Response to review by Anonymous Referee #3

> *This article reports an intercomparison of atmospheric ice mass, first between different retrievals, leading to an estimated uncertainty within the observations, and then between global climate model simulations as well as global storm-resolving model simulations in reference to these observations. The article is very rich in results. The comparison of atmospheric ice mass is not easy, as the range spans several orders of magnitude, and mean values over such a large range are not enough to fully understand differences. Therefore, distributions of atmospheric ice mass are also shown in observations and in global storm-resolved model simulations. In that way, four datasets, all based (more or less directly) on radar-lidar observations from space, and nine global storm-resolving model simulations are compared.*
> *The comparison between the models is more reported than understood, in particular for the global climate models, but this reporting is a first step so that the different model teams can work on further improvement of the parameterizations.*
> *As the outcome of this assessment is important, I recommend publishing, but only after a major revision. This revision is mostly to improve the structure of the article and to clarify certain points, so it should not be too difficult for the authors.*

- Many thanks for finding our assessment important, pointing out some of the challenges, and providing helpful comments. We have done our best to clarify uncertain points. However, we have not changed the structure of the manuscript, a decision motivated below.
- 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

> *The introduction explains well the problems in the definition of cloud ice. However, it is not completely clear if the final term 'frozen water path (FWP) corresponds to a grid-box average or an in-cloud average. From Pfreundschuh et al (2025) I deduce that it is the grid average. This should be clearly stated. The in-cloud FWP, which is directly retrieved by DARDAR and 2C-ICE, is also interesting to compare, as it is used in the cloud radiative transfer in the models.*

- Many thanks for this important remark. We missed this aspect despite ourselves occasionally struggling, when reading journal articles, to understand if "in-cloud" or "all-weather" averages are displayed. We use all-weather (grid-box) averages. This is now clarified, just after defining FWP (towards the end of the Introduction), and is repeated in the Conclusions.

> *The overarching goal of this article seems to be the assessment of the simulated atmospheric ice mass, but the authors also took the effort to intercompare four different datasets, based on satellite retrievals using radar-lidar or radar-only observations. In particular, they present results of the relatively new CCIC dataset, which is based on Machine Learning (ML) techniques trained on the CloudSat-lidar 2C-ICE product. So, the goal is actually two-fold. This should be more clearly formulated in the abstract and in the introduction.*

- We really had the two-fold goal in mind, to both assess observations and models. To make this clearer, observations is now placed before models in title, the start of the abstract is rewritten, and additions have been made at the end of the Introduction.

> *Though several publications exist about this new dataset, it would be very helpful to clarify the description of this dataset and to show the uncertainty of this dataset which comes out of the applied Machine Learning technique, as explained in one of the earlier publications. More detailed questions and comments on this issue:*
> *(a) In Pfreundschuh et al. (2025), the used ML technique for the CCIC is a convolutional neural network (CNN), while in this article the retrieval is given as quantile regression neural networks (QRNN). This is confusing. Indeed, the authors cite several articles which describe the retrieval, but it would help to give a more detailed overview of this retrieval. I am very surprised how with only the use of one 11 micron brightness temperature (TB) together with the structure of the TB variability over regions of about 900 x 900 km2 (with 256 x 256 pixels) allows for such accurate prediction of IWP of the CloudSat-lidar 2C-ICE product, the latter given on a spatial resolution of about 1.5 x 2.5 km2. The TB depends on cloud height, on ice crystal habit and size distribution and on IWP. The TB also depends on season and daytime. How is this taken into account, in particular when the data are also expanded to other observational times than 1:30 AM and 1:30 PM LT?*

- Sec 2.1.3 has been expanded to provide more information on CCIC. For example, it shall now hopefully be clear that CCIC combines CNN and QRNN. 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.
  We can not really explain why machine learning is managing so well in these retrievals, except concluding that there must be patterns that we humans still are unaware of, but the neural network has managed to identify. On the other hand, in the new text we indicate why CCIC also works outside of the CloudSat local times. In short, the retrievals should work acceptably at e.g. 6:00 PM if something resembling the local cloud situation is found in the training data, obtained at 1:30AM or 1:30 PM. Please remember that we do not make use of any optical or NIR channels and thus are not sensitive to the solar angle.

