Reviewer 1 – round 2

We thank the reviewer for their careful reading of the manuscript, and for noticing the problem with the different units. We have answered all comments below. The reviewer comments appear in bold, our answers in normal font, and changes to the manuscript in italics.

- the problem of using different units (m W.E or GT) remains. For example, in the abstract (lines 10-15, in the authors_tracked_changes_file) SMB is given in mWE, runoff in GT and refreezing in m WE. All the SMB components should be listed in the same units as SMB is the sum of them. We usually use GT as unit for integrated value of SMB and TMB but I'm OK if the authors prefer to use m WE/yr bt in this case, it should be also the case for runoff.

Thanks for noticing this, you are right that we did not change the units in the abstract and conclusion. This has now been fixed.

- line 109: what if the difference between CMB (climatic mass balance), total mass balance and SMB (surface mass balance). It is the 1st time that I see a mention of CMB. We usually use total mass balance (TMB) for SMB + ice dynamics mass balance.

As mentioned in the first review, we are following the recommended terminology of climatic mass balance, surface mass balance, and total mass balance as given by Cogley et al (2011). Previous papers on Svalbard have made a similar distinction (e.g. Østby et al (2017), van Pelt et al (2019 and 2021), Schuler et al (2020)).

The surface mass balance (SMB) quantifies the mass fluxes between the atmosphere and the glacier at the surface, and within the current year's snow layer (refreezing within the annual layer). This is what is measured by in-situ glaciological observations.

The climatic mass balance (CMB) additionally accounts for mass changes below the last summer surface (in the firn layer) and can therefore be simulated with a model like the one used in this study.

Total glacier mass balance is the sum of CMB, basal mass balance and frontal ablation (i.e., subaqueous melting and calving).

We had already described this difference in the introduction in the revised manuscript:

L82-88: CryoGrid simulates both the surface mass balance (SMB) and the climatic mass balance (CMB). The surface mass balance quantifies the mass fluxes between the atmosphere and the glacier at the surface, as well as refreezing within the annual layer. The SMB is what is measured by in-situ glaciological observations. The climatic mass balance additionally accounts for mass changes below the last summer surface. For tidewater glaciers, CryoGrid, however, cannot calculate the total glacier mass balance, as this is the sum of CMB, basal mass balance and frontal ablation, and the latter cannot be determined from an energy-balance model.

But have tried to clarify further by changing the text slightly to the following:

L82-88: CryoGrid simulates both the surface mass balance (SMB) and the climatic mass balance (CMB). The SMB quantifies the mass fluxes between the atmosphere and the glacier at the surface, as well as *refreezing within the annual layer. The SMB is what is measured by in-situ glaciological observations. The CMB additionally accounts for mass changes below the last summer surface, e.g. in the deeper firn layers. The total mass balance, the sum of CMB, basal mass balance and frontal ablation, cannot be calculated for tidewater glaciers by an energy-balance model like CryoGrid, as glacier dynamics are not included. This terminology follows that suggested by Cogley et al. (2011).*

- line 214, for future validation, when several observation points fall within one 2.5 x 2.5 km model grid, I suggest rather to use the one the closest in altitude of the model or the average of all of them.

Thanks for the suggestion, using the altitude is indeed better, particularly when there are differences between the modeled and measured altitude. In these cases, however, the point closest to the midpoint did correspond well to the model elevation.

- line 312: what is the value used for the irreducible water content Wex?

The irreducible water saturation is 0.05, following Vionnet et al (2012), and Wex is thus 0.05*porespace. We have clarified this in the text:

L294-295: *The irreducible water saturation is 0.05, following Vionnet et al. (2012), and the irreducible water content is thus 5% of the total porespace.*

- Fig 9 showing times series in m WE and in GT is not relevant here. Mapped value should be in mmWE/yr or mWE/yr and integrated value in GT/yr or everything in mWE/yr
We think it is relevant to also show the integrated values in both GT/yr and mWE/yr - in mWE/yr to be consistent with the CMB and refreezing, as you have previously pointed out, and in GT/yr so it is easier to compare with previous studies which provides similar figures in Gt/yr (e.g. van pelt et al, 2019)

- in Table3 adding the 2016-2021 statistics of both CARRA and AROME models will be relevant.
Table 3 shows comparison to previous studies, which at maximum includes results until 2018. Comparing to CARRA and AROME simulations would only be for 2016-2018, and not add much to the comparison. If the reviewer means table S3, we already have the statistics for both models in table S4.

