Answer to referee 1

The manuscript submitted by Buchmann et al. is a nice contribution that complements the analysis performed by the same group of authors in the past years. I consider the manuscript of interest for the community, however I believe the methodology should be significantly better described. In particular, the fundamental step of the calculation of the adjustment factors and their comparison in terms of applied equations is missing. Moreover, the role played by autocorrelation in the analysis of the time series is not discussed. Additional comments are contained in the pdf file.
Thank you for your time to work through the manuscript and for sharing your thoughts to improve it. We really appreciate it! We reworked the methodology-description for making it easier to understand in the new version. Regarding the autocorrelation: We used seasonal derived indicators (e.g. seasonal number of days with snow depth > 5 cm), so autocorrelation does not play a role here as it would of course do to daily snow depth series. Some examples for this are shown in table 1 below.

50: I think the authors here have chosen the wrong publication. It should be Marcolini et al. 2017 (Marcolini, Giorgia, Alberto Bellin, and Gabriele Chiogna. "Performance of the standard normal homogeneity test for the homogenization of mean seasonal snow depth time series." International Journal of Climatology 37 (2017): 1267-1277.)
We added Marcolini et al. 2017 to the list.

52: Probaly the work of Marcolini et al., 2019 is more appropriate here, since it already investigated the impact of different homogenization approaches on trends.
Done. We rewrote the paragraph to give a better overview of recent snow depth homogenisation efforts and changed the citations.
“In spite of recent efforts of studying the homogenization of snow depth (Marcolini et al., 2019; Schöner et al., 2019; Buchmann et al., 2022; Resch et al., 2022), it is still an open question whether the use of one of the methods is more advantageous than the others, and how the methods impact trends and extreme values. So far, Swiss snow depth time series have not been homogenized and an impact assessment has not been carried out neither for Switzerland or any other region.”
was changed to
“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.

79: Some steps in the methodology are not clear to me:

1. I assume that for some time series all three methods detect a breackpoint (let's call it set A) and for some time series only two methods detect a breakpoint (let's call it set B). I think it should be somehow indicated in the presentation of the results which stations belong to set A and which to set B in order to see if any difference is present.

This is described in Buchmann et al. 2022, Table 3. (https://doi.org/10.5194/tc-16-2147-2022). As a better description of the break detection results was added to the text: “For details, see Buchmann et al. 2022.” was changed to “For details, e.g. on the differences between the methods in the detected breaks, see Buchmann et al. 2022.”

2. Is the set of stations used to detect the breakpoint different from the set of stations used to adjust the breakpoint and compute the correction factor? I think they should remain the same and if they change, I would expect some more discussion about the implication and the mathematical soundness of this choice.

We used the same set for both detecting and calculating the adjustments, but the chosen stations used for calculating the adjustments is unknown except for interpQM. Given your comment we reworked the methodology section to make this more clear.

81: So, probably autocorrelation is present. The authors should discuss this point both in the analysis of the following results as well as in the methodology.

Thanks for bringing this up as autocorrelation is an important topic in time series analysis. We analyze seasonal values, therefore autocorrelation would only be present in the case that snow did not completely melt between seasons. This is not the case in any of our stations. We performed an autocorrelation-analysis prior our analysis for all indicators for the lags 1 - 10. It showed that the correlations are generally low and therefore a pre-whitening, e.g. from Yue et al. (2002, DOI: 10.1002/hyp.1095) is not necessary. A short summary of the correlations for lag 1 of each indicator analyzed is shown in the table below (n = 42 stations):

Table 1: Summary of autocorrelations for lag (year) 1 for all stations (n = 42)

Indicator	Minimum	Maximum	range 25. - 75. percentile (IQR)	Mean	Median
HSmean	-0.34	0.36	0.22	0.1	0.1
HSmax	-0.25	0.42	0.24	0.03	0
dHS5	-0.1	0.48	0.14	0.18	0.21
dHS30	-0.29	0.36	0.15	0.06	0.06
dHS50	-0.3	0.26	0.16	0.03	0.06

