The review comments are shown in black, the author responses in blue and text from revised manuscript in *blue italics*.

We would like to thank the reviewer for reviewing our manuscript. We appreciate all the comments and have revised our manuscript according to them. Please find below our responses to the comments.

Referee #1

Major comments

Model selection: The decision to limit analysis to a subset of high resolution GCMs seems somewhat arbitrary and limits the paper's value. This decision should be better justified in the text. For example, the authors could show a comparison of winter SWE in high vs low resolution models as supplemental material. Otherwise, the authors should consider adding a few of the HighResMIP historical simulations (https://gmd.copernicus.org/articles/9/4185/2016/gmd-9-4185-2016.pdf) to their analysis so as to increase the ensemble size.

We have now downloaded all the available CMIP6 models and compared SWE in the high resolution (100 km) and low resolution (250 or 500 km) models. We compared the monthly SWE sums over the entire study area (Fig. 1), following Fig. 2 in the submitted manuscript. Fig. 1 shows that the low-resolution models (thin lines) do not significantly differ from the high-resolution models (dashed lines); the SWE sum is mostly in the same range in both resolution groups. Also, the mean values for both high-resolution and low-resolution models are very close to each other. Therefore, we have performed the subsequent detailed analysis only for the high-resolution models to keep the number of models reasonable and to only consider models with spatial resolutions which are reasonably comparable with e.g., the SnowCCI data. The number of high-resolution models in the analysis has also slightly increased from the previous version, as we now included all models that were available for downloading in August 2021. Fig. 1 shows that there are two low-resolution models that show very high SWE sum values in every month, which are clear outliers. These outliers are “GISS-E2-1-G historical” and “GISS-E2-1-G-CC esm-hist” and we have found that the anomalous values are due to very high SWE in the mountainous areas.

Following Referee #2’s comments, we have also added mountainous regions to the analysis and, therefore, the SWE sum values are somewhat larger than in the previously submitted version. As the SnowCCI data are available only for non-mountainous areas, we have used MERRA-2, Brown and Crocus v7 datasets to fill the missing SWE over mountainous areas. We calculated the mean SWE of these three datasets for each grid cell that were defined as mountainous in SnowCCI.

