Comments on 'Advancements and continued challenges in global modelling and observations of atmospheric ice masses'

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.

Major comments

1. 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.

2. 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.
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?

(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?

(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***.

3. 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.

(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.

(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.

(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.

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?

(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.

Minor comments

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

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.

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).

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

p 4, l 116: take out 'retrieval'

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

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

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

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?

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.

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

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.

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.

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.

Conclusions:
According to the questions and comments before, some parts need to be rewritten, in particular the 2. 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).

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?

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

Typos

p 2, l 32: 'observations' instead of 'observation'

p 2, l 39: 'GSRMs' instead of 'GSMRs'

p 3, l 67: take out 'the' in 'the the'

p 5, l 144: 'a unit' instead of 'an unit'

p 14, legend of Table 3, last line: 'it is stressed' instead of 'it stressed'

p 14, l 313: 'two datasets' instead of 'two dataset'

p 17, l 386: 'DARDAR instead of 'DARADR'

p 18, l 413: add 'to' between 'up 664 GHz'

p 19, legend of Fig. 5: 'based on the zonal mean' instead of 'based the zonal mean'

p 26, l 552: change the first 'between' to 'in' would be easier to understand

p 26, l 553: 'latter' instead of 'later'

p 31, l 605: 'organizational' instead of 'ogrnizational'

p 33, l 685: in 'Data availability': 'Center' instead of 'Cente'