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 #2

Major comments

Introduction

The current section has extremely limited information about the previous studies for climate model-driven snow products in the Introduction section (such as the general performance of SWE products from earth system models within the CMIP, and what are the previous findings of the differences in CMIP6 as compared to CMIP5 snow products, etc). I would strongly recommend including a further description about climate modeldriven snow products and comparison studies (CMIP5 & 6, and statistical or physically downscaled products e.g. CORDEX) with its reliability and uncertainties in Introduction section. Also, the authors should provide much more sufficient information about a recent progress of the SnowCCI products from Luojus et al., (2021) and Pulliainen et al. (2020) [this manuscript should provide that information as a standalone work]. I'm sure this will draw potential readers' attention to the necessity of this study.

We have revised the Introduction section according to the comment. We have, for example, included more information about model/reanalysis snow products and, also, added comparison studies between CMIP5 and CMIP6 models.

Non-mountainous regions

The authors clearly stated that a main differentiation of the current study from one previous study comparing SWE in CMIP6 models (Mudryk et al., 2020) is to consider both temperature and precipitation to explain the differences in SWE. However, I would note that, unlike Mudryk et al. (2020), this study was conducted only for non-mountainous regions because of the unavailability of the SnowCCI SWE product over complex topography. This is crucial for SWE because a large portion of the seasonal snow exists mountains (for example, 40 to 60% in North America; Wrzesien et al., 2018; Kim et al., 2021). To achieve the comprehensive results across the NH, thus, I strongly suggest that the authors would consider adapting the weight-based blending approach used in Mudryk et al. (2020) with one or more additional reliable SWE products to include mountainous regions in this study. They used this approach to overcome the unavailability of the Globsnow SWE in complex terrains. The approach allowed them to merge multiple observations and reanalysis products to be able to evaluate CMIP6 SWE over the entire NH domain (not just non-mountainous areas). As the authors may know, the method is that a weight given to the GlobSnow data is linearly reduced with increasing the fraction of mountainous terrain, reaching zero for grid cells containing only mountainous terrain. Regarding dominant portions of the seasonal snow in NH exist in mountain regions, this will surely strengthen the results. Otherwise, it should be clearly stated that this study focuses on non-mountainous regions.

- Wrzesien, M. L., Durand, M. T., Pavelsky, T. M., Kapnick, S. B., Zhang, Y., Guo, J., and Shum, C. K.: A new estimate of North American mountain snow accumulation from regional climate model simulations, Geophys. Res. Lett., 45, 1423–1432, 2018.
- Kim, R. S., Kumar, S., Vuyovich, C., Houser, P., Lundquist, J., Mudryk, L., Durand, M., Barros, A., Kim, E. J., Forman, B. A., Gutmann, E. D., Wrzesien, M. L., Garnaud, C., Sandells, M., Marshall, H.-P., Cristea, N., Pflug, J. M., Johnston, J., Cao, Y., Mocko, D., and Wang, S.: Snow Ensemble Uncertainty Project (SEUP): quantification of snow water equivalent uncertainty across North America via ensemble land surface modeling, The Cryosphere, 15, 771–791, https://doi.org/10.5194/tc-15-771-2021, 2021.
- Mudryk, L., Santolaria-Otín, M., Krinner, G., Ménégoz, M., Derksen, C., Brutel-Vuilmet, C., ... & Essery, R. (2020). Historical Northern Hemisphere snow cover trends and projected changes in the CMIP6 multi-model ensemble. *The Cryosphere*, *14*(7), 2495-2514.

We have added the mountainous regions to the analysis to strengthen our results. As the SnowCCI data are only available over non-mountainous areas, we used MERRA-2, Brown and Crocus v7 datasets to fill the mountainous grid cells. These same datasets were used in Pulliainen et al. (2020). We calculated the mean SWE of these three datasets for each grid cell that were defined as mountainous area in SnowCCI. Thus, we have now included also the mountainous regions in the analysis, but the results and the conclusions remained quite similar.

