Dear Ross Brown

Thank you for your comments. Please find our responses below in blue.

Dear Authors, a timely and practical piece of work that will help guide folk who wish to examine the homogeneity of in situ snow cover series. My comments are manly minor and editorical in nature, and are included in the attached annotated pdf file.

See detailed comments below

My only criticism of the paper is the lack of independent testing of the three break point methods with synthetic series. Without this, one of the main conclusions to use multiple break point detection methods seems a little weak given that one method was shown to greatly over-identify break points.

We decided against the compilation of a synthetic data set as the magnitude of real breaks observed in the Swiss snow series were unknown. Our aim is to see what these established methods are capable of when applied to real world data. Admittedly, this makes the interpretation of the results more complicated, but also more practicable.

A simple synthetic data set would not be of any particular use, and the compilation of a more sophisticated one would be well beyond the scope of this study. As such benchmark data sets exist for temperature and precipitation it would, without a doubt, be beneficial to have one for snow in the future as well.

A small set of synthetic series was used by our Austrian project partners to assess the performances of two adjustment methods (INTERP and InterpQM), which is currently under review for the International Journal of Climatology (Resch et al. 2022).

The authors make a point to avoid going very far in the correction side of the homogenization process, but this diminishes the interest factor. For example, it would be interesting to show some preliminary metric of the impact of correcting series e.g. on trends in regionally-averaged snow cover series.

We are already working on the follow-up paper focusing on the impact of homogenisation on trends. That is why we intentionally wanted to focus on the detection of break points, rather than going for the corrections and impacts as well. Furthermore, a new correction method is investigated for Austria in Resch et al. (2022), which is currently under review for JOC. The results of that study will be incorporated in the follow-up paper.

The paper was well-written, easy to follow and a pleasure to review. Best regards, Ross D Brown, Canada.

Detailed comments referring to the annotated PDF

L3:
rephrased accordingly to: have rarely been applied to snow cover related time series.

L5:
added: monthly

L5/6:
rephrased accordingly: The multi-method approach allowed us to compare the different methods and to establish more robust results using a consensus of at least two change points in close proximity to each other.

L8:
Sentence rephrased: of which

L10: This suggests break points are mostly random features. Did you apply any method to assess the probability distribution of shifts to get some idea of their statistical significance? e.g. shift detections in randomly generated series versus observed series using a bootstrapping approach.
No, we did not. However, Figure 3 shows the applied corrections as a metric for the size of the detected breaks. The impact on significance (as well as trends) will be addressed in a follow-up study already in preparation.

L17: This seems a rather round about way to say that the data series have to be adjusted to remove the inhomogeneities - that is the real challenge; detecting the changes is the easy bit!
We do not agree that detecting break points is easy. From a strictly technical point of view that maybe true, however, the interpretation of the results from these methods is far from easy as this study shows. We highlighted that even the "easy" part of the homogenisation process is actually far from it. Furthermore, this is the reason why we only focused on break point detection and not included trends and further impact analyses, which are investigated in a follow-up study.

L66: Hard to see why there would be an elevation dependence in break point detection. Can you elaborate a bit on this.
The availability (and quality) of reference series is key for proper break point detections. Availability of snow in Switzerland is highly elevation dependent and our stations range from 200 to 2500 m a.s.l. To test the hypothesis that lower stations might not have enough good-quality reference series to properly detect break points, elevation dependence is an easy way to do so.

L67: Are the detected break points consistent with changes inferred from available metadata?
Sentence rephrased accordingly

L69: This sentence is a bit confusing. I think what you are trying to say is "Are break point results similar for different snow cover variables".
Rephrased as suggested

L83: Table 3?
Yes, table 3, rectified.

L85: This is confusing as annual series were used in the break point detection analysis.
Yes, break points are detected using annual means, but the methods themselves require monthly input data to run properly.

