We are very grateful for the constructive comments and suggestions provided by the reviewer that have significantly improved our manuscript. We agree that including the entire Barents sea area would have been interesting, but since the paper was already very long, we have saved that comparison for another manuscript. We have included all suggestions for changes and have outlined the major comments below.

The reviewer comments appear in bold, our answers in normal font, and changes to the manuscript in italics.

L9: Might be better to use CMB instead as capital letters are more easy to recognize as an abbreviation

That is a good point, we have changed this in the text and figures.

L14: It should not be forgotten that CARRA has its uncertainties as well, probably of similar magnitude AROME-ARCTIC, as both are using some climate model constrained with (largely the same) sets of observational data. Just saying that small differences between the two datasets do not necessarily mean good performance of one of the two

Good point, we have made it clearer in that this only means they will provide similar quality predictions:

L17-19: This indicates that AROME-ARCTIC may provide similar high-quality predictions of the total mass balance of Svalbard as CARRA, but regional uncertainties should be taken into consideration.

L68: Why is this mentioned here? It almost gives me the impression that the Russian Arctic mass balance will be presented as well in this study, but that does not seem to be the case.

The idea was to first test the model for Svalbard, where many observations are available, and show that the model setup works well in the Barents Sea area. But you are right, it is misleading, and we have removed the section now.

L89: It could be good to modify figure 1 so that it also includes geographic names that are used in this study (oceans, places and regions in Svalbard etc). For example, many readers will not know the difference between Spitsbergen and Svalbard.

Good point, geographic names of oceans, places and regions used have been added to Figure 1.

L104: Could it be summarized here in a few sentences?

We have added the following description:

First, both the CARRA reanalysis and AROME-ARCTIC forecasts are evaluated against available observations from automatic weather stations (AWSs). Unsurprisingly, both products performed well when compared to AWS data which had been assimilated into the forcing products but had larger biases when compared to glacier measurements which had not been assimilated. The comparison of AROME-ARCTIC and CARRA at the AWS locations show that both products were similar, albeit with larger biases and root-mean-square-errors for AROME-ARCTIC. In addition, the consistency between the two forcings is evaluated for the overlap period (2016-2021). We found that AROME-ARCTIC is on average colder than CARRA, particularly in NW Spitsbergen where the average yearly temperature was -2◦C colder in AROME-ARCTIC. The full results of this analysis are described in Supplement S2.

L105: It could be worth mentioning the maximum depth of the subsurface model and the vertical resolution already here.

The following sentence has been added here to the text:

L119-20: The model is initialised with 47 layers of ice with a thickness between 0.1 and 1 m, totalling 20 m of glacier ice.

L118: I support this approach. Still, it is good to realize (and possibly discuss) that using the same initial conditions will probably reduce differences in calculated mass balance with both forcings

This is a good point; it reduces the differences (which is also one of the reasons we use the same initialisation). We have added the following lines to make this clearer to the reader:

L133-34: This most likely will reduce the difference in CMB calculated using the two products, compared to if different spin-ups were produced.

L126: Could be good to specify the "final results" here. I suppose runoff and snow depth can simply be weighted. But what is done with glacier-specific variables like cmb?

We do not add the land and glacier components together (although it could be done for runoff like you said). What is meant is that to calculate the average or sum of a variable for glaciers, we do a weighted average/sum based on the glaciated area in a point. Similar for non-glaciated points, when we are calculating the total value for all of Svalbard, a weighted average or sum is done based on the area-fraction. We have clarified this in the text as:

To calculate the average or sum of a variable for a specific region or all of Svalbard, the results are weighted based on the fractional glacier coverage.

Figure 1: It would be nice with a bit larger and detailed map, with some place names. Also elevation contours could be useful

The figure has been made larger and elevation contours and place names mentioned in the text have been added.

L158: Maybe it could be explained in this paragraph how 2.5x2.5 km results are compared to point observations. What was done when several mass balance observation points fell in one grid cell of the model?

The following sentence has been added to explain this:

L175-77: When several observation points fall within one 2.5 x 2.5 km model grid, only the measurement point closest to the center of the gridpoint is used.

L183: Has CryoGrid been calibrated in any way? Have observations been used to optimize uncertain parameters in the energy balance and snow routines? In case this is described in Westermann et al. (2022) please indicate that here.

The only calibration that has been done against observations is using the mass balance and albedo observations. Most values are taken from other studies, and not many variables needed to be altered from the standard (we e.g. tested the snow/rain temperature threshold, but the best results were found

using the CryoGrid default of 2 degrees). We did calibrate the ice albedo and found that 0.4 provided the best results. This has been added to the text:

L222-24: Previous mass balance studies of Svalbard have used ice albedo values in the range of 0.3-0.4 (Østby et al, 2017; van Pelt et al, 2019) for all of Svalbard. From calibration with available mass balance and albedo observations, we found the best results using an ice albedo of 0.4.

L233: any reason for using snow water equivalent rather than snow thickness as a threshold? I am just curious because it is mostly the (minimum) snow depth of layers that affects numerical stability, e.g. of the heat diffusion equation, rather than minimum snow mass.

The stability is affected by both the minimum snow depth and the heat capacity: if the heat capacity is low, the same energy input will lead to a stronger temperature change which leads to instabilities. Using the SWE as a threshold captures both these sources of instability – it keeps the grid cells from getting too thin, and it keeps the heat capacity from fluctuating.

L237: How is the fresh snow density set? Is it constant or variable?

The fresh snow density is variable and depends on the air temperature and wind speed. The equation is given in the supplemental material (Eq S5).