Daily snow depth time series inherit a very strong autocorrelation. In this study, seasonal indicators were analysed, which imply that autocorrelation would be present in cases that the snowcover did not completely melt between seasons. However, this is not the case in any of the selected stations and for any of the seasons analysed. In order to prove our hypothesis, autocorrelation was investigated for the lags 1 - 10 years. A summary for lag 1, where it would be strongest, is presented in table 1. The results showed that autocorrelation is very low and a Trend-Free-prewhitening (e.g. Yue_2002) was not necessary.

83: I found this sentence more confusing than explanator. I suggest to remove it.
Changed from “The analysis described in Buchmann et al. (2022) is re-run with the three additional Austrian data series along the eastern Swiss border to increase the number of available reference stations for Eastern Swiss series, thus increasing the 85 valid break points to 45 (from the original 43 mentioned in Buchmann et al. (2022)).” to “To improve the station density near the eastern Swiss border, three Austrian stations were added.”

Figure 1: Please add this information in a legend as well.
Done.

95: It is fundamental in my view to add the equations used for the calculation of the adjusment factors either here, in an appendix or in the supporting information.
Done. We decided that it’s most useful to add the equations and short explanations directly to the main text and slightly improve the corresponding description-paragraphs.

118: Unclear sentence:
“Since it was decided that interpQM does not change the total number of snow days (HS > 0), the thresholds of 5, 30 and 50 cm (dHS5, dHS30, dHS50) were used and summed up for each hydrological year between November and April.”
Changed to: Since interpQM does not change the number of days with snow (HS > 0), no changes at the 1 cm threshold were expected for the homogenized series. Instead, the threshold values 5, 30 and 50 cm were

used and the number of days between November and April were summed up for each hydrological year.

122: Please indicate if autocorrelation is still present (I expect there is still significant autororrelation if seasonal time series are used). Why did not you use a trend test which can handle autocorrelation instead of the standard Mann-Kendall test?
As mentioned (answer to line 81) from the physical standpoint seasonal values of snow are usually not autocorrelated as the snow cover disappears in summer. Nevertheless, we performed an analysis for lags 1 - 10, which showed that autocorrelation in the seasonal time series of the indicators analyzed was very low, so we did not account for it, e.g. with a pre-whitening or a modified MK test. To clarify this in the article, “To avoid the influence of autocorrelation on the seasonal trend analysis, it was examined for all indicators and time series. It was found that the values were low enough (mean 0.03 - 0.18, interquartile range 0.14 - 0.24) that trend-free-pre-whitening or the application of a modified MK-test was not necessary.” was added.

128: Please provide a reference for this choice
Done. Buchmann et al. 2021 (https://doi.org/10.5194/tc-15-4625-2021) and Marty and Blanchet 2012 (https://doi.org/10.1007/s10584-011-0159-9) (100 year return period).

130: Sentence unclear (To assess the significance of the homogenization performed, a Kolmogorov-Smirnov test was conducted with seasonal data.)
Changed the sentence to make it more clear: “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 can be achieved by different tests, e.g. via a two-sample KS-test or a Wilcoxon rank sum test followed by Dunn’s test. Here, we have chosen the two-sample KS-test for this task.

137: why only for these and not also for the aggegated measured time series with HOMER and Climatol?
Because HOMER and Climatol only provide monthly data, but daily data is needed to calculate the number of snow days. We changed the corresponding sentence to make this important detail more clear:
“Trends of days per seson with a certain snow depth (snow days) are only compared for the original data and interpQM, as Climatol and HOMER only provide monthly data.”