Forested areas

I am not fully sure about the reliability of the SnowCCI product is enough as a single reference dataset to evaluate the CMIP6 SWE product to achieve a general conclusion, particularly in not only mountainous areas (which were already masked), but also vegetated (or forested) areas in this study. There are well-known limitations of satellitebased passive microwave (PMW) sensors for snow remote sensing which have been used to develop the GlobSnow product as the main component. Numerous previous studies have found that the passive microwave SWE products are problematic due to many issues (e.g. deep snow “saturation effect”, wet snow, forest canopy, terrain heterogeneity, etc.; Dong et al., 2005; Derksen et al., 2010). I believe many readers may also concern about the issues regarding the reliability of the SnowCCI product, particularly in snow hydrology community (Larue et al., 2017).

To address the issue of the product in forested areas, ideally, employing a model/reanalysis SWE product could mitigate it (such as MERRA2 or ERA5; Colleen et al., 2019). Also, it might be helpful to discuss about recent findings in the Introduction or Discussion sections. For example, a recent study from an independent group found that there were better performances of the GlobSnow SWE product as compared to the passive microwave alone SWE retrievals, particularly in maritime and warm forest environments (Cho et al., 2020; this study used the previous version; GlobSnow v2). I strongly recommend providing clear descriptions how (not) to deal with the issues with sufficient literatures.

Dong, J.P. Walker, P.R. Houser, Factors affecting remotely sensed snow water equivalent uncertainty, Remote Sens. Environ., 97 (1) (2005), pp. 68-82

Derksen, P. Toose, A. Rees, L. Wang, M. English, A. Walker, M. Sturm Development of a tundra-specific snow water equivalent retrieval algorithm for satellite passive microwave data, Remote Sens. Environ., 114 (8) (2010), pp. 1699-1709

Larue, F., Royer, A., De Sève, D., Langlois, A., Roy, A., & Brucker, L. (2017). Validation of GlobSnow-2 snow water equivalent over Eastern Canada. *Remote sensing of environment*, *194*, 264-277.

Cho, E., Jacobs, J. M., & Vuyovich, C. M. (2020). The value of longâ€• term (40 years) airborne gamma radiation SWE record for evaluating three observationâ€• based gridded SWE data sets by seasonal snow and land cover classifications. Water resources research, 56(1).

It is true that satellite-based SWE estimates have had problems with several issues. However, all the studies mentioned here are conducted using GlobSnow v2 product, which is not bias-corrected. The bias-correction method has been found to clearly improve the SWE estimation and to solve many issues, which were previously associated with satellite-based SWE estimates. The bias-correction has removed the deep snow saturation effect that was previously an issue for the satellite-based SWE estimates. Figure 1a in the Extended Data in Pulliainen et al. (2020) shows that the bias-correction method clearly improves the SWE estimates when SWE >150 kg m^{-2}. We have also added fractional forest cover to the analysis and studied the effect of fractional forest cover on the residual term (Fig. 1).

Reorganization of the structure of the manuscript

I think the current manuscript should be re-organized. There exist many statements in Discussion section which are supposed to be in "Result" section (or already mentioned here). There is a limited discussion in the current manuscript which should have been here such as "comparison to previous findings and why they are similar/different", "Limitations in the methods and results", and "future perspectives". To make a more structured manuscript, I would recommend separating Data and Method and making subsections within “Data” section such as “SnowCCI”, “MERRA-2 temperature”, “GPCC precipitation”, and “CMIP6”. Also for “Discussion” section, I suggest separating the current form into subsections based on the major findings such as “CMIP6 performance”, “Relative contribution of P and T to SWE”, and “Limitations and future perspectives”, something like them. This would help readers explicitly find and understand this work.

We have now reorganized the manuscript and added subsections according to the comment to make it more clear and easier for readers to understand our work.

The residual term

There are many parts that just speculated the reasons of the residual term without supporting explanation based on previous findings or sensitivity analysis (e.g. L254-255, L413-414), even though the portion of the term was considerable. (1) Please provide reasonable rationales to support the author’s statements. Regarding this, I think land characteristics such as forest fraction and/or spatial heterogeneity also can impact on generating the residual. To examine this, (2) I would suggest that the authors conduct some sensitivity analysis to provide useful information to be able to explain regional differences in residual from Figures 7 and 12.

