Reply to Anonymous Referee #1 comments on

"Bias in modeled Greenland ice sheet melt revealed by ASCAT"

by Anna Puggaard, Nicolaj Hansen, Ruth Mottram, Thomas Nagler, Stefan Scheiblauer, Sebastian B. Simonsen, Louise S. Sørensen, Jan Wuite, and Anne M. Solgaard

Dear Anonymous Referee #1,

We first and foremost would like to thank you for your insightful comments on our manuscript. In the following, we try to follow and implement your suggestions to the best of our ability, and we sincerely believe that your review/comments have improved the manuscript. Below is a point-by-point response. To ease following the reply, we have your comments in black and our responses highlighted in **Blue**, and suggest changes to the manuscript in **Red.** Moreover, line numbers in our replies to comment refer to the updated MS.

 # General comments

1. It seems to me to be rather poorly considered that the data derived from ASCAT don't give an indication of meltwater in particular, but of all the water present on the surface and in the first layers of the snowpack. This is well presented in the introduction, but in some other parts of your text, this is a bit more confusing. Does it make sense, then, to compare this extent with melt extent in the models? And not the liquid water content in the first snow layers (see for instance Dethinne *et al*. 2023 and Picard *et al.*, 2022)?

We agree that this fact/deficiency of ASCAT melt detection should be clearly addressed throughout the manuscript. The ASCAT algorithm detects the presence of liquid (melt)water in the surface snowpack and not the process of melt. It is currently not possible to derive the quantity of meltwater with EO data alone, as stated in L39. For this reason, several simplifying assumptions and modelling are needed to derive the meltwater volume from ASCAT observations. This is, of course, a limitation in the ASCAT melt observations. In the manuscript, we show, however, that the ASCAT dataset is a useful addition to understanding biases in models better, especially as it is independent of the modelling output. The methods of Dethinne et al. (2023) and Picard et al. (2022) involve assimilating remote sensing datasets into more detailed modelling of surface melt processes to derive the meltwater volume. By including more observational data generally tends to improve the representation of present-day surface conditions (see Dethinne et al. (2023) and Langen et al. (2017)), there's still a great importance of having independent datasets to assess the models model output, in this case the ASCAT melt maps, which can help guide future improvements in model development. Together with the proposed revisions to the below general comments we propose to add to the introduction (L 43):

"Recent approaches, such as Dethinne et al. (2023) and Picard et al. (2022), have assimilated remote sensing datasets into more detailed modeling of surface melt processes to estimate

meltwater volume. Although including more observational data generally improves the representation of current surface conditions (Dethinne et al. (2023) and Langen et al. (2017)) there remains a strong need for independent observational dataset to assess model outputs."

Are we sure that the variables selected in the models are indeed representative of what the *satellite observes?*

While we acknowledge that the surface melt simulated by RCMs is not what ASCAT directly observes, there is a strong relationship between liquid water in the snowpack, to which the satellite is highly sensitive, and surface melt. This makes surface melt a meaningful and relevant model output to compare with the satellite data. Furthermore, as described in detail in section 2.3, we have applied a hierarchical decision tree using dynamic thresholds based on the conditions from the previous winter months to the ASCAT SIR product. This method allows us not only to distinguish between liquid water presence and absence but also to classify periods when liquid water starts to refreeze. In our comparisons between ASCAT melt maps and RCM melt extent, we focus only on instances where ASCAT detects a decrease in backscatter signal or when the signal is fully saturated, both indicating active surface melt. We do not include periods of refreezing in this comparison, as seen in Figure 4. This approach ensures that we are comparing relevant melt periods.

To convey this throughout the MS we suggest adding (Line xx):

"Using this method, we can distinguish between different stages of the melt water in the snow pack associated with melt water instead of only providing a binary melt extent"

We also suggest adding to line XX:

"Further, we combine label ST-2A and ST-2B into a binary active melt label"

We use the "ASCAT melt maps", which show the presence of liquid melt water on the ice sheet. In the MS we refrain from using "ASCAT observes melt" and will instead use "ASCAT observes presence of liquid water".