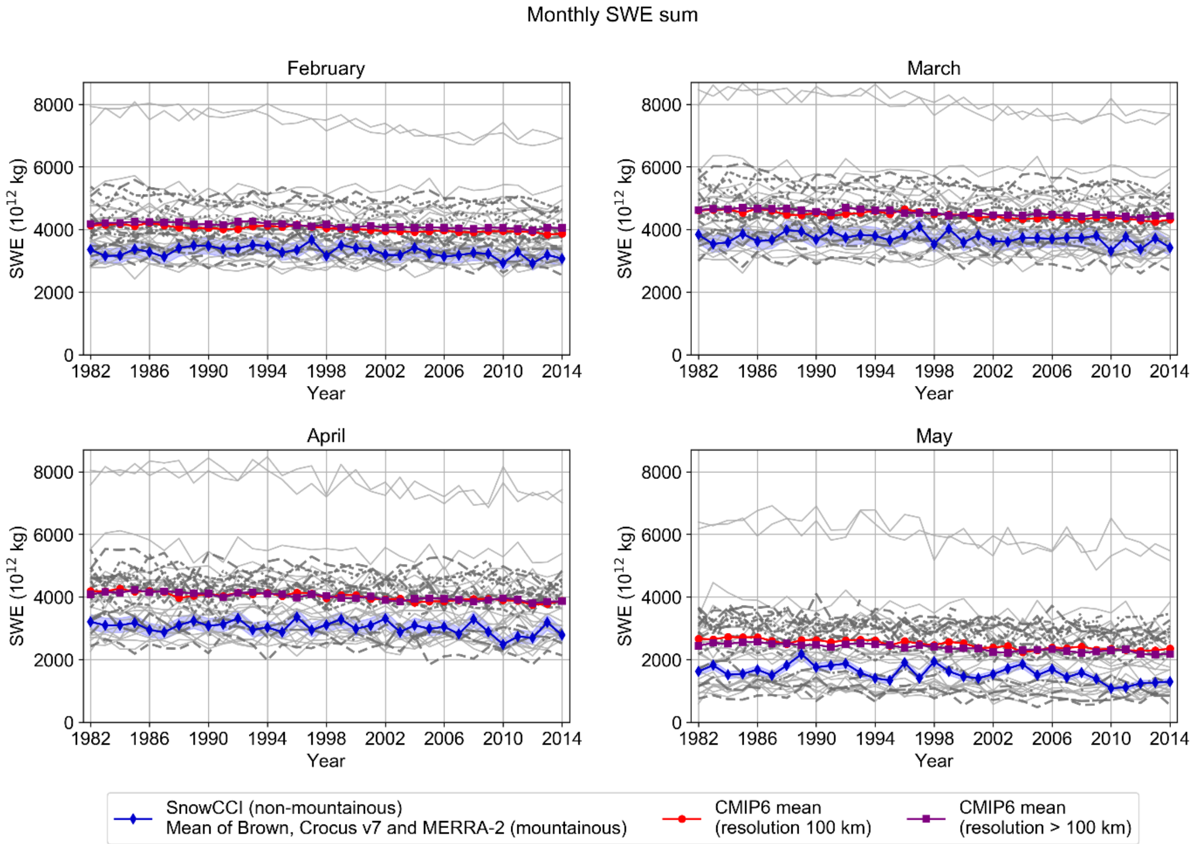

Figure 1. Monthly SWE sum over the entire study area in February, March, April, and May separately for each high-resolution (100 km) CMIP6 model (grey dashed lines), for each low-resolution (> 100 km) CMIP6 model (thin solid lines), for the high-resolution CMIP6 multi-model ensemble mean (red markers), for the low-resolution CMIP6 multi-model ensemble mean (purple markers) and for the SWE reference data (blue markers). The blue shaded area indicates the 7.4% uncertainty range of the SWE reference data.

Interpretation of results: The authors point out discrepancies between models and observations but offer little commentary on what could be driving biases in specific GCMs. For example, they discuss a cold bias in the EC-Earth models as unique to the ensemble but fail to connect this to the fact that EC-Earth is the largest outlier in terms of snow cover extent among CMIP6 models (Mudryk et al. 2020). More insight could also be added when discussing the CESM models, which feature anomalous winter SWE.

We have rewritten the Discussion and discuss more about both EC-Earth3 and the CESM2 models. We added discussion about the snow cover extent in EC-Earth3 models. We also contacted the CESM2 model developers to get more information about the anomalous SWE in the CESM2 models. A unique feature within CESM2, which should be considered, is that the model allows for a very large maximum SWE (10 m). This is done to enable the simulation of firn production over ice sheet regions, but it is possible that SWE can get very high in other cold regions as well (van Kampenhout et al., 2017; D. Lawrence, personal correspondence). We have also added a subsection “Residual” in the Discussion, where we discuss more about the biases that cannot be explained by biases in T or P.

Van Kampenhout, L., J.T.M. Lenaerts, W.H. Lipscomb, W.J. Sacks, D.M. Lawrence, A.G. Slater, and M.R. van den Broeke, 2017. Improving the representation of polar snow and firn in the Community Earth System Model. JAMES, 9, 2583-2600.

Readability: There are also several notations used throughout which can be improved to help the reader. For example, the "model-minus-observations difference" can simply be referred to as model bias. The results section can also be better tied together. Most paragraphs in Section 3 start with "Figure __ shows ...", which becomes very repetitive and causes the paper to lack flow.

We have now used the term "model bias" throughout the paper. We have also revised Results section according to the comment to reduce repetition and make the flow better.

