Answer to referee 2

The paper by Buchmann et al. studies the impact of homogenization on snow series in Switzerland. It is a continuation of previous work on break detections. The application and testing of different homogenization approach for in-situ snow cover series is important for many climatological applications, but yet little researched. The authors have made a considerable contribution in this area previously and also with this paper.

The manuscript is well organized and, generally, well written. Figures focus strongly on displaying trend magnitudes as colours and not as x- or y-axes, which sometimes only allows a rough interpretation. Finally, there are two major concerns that would need be addressed before publication (see below). One is related to the methodology, which is not well described. Also it is difficult to understand how the authors arrived at some of the conclusion (e.g., in the abstract). The other concerns the non-random changes in trends.
Thank you for taking the time to read, analyse and try to improve our manuscript. In the following, we addressed each of your points.

 # Major comments

- The description of the different homogenization approaches is very explanatory and vague. I find it hard to distinguish between what the methods do and need in general, and how the authors applied them. Maybe a table would help? I think it is important to understand if the approach works daily or monthly, what thresholds on distance (h and v) were used – btw, are they the same for all methods, it is unclear to me – what thresholds on correlations, etc. Are the adjustment factors applied additive or multiplicative? I think the authors conclusion in the abstract are based on their understanding, but for readers it is impossible to follow.

Since the criteria for selecting the reference series are different for each method, we added them to the appropriate paragraph for each method. As suggested by R1, we also added the formulae calculating the adjustment factors to improve the understanding of the methods. The thresholds were already there for interpQM, we added them for HOMER and Climatol.

Climatol:
"… It uses composite reference series that are constructed as a weighted mean, using the horizontal and vertical distance between suitable reference and the candidate series as weight. We used the standard settings, which are 100 km at which the weight of the horizontal distance is halved and 0.1 for the scaling parameter of the vertical distance."

HOMER:
"The adjustment factor is determined by means of variance analysis ANOVA (Caussinus and Mestre, 2004; Mestre et al., 2013) based on the selected reference stations. These are chosen either based on the horizontal distance or first-difference correlations. Due to the large vertical distances between stations, even within short horizontal distances, the latter was chosen with a minimum ρ of 0.8."

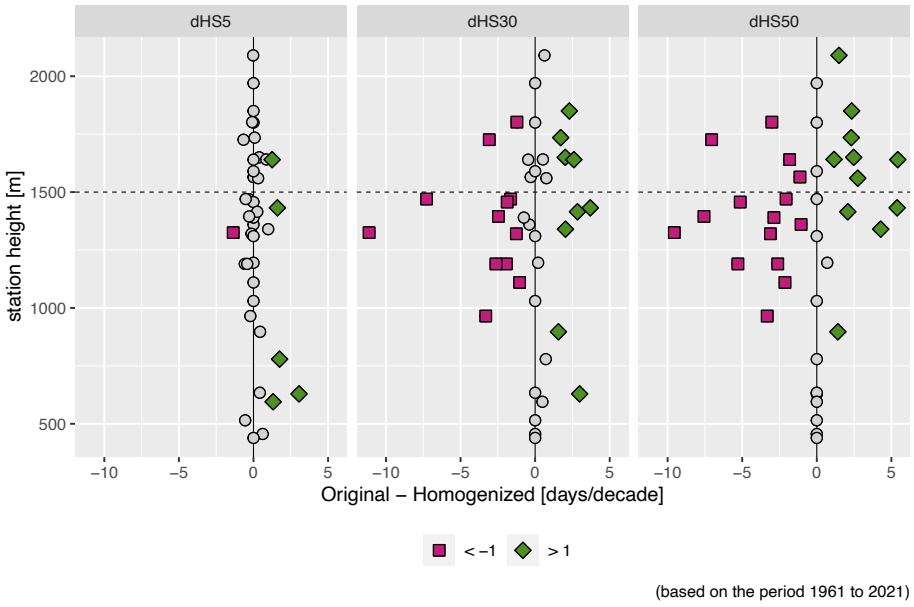
To improve the general understanding of how the methods work, we added a paragraph prior to the method-descriptions, explaining topics that apply for all methods, e.g. that all use multiplicative adjustment factors:
"Each breakpoint of the station to be adjusted (candidate series) is adjusted by multiplying the values series between the start of the measurements or an earlier breakpoint and the breakpoint to be corrected with an adjustment factor. This is calculated using one or more nearby reference series. All methods compared use the same data set to select suitable reference stations for the calculation of the adjustment factors. These are then multiplicatively applied to the daily (interpQM) or monthly (Climatol, HOMER) values of the candidate series. Although it is of course possible to manually select suitable reference stations for each series and use only these for each method, we have chosen to let the methods themselves select their reference stations based on their built-in criteria."