> *(b) In general, regression neural networks give the right average compared to the dataset they are trained on, but scene-dependent biases exist when scenes with very large IWP are rare, as is the case in the tropics. This effect is even larger when the retrieved variable spreads out over several orders of magnitude. This reduces then the range of the ML-derived variable, as can be seen in Fig. 6 of Amell et al. (2025) or in Fig. 5 of Pfreundschuh et al. 2025. Somehow these biases show in the difference between the distributions in Figures 3. There is a large part with very small FWP, can the authors explain these cases?*

- We agree with these observations. CCIC tends to over/under-estimate when the true FWP is below/above about 1 kg/m2, as shown in earlier works and here in Fig 2. And these tendencies give CCIC the highest values in Fig 3 at lower FWP. As discussed just below, still CCIC manages to represent the retrieval uncertainty through providing data to establish confidence intervals.

> *(c) Indeed, the results are much better than those using only passive remote sensing, but it would be interesting to see the uncertainty of the ML retrieval. In Pfreundschuh et al. (2018), it is written that QRNNs also provide the uncertainty, but I do not see this uncertainty quantified or presented in the current manuscript.*

- CCIC uncertainty estimate is now exemplified in Fig. 1. Please note the start of new text in Sec. 3.6, explaining that the uncertainty estimates are difficult to apply for averages, independently of whether CCIC, DARDAR, or 2C-ICE is used.

> *The structure of the article:*
> *(a) After section 2 (Data) which presents Satellite retrievals and models, it is confusing to see sections 3 Satellite retrievals, 4 GCMs and 5 GSRMs. I would include section 3 'Intercomparisons' and then put the initial sections 3-5 as subsections: 3.1 Satellite retrievals, 3.2 GCMs and 3.3 GSRMs.*

- We have not changed the structure as we see this as two-stage process, that we first establish the retrieval trueness in Sec. 3, and then move on to assess the models. The start of the abstract has been rewritten and a paragraph has been added at the end of the Introduction, to be clearer about that the second part builds upon the first. To this we add that none of the other referees have suggested changes in this direction.

> *(b) Furthermore, it is very confusing to see an outlook (section 3.7) in the middle of an article. Normally the outlook comes after the conclusion of a scientific article, which itself presents scientific results and their interpretation.*

- Using "Outlook" was a poor choice of name for the section. Now called "Emerging satellite retrievals".

> *(c) The interpretation of Figure 5 needs some clarifications: The CCIC results are now shown for 10:30 and 22:30 LT, while they have been obtained via ML with a training at 1:30 and 13:30 LT. There is not one sentence on the reliability of this expansion in time. Also, what exactly is the satellite uncertainty shown in gray in Fig. 5? Another interesting point is that CCIC and SPARE-ICE show very similar zonal averages (except NH subtropics). Does this mean that the microwave information is useless in the retrieval of IWP (as CCIC only uses one IR channel)? Is it possible to give some explanations? Also, the authors state that the EarthCARE sensor and retrieval are improved. As the EarthCARE zonal mean is quite low in the tropics, does this mean that the high peaks in CCIC and SPARE-ICA and AWS are due to not-detection of thinner cirrus? This seems to be a huge effect.*

- The low mean of EarthCARE CPR_CLD_2A in the tropics was explained in the text, but could easily be missed. Anyhow, no EarthCARE product is now included, following a suggestion by Karol Ćorko. The figure is thus updated, and now including CCIC for both July and August.
- The validity of CCIC outside of 1:30 and 13:30 LT is commented on above. We also want to draw the attention to that we cited Leko (2025), a master thesis report showing that CCIC indeed provides realistic diurnal variations of FWP, that compare well with the DYAMOND models (or the reversed?). Based on these promising results, we have now started a manuscript on the subject, including comparison to the diurnal variation of precipitation (as given by IMERG), that we hope to submit to a journal relatively soon.
- Regarding the role of microwave data in SPARE-ICE, we remind that a retrieval can have poor precision but still have a good trueness (like CCIC). That is, a less precise retrieval can still provide good averages. We have not looked at this specifically for SPARE-ICE, but we see this clearly when comparing local values from AWS and CCIC. The AWS retrievals contain more horizontal structures and has a higher "dynamical range". So yes, we strongly believe that also passive microwave observations can contribute to FWP measurements.