From another point of view, if we look at the problem the other way round, we could ask ourselves whether an additional ASCAT data processing step could be applied to translate them more specifically into meltwater.

As highlighted in the introduction, this is an active area of research but out of the scope of this MS. We believe that the current version of ASCAT water detection still is beneficial for informing on model biases. Hence, the data is mature to use as they are, despite the lack of quantifying melt volumes.

2. The comparison between RCMs is original for this type of comparison (not included in a "MIP" exercise) with remote sensed data but is also audacious and complicated. This kind of work require deep knowledge of how processes inside the model are modelized/parametrized to be able to compare it with other products. I recommend to deeply double check the way that components of meltwater production or water in the snowpack are determine in the 3 different models and be sure that everything is comparable. In that aspect, not including MAR and RACMO people is a risky move, or at least limit the study to HIRHAM alone. The idea is not to transform this study into an intercomparison of these 3 models, but to better understand it to highlight why the biases you mention are present.

The aim of this paper is to present a framework for evaluating the performance of RCMs using independent satellite observations, in this case using ASCAT melt maps to assess how well RCMs simulate the temporal variability of present-day melt extent. We understand the

reviewer's concern of the potential perception of the work as an intercomparison exercise. However, the emphasis of this study is on applying the same observations dataset to assess the model outputs. Although the internal processes related to meltwater production are undoubtedly important, however, we do not see the scope of the paper to explain the origins of the biases within the models' parametrizations but to assess how well each model captures melt patterns. While the suggestion to limit the study to HIRHAM alone is appreciated, we believe that including MAR and RACMO adds significant value by providing a broader context for evaluating model performance. This enables us to assess how well the intercomparison setup works across multiple models and see how each model captures melt patterns to highlight potential biases and strengths in the models' abilities to represent present-day melt, which could inform future model development and improvements. This clearly strengthens the argument for incorporate this approach in future MIP efforts.
To clarify this, we suggest adding to the introduction (at L. 67-71):
"By using ASCAT melt maps we aim to establish a framework for evaluating the performance of RCMs in simulating the temporal variability of present-day melt extent. As RCMs are often calibrated with respect to basin-wide surface mass balance, incorporating an independent satellite dataset like ASCAT melt maps enables a more comprehensive assessment of model performance. By including HIRHAM5, RACMO2.3p2, and MARv3.12, we assess how well each model captures surface melt patterns, focusing not on internal model parameterizations, e.g. albedo and near-surface temperature, but on the representation of melt extent"
And also, in the conclusion (L. 386) we suggest to add:
"Our analysis demonstrates the value of using independent datasets like ASCAT to identify the spatiotemporal variability of RCM simulated melt. This approach complements traditional model validation methods and intercomparison exercises, which can inform future model development to better simulate ice sheet surface melt and possibly be incorporated in future MIP efforts."

Specific comments

Data

L119: For MAR add the same detail-level than the 2 other models (at least how albedo is retrieved).

We have added the name of the snow module, namely Crocus, and added a sentence about the internal computation of albedo.

L134-140 "MARv3.12 includes the snow model Crocus (Brun et al., 1992), that simulates a number of layers of snow, ice, or firn of variable thickness and energy- and mass-transports between each layer. The snow model also provides snow grain properties, which are used in combination with density, age, and type to simulate snow albedo (Brun et al., 1992; Fettweis et al., 2017; Antwerpen et al., 2022), MARv3.12 also have an albedo range for bare ice between 0.4 and 0.55 depending on the cleanliness of the ice (Fettweis et al., 2017). While both RACMO2.3p2 and HIRHAM5-ERA5 incorporate MODIS observations into the albedo computation, the surface albedo in MARv3.12 is only based on the internally computed broadband albedo (Brun et al., 1992)."

L136-137: Could you a bit more detail the SIR algorithm? As you need more measurements, is the reconstruction constant in time and in space? Are there any supplementary uncertainties bring with this method?