We have revised our manuscript according to the comment. We have added a new subsection “Residual” under the Discussion section, where we discuss more the residual term. We have also added forest cover data to the analysis and studied the effect of fractional forest cover on the residual term. Figure 3 shows the dependency between fractional forest cover and the residual for winter and spring. We have calculated the dependency between residual term and fractional forest cover for the entire study area (top row in Fig. 1) and for the non-mountainous area (bottom row). We will add this figure also in the manuscript.

While some correlation between forest cover and residual seems apparent in some models such as CESM2 and NorESM2 (in winter), inter-model differences are still the primary cause of the large spread in the residual term (Fig. 1). The spring-period residual correlations with forest cover, if any, could conceivably result from, for example, snow surface albedo treatment differences in the models. The treatment of snow surface albedo should manifest strongest over open tundra regions and less so over dense forest cover, as we now note elsewhere and have added to the discussion section. However, we emphasize that more detailed per-model investigations on this topic for all participating CMIP6 models in this study are not feasible within the present scope and purpose of the study.

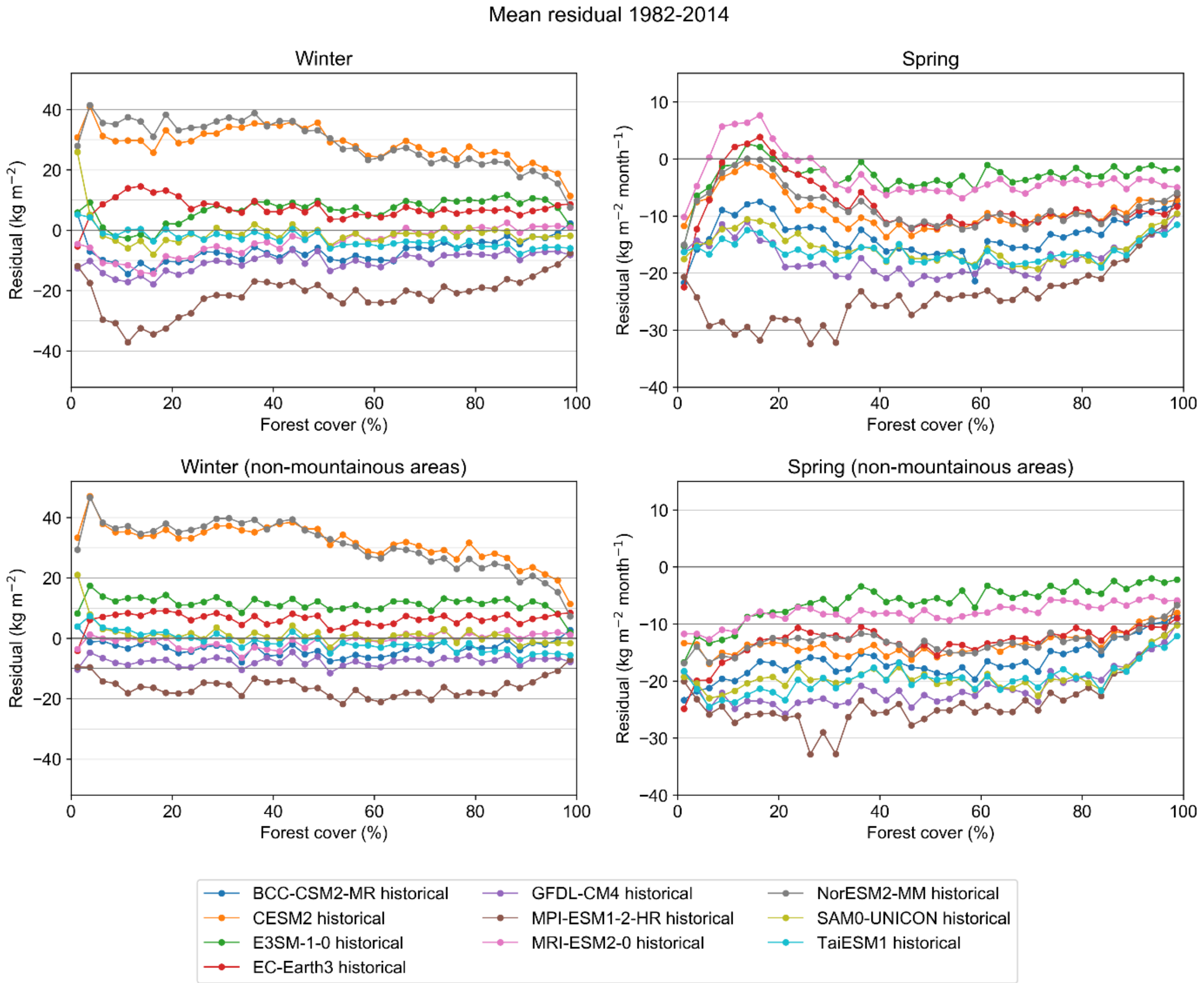

Figure 1. The dependency between the fractional forest cover and the residual term. Here, we show only one model from each modeling group to keep the number of the models more reasonable.

Specific comments

L13 Specify in-situ "snow depth"

We have edited the text as suggested.

L54 Even though a satellite remote sensing technique is the only option for "observing" SWE at continental scale, state-of-the-art model/reanalysis SWE products have been successfully estimated, and they have been widely used for hydrological and climate research rather than satellite-based approach (mostly passive microwave) probably due to its limitations above. I would suggest rewriting this part covering not only remote sensing approach but also model/reanalysis products for NH SWE.