- Throughout the manuscript I was surprised to see that homogenization drove trends to be more negative. What are the reasons for this? In theory, I would expect non-climatic breaks to be random, especially since I think manual observing procedures have remained the same for snow depth for the past, or? So there should be no instrument bias. Inhomogeneities are then mostly related to observer changes or relocations, which should not be biased towards less snow consistently. But please correct me, if I'm wrong here. The authors also confirm this supposition in L194, where the percentages they give show more or less similar strengthening and weakening of trends. I think this issues deserves some more consideration. It might be partly related to how strength and direction of trends interact, especially when close to zero, where small changes in trends can lead to a change in sign. Could the authors provide a figure where they show explicitly how trends have change in a numerical fashion? Currently, all plots show trend magnitude with colours, where it is hard to judge values from. Also, since the authors have a larger dataset in the background (the homogeneous series they did not use here, but in previous works), they could compare inhomogenous vs homogeneous series in their fraction of positive/negative trends to check whether inhomogeneities were really biased to more snow.

We have added a paragraph to the discussion dealing with this topic, as it is very important: *"So far, the homogenised snow depth time series, show no evidence of a bias in the methods towards increasing or decreasing snow depths due to the adjustments, neither in Austria nor in Switzerland. In this study, depending on the method, the mean snow depth was increased at about 45 - 57 %' of the stations and decreased at*

between 46 - 43 % before a break. 95 % of the 40 inhomogeneous stations showed a negative trend, and for 58 % it was significant. With 78 % negative and 50 % significant, these figures were lower for the 144 homogeneous stations."

Regarding your suggestion to show the change of trends in numerical form, we have adapted Figure 2. It now presents the influence of the adjustments to the snow day indices more clearly on the x-axis.



 # Minor comments

- L16: A station cannot be significant, but its trend. Please adjust.

Changed to:
"In addition, the number of stations with a significant negative trend was increased …"

- L19: Do you have a reference for that 50%? Does it refer to all of NH?

We added the reference in question: Armstrong & Brun (2008): https://doi.org/10.3189/002214309788608741

- L41: Would be good to include Marcolini here, too.

Done. We added both Marcolini-papers (2017 and 2019), changed the corresponding sentence
"This is a standard process for climate data like temperature and precipitation but has only recently been adopted for snow depth time series (Schöner et al., 2019; Resch et al., 2022)."
with a paragraph below. It should be a better introduction now and includes a short summary of the articles mentioned:
"First steps towards detection and adjustment of breaks were made by Marcolini et al. (2017). Schöner et al. (2019) used homogenised seasonal time series for calculating trends and identifying homogeneous regions of snow depth in Austria and Switzerland. Marcolini et al. (2019)

compared the applicability of two homogenisation methods and their effects on trends in seasonal mean snow depth. These analyses showed the need for improved adjustment methods to enable both the application to data with higher temporal resolution, such as daily data, and make the adjustments of large values more robust. For this, an implementation of an adjustment method using quantile matching was proposed by Resch et al. (2022). Buchmann et al. (2022) analysed the break detection capabilities of three widely used semi-automatic detection methods. The question remained open as to how well other widely used adjustment methods are applicable and how they affect other indicators of interest besides the seasonal mean snow depth, e.g. extreme values. In order to consider the homogenised part of time series as more credible and statistically robust, a more detailed study of the effects of homogenisation methods on snow depth and derived metrics was necessary."

- L53: If I remember correctly, Marcolini did that.
You remembered correctly. We rewrote the whole paragraph, merged it with the sentence you mentioned in L41 and moved it there to serve as a introduction to the already performed snow depth homogenization work.

- L106: IQR is a widely used term in statistics for inter-quartile-range (note R not N in quantile/quartile), please use another abbreviation for clearness. And I think you mean that you split into different bins based on quantiles, and not split into inter quantile ranges, which are only numbers (differences between the quantiles).
Good point! It is indeed misleading to use a common abbreviation for two different contexts. The "range" in IQR was indeed to be understood as bins. To avoid further confusion, IQR was changed to "interquantile subset" and we introduce "IQS" as the abbreviation.