Minor comments

L13-14 and throughout: change "SWE change rate in spring" to "spring SWE loss" or similar since the February to May SWE should decrease everywhere.

It is true that SWE decreases everywhere from February to May. However, we are studying the monthly changes and there are regions where SWE can temporarily increase, e.g., from February to March. Therefore, after consideration, we have decided to keep using the term "SWE change rate in spring".

L16: I don't understand what point is being made here: "Even too cold temperatures cannot cause too high SWE without precipitation".

We have removed this statement from the abstract.

L47: State that this is largely because of the increased atmospheric moisture holding capacity.

We have revised the text as suggested.

L48: "Trends in seasonal snow also vary seasonally" awkward wording.

We have revised the text according to the comment as follows:

Trends in snow cover also vary seasonally

L48-49: State why spring snow is especially sensitive to warming (e.g., surface albedo feedback is strongest during spring).

We have revised the text according to the comment.

L50: Clarify what is meant by "early-winter"?

We have revised the text according to the comment as follows:

Early winter from October to December shows even slightly positive trends in both Eurasia and North America, while in January and February, there are no significant trends.

L70: Change “the difference” to “the model bias”

We have revised the text as suggested throughout the paper.

L72-73: They stated that analysis is needed to understand SWE trends, but this paper only looks at climatological values.

In general, simulated trends of SWE can be considered more reliable if the current climatological distribution of SWE is simulated accurately. We have clarified this in the text.

L88-89: Could be worth showing this for one GCM in the supplement. E.g. a version of Figure 2 where the grey lines represent internal variability rather than intermodal variability.

We have added a version of Fig. 2 in the Supplementary material showing all realizations of three different models (CESM2, MPI-ESM1-2-HR, and EC-Earth3).

