Response to the referee 1

Thank you very much for a helpful and constructive review. Please see below for a point-by-point response on your corrections and suggestions (with our replies marked with blue colour and indentation).

The paper is well written and very timely with the recent field work (MOSAiC) that has been undertaken within the Arctic. The paper provides insight into deficiencies with the global reanalysis product for the region, ERA5, and highlights where the regional analysis can provide improved estimates of the surface properties of the sea ice. This is especially useful resource for the modelling community when trying to improve their models of the region. The conclusions provide a good summary of the results and the limitations of the current system which are important for users of this reanalysis. It also gives insight into how future systems can be improved in the future. I think the paper is close to being ready for publication but have a few queries about some details that I think may help to clarify what has been done.

I have very few detailed comments for the authors but a general question I had that made me hesitant to accept as is, was based on understanding the impacts of: the production streams for the reanalysis and the boundary conditions on the results shown.

Two questions come to mind when considering the East and West portions of the reanalysis with multiple production streams and different boundaries. One element is the spin up when you start the production and the impact that has. If I understand correctly, the production streams are not initialised on the same dates for the two different regions, and I wonder whether this provides additional information (or introduces errors in determining the state) in the regions of geographical overlap. The second element is that within the regions of overlap whether particular areas/cells are more constrained by the boundary conditions imposed at the edges of the domain and whether you also see some signature of this comparing similar geographic points in the two analyses. Is this something you looked at all? Can it help provide insight into the performance/constraints of the system? I wondered if when the ice extent was smallest in late summer early autumn and then closer to the edges of the analysis domain is this why the ERA5 and CARRA systems are more similar or if it is a function of the physical processes governing the errors?

> Indeed, due to timeliness constraints the western CARRA domain used more production streams than the eastern one. As a consequence, there are two extra production streams with different initialisation dates. Apart from those two, productions streams in the two CARRA domain were initialised at the same dates. As we mention in our reply to the next comment, across-stream discrepancies are small for all the variables except ice thickness over perennial ice. Therefore, taking into account that ice thickness is a rather peculiar variable in the CARRA product that should be used with great care even outside the stream transition periods, we believe that the two extra production streams in the western CARRA model domain do not significantly affect the representation of sea ice state in the reanalysis.
>
> As for the impact of boundary effects on the overlapping region and in general, we have not investigated it in detail when preparing the original version of the manuscript. It is true that boundary effects in CARRA are not negligible: firstly, due to the coupling strategy applied in the CARRA system, which does not use hydrometeors from the host model and thus requires some spin up to build up cloud water and ice; secondly, near-surface temperatures over sea ice in ERA5 are generally warmer than in CARRA (in winter time) which also introduces a discrepancy. Assuming that snow depth over sea ice is a variable the most affected by the boundary effects (since it directly depends on model precipitation), and following referee's

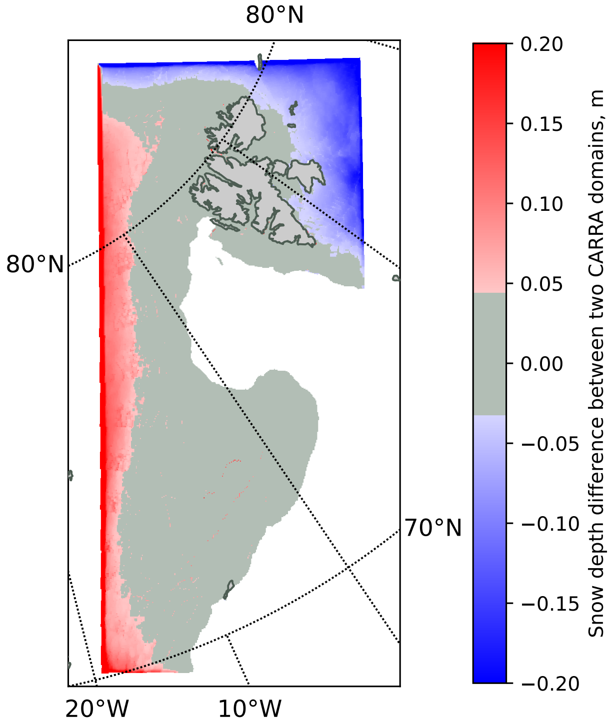

Figure 1: Average snow depth difference between western and eastern CARRA model domains as of 1 April, computed throughout the period from 1990 to 2021. Positive values mean that western CARRA domain has thicker snow layer than eastern CARRA domain, and vice versa. Means falling within the interquartile range of snow depth differences are masked out.

