 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)*

- Thanks for the input. It made us spot a resemblance with some retrievals in development. We certainly agree that this is a good way forward, and now bring this up in the Conclusions.

> 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.*

- Suggestion adopted.

> 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.*

- This is correct, the mean FWP is not very relevant for CRE, but its distribution has some relevance (see citation below). However, we see a broader motivation. We wrote (towards the end of the Introduction): "We focus the comparisons on FWP for several reasons. FWP is directly proportional to the latent heat released during the conversion of water vapour to ice. It can be used to estimate precipitation efficiencies, the fraction of hydrometeor condensation that ends up as surface precipitation (Kukulies et al., 2024). The impact on longwave radiation at the top of the atmosphere is, on average, proportional to the logarithm of FWP (Deutloff et al., 2025)."
  To be even clearer, we have now incorporated Berry and Mace (2014) in this text, has added a citation to put more emphasis on the importance of latent heating and another one referring to tracking of convective storms.

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.*

- The IPCC list on Standard Output from Coupled Ocean-Atmosphere GCMs (see https://pcmdi.llnl.gov/ipcc/standard_output.html) includes the following variables that are relevant in the context of this study: atmosphere_cloud_ice_content (clivi), precipitation_flux (pr) and convective_precipitation_flux (prc), the latter both including liquid and solid phase. There is no specific standard variable that can be used to estimate the precipitating ice hydrometeors. Besides, many models do remove precipitating ice crystals within one time step, which makes accounting for that contribution to the ice mass difficult.

> *Line 39: GSMRs -> GSRMs*

- Changed.

> *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.*

- Changed.

> *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)*

- A good point. The text has been changed to incorporate this information.

> *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.*

- Sec 2.1.3 has been expanded to provide more information on CCIC. However, to not introduce a poor balance with respect to the other retrievals, we still try to keep the description of CCIC compact. That is, we don't want the CCIC to overshadow DARDAR and 2C-ICE. See also our reply to referee #3.

> *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.*

- We have looked at this. It is not a question of the colourbar cutoff, it's the resolution of the plot. There is indeed a tiny amount of ice in DARDAR at latitude ~3.3 N, that extends slightly under 500m in height but only 0.1 degrees in latitude. Since we plot over quite a wide range of latitudes, this pixel-wide area of IWC doesn't appear in Fig. 1a. However, since we use a line plot in Fig. 1c, it appears. However, there are very few of these pixel-wide regions of IWC (only around 5 of them), so the statement that DARDAR detects no cloud ice for these thin cirrus is still valid.

> *Line 370: sensitive -> sensitivity*

- Changed.

> *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.*

- It is a valid point, that has been added to the text.

> *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.*

- The closer agreement to DARDAR can partly be explained by the common 0C threshold, a similarity that is now explicitly mentioned, but we can still not explain why their zonal means end up to be so similar. We are surprised and remind that individual AOP and DARDAR retrievals can disagree significantly (see Figs 1-3).

> *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*

- This article was published after our submission. A sentence referring to IceCloudNet has now been added.
- We use IceCloudNet to exemplify the trend of increased usage of machine learning (ML). When looking at the manuscript with fresh eyes, we felt a need to give a more balanced view on ML. For this reason we have added text to the Conclusions, with a critical discussion of the uncertainty estimates provided. There is here room for improvements in "traditional" retrievals, but, unfortunately, ML in general constitutes a step backwards when it comes to error reporting and comparison to other retrievals (beside the data used for training).

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

- Here we tried to be very formal, pointing out that there can be non-reported ice masses (even when including "snow" in IWP) that theoretically could move the models up to the satellite range, but this is very unlikely. Anyhow, the topic is in fact discussed elsewhere. Accordingly, we have simply removed the paragraph, in order to not cause confusion and make the manuscript a bit shorter.

> *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.*

- It is the same in the zonal-mean FWP picture - there is no systematic difference between the models including falling ice or not. There is not much of a latitudinal pattern either. The models including falling ice that have a low global mean FWP (for example the CMCC, EC-Earth, NorESM and IFS model family) have also overall lower zonal means over all latitudes. The FGOALS model that have a very high global mean FWP show also higher zonal means over all latitudes that are in the range of the satellite observations. One exception is the model KACE-1-0-G which has a comparable high global mean FWP which results from a high latitudinal FWP at the poles (the FWP value at the equator is comparable low). We added this additional information to the manuscript: "Again, no systematic difference between the models including falling ice or not is found and for most models a small relative latitudinal difference in low/high FWP values can be seen, i.e. a model with low quasi-global monthly FWP means also shows low zomal mean monthly FWP values over the whole latitude range.".

> *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*

- The final comments in the paragraph in fact refer to all latitudes. To make this clearer, this summary part is now placed in a separate paragraph. We also removed a comparison to the satellite observations, as it also could add to the confusion.

> *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*

- Renaming done. Beside using FV3, GSAM has been changed to SAM in the same spirit.