141: could you please be more specific? it is hard to follow.
"Climatol automatically fills in missing values, thus artificially increasing the length of the series. However, for the trend analysis, only the time period available in the original series are considered for the homogenized data in order to make it comparable."
changed to
"As the series need to have the same length for Climatol to work, missing dates and their values are automatically interpolated, thus leading to an artificially increased length these series. However, for our analyses we only considered the original length of the series in question."

169: not sure it is the right wording in English.
"This may be irritating, but it has several reasons:…"
changed to:
"There are several reasons for this:"

Figure 3: Here I suggest the use of a better colorbar. In fact, low elevation sites have lower values for the trend and with this colors it seems there is no trend. Maybe one possibility could be to normalize the trend values by the long term mean value.
We adjusted the colorbar to make the colors stronger and differences clearly visible. As we want to highlight the difference between HSavg and HSmax, identical color code for the trend was our choice. Different color sets for HSavg and HSmax, would have been a possibility. But as especially the lower elevation sites (<1000 m) sites only have small trends for HSavg, varying between -4.3 - 0.5 cm/decade, in our view, these small trends would not distinguish better from larger values as with the current color choice. Your suggestion using normalized values is also a good idea dealing with a large span of values, but with the reworked colorbar we think that this should not be necessary.

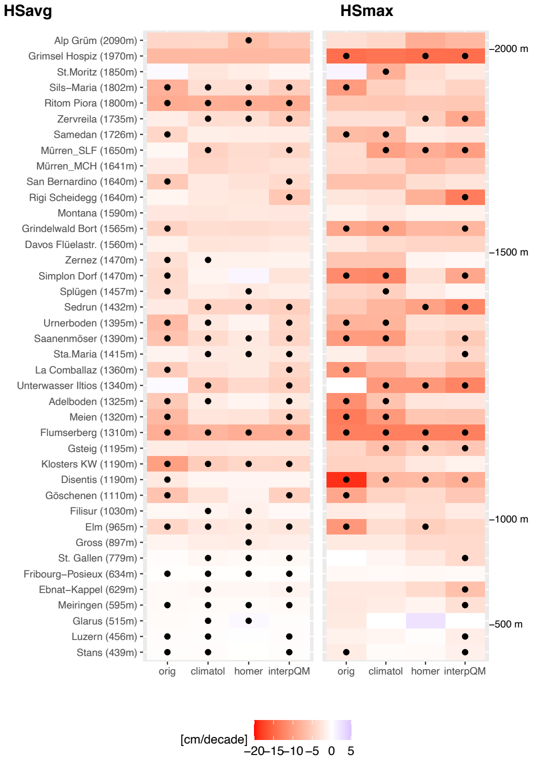
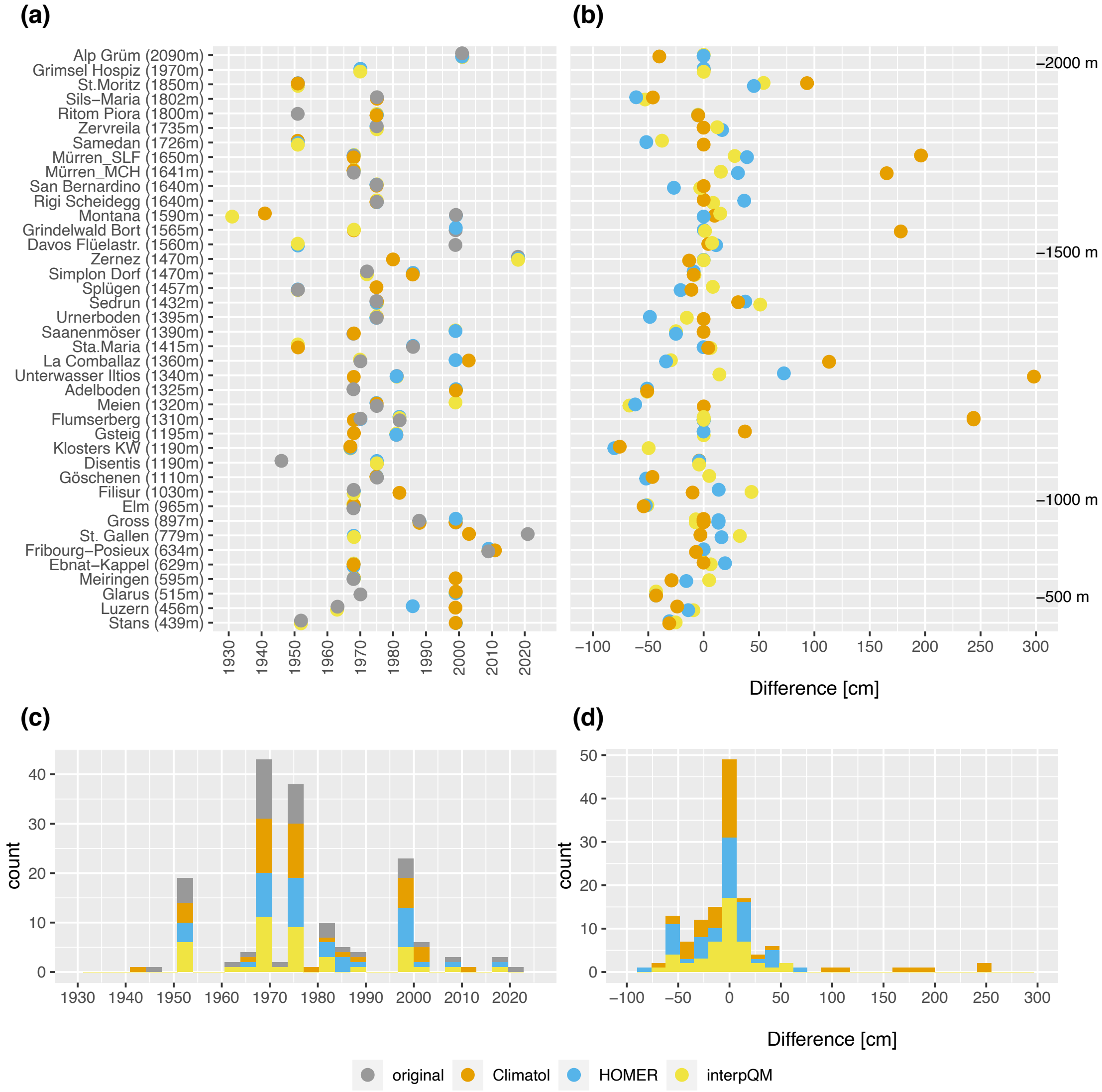