suggestion, we assessed them by comparing spring-time snow depths in both CARRA domains over the region of geographical overlap. Figure 1 shows consistent differences in the snow depth and indicates areas of the domains where boundary effects are more pronounced. We believe this information would be of considerable interest to readers and users of the CARRA product and we will provide it in the updated version of the manuscript.

Concerning the late-summer/early-autumn performance of the CARRA system, when differences between the regional reanalysis and ERA5 are smallest, we attribute it to sea ice being in a more constrained state in general rather than to boundary effects. At that time snow cover over sea ice in CARRA has already melted and in absence of the snow layer the sea ice schemes of CARRA and ERA5 become quite similar. Moreover, ice temperature is essentially a bounded variable, and summer ice surface has temperature close to melting point with much less variability compared to cold winter-time ice cover, which further reduces differences between CARRA and ERA5. Similarly, for the ice surface albedo fields, CARRA and ERA5 have less discrepancy over warm summer-time snow-free ice.

In lines 118-119 you talk about the spin up of the snow loading and later in figure 7 you highlight the different production streams used for the E and W components of the reanalysis. Do you have overlapping time periods for the streams that you can compare to determine if the system is spun up? Do you assess the difference where there are overlapping geographical points between the West and East analysis areas given that the production streams restart more frequently in the West - is that also a way of determining if there is any significant impact?. if you have determined that the model has "spun up" within a year please add this to the text. Do the different start dates in the production between the W and E regions means that you need to do more complex averaging e.g. weighted mean for the points which are both in the E and W regions for zones B,C or D when