L278: Nordenskiöldbreen is tricky! Especially because of high wind speeds and snow drift at low elevations, whereas higher elevations have much calmer conditions. This generates a very strong accumulation - elevation gradient.

Good point! We have added a description of this to explain the higher biases for this glacier.

L293: And over what periods are summer and winter balance defined? Maybe I missed it...

The summer balance is defined as from April 1- August 31, and the winter balance is from September 1st – March 31st. We have added this to the text:

L351-52: For calculations of the winter and summer mass balance, we use fixed dates of 1 April and 1 September.

Figure 6: It is interesting that AROME-ARCTIC simultaneously has a more positive summer balance and a more negative winter balance. It would have been more likely that a negative winter balance anomaly would also give a negative summer balance anomaly since less snow in winter usually means more melt in summer. Why is this not the case here? Maybe other weather variables, e.g. cloud cover, provide an explanation? It would be interesting to add a brief discussion on this in the manuscript.

This is indeed different than what you normally expect. AROME-ARCTIC generally has lower temperatures than CARRA, also in the summer, which of course leads to less melting. In addition, although the precipitation over the year is very similar in the two products, there is a bit less precipitation in the winter in AROME-ARCTIC and a bit more in the summer. This is currently discussed in the Supplement S2. A combination of lower temperatures and slightly different precipitation patterns is probably what causes the underestimation of both the summer and the winter balance.

L314: Please be consistent with the units (m w.e. or m w.e. yr-1)

We have changed the units to always be m w.e. yr-1

L328: Was this always the case? I suppose in cold years this may not be true everywhere. I can also imagine that there may be points in Svalbard for which the model simulates a positive mass balance, but which are not part of a glacier (just because of model uncertainty). It could be good to mention somewhere how such points were dealt with (in case they occurred) when calculating average snow onset / disapearance dates

This section was removed at the request of another reviewer. But this is true, these points did occur – both due to model uncertainty and during cold years. This is only an issue during a few years, and on average only affects 4% of the land grid points. However, in extreme years this can be up to 30%. To avoid this, the model removes all perennial snow on September 1st, and thus for the calculation of snow disappearance day, this value is used.

L356: Could it also play a role that the AROME-ARCTIC simulation (presumably) starts from subsurface conditions that were initialized with the CARRA forcing?

Yes, this could definitely also be a factor.

L364: The fact that there are no points with 0 runoff may imply that there is no deep firn in the model results that is still cold (<0 deg C). Is that indeed the case?

Yes, this is correct. The large melt years of 2013 and 2020 caused the deep firn to heat up and become temperate. This is also partly related to the spin-up – the period we use for spin-up is relatively warm, which means that the firn likely is too warm in the beginning of our simulations, and thus heat up too fast.

L367: The time-series in Fig. 10 seem to agree really well with what was presented in Van Pelt et al. (2019) in a similar figure. Both the absolute values and year-to-year variations match very well.

This is true, they do match up very well! We also mention this in section 6.4.

L392: Right now only a range in calving rates is given for 2005-2009. To be consistent it would be good to indicate a similar range for all presented periods. Furthermore, it could be mentioned that high calving since 2010 is likely the result of the surge of Basin-3.

We have added the ranges for all presented periods, and have added the following sentence about Basin-3:

L456-57: The calving after 2010 is likely increased due to the surge of Basin-3, the largest outlet basin of the Austfonna ice cap, which significantly increased the calving from the ice cap (e.g. Dunse et al, 2015).

L394: Please note that also CARRA reanalysis data will from 2023 onwards be updated on a monthly basis (https://climate.copernicus.eu/copernicus-arctic-regional-reanalysis-service). That could be sufficiently "real-time" for many applications.

This is true, we have added a line mentioning this in the conclusion:

L575-76: For many applications, however, using CARRA forcing may soon be enough, as it will in the future be updated on a monthly basis.

L400-408: I am not fully sure what the main point here is. Maybe it could be clarified. It does nicely show, including error bars, that 2021/22 was indeed the most negative mass balance year since (at least) 1991. Maybe that could be highlighted.

We have now removed this section, and instead it is a part of a specific results section on AROME-ARCTIC (section 5.4). We now highlight that 2021/22 was the most negative mass balance year more clearly in the text.

L442-44: Based on AROME-ARCTIC, 2021/22 is a record negative mass balance year for Svalbard, with a CMB of -0.86 m w.e. yr^{-1}. There is a highly negative mass balance in all regions in Svalbard. The runoff from glaciers is also the highest over the simulation period, of 1.6 m w.e. yr^{-1} (58Gt yr^{-1})

L423: Does this apply to only the snow and firn parameters or also the energy balance parameters?

This is referring to the snow/firn parameters in the Vionnet/Royet paper. We have clarified this in the text:

L468-69: However, most of the model parameters used by the snow and firn scheme are based on recommendations from previous studies and have not been tuned for the conditions of Svalbard.

L461: Please note that the effects of ignoring elevation and mask change were also investigated for a future projection for Svalbard in Van Pelt et al. (2021; https://doi.org/10.1017/jog.2021.2). See Fig. 13 and the related discussion in that study. It is found that the elevation and outline changes have counteracting mass balance effects that are approximately in balance.

Thanks for pointing this out, it very relevant for the discussion. We have added the following to the text:

L527-30: In addition, van Pelt et al (2021) investigated the effect of ignoring both elevation and glacier mask changes on future projections for Svalbard from 2018-2060. Over this time period, the authors found that the increased melt due to a lowering of the glacier surface was nearly balanced by the melt reduction due to a changing glacier mask, and thus the introduced error in the runoff and CMB was small.