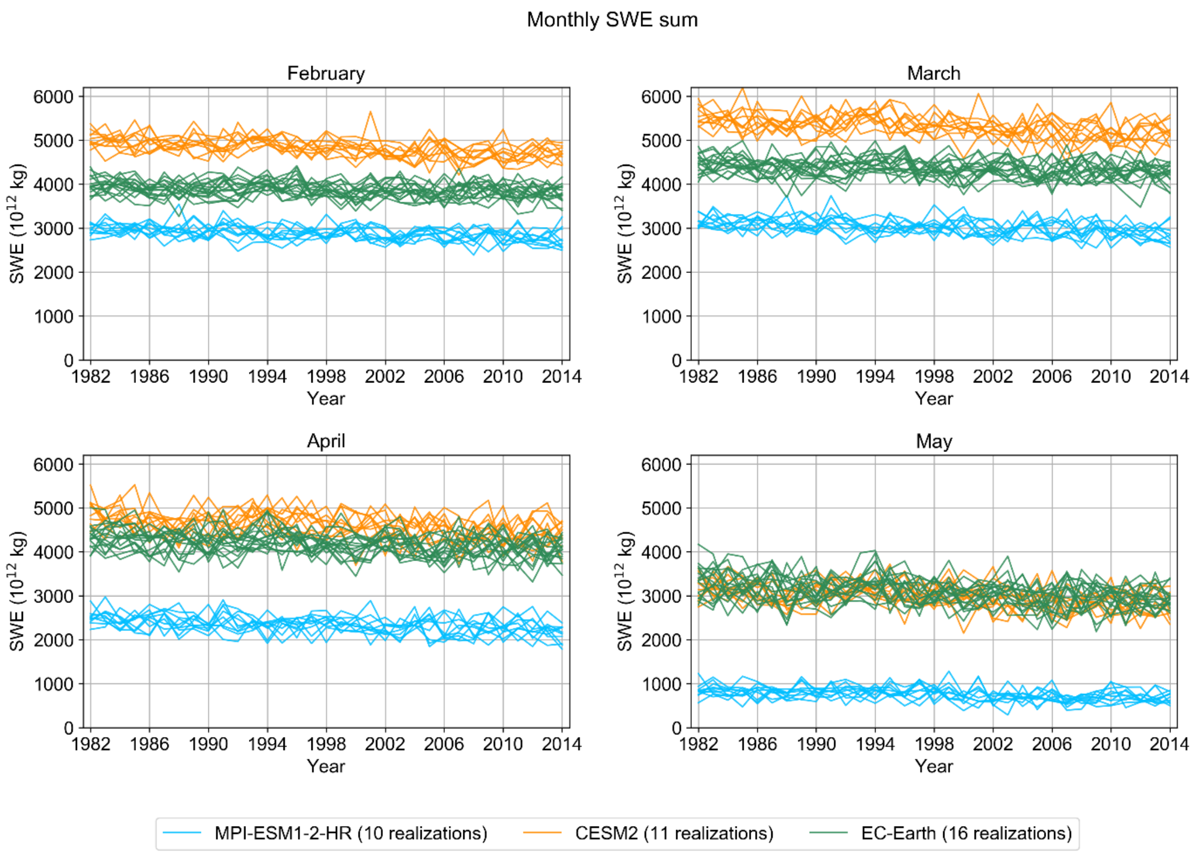

Figure 2. Monthly SWE sum over the entire study area in February, March, April, and May for all realizations of three CMIP6 models.

Table 1: Add model resolution as a column since that is one of the requirements for this study.

We have added a column showing the resolution in the table.

L109: Remove “year”

We have edited the text as suggested.

L109-110: Awkward wording, rephrase: “cover non-mountainous regions, and glaciers and ice sheets are excluded.”

We have revised the text according to the comment as follows:

Mountainous regions, glaciers, and ice sheets are excluded from the data.

L119: Sun et al 2018 (doi: 10.1002/2017RG000574) is a good reference for this statement.

We have added this reference in the manuscript.

L120: Why not convert it to mm/month so they are directly comparable?

The units mm month^{-1} and kg m^{-2} month^{-1} are equivalent to each other. For precipitation, we have used kg m^{-2} month^{-1} instead of mm month^{-1}, because we have used the unit kg m^{-2} for SWE. We have revised the text as follows:

All precipitation data are presented here in units of kg m^{-2} month^{-1} which is equivalent to mm month^{-1}.

L125: Citation needed for this statement.

We have revised the text according to the comment as follows:

In this study, we have used the monthly mean 2 m air temperature product, which agrees well with other datasets in the arctic (Gelaro et al., 2017; Simmons et al., 2017) and the mean values show very small biases (Bosilovich et al., 2015).

L133: Is there any downside to comparing the models at the observational resolution rather than regridding the observations to match the GCMs?

Resampling data can cause uncertainties in the results regardless of which resolution the data is resampled to. In this case, we decided to use a finer resolution because the resolution of the model is too coarse especially at marginal zone of seasonal snow cover. However, a disadvantage of the resampling approach is that small-scale differences between simulated and observed SWE are interpreted as model biases, even if the grid-mean SWE is simulated correctly. The resampling doesn’t however impact the larger-scale biases (i.e., biases at scales resolved by the models). Also, the large amount of data used in this study decreases uncertainties caused by resampling the data.

L138: Is this snow covered area calculated for each GCM or is a common snow covered area used across all models? We know from Mudryk et al. (2020) that snow cover extent is highly variable across CMIP6 models.

The snow-covered area is calculated individually for each model. We have clarified this in the text.

L144: Shouldn't February be included in this as well since you are assessing the February mean rather than Feb 1 SWE?

We could include February as well, however, as we are using the mean SWE of the whole month and not Feb 28 SWE, there would still be a slight mismatch in the variables.

L159 and throughout: change "model-minus-observation difference" to "model bias".

We have revised the text according to the comment.

L188: The precipitation and temperature biases seem fairly important to the overall story so it might be worth promoting this material to the main text.