- lines 1673-283: there is the same problem with the units. SMB is given in mWE and runoff in GT/yr
Thank you for noticing this, we have changed the runoff to mWE/yr.

- line 700: this sentence should be in the “data availability” section (line 709

The sentence has been moved as suggested

Reviewer 3 – round 2

We thank the reviewer for reading the manuscript again and their helpful comments. Sorry that we missed some of the comments last round, we have now answered the ones we missed.

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

The authors have addressed all my three major comments from the previous round of revisions, but have not responded to the last 4 (out of 10) minor comments. These comments are meant to better understand the methods and possibly provide the interested reader with some additional discussion points. In particular I believe it would be very helpful for future work to further quantify and discuss the errors of the model in part 5.2.1, especially since the main goal of this research is to have accurate estimates of surface mass balance.

I am sorry; I must somehow have missed the last of your minor comments when I copied everything into a word document for the replies. We have answered all the comments below, including those we missed last round:

Also, if ”Water is not allowed to flow into an impermeable layer, here defined as layers with a density higher than 830 kgm3” (L219), it is interesting and perhaps counter-intuitive to see in the results that ”the average annual internal accumulation is 0.11 mw.e., and thus accounts for almost half of the total refreezing (Fig. 9c)”. Perhaps the authors could comment on this in the discussion ?

Due to the re-gridding of model layers, it is quite rare that an impermeable layer occurs high up in the snow pack (or if it does, it often disappears as layers are merged). Impermeable layers mostly occur in the deep firn in the model, and thus refreezing can occur below the yearly layer, leading to a high internal accumulation. For model simulations where more thin layers are used than here, we would allow water to penetrate thin layers as it is unlikely that e.g. a 1 mm layer would completely stop water percolation.

L209: Consider replacing ”water saturation” by ” volumetric water content in the snow” to be consistent with other snow models.

Changed

L225 How is the ”water in excess of the field capacity” defined ?

The irreducible water saturation is set to 0.05, following Vionnet et al (2012). The water in excess of the field capacity is thus any water quantity exceeding 0.05*total_porespace. We have now added the following sentence to clarify this:

L 294-295*: The irreducible water saturation is 0.05, following Vionnet et al. (2012), and the irreducible water content is thus 5% of the total porespace.*

L256-259. I acknowledge that the manuscript will be a bit long if the AWS evaluation is further detailed here. Yet I find the statement ”both models generally fit well with observations” a bit simplistic, not very specific and possibly also inaccurate. It would be very useful for future studies to better understand what causes the limitations of a SEB model forced by CARRA. For instance it seems

that summer ablation is overestimated at Nordenski¨oldbreen (figure 4, red lines), which could partly be explained by an underestimated albedo by 0.06 (Table S1). Also, it appears that there is a systematic underestimation in both incoming and outgoing longwave radiation components at all glacier AWS

We have now added the following text about the evaluation of the CARRA forcing:

L329-344: The comparison of the CARRA forcing against observations from automatic weather stations shows a general good agreement. The MET Norway stations have been assimilated into the CARRA product, and it is therefore not surprising that there is a good agreement between the two. The largest differences in temperature are found for the Sveagruve II station (ΔT = −1.8C), but for most of the MET Norway stations the mean temperature difference is below 1C. The largest differences in relative humidity and wind speed are found at Kvitøya (ΔRH = 6.4%) and Pyramiden (ΔWS = −1.9 m s−1), respectively.

The 2m temperature at the glacier stations, which were not assimilated into the CARRA product, is generally well represented, with biases generally smaller than 1◦C. The exception is at the Etonbreen AWS, where CARRA has a cold bias. This can, however, partly be attributed to a warm bias in the AWS observations over time at this station due to sensor drift, before redundancy has been installed in 2016. The relative humidity has a maximum bias of 6.2%, while the wind speed bias ranges between -1.3 and 1.5m s−1. The incoming longwave and shortwave radiation in CARRA generally fits well with the observations, albeit with a small negative bias in the longwave radiation for most of the stations (ranging between -1.6 and -14W m−2).