- Huning, L. S., & AghaKouchak, A. (2020). Global snow drought hot spots and characteristics. Proceedings of the National Academy of Sciences, 117(33), 19753-19759.

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

Observing SWE at continental scale is only possible from satellites, but also model and reanalysis SWE products provide gridded SWE estimates and have been widely used in hydrological and climate research (e.g. Huning and AghaKouchak, 2020; Mortimer et al., 2020; Mudryk et al., 2020). Previously, substantial uncertainties have been reported in NH SWE estimates (Bormann et al., 2018; Mudryk et al., 2015). However, our knowledge of the NH SWE has recently improved considerably, with new bias corrections which reduce the uncertainty of the SWE estimate integrated over NH from 33% to 7.4% (Pulliainen et al., 2020). The bias-correction method, for example, considerably improves SWE estimates in the moderate and deep SWE range (Pulliainen et al., 2020), which has previously caused low bias in SWE estimates (Cho et al., 2020). However, limitations still exist: the bias-correction method cannot be applied in mountainous regions due to the lack of snow course measurements and the large SWE variability in complex terrain (Pulliainen et al., 2020). Even though the area of mountainous regions is limited, these regions store a considerable portion of the seasonal snow (Kim et al., 2021). The bias-correction method mostly increases SWE, and it is therefore likely that without the bias-correction, SWE in mountainous areas is biased low (Pulliainen et al., 2020; Wrzesien et al., 2018).

L69 They have used four model/reanalysis and satellite SWE datasets and combined them using a blend approach, not just satellite-based data.

We have revised the text according to the comment.

L87 – 89 I think presenting the results from the brief analysis (even in supplementary info) should be helpful for keen reader. Also please provide the detailed description of how the difference among the ensemble members are quantitively smaller than that of models.

We have added a figure (Fig. 2) to the Supplementary material showing all realizations of three different models (CESM2, MPI-ESM1-2-HR, and EC-Earth3). The figure shows that internal variability of each model is smaller than the intermodel variability.