- L118: "Since it was decided…" : I do not understand what you mean, or who decided.
We rephrased the sentence from
"Since it was decided that interpQM does not change …"
to
"Since interpQM does not change …"

- L127: The authors should explain why they choose exactly this specific approach to define "better" performance for R50HSmax, since this is not a trivial task.
We added
"This approach was chosen because the international standards for maximum snow load on buildings are based on R50HSmax (see e.g. Schellander_2021)."

- L146: For what exactly did you use the KS-test?

The corresponding sentence was deleted as it was mentioned already earlier (L130) and did not make sense here. We extended the reasoning behing the two-sample KS-test to

"In order to determine to what extent homogenised and original time series differ, as well as to assess the differences between the results of the applied adjustment methods, a two-sample Kolmogorov-Smirnov test was conducted with seasonal data."

This comparison can be done by different tests, e.g. the two-sample KS-test or a Wilcoxon rank sum test. Here we have chosen the former.

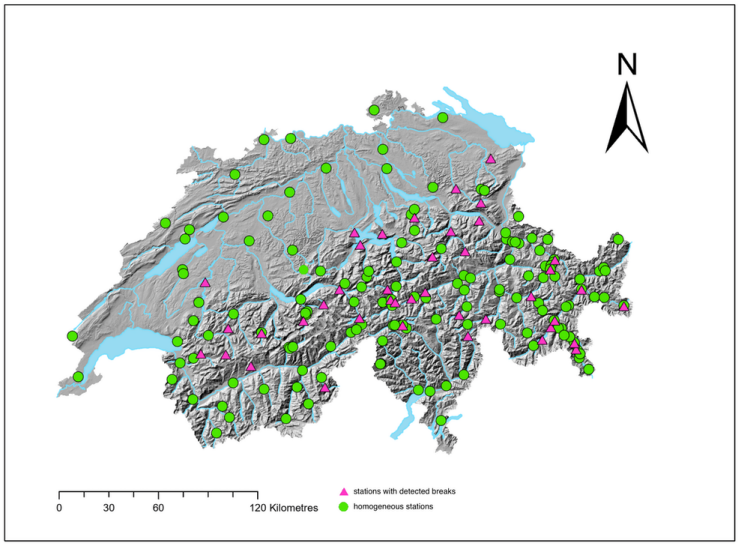
- L174(paragraph): I find the reasoning irritating in this paragraph. Are the trends in snow driven by absolute temperatures? Which the authors imply in L177. And then the statement in L179 is based on what source? Is it stating theory or factual evidence?

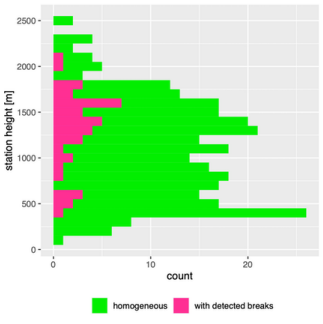
We found that the whole subchapter was a bit inconsistent and difficult to read. It was therefore rewritten. Regarding the argumentation: It was moved from results to discussion in order to achieve a clearer separation of results and interpretation. The statement mentioned was deleted and changed to

"The altitude-dependent pattern with the strongest adjustment effects for dHS30 and dHS50 between 1000 - 1700 m can be explained by the fact that, firstly, at stations below 1000 m a.s.l. there are few days with a snow depth of 30 cm (or more) due to the generally warmer temperature and lower snowfall amount and, secondly, that above 1700 m winter temperatures are lower and therefore less sensitive to warming in winter that the trends are smaller. A similar pattern can be seen in the absolute values (Appendix 1)."

- Figure 1: Could you please add an elevation distribution of the stations? And highlight in the distribution homogeneous and not.

Done.





- Table1: Would be good to know how many stations are below 1500m and above 1500m. Also is the number constant across dHS5, 30 and 50? Also the numbers seem to not add up: Shouldn’t positive plus negative always be 100% for each column? It is for some columns, but not for all. And these should be consistent to the rows “pos to neg” and “neg to pos”, no?

We added the number of stations in the two altitudinal subsets below the table and changed

“The snow days of each season were examined in two subsets of stations below and above 1500 m, in this chapter called "lower altitude" and "higher altitude" stations.”

to

“The snow days of each season were examined in two subsets of stations below (n = 26) and above (n = 14) 1500 m, in this chapter called "lower altitude" and "higher altitude" stations.”

This number is constant for the three indices. The reason why the numbers in a column sometimes did not add up to 100% was that the table only showed positive and negative trends and not stations with no trend. Initially we thought that only positive and negative trends would be sufficient, but in view of your question we decided to also include "No trend" and trend changes from "Positive to no trend" and "Negative to no trend" in the table and the analysis.