Figure 4: fonts too small.
Font size increased.



250: Here a reference should be added or some additional analysis should support this statement
“Moreover, the size of the adjustment factor is not directly responsible for these large differences, as the impact on the trends is more sensitive on the length and location of the adjusted sub-period than the magnitude.”
We decided to delete this as we neither found reference nor could do additional analysis. It is also not relevant for our results.

252: I do not get this sentence. A multiplicative adjustment factor of 0.1 for instance is much more relevant than a multiplicative adjustment factor of 1.1.
This is of course true. With “smaller than one” we didn’t mean much smaller than 1. The mentioned examples have adjustment factors around 0.6 - 1: Samedan and Sils-Maria (~0.8), Zernez (0.6 - 0.8) and Adelboden (0.8 - 1). Here, we also decided that this sentence is not relevant for our results and therefore deleted it.

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."

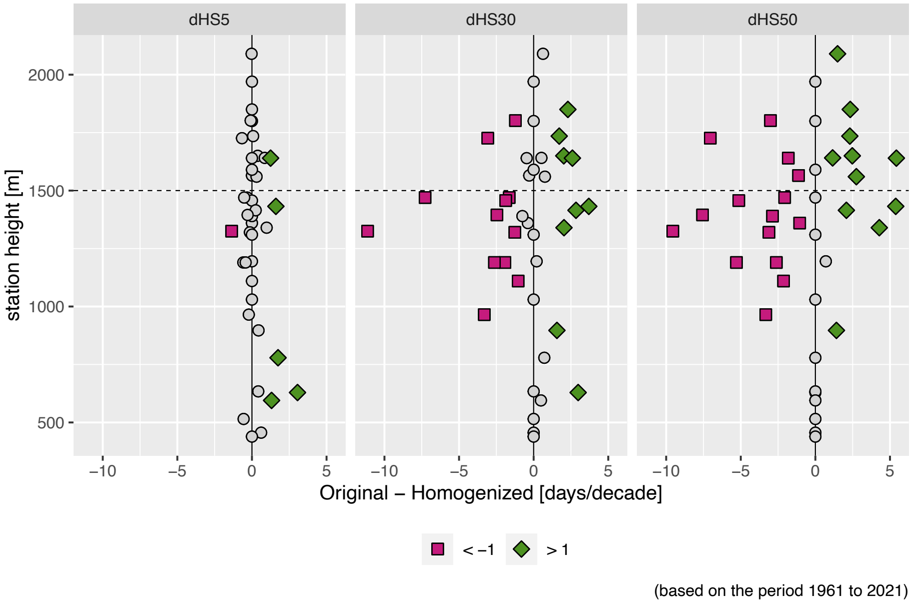
HOMER:
“The adjustment factor is determined by means of variance analysis ANOVA (Caussinus and Mestre, 2004; Mestre et al., 2013) based on the selected reference stations. These are chosen either based on the horizontal distance or first-difference correlations. Due to the large vertical distances between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

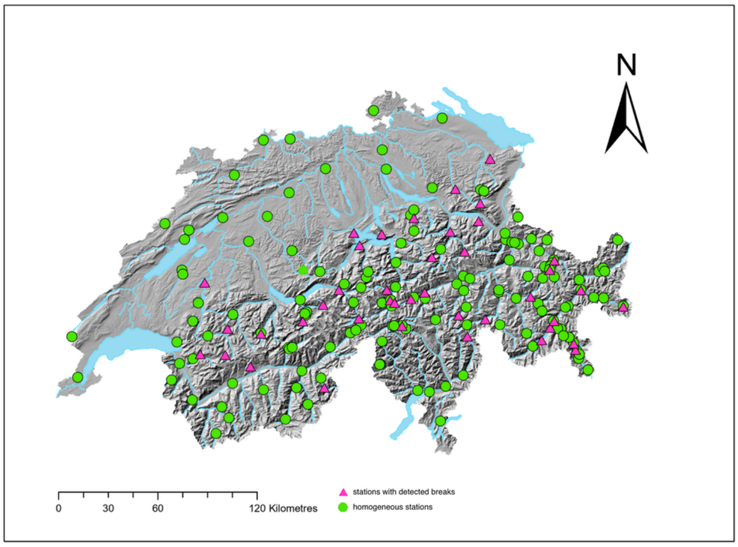
- L174(paragraph): I find the reasoning irritating in this paragraph. Are the trends in snow driven by absolute temperatures? Which the authors imply in L177. And then the statement in L179 is based on what source? Is it stating theory or factual evidence?

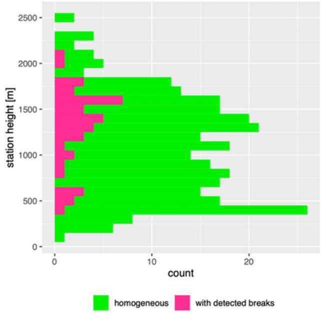
We found that the whole subchapter was a bit inconsistent and difficult to read. It was therefore rewritten. Regarding the argumentation: It was moved from results to discussion in order to achieve a clearer separation of results and interpretation. The statement mentioned was deleted and changed to

"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.