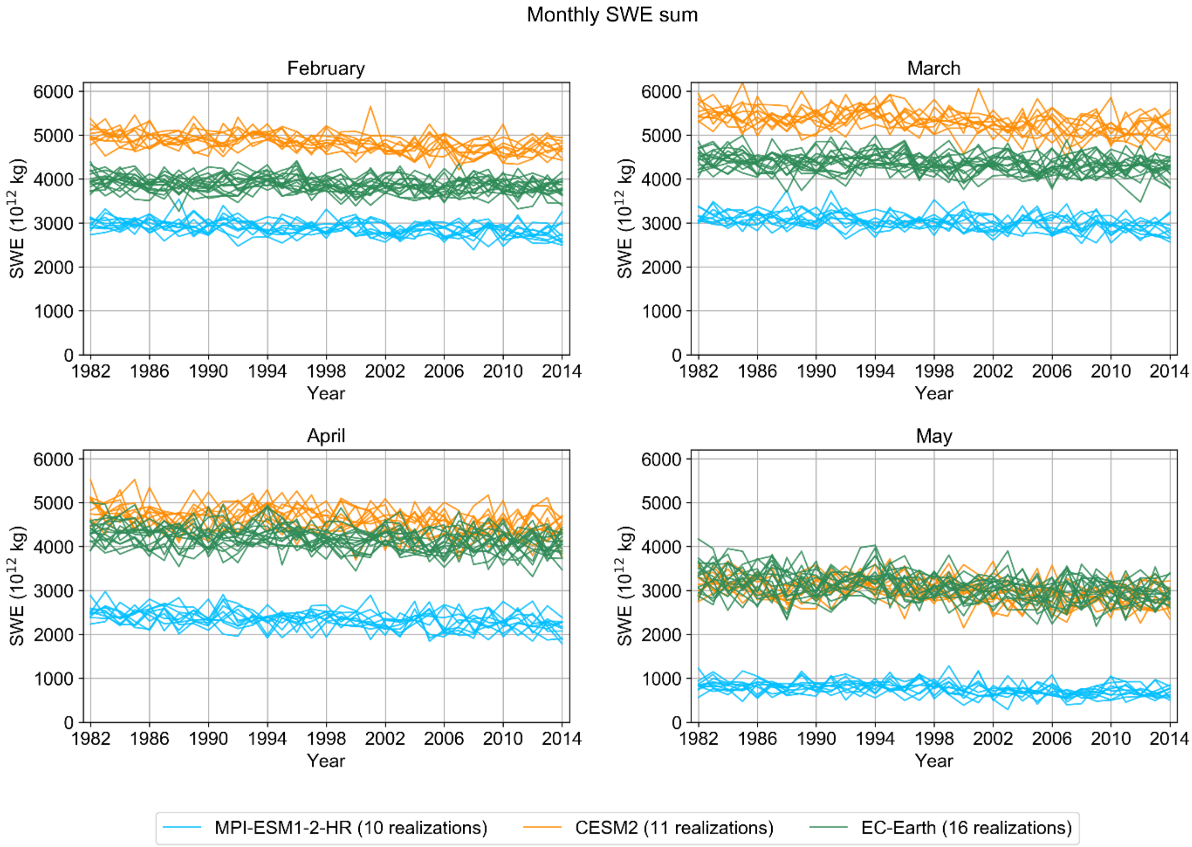

Figure 2. Monthly SWE sum over the entire study area in February, March, April, and May for all realizations of three CMIP6 models.

L91-92 Is the GlobSnow v3.0 the same product as SnowCCI used in this study? If not, please add the differences.

Yes, it is the same product. This is mentioned in the revised manuscript.

L100 – 102 Even though the GlobSnow retrieval was improved by combining in-situ snow depth observations as compared to a satellite-only retrieval SWE, there was still large uncertainties for moderate and deep SWE range (about > 150 mm) which is probably due to the "saturation effect" of the volume scattering approach (Derksen et al., 2010; Cho et al., 2020). Was the SnowCCI improve these limitations as compared to the previous version of the GlobSnow? Based on the SWE assessment in Luojus et al. (2021), the overall RMSE for all samples and for shallow to moderate snow conditions only (SWE below 150 mm) is 52.6 mm and 32.7 mm, respectively.

The bias-correction method improves the SWE estimates significantly when SWE > 150 mm. Please see Fig. 1a of Extended Data in Pulliainen et al. (2020).

L109-110 What percentage of the seasonal snow-covered area is non-mountainous area over NH? It would be helpful for reader to get the conclusion from this study within nonmountainous areas (if the authors adhere to non-mountainous area).