The evaluation of both forcing products against available AWS observations shows that the two products often provide similar results, but that the bias and root-mean-square-error of the CARRA product is generally smaller than for AROME-ARCTIC. For detailed evaluation of the model forcing against available AWS observations for both CARRA and AROME- ARCTIC, in addition to a discussion on the inter-comparison, we refer to Supplement S2, S3.

Regarding Nordenskiöldbreen, this is generally tricky to simulate. This is especially because of high wind speeds and snow drift at low elevations, whereas higher elevations have much calmer conditions. This created a very strong accumulation – elevation gradient, which affects e.g. the albedo and thus the ablation. We added the following lines in the manuscript describing this during the last review round:

L366-368: Nordenskiöldbreen experiences a very strong accumulation-elevation gradient, due to high wind speeds and snow drift at lower elevations and calmer conditions at higher elevations. It is therefore difficult to accurately simulate this glacier without including snow re-distribution between grid points.

Regarding my 2nd major comment; it would still be useful for future evaluations, and technically more correct, to mention at L209-210 from the tracked changes document that the temperature and wind speed are not always measured at 2m and 10m respectively, especially not on the glaciers. For the IMAU AWS at Nordenskiöldbreen and Ulvebreen the sensor boom is located between 3m and 4m depending on the tripod design. The errors in observed q, T and WS should not affect the modelled turbulent fluxes since CARRA is used as forcing. The errors will however affect the statistics of the evaluation presented in the supplementary material.

You are right that we should mention how we deal with different sensor heights. We have now mentioned how we handle that the measurements are not at exactly two or ten meters in the text. For windspeed, we assume neutral stratification and a roughness length of 1 mm. For temperature and relative humidity, we use the temperature gradient between two model layers.:

L176-182: When available, daily mean observations of the 2m temperature, 2m relative humidity, 10m wind speed, and incoming and outgoing longwave and shortwave radiation is used for the evaluation. When windspeed is only available below 10 m, as is the case for most of the glacier stations, the windspeed at 10 m is calculated using a logarithmic wind profile (assuming neutral stratification) with a roughness length of 1 mm. The assumption of neutral stratification, however, is a limitation, potentially having larger impact on the wind speed correction than sensor level alone. For Nordenskiöldbreen and Ulvebreen, measurements were conducted at ~4m above the surface, and the CARRA humidity and temperature is therefore interpolated to the measurement height by interpolation between the lowest model level (15 m) and 2 m.

It is difficult to take into account the effect of snow accumulation, as many of the stations used do not measure this. We have instead added a section on the uncertainty introduced by adding 1 m of snow below the sensors, based on downscaling CARRA data and assuming a log-log profile of wind speeds:

L187-200: The snowdepth is not measured at the majority of the used stations, and we therefore do not apply any correction factor due to changes in height after snow accumulation. The uncertainty associated with ignoring this effect depends on the specific variable (temperature, humidity, windspeed) and the measurement height. These uncertainties only affect the evaluation statistics, and not the model results.

Snow depths on Svalbard are modest and seldom amount to more than 1 m at most AWS sites. Assuming a snow depth of 1 m, a roughness length of snow of 1 mm, and that the windspeed can be approximated by a logarithmic profile (neutral stratification), the windspeed at 1 m above the surface is 7% lower than the windspeed at 2 m. For windspeeds measured at 10 m, decreasing the height by 1 m only amounts to a 1% decrease in windspeed. The windspeeds measured at the MET Norway stations and Kongsvegen, which are measured at 10 m, are therefore more robust to the effect of snow accumulation. The study by Østby et al. (2013) suggests a roughness length smaller than 1 mm which in turn would decrease the effect on wind speed.

It is trickier to estimate the uncertainties for temperature and relative humidity. Here, we use CARRA estimates of the temperature, pressure, wind speed, and humidity at the lowest model level (15 m) and at surface level (0 m) to interpolate the temperature and specific humidity, taking into account the stability of the atmosphere. The same method and parameters are used within CARRA to calculate variables at 2m height and is described in detail in the CARRA product user guide (Schyberg et al, 2020). The difference in temperature and humidity for all station locations is simulated for 2 m and 1 m above the surface over two different years (1994, a low melt year, and 2020, a high melt year). Even assuming the snowpack lasted the full year, the yearly average deviation was < 0.2◦C. The specific humidity at surface level was not available in CARRA, so for simplicity we assume fully saturated conditions. The yearly average difference in the results was always below 1%.