Multiple passes of ASCAT data are used to enhance image resolution and improve spatial coverage. Temporally combining these passes results in an averaged representation of the region. For Greenland, as mentioned in the manuscript, data from 4 days are combined (this is constant) to produce a single SIR (all-pass) product. These products average out any diurnal variations in sigma nought over ice (i.e. melt during day, refreezing during night). This process introduces additional uncertainty. Furthermore, potential azimuth angle dependencies are not considered in the construction of the SIR products.
We have added the following:
L152-154: "With the 4-day time interval diurnal variations in backscatter signal over ice, such as melt during day and refreezing during night, is averaged out, which introduces additional uncertainty. Additionally, the resolution enhanced ASCAT product may not capture short melt events in the spring and intense precipitation events, as these signals are averaged. Furthermore, potential azimuth angle dependencies are not considered in the construction of the SIR products."

Methods

- L170-171 : "We compare the RCM output of surface melt with observed 2 m temperature data", even if it could be obvious, if think this kind of sentence could be not so easy to understand at first read. I suggest switching some parts of the second paragraph of your method with the first one to be more readable.

We have rewritten the first part of the methods (L184-188): "To evaluate temperature biases in the RCMs, we compare the modeled output with observations from PROMICE GC-net AWS. Since melt is not directly measured at the AWS stations, we use 2m air temperature as a proxy for melt conditions, as near-surface air temperature is closely linked to melt processes. This approach allows us to identify and quantify temperature biases in each of the RCMs and assess how well the models simulates melt compared to in situ observations."

- Could you more detail how you construct the ROC- and PR-curves. Specifically, how do you determine if it is a true or false melt-day compared to AWS if you already calculate it for different temperature rate. If I understand well, you don't have directly melt rate observed at the AWS, but a guessing relates to the temperature measured, transform to melt through a lapse rate correction (which should be explained here). I think something is unclear for me here because of a lack of details.

To make it more clear how true positives, true negatives, false positives and false negative as constructed we have added the following:
L212-214 "Here we define the AWS as true. Thus, we define the TP when the RCMs and AWS agrees when melt is present, and FP is when no melt is observed both by the AWS and RCM. When melt is only observed at the AWS but not the RCM, it's defined as a FN and vice versa for FN."

- Please, explain here your method to choose the grid cell(s) corresponding to your AWSs. Is it the nearest neighbor, or the 4-nearest, ...?

It is only the nearest grid cell with the center closest to the AWSs that is selected. To make it even more clear for the reader we have added the following sentence:
L185: "We compare the AWS to the RCM grid cell which has the closest center point to the AWS location"

- L195: You should reefer earlier to Fig. 5 to illustrate how understand the ROC- and PR-curves.
We have added the following to the sentence in L195 to refer earlier to the ROC-curves.
L215 "The ROC curve provides the total performance measure across all potential classification thresholds where a random model will produce a diagonal line, see Figure 5 as example"

- L207: You're talking about a threshold of -2°C (also in Table 1), but in your Figure 5, your curves are only from -1 to 1 °C. How did you determine this threshold?
We explored a wider range of temperature thresholds not only limited to what is shown on the plot. We have now updated the figures to include more temperature thresholds from -2 to 1 C. Based on referee #2 comments we have plotted less temperature thresholds to make the plot more concise.

- Table 1: Why only one melt threshold and one per AWS for temperature? Is it an average from all melt rate retrieved at each AWS? Your method needs to be better described for that too. Moreover, are you only using these few AWSs as presented in the Table 1? Why only these ones? In the AWS description section (2.1), you mentioned that you will use all the AWS, at least much more than presented in Table 1.
We use all PROMICE GC-net AWS stations located on the ice sheet that have data in the period of ASCAT. The included stations are shown in Figure 1. To make it clearer, we have made a correction to L186:
"To evaluate temperature biases in the RCMs, we compare the modeled output with observations from PROMICE GC-net AWS shown in Figure 1. "
Table 1 only gives example of temperatures at six selected AWS to showcase the spatial variability for temperature differences across the ice sheet. To make it even more clear from the figure text:
"Table 1 Melting thresholds for the different RCMs based on in situ PROMICE AWS observations of air temperature. The table also gives examples of the mean air temperature for August and July at six selected AWS and mean across all stations shown in Figure 1. The corresponding mean air temperature for July and August are also showcased for each RCM. Figure 3a-d illustrates the mean JJA air temperature for each RCM. "