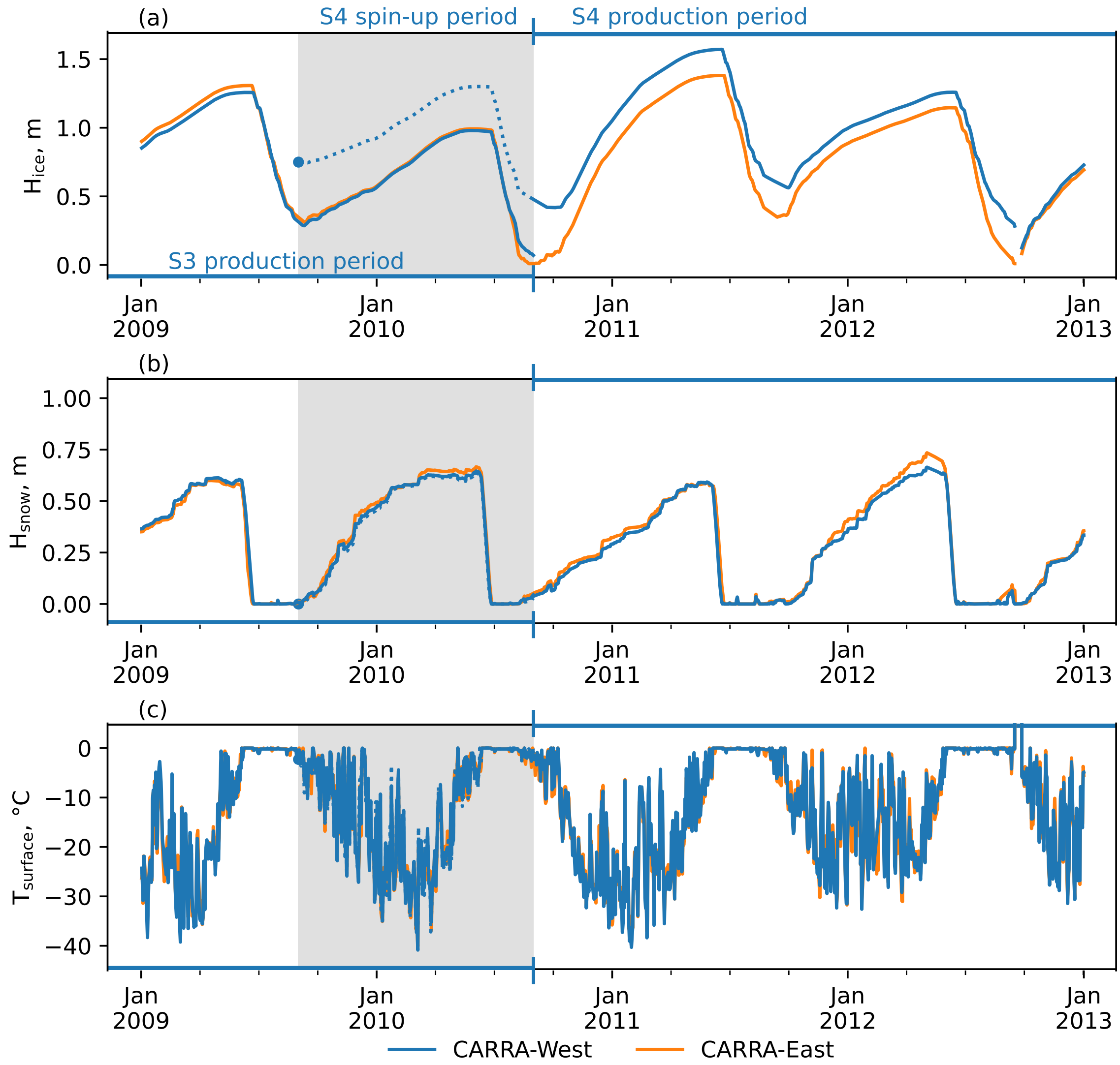

Figure 2: Evolution of the sea ice state in the CARRA system over a transition period between production streams for a single selected point (82.4°N, 16°E) within the region of geographical overlap. The stream transition occurred for the western CARRA model domain, all values extracted for the eastern model domain come from the same production stream. (a) Ice thickness; (b) snow depth; (c) ice surface temperature.

calculating trends etc. Have you assessed this spin up impact on sea ice thickness as well? Are the production streams always started at the same point in the year (eg. Jan 1st)? Could this have any effect on the climatology estimates shown in figure 8? I wondered whether it might be helpful to expand on this in section 3.1 or in the results section.

> The released CARRA data set does not include model output from the spin up parts of production stream and therefore provides no overlapping periods for a selected CARRA model domain. Thus, in the present study we treat the CARRA product as a flat data set and do not try to compensate the differences in productions streams of two CARRA domains by applying weighted averaging over the region of model domain overlap. For most of the sea ice variables discussed in our paper the across-stream difference is rather small since these variables are either 'fast' and considerably influenced by the atmospheric forcing or have a pronounced annual cycle and disappear by the end of summer (like snow cover, or annual ice cover). The only variable which can show a consistently high across-stream discrepancy is the ice thickness, for cases when ice persists for multiple years in a grid cell. Typical performance of the CARRA system across the production streams is shown in Fig. 2. The figure might suggest, that a one

year spin up period used in CARRA production is too short for properly initializing perennial ice cover within the model domain. However, this claim is difficult to validate due to several factors: firstly, sea ice scheme of CARRA is not constrained by observations and may start drifting after years of integration in a production stream, which would introduce across-stream discrepancy even if the spin up period yielded perfect initial state of the ice cover; secondly, the overlapping part of the CARRA model domains, which can be used to assess the spin up effects contains only a minor amount of perennial ice close to the domain boundary, where boundary effects are the most pronounced. Therefore, we believe that in these circumstances, and taking into account limited computational resources available and time constraints of the production phase of the reanalysis project, having a one year spin up period results in adequate initial state of the sea ice thickness field. We assessed potential impacts of across-stream discrepancy on the long-term trends in the ice thickness series (presented in Fig. 7 of the original manuscript) by excluding an extra year of data at the beginning of each production stream, however computed scores were not significantly different from the values reported in the original manuscript. We believe that similarly it would have a minor impact on the annual evolution of errors in modelled ice thickness of CARRA (Fig. 8 in the original manuscript).