> *(d) Many intercomparison results are shown, but for example to compare the global mean of a variable which spans several orders of magnitude is not a strong assessment. One interesting point here is that the IFS distribution (Fig. 11) does not agree with the observations, but the near-global mean does! Since the intercomparison sections are quite long, one could probably take the comparison of the global means to the supplementary material and include the global mean values to Table 1 which could also be moved to a supplement, and then one starts this section with the comparison of zonal means. The same for the global means of the GSRMs: I suggest combining Table 6 with Table 2 and moving them to the supplementary material.*

- As argued above, we see this is a two-stage process, starting with the observations and moving on to the models. Unfortunately, we failed in the original manuscript to make that clear. Anyhow, a rearrangement as suggested would indicate that we would look on observations and model results equally. We strongly argue for that the observations shall be seen as a reference for the models.
- We think that global means provide a broad overview of the models performance on FWP, and thus constitute a good start in both Secs. 5 and 6. But yes, global means do not give the complete picture, as IFS exemplifies. However, good correct zonal means can be achieved by an incorrect distribution of FWP. If the FWP distribution behind a zonal mean is fair, it can still be bad for certain longitudes ... We start at the broadest level and zoom in one or two steps, that is all that can accommodated in a single journal article. We encourage others to take the assessment further.

> *4. Retrieval trueness and estimated uncertainty in section 3.6:*
> *(a) I have difficulties to follow the argumentation. From Fig. 1 it looks like 2C-ICE seems to be more sensitive to thin Cirrus and therefore the distribution in Fig. 3 shows two peaks. 2C-ICE also seems to have a larger range in FWP towards larger FWP. Since the range towards the larger FWP counts more in the mean than the larger range towards smaller FWP, the authors find a 24% larger mean. Why should you put more weight on DARDAR and AOP, the latter only using CloudSat data?*

- Yes, Sec. 3.6 was far from clear. Section 3.6 is totally rewritten and some critical additions have been made to Sec. 3.4.

> *(b) The uncertainty range of 40% is assumed without any further explanation, and this is highlighted as result in the abstract. Why do you not show the uncertainty of CCIC which you claim in earlier articles can be obtained via QRNN? The sensitivity studies in section 3.5 show another part of uncertainty, based on the microphysical assumptions. You could base your argumentation on these findings.*

- We did not involve CCIC in Sec. 3.6, the errors stated in the DARDAR and 2C-ICE products would have been more relevant. Anyhow, we now start the section by pointing out that local uncertainty estimates (as provided today) unfortunately are hard to map to errors for averages. Section 3.5 was added with the "retrieval trueness" discussion in mind. We ended up to using it more as background information. The new version of Sec. 3.6 refers more directly to the results in Sec. 3.5.

Minor comments

> *Title: 'ice mass' instead of 'ice masses'?, same in line 11*

- A good suggestion. We did not consider that mass here can be pluralized as mass. Title changed. We have also changed in the text, but kept masses in some places where it can help the understanding.

> *p 1, l 6 -7: 'but its accuracy is limited by biases inherited from its training dataset' : it is true that ML can as best be the same as the training dataset and therefore naturally includes its biases. However, this is trivial, and I would like to see in the abstract also mentioned the additional biases and uncertainty linked to the reduced input.*

- Correct, that phrasing was not very informative. The sentence now reads: "A recently developed machine learning product based on passive thermal infrared observations highly extends spatial and temporal coverage for comparisons, but its local precision is limited compared to radar-based retrievals."

> *p 4, l 93: you may add Vidot et al. 2015 (DOI: 10.1002/2015JD023462), they compared IWC profiles for small and large COD (Fig. 4).*

- We have changed the text (e.g. adding "in depth"), and as DARDAR and 2C-ICE are used together in many studies the word "surprising" has been removed. To reflect this we have also added some references. To do this in "objective" manner, we asked Google Scholar Labs to list the three studies that have performed the most in-depth comparison of DARDAR and 2C-ICE, and we added the two we had not already cited.

> *Section 2.1.2: I would move the second paragraph (p 5, l 120-123) to the front of this section*

- Yes, a better order. The suggestion has been implemented.

> *p 4, l 116: take out 'retrieval'*

- Done.

> *p 5, l 143: 'radar bin' perhaps 'radar vertical segment' ? is each bin or vertical segment about 0.5 km ?*

- Yes, "bin" not the best word. We have changed to use "range gate", or just "gate". Yes, the size of the ranges are 500 m.

> *p 7, l 6 & 7: please add 'boreal' in front of 'Summer' and 'Winter'*

- Boreal has been added.

> *p 7, l 210-211: we sum up … (IWP, GWP, SWP): is this weighted by their fraction within the grid ?*

- With the addition to the Introduction, that we solely use all-weather means, this should now be clear. Adding a remark here could rather have the opposite effects. It could be taken as that in-cloud averages are used in some place.