- Still concerning observation from AWS, how do you manage the fact that most of the AWSs are in ablation area, and probably in the area where ASCAT cannot correctly detect the presence of liquid water, that you even mask outed? Or do you only consider AWSs inside the ASCAT mask? Otherwise, you decide to correct RCMs' melt with a threshold determine with comparison outside your area of melt comparison, which is not 100% valuable. Also, could please consider adding this mask in your first figure to well situated it compared to AWSs' localization as well as a more detailed comment on how you retrieved it.
The evaluation of RCMs vs AWS is done independently of ASCAT. Thus, AWS stations outside the ASCAT snowline mask are included since we want the RCMs to align with in-situ measurements across the entire ice sheet. The uneven distribution of weather stations are stated in the discussion L290-293 as a cause for a uneven representation. We further want to point that even if the stations in the ablation zone were excluded there's still an uneven distribution of weather stations on the ice sheet.

Results

- L242-243: "ASCAT detects the increase in melt extent earlier compared to RCMs" Isn't it due to the detection of water by satellite and not directly melt (cf. 2nd general comment)?

It's true that ASCAT does not detect melt water directly, but rather the presence of liquid water. However, the satellite should not be able to detect firn aquifers as the penetration depth is not more than 1-2 m in dry snow conditions. Further to account for changes in the snowpack associated with melting form the previous melt cycles the algorithm applies a "recalibration of the winter signal" to account for these. This means that we ASCAT detects a decrease in backscatter we know it can only be associated with a increase in liquid water in the snowpack. We propose the following revision to line L272-275:

"At the beginning of the melt season, ASCAT detects the increase in the extent of liquid water 10-15 days earlier compared to when the RCMs simulates an increase in the melt extent. However, the decrease in extent of liquid water at the end of the melt season corresponds well with the modeled melt extent"

- L249: can we talk about 'prediction' here as RACMO is prescribed by reanalyses at its lateral boundaries? Please rephrase with another verb.

We use simulates instead. L281: "Results show that using the in situ informed thresholds, only RACMO2p2.3 simulates melting of…"

- Table 2: Could you add the mean number of melt day for both observation and RCMs to compare your RMSE. You should also compare the difference between your two methods (uniform or in situ informed threshold) to determine if the gain with one or the other is significant (or real statistical test, it's even better).

We have now performed Wilcoxon signed-rank test to test if there's significant difference between applying the two thresholds. We use the Wilcoxon signed-rank test since we have a paired sample, and we cannot assume a normal distributed. Thus, we cannot use the standard t-test to test if there's a significant difference between applying the two thresholds. For all RCMs the p-values shows a significant difference between the two thresholds and using the Rank-Biserial correlation we also get an indication that the magnitude of difference is significantly large. Further, we have added a new table to the MS stating the mean duration of melt period, mean number of melt days and the p-values when applying both the uniform threshold of 0.01 mmwe/day and the in situ informed melt threshold:

Table 2. Mean annual melt days and mean duration of the melt season for each model using two different thresholds and for ASCAT. The melt season duration is defined as starting when at least one grid point experiences melting. A Mann-Whitney U test was applied to assess whether there is no effect of using an in situ-informed threshold. Additionally, the Rank-Biserial Correlation (r-value) was computed to indicate the magnitude of the difference between thresholds.

	Mean melt days		**Duration of melt season**			
Threshold method	uniform	in situ	uniform	in situ	**p-value**	**r-value**
HIRHAM5-ERA5	25	21	225	204	> 0.001	0.48
HIRHAM5-ERAI	16	16	205	201	> 0.001	0.49
RACMO2.3p2	24	18	365	344	> 0.001	0.47
MARv3.12	25	18	274	166	> 0.001	0.45
ASCAT	18		154		-	-

Further we have added to the result section:

"Table 3 shows the mean annual number of melt days and the mean duration of the melt season for both the RCMs with the two thresholds applied and ASCAT. We define the start of the melt season when at least one grid point experiences melting. The melt extent using the in situ informed thresholds tends to align better with ASCAT observed mean number of melt days and the duration of the melt season. Furthermore, when the uniform threshold of 0.1 mm w.e. day^{-1} is applied melting occurs in parts of the SW basin all year. We apply a Mann-Whitney U test to test if there is no effect of using an in situ-informed threshold. The p-value in Table 3suggests that it is very unlikely that there's no effect of using an in situ informed threshold."