Considering your question about the starting date of the CARRA production streams. All but one production streams in CARRA start on 1 September, and one production stream (the first production stream of the initial CARRA reanalysis period, before back extension) starts on 1 July.

We will update the manuscript to document and discuss sea ice spin up and discrepancies across the production streams in CARRA.

Finally, is it possible to show a comparison with the ice surface observations that were made during the MOSAIC drift campaign to show an example of how the reanalysis compares, particularly as there is already a published comparison with the ERA5 and MODIS data by Herrmannsdörfer et al. (2023) for IST.

Indeed, the MOSAiC drift campaign provides very valuable information on the sea ice state in the modern-day Arctic. However, due to the location of the CARRA model domains the drifting station entered them at the late stage of the drift in May 2020, when temperatures rise and melting season starts. In a sense, this period is of somewhat less interest compared to the winter-time one discussed by Herrmannsdörfer et al. (2023) when assessing the model performance, since ice surface temperature in the model (and in observations) becomes close to the melting point of snow/ice and does not wary that much. Therefore, we decided to not include comparisons against MOSAiC in the present study.

Specific comments: Section 2: Lines 110-111: "new ice is always snow free" - I assume that if the ice concentration is updated from a non zero value to a different non zero value (increased/decreased) the snow depth remains the same? or is the snow volume conserved?

Yes, when ice concentration is updated from a non-zero value to another non-zero value, the snow depth within that grid cell remains unchanged. In the updated version of the manuscript we will explicitly mention that 'new ice' there implies the case when ice concentration is changed from zero to a non-zero value and that snow volume is not conserved when ice concentration is being adjusted.

Line 125: IFS-HRES - this acronym is not well explained - perhaps the more relevant information is that it makes use of the ECMWF 4D-Var data assimilation and forecast model

We agree that right now the text assumes that readers are familiar with the IFS-HRES NWP system and lacks details. We will update the manuscript to provide more details as suggested.

Section 3: When you make comparisons with the sea ice thickness and snow depth data have you taken into account the observational error? For the sea ice thickness data the errors can be quite large; the ice thickness retrievals may also be limited by the assumptions made about snow loading.

Indeed, satellite retrievals of ice thickness and especially snow depth can have quite high uncertainty. For this reason, we focused mainly on the qualitative assessment when using these products. However, we have also performed additional checks by comparing the differences between the satellite products and reanalyses against the uncertainty estimates provided with these retrievals. We found that on average the observed differences are above the uncertainty level for most of the area represented in CARRA. We will update our manuscript to explicitly mention that fact.

Section 4 (and figures) Fig 5: Is it possible to plot the ice cover amount somehow with the 4 zones in figure 5? could it be that we are looking at a fairly small area in zones A, B and C near the ice minimum? do you know how much the ERA5 boundary conditions may influence the behaviour of the CARRA as the ice retreats in these regions?

Thank you for a nice idea! You are right, ice-covered area within the four zones of interest can vary considerably throughout the year, and we agree that having this information could be beneficial. We will try to incorporate it in some form into the updated version of the manuscript.

Considering your second question, we believe that ERA5 boundaries do not significantly affect sea ice state in CARRA when ice retreats since it happens during the summer melt season. At that time (when ice have retreated considerably enough to remain only in the areas close to the model domain boundary) sea ice and snow cover on top are well into the melting regime with temperatures close to 0°C in both CARRA and ERA5 which reduces the difference between two products and thus potential influence of the boundary conditions.

Fig 13: What are the whiskers representing in the box and whisker plots (standard deviation, max/min values?)

For this figure in the original version of the manuscript we used a common form of a box plot where whiskers represent the distance to the farthest data point still located within the $1.5 \times \mathrm{IQR}$ distance (where IQR stands for "interquartile range") counted from the box's top/bottom. In this specific case the applied procedure yielded no outliers for all the box plots except the one representing the standard deviation of ice thickness errors in ERA5, which means that in these box plots with no outliers whiskers show the min/max range of the processed sets. To make the figure less cumbersome we will update the box plots to always represent the min/max range of input data and explicitly mention that in the figure caption.