> *p 19, l 434-435 Section 4: 'the overall assessment is based on global means':*
> *This is really a pity, but probably CMIP6 results only provide the monthly means? It would be important to add in the conclusions that distributions should be added as output for CMIP7. However, you need also to mention that the distributions in Figures 3 may change their shape when reducing the spatial resolution to 100 or 250 km. Did you have a look how they would change?*

- CMIP6 provides also daily and 3-hourly clivi outputs, but these are available for a smaller subset of models compared with the monthly mean outputs. Therefore we consider the monthly mean outputs to be more suitable for our work, since we wanted to include as many models as possible. Furthermore, the usage of 3-hourly, daily or monthly outputs should not produce a remarkable change in the results. Regarding Figure 3, it does not include CMIP6 models. For this reason it is unclear to us why it is referenced in this context.

> *Comparison of zonal means (Fig. 7): the authors compare grid averages of FWP from the model simulations to the range in satellite observations coming from nadir tracks; how do the authors build grid averages if there is only a narrow track within a grid of the GCM spatial resolution? Here, actually the CCIC dataset may be useful as it is expanded to fill a whole grid, even though additional uncertainty is added due to ML expansion.*

- The narrow across-track coverage of the CloudSat results in relativly "noisy" local averages (for e.g. 1x1 lat/lon grids), but we are not showing such averages. For annual zonal means each latitude bin contains enough of samples to provide relatively stable statistics (otherwise the zonal means would be less smooth). The CloudSat measurements are equally distributed in longitude, and averaged they will give a fair approximation of the true zonal means.

> *Section 5: Why do you limit the GSRM means to 60N-60S while the GCMs are averaged over 90N-90S?*

- We limit the analysis of the GSRMs to 60N-60S because our goal is to compare them with CCIC, which is limited to these latitudes. For consistency we have now updated the GCMs analysis (including figures) to also cover only the 60N-60S latitude range.

> *Figure 10: instead of (or in addition to) comparing the mean FWP of CCIC and the 9 models, one could show the difference map between both estimations in order to see where there may be differences.*

- The aim of the figure is to introduce the general capacity of the GSMR models, that it is high in comparison to the GCMs. Representing the GSRMs by their difference to CCIC would instead put emphasis on the differences. In any case, the differences to CCIC are hard to analyze for the ensemble mean, they are more relevant for individual models.

> *Figure 16: It is known that the diurnal cycle of convection differs over ocean and over land, therefore a comparison seems only to make sense when ocean and land are separated.*

- Thanks for this suggestion, that we have followed. The figure and text have been updated accordingly.

> *Section 5.3:*
> *It is interesting that the authors also explore convective indices, but it is difficult to follow this section.*
> *p 31, l 604-605: 'In particular, they conclude that the use of multiple indices is advantageous to successfully characterize the underlying organizational structure.' For me, it looks like they concluded first that several of these indices don't fulfil certain quality criteria, like sensitivity to noise under certain conditions, to spatial resolution etc. and this can explain differences in conclusions about convective organization when using different indices; and second that these indices may not be enough to completely characterize organization. Another conclusion was that some indices are highly correlated with one simple variable, like ABCOP reflects the total area of convective objects, while ROME is very strongly correlated with the mean size of the objects. The latter can be seen by comparing Fig. 15 with Table 8. SCAI and MSCAI strongly depend on the number of objects, which may be very noisy. Since Iorg does not consider the size of the objects, the conclusion on organization using Iorg and ROME does not agree. Perhaps one can add in the table the correlations with the corresponding variable, and it would also be good to highlight in bold or italic the largest and smallest for each index.*

- We have added highlighting in the table, as suggested. The cited sentence seems to have given a wrong impression of the sub-section's aim. It is now removed, but other text has been added to be clearer about the aim, as well as limitations. Our purpose was not to identify a best index. Rather the opposite, despite testing multiple indices, we find none that gives an agreement between the models. We think this is an interesting finding, considering the present strong interest in convective organization.

> *Conclusions:*
> *According to the questions and comments before, some parts need to be rewritten, in particular the2. paragraph p 32, l 652 – p33, l 654.*
> *p 33, l 655: 'Global climate models remain biased low': it may be good to add here that this is on grid average; it can be different in in-cloud IWP (see above).*

- We now remind about the all-weather averaging at the beginning of the Conclusions.

> *p 35, 653-654: I do not understand the last sentence of the paragraph. Why indicate preliminary results from new sensors, for which the retrieval is also based on assumptions indicate the pessimistic view?*

- Yes, that argumentation was not clear. The sentence has been removed.

> *p 35, l 721-722: I do not see any Figures B2 and B3 in the newest manuscript.*

- Please note the hint in our comment (AC1), that some action can be needed to make the browser to download the new version.

Typos

All "typos" have been fixed. Thanks for spotting these mistakes.