Discussion

- L258-259: "Tab. 2 shows that by ensuring that the RCMs align with in situ measurements at specific locations." You cannot claim that RCMs align with measurement only by considering the RMSE. You need deeper statistical analyze to claim this.

We have now added an additional table, see above. We now see that applying a in situ informed threshold also aligns the mean number of melt days and the duration of melt season more closely.

- L274-276: Apply different melt threshold (based on the same in situ observation) for the different RCMs is also revealing a certain kind of bias in the model. Could you discuss that too in your Discussion?

We have re-written the start of the section to:

L308: "To get the most valid comparison between each RCM and ASCAT, we utilize in situ observations to assess biases and to determine an appropriate threshold for the melt extent in RCMs. By fitting each RCM to in situ observations we minimize the differences that is introduced due to model set-ups like resolution, parameterization etc.. Thus, we reduce..."

- L297-298: First you say that RACMO present the lowest albedo, then you also explain that MAR and HIRHAM have a lower albedo. Could you rephrase to better emphasize what are the differences and key features for each model/group of models?

Now corrected to:

L335: "RACMO2.3p2 is characterized by the highest surface albedo across the entire ice sheet, while MARv3.12 and HIRHAM5-ERAI are dominated by lower surface albedo, especially in the accumulation zone."

- L300: Could you investigate a bit more why you have such differences in RACMO and HIRHAM MODIS-based albedo in the ablation area? If it's possible, it could be nice too also compare the different albedo of the models to the MODIS albedo.

The scope of this paper is to showcase how ASCAT melt extent can be used to evaluate RCMs' performance in simulating melt extent. While albedo is an important factor influencing surface melt, our focus is on assessing melt extent directly and not on investigating albedo variability across models. Further studies could build on this work by exploring the impact of albedo differences on model performance.

Based on this comment and the 2. general comment we suggest adding a paragraph to the introduction, see 2. General comment.

- You do not talk about the differences on how the models represent the firn layer, whereas in you introduction you mention that "*The magnitude of the decrease in backscatter varies due to factors such as the snow water content and the specific properties of the snow, such as grain size and the presence of ice layers and lenses, which influence the dielectric properties and roughness geometries (Wismann, 2000; Long, 2017).* " I heard there that the signal to detect (melt)water at the surface is dependent of the snowpack conditions which are not represented/modelled in the same way in the 3 models. It should be a supplementary discussion point as melt event, and presence of water, could be delay, or more or less important, due these different way to model/parametrize the firn layers, then lead to difference when compared to ASCAT.

We agree that the model has different implementations of the surface snow/firn properties, including grain size, which could inform us about why they act differently in the evaluation. However, as the scope of the MS is not to conduct a full MIP, we here try to stay objective and see how state-of-the-are processing of ASCAT data can be utilized. It will complicate matters significantly if we model snow properties that need to be introduced in the ASCAT data, and we would not be able to separate the model biases from the observational biases.

We suggest adding at L366: "Finally, the ASCAT backscatter varies due to additional factors such as specific properties of the snow, e.g. grain size and the presence of ice, due to its influence on the dielectric properties and roughness geometries. Here, two possibilities consist in progressing the melt retrievals of ASCAT; we could use the surface properties from the individual RCMs or in situ observations. For the latter, there is a seasonal bias in the in situ observations and a lack of spatial coverage, making this difficult to use for ice-sheet-wide earth observation data production. As for the RCMs implementation, this would hamper the ASCAT data as an independent data record."

- Figure 3 a-d: center your color bar to 0, it's misleading as it is. And please use only 2 varying colors, one for positive and the other for negative values. Also, please avoid yellow at pivotal value.