Lines 565-566: you comment on the accumulation of the snow - do you have a sense of where the deficiency is in the model - is it the lack of the advective processes in the model?

We believe that in this specific case, the observed underestimation of snow depth in CARRA is more likely caused by boundary effects rather than by the lack of the advective processes in the sea ice parameterisation scheme. In the beginning of August 2012, when the IMB reports the onset of new snow layer accumulation, the buoy was located relatively close to the edge of the model domain where boundary effects (such as precipitation spin up) are not negligible. This can be illustrated by Fig. 3 which shows a considerably later start of snow accumulation in CARRA compared to ERA5. On the other hand, representing sea ice dynamics in CARRA would have had only a limited effect on the evolution of snow layer in this case as suggested by relatively small difference between snowfall accumulated along the IMB drift trajectory and

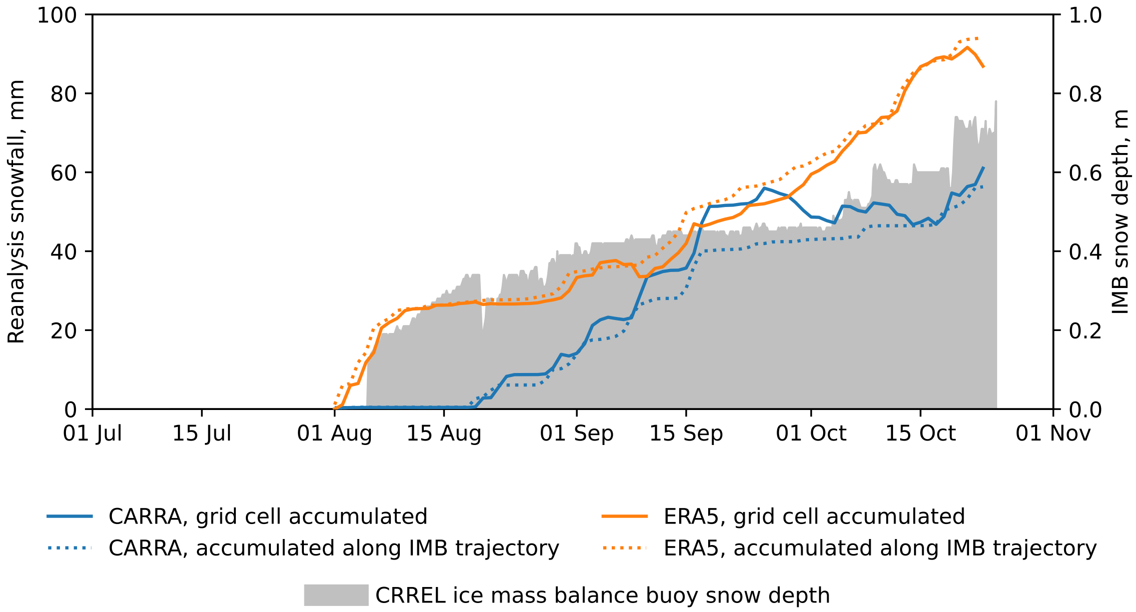

Figure 3: Accumulated snowfall in CARRA and ERA5 over a period from 1 August 2012 to 24 October 2012, extracted along the IMB drift trajectory. Solid lines represent total snowfall amount accumulated in a grid cell between 1 August and the date when IMB entered that grid cell. Dashed lines show snowfall accumulated along the IMB drift trajectory. Also in the figure, snow depth reported by IMB is outlined.

in stationary model grid cells (see Fig. 3). However, individual IMBs show a lot of variability and provide highly local observations therefore in our manuscript we decided to focus on the summary statistics across a set of buoys and to not provide detailed case studies on individual IMBs. We will update the manuscript to properly mention potential effects induced by the model boundaries in CARRA.

References

Herrmannsdörfer, L., Müller, M., Shupe, M. D., and Rostosky, P.: Surface temperature comparison of the Arctic winter MOSAiC observations, ERA5 reanalysis, and MODIS satellite retrieval, Elementa: Science of the Anthropocene, 11, 00 085, https://doi.org/10.1525/elementa.2022.00085, 2023.