We have added the mountainous regions to the analysis.

L112-113, L361-362 Overall, I felt that the paper is overvaluing the accuracy of the SnowCCI product as reference dataset. Please tone down.

We have added three other datasets to cover the mountainous areas, and added also more discussion about the uncertainties of the datasets in the Introduction and Discussion sections.

L254 What does “model structural factors” mean? Be specific.

With model structural factors we refer to modelling deficiencies that cause errors in the simulated SWE, other than biased simulation of T or P. In particular, these factors may include deficiencies in the parameterization of surface energy budget and other snow-related physical processes (phase of precipitation, snow cover fraction, snow albedo, heat conduction in snow, etc.). This is discussed in more depth in the subsection “Residual term” of the revised manuscript. We have edited this sentence as follows:

The large residual term implies that observational uncertainty or model structural factors, such as deficiencies in the parameterization of surface energy budget and other snow-related physical processes, play a considerable part in the observed SWE biases.

L254-255 This is speculation for me. Please provide rationale based on literatures related to this statement.

In our understanding, this statement is not speculation - in fact, there aren’t really other possibilities. The results clearly indicate that the residual term is substantial and, also, that it varies between the models. Thus, even if the models were able to simulate temperature and precipitation correctly, the simulated SWE would differ from observations. Obviously, any errors in the observations would impact this residual. However, since the residual varies substantially between the models, it is also clear that the residual is influenced by modelling inaccuracy (other than errors in temperature and precipitation).

L259 I do not think R^2 is a “parameter” of linear regression.

We have edited the text according to the comment.

Figure 11 To me, the residual terms overwhelmed the contribution of P and T. In this case, are the contributions of P and T still statistically significant?

It is true that the residual term is considerably larger than the contributions of T and P. Also, R^2 values are quite low in spring, suggesting that bias in SWE change rate does not depend much on bias in T or P. We have added figures showing the statistical significance of the terms in the Supplementary material.

L337 Please add further discussion "other factors" particularly in spring season. Do you think mismatching of the spatial resolution among the data sets can be one of the reasons? If so, please add some discussion about this. Regarding this, how do you think of the resampling method (nearest neighbor)?

We have added more discussion about the residual term in the Discussion section. The resampling can influence the small-scale (i.e., model subgrid-scale) features of the residual, but not the larger-scale features. Since the residual clearly shows large-scale structures in many cases, the resampling does not appear to be a major factor explaining the residual.

Figure S6 There are areas where the R^2 values are extremely low. I think it would be good to show the beta_P and beta_T for regions only where there are statistically significant. Please consider applying this throughout all figures.

We have added figures showing β_P and β_T only for regions, where there are statistically significant, in the Supplementary material.

L342 Be consistent either "Fig" or "Figure"

We have edited the text according to the comment and used "Fig" throughout the text.

L360-361 This sentence is redundant as the authors already mentioned. I would suggest rephrasing something like "while …, our study focuses on analyzing the CMIP6 SWE responses to both temperature and precipitation"

We have revised the text according to the comment.

L362-364 I am not sure if the statements are needed here, which were already mentioned several times.

We have removed this statement from the text.

L373 Figs.

We have revised the text according to the comment.

L360-364 & 376-380 To me, it seems like the summary, not discussion. I would strongly recommend using here for the detailed discussion, such as what are similar/different and what are new findings from this study as compared to previous studies?

We have edited Discussion according to the comment. We have, for example, added more discussion about the residual term, the uncertainties of the reference datasets and compared our results more with previous studies.

L388 Figs. If you refer more than two figures, please use Fig"s"

We have edited the text according to the comment.

L430 I suggest providing much more details of the limitations/uncertainties from the SnowCCI and others to provide sufficient information for those who would use the data sets for their own research, particularly for the issues that I provided in the major comment (such as uncertainties in forested areas which have been challenging areas in snow community). What would the authors expect potential uncertainties in GPCC? Please add discussion sufficiently.

We have edited Discussion according to the comment and added more details of the limitations and uncertainties from all the datasets used in this study.