We have revised accordingly. For figure 3 a-d the colorbar is now centered at zero. Further, we have chosen another colorbar without yellow as a pivotal color for Figure 3, 6 and A1. All colorbars are now segmented colorbars, as suggested by referee #2. We refer to the updated MS for the updated figures.

- Figure 3, RCMs' Albedo: concerning the MAR model, you plot albedo for entire land areas and not only what looks like an ice mask in the 2 other models. Are you sure you plot the albedo used in the melt calculation, meaning the one for the ice grid points? Concerning the albedo from RACMO, considering the intercomparison and preliminary feedbacks from the PROTECT project, the albedo from RACMO presented here suggests high values, hinting a potential error when choosing which albedo plot.

We have updated all plots in figure 3 and A1 to only include data where we have ASCAT observations. Concerning the RACMO albedo, we have chosen to use the albedo provided directly by the RACMO development team. This ensures consistency with the model's intended output. Regarding the preliminary feedback from the PROTECT project, as they are not yet published, we are unable to comment or incorporate these findings into the current analysis. Once published, future studies may consider these insights for further comparison and refinement.

Conclusions

- L340: "[...] *can lead to more accurate simulations of surface energy balance.*" Could you rephrase, as you don't actually look at the entire surface energy balance, but only some components.

We acknowledge that we do not assess all components of the surface energy balance and how they affect the meltwater production. Instead, we focus on key variables such as radiation and albedo. Specifically, we show that by ensuring the variability of albedo is accurately simulated by the models—such as through the incorporation of MODIS—it can contribute to a more accurate representation of the surface energy balance overall. To make this more clear we have rephrased the sentence to the following:

L383: "By ensuring that the models accurately simulate the variability in albedo, such as through incorporating MODIS bare ice data, it can lead to a more accurate representation of the surface energy balance, and consequently, meltwater production."

Appendix

Figure A1 is exactly the same than Figure 6. Is it necessary as the appendix are in the continuity of the text and not in another document as Supplements?

This is a typo. Figure A1 is not the same as Figure 6, but rather the melt extent when a uniform threshold of 0.01 mmeq/day is applied to all models. The figure text is now updated to:

"The mean annual number of melt days modeled by the RCMs using an in uniform melt threshold of 0.01 mmeq/day to define days with significant melt. Pixels with <1 day of melt on average are marked as white, showcasing areas where melt rarely occurs. (e-h) The mean annual difference between the number of melt days in ASCAT and RCMs areas above the 2007-2020 maximum snowline elevation (Fig. 4d). Red areas correspond to more melt days in ASCAT on average and blue areas correspond to more melt days in the RCM on average. Melt in ASCAT is defined as Label ST-2A and ST-2B."

Technical corrections

- L109 ACMO2.3p à RACMO2.3p2;

We have now corrected accordingly.

- L111 2x in a row "On the lateral boundary,";

We have now corrected accordingly.

- L115 2x "." in a row;

We have now corrected accordingly.

- L144-145 : 2 times in a row : "the first and second";

We have now corrected accordingly.

- Caption of Table 1: There is something wrong in this sentence: "Melting thresholds for the different RCMs based on in situ PROMICE AWS observations of 2m temperature and mean air temperature for August and July simulated by the RCMs at AWS stations and observed by the AWS stations using a lapse rate correction." I think you need to remove 'and observed by the AWS stations'.

Agree, we revised accordingly. See above response in the section about methods.

- L232: close the bracket here: "(Fig. 6.";

We have now corrected accordingly.

- L357: HIMHAM5 dataàHIRHAM5 data.

We have now corrected accordingly.

References

Dethinne, T., Glaude, Q., Picard, G., Kittel, C., Alexander, P., Orban, A., & Fettweis, X. (2023). Sensitivity of the MAR regional climate model snowpack to the parameterization of the assimilation of satellite-derived wet-snow masks on the Antarctic Peninsula. The Cryosphere, 17(10), 4267-4288.

Picard, G., Leduc-Leballeur, M., Banwell, A. F., Brucker, L., & Macelloni, G. (2022). The sensitivity of satellite microwave observations to liquid water in the Antarctic snowpack. The Cryosphere, 16(12), 5061-5083.