Figure 1 in the submitted manuscript is meant to only show as an example the spatial distribution of SWE in CMIP6 models and in the SWE reference data. We have considered adding P and T also in Fig. 1 but decided to show them only in the Supplementary material to keep the number of figures more reasonable. The P and T biases for each model separately for winter and spring are shown in Figs. 4, 5, 9 and 10 in the manuscript.

Fig 3: "Mean difference in SWE" should be referred to as "SWE Bias" throughout

We have revised the text according to the comment.

Fig 3-4: Slightly confusing how "SWE in winter" refers to February, but "Mean P in winter" refers to the Nov-Jan mean.

We have replaced "Mean P in winter" with "Mean P in Nov-Jan".

L221-222: Can you quantify this bias in terms of a percent of the climatology?

We calculated the bias in terms of a percent of the climatology for these models in the northern regions and revised the text as follows:

For both models, the bias is very high in large regions in northern parts of North America and Eurasia. In these areas, the relative bias is typically 150- 200%.

L225: "Overall, the GFDL models are the most consistent with the SnowCCI data" – add "during February" after this statement.

We have revised the text according to the comment.

L230: add "NH extratropical" between overestimate and precipitation.

We have revised the text according to the comment.

L231: remove "dotted"

We have revised the text according to the comment.

L237: reword "either too high SWE and too low T or too low SWE and too high T"

We have revised the text according to the comment as follows:

either cold bias and positive SWE bias or warm bias and negative SWE bias, i.e., the areas where T bias could logically explain the SWE bias.

L251-252: "whereas in other models, deltaSWE is clearly smaller." This is not the most meaningful insight, can you be more detailed.

We have revised the text according to the comment as follows:

The mean ΔSWE varies from under 30 kg m^{-2} in the GFDL-CM4 model to around 50 kg m^{-2} in the CESM2 and NorESM2-MM models.

L277: Is it realistic to treat T and P as independent variables?

T and P are not fully independent but are linked to each other through several mechanisms. Especially in large scale and in long time periods T and P are linked to each other, but in shorter time periods and small scale (a grid cell), the dependency between these variables is complicated and varies spatially and seasonally. We have now revised this part to bring up this limitation:

In fact, this exposes a limitation of the regression Eq. (1): it treats ΔT_{cum} and ΔP_{cum} as independent variables, which is not fully realistic. When these variables are correlated, their contributions to ΔSWE cannot be fully separated.

L280-284: Hypothesize what is unique about these models that could be driving this.

We have added a subsection "Residual" in the Discussion, where we discuss about the possible factors behind these large positive SWE biases.

Prior to Figure 8: it seems like there should be a figure showing spring SWE change from OBS and models before showing the biases.

We have added a figure to the Supplementary material showing the SWE_{change} in models and in SWE reference data.

L294: DeltaSWEchange is confusing notation. Consider alternatives such as DeltaSWEmelt?

We considered this but decided to keep using DeltaSWEchange.

L298-299 and elsewhere: change "melts more slowly" to "there is less snowmelt". What is shown does not necessarily mean snow is melting faster because they all have different SWEmax values.

We have revised the text as suggested.

L316: change "mutually biases" to "mutual biases"

We have revised the text as suggested.

L327-348: Discussion of EC-Earth biases could mention that these models drastically overestimate NH snow cover extent.

We have revised the text as suggested. We have also added a subsection "Residual" in the Discussion, where we discuss more about the possible factors behind the EC-Earth biases.

L337: "biases in snow melt rate in spring are dominated by other factors than T or P" – further discuss some possible factors in the text (e.g. snow-covered surface albedo biases, which have been documented by numerous studies, albedo feedback strength).

We have added more discussion about the possible factors in the new subsection "Residual". Parameterization of surface albedo and the representation of the albedo feedback are among these factors. This is discussed based on the papers by Thackeray et al. (2015, 2018, 2021).