We are very grateful for the constructive comments and suggestions provided by the reviewer that have significantly improved our manuscript. We agree that the manuscript would also have fitted well within a journal like ESSD, but unfortunately they do not publish outputs from climate models or land surface models. We have restructured the introduction based on your helpful suggestions, and the research aim of the paper is now hopefully clear.

In general, we have included all suggestions for changes and have outlined our answers to all comments below. The reviewer comments appear in bold, our answers in normal font, and changes to the manuscript in italics.

Major comments

The introduction does a good job in describing the scientific literature, yet as mentionned above it could be more clear in describing the scientific challenge, and also more clear in explaining why using CryoGrid forced by CARRA offers a solution to the scientific challenge.

Thank you very much for your helpful comment! We agree that the scientific challenges should be mentioned more clearly. We have now restructured the introduction following your above suggestions as:

- Motivation: accurate estimates of the past evolution of mass balance and runoff
- Obstacle: global reanalysis products are too coarse, and statistical downscaling does not resolve all physical processes
- Solution: regional high-analysis productions which assimilate local observations, such as CARRA or AROME-ARCTIC. We evaluate these products for use in mass balance simulations and use the output to investigate the climatic changes in Svalbard over the last 3 decades.

It is not clear why AROME-ARCTIC is also used as a forcing for CryoGrid. It seems that CARRA and AROME-ARCTIC differ quite significantly, which is interesting yet not very surprising given the fact that CARRA assimilates some observations and that AROME is a weather forecast model (as mentioned by the authors in the conclusion as well). Regional difference between the two products may average out, but this still reduces the applicability/usefulness of a near-real time forecast of mass balance. Perhaps the authors could more clearly motivate the choice of using AROME-ARCTIC, or just briefly mention this possibly in the discussion and then remove section 6.2.

We have tried to motivate this choice more in the introduction:

L69-72: In addition, we investigate if the forecast product AROME-ARCTIC, which uses the same model and similar observations as CARRA, can be used to extend the CARRA product, thus providing almost real-time updates of the mass balance and runoff. Almost real time-simulations could provide valuable information for e.g. fieldwork planning (to check the current conditions) and public outreach.

In addition, we have tried to have less focus on AROME-ARCTIC in the paper and more on the CARRA simulations. The AROME-ARCTIC simulations are now mostly described in a separate section in the results section (Section 5.4), and section 6.2 has been removed.

How did the authors compute the temperature and relative humidity from automatic weather

stations at 2 m, and wind speed at 10 m ? These variables are often measured at some height above the surface that changes between each maintenance, and also due to snowfall. Because of the large vertical gradients on a glacier, the deviation of T,WS and q in using a wrong sensor height can be quite large, which could make the evaluation with CARRA or ARMOE-ARCTIC inaccurate.

AWS on glaciers are set-up as so-called floating stations, i.e. they are based on a tripod standing on the ice surface, rather than fixed to a mast drilled into the ice. That way, sensors keep constant height above the ice surface during the snow-free period, as the station moves together with the ice surface during melting. During the snow cover period, the height of sensors is reduced by the snow height. Snow depths on Svalbard are modest and seldom amount to more than 1 m at AWS sites. We have not attempted to correct AWS data for this effect for two reasons: i) the largest effects of incorrect T,WS and q are on the calculation of turbulent fluxes which in turn have moderate significance for melting that predominantly occurs when the sensor height problem is smallest/ does not exist; ii) variations in sensor height due to snow cover do equally occur at AWS outside glaciers; where this typically is not either taken into account.

Minor comments

Figure 1: Perhaps it would be useful for the non-expert reader to add the locations of each glacier/region mentioned in the text (Etonbreen, Austre Brøggerbreen, etc ...). It would also be very useful to add the coordinates and elevation of each AWS, or to refer to studies where this information is available.

This is a good point; we have names of regions and glaciers mentioned in the text. We have also added the coordinates and elevations of each AWS in Table 1. Studies where this information is available are also mentioned in the text.

L69-71, It is not clear why the Russian Arctic is mentioned here since it does not re-appearin the methods or in the results. This interesting idea should be better mentioned in the discussion.

This has been deleted from the introduction and is instead only mentioned at the end of the conclusion as an outlook to where this method could be used in the future.

Lines L83-84 can be made more consistent with L27-28.

The sentence has been changed to be more consistent with L27-28:

L91-93: Located in the Norwegian Arctic between 75 and 81°N, the Svalbard archipelago is in one of the currently fastest warming regions in the world, the Barents Sea region (Screen and Simmonds, 2010; Lind et al, 2018}, and has the strongest observed warming in Europe since the 1960's (Nordli et al, 2014),

L186: Please clarify what modules of CryoGrid are part of Westermann et al (2022), and what modules have been specifically added for this study. Also, please explain how the 'water percolation and runoff modules' (4.2.1) differ from the original CryoGrid setup.

We have now made this clearer in the text in several locations:

L220-21: The glacier module consists of layers of pure-ice with a user-defined constant ice thickness. This module has not been altered compared to the one described in Westermann et al (2022)

L231-38: *These modules have been specifically added to the model for this study.*

The snow and firn modules follow a slightly altered CROCUS (Vionnet et al, 2012) snow scheme as described in Westermann et al (2023). Some of the main differences to the snow schemes presented in Westermann et al (2023) are:

- *Additional output variables, including refreezing, internal accumulation, CMB, SMB.*
- *Updated water percolation and runoff scheme, including a parameterisation for the hydraulic conductivity and a runoff timescale (described in Section 4.2.1)*
- *Regridding of layers below the surface (described in Section 4.2.2)*

L246-47: In Westermann et al (2023), a constant user-defined hydraulic conductivity is used. Here, the hydraulic conductivity is parameterized in terms of […]

L192: Why was a constant bare ice albedo of 0.4 chosen ? This seems very simplistic yet it appears that the errors in albedo are very small nonetheless (Table S2.).

We decided to use an ice albedo of 0.4, as this gave the best results when calibrated against observations of mass balance/albedo. We have now added a sentence about the calibration:

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.*

Previous studies have used a constant ice albedo ranging from 0.3 to 0.4 (e.g. Østby et al, 2017; van Pelt et al, 2019) with good results, but this is of course a simplification, as the ice albedo varies from 0.15 to 0.44 across Svalbard (Greuell et al, 2007). Using albedo observations, e.g., from MODIS, to create an ice albedo map would probably give better results and could be included in the model in the future. We have added a few sentences on this in the discussion:

L471-74: In addition, we use a constant ice albedo in the model, which could be a major simplification given that the ice albedo varies across Svalbard from 0.15 to 0.44 (Greuell et al, 2007). In future work, this could be improved by using estimates of the ice albedo from e.g. MODIS observations to create a map of the ice albedo (Schmidt et al, 2017) and/or updating the albedo parameterisation to account for dust and impurity content.

L205 I believe that this is commonly called a "bucket scheme". Then the variable θf c in Eq (1) is the irreducible water content (or " maximum liquid water-holding capacity " as definedin CROCUS by Vionnet et al 2012.), which is not necessarily the same as the field capacity(typically defined for soils, as far as I know).

Thank you for noticing this, we have changed “field capacity” to “irreducible water content”