L86: The final outputs used in the break point analysis were annual time series of winter mean snow depth (HSavg) and snow cover duration (dHS1).
Rephrased as suggested

Fig2: Is this the combined result for HSavg and dHS1?
No, only HSavg. Changed caption

Table 2:
Rephrased caption to: [..] A valid break point means it was detected by 2 out of 3 methods. The complementary category uses break points from both dHS1 and HSavg (six break points are identical).

L193: Insert „In“
„In“ is already there. L192, last word

L195: Do HOMER breaks agree better with the metadata? I would be suspicious about methods that yield high frequencies of small shifts. You could use synthetic series with known shifts to test each method and the performance of a method consensus.
As the metadata is not perfect (i.e. neither complete nor perfectly accurate) (see Section 2.1) any ranking efforts concerning the individual methods are not robust.
Our reasons for not using synthetic series can be found below in the response to L368.

L207: Is this related to the number of stations in each elevation band? I wonder if the interannual variability is higher in the 1300-1700 m elevation range which may yield more frequent breaks.
The reason why we found more breaks in the 1300-1700 m a.s.l. elevation band than below or above is, in our opinion, due to the availability of potential and suitable reference series.

L210: Figure 5 is rather unremarkable as expected - not entirely clear how elevation would play a role in break points.
Since our network spans from 200 to 2500 m a.s.l., elevation could be an issue. See the reply given to L66 above. Especially since the availability (or lack) of reference series is key to a proper break point detection.

Fig 4: interesting that the break points are normally distributed which suggests the shifts are pretty random without any evidence of important systematic ruptures. What do the averaged time series of normalized corrections look like for each method? This would also give some indication of the potential impact of the corrections on regionally-averaged snow cover time series.
We purposely avoided to cover potential impacts and go into too much detail about the corrections, as that is the topic of the aforementioned follow-up study.

L368: An argument for using synthetic series to test the methods.
Yes, synthetic series could help, but only if they are close enough to reality. A set of very simple synthetic series would not separate the wheat from the chaff, as all the methods would probably detect simple single shifts. The synthetic compilation of complex real-world series is far beyond the scope of this study.

Dear reviewer, thank you for your comments and suggestions. Please find our replies below in blue.

The authors have performed homogeneity testing on snow series using three different methods on two different snow variables. The paper is well written with good structure and nicely presented results. It is a good addition to the scientific literature on homogenisation of snow depth series. However, I do think it would benefit the paper to include some results that shows the impacts of the homogenisation (difference between raw and homogenised series, trends…).

We understand the interest for trends and differences between raw and homogenised data and agree that such an analysis is relevant and important. Our reasons for not including these analyses in the current study are three-fold:
- In order to obtain homogenised data, the series have to be adjusted at the break points. The adjustment process itself is far from straightforward (and so far, not really looked at for snow) and including it in this study would not have done it justice.
- We considered the topic too important and complicated for a mere paragraph, so we decided to give it its own dedicated article, which is currently being prepared. In undertaking this process, we are currently experiencing large method dependent differences, which merit further investigation and more detailed reporting to share with the community.
- Including trends or differences of homogenised series would also have shifted the focus away from the break point detection problem.

I agree with reviewer 1 about it not being clear why break detection depends on elevation. I suggest adding a version of what you answered reviewer 1 to ch. 4.1.3 or another suiting place in the paper to explain your motivation for looking into this.

Added information to 4.1.3 and rephrased paragraph to:
As the availability (and quality) of suitable reference series for each candidate station is key for a proper break point detection and our stations range from 200 to 2500 m a.s.l. To test the hypothesis that lower stations might not have enough suitable reference series for proper break point detections a possible elevation dependence is investigated. To do so, a possible altitudinal or amount-of-snow influence on the break detection capability of the methods, the break [....]

Fig. 8 could use some refinement:

What I understand from the figure text is that valid breakpoints detected by two methods are shown in grey but breaks only detected by one method are also shown in grey. This is confusing. And are breaks detected by only one method shown in the figure?
Thanks, the figure caption was wrong, only valid breaks are shown.
Comparison of valid break points found for HSavg and dHS1. Valid break points, detected by at least two methods, are coloured grey (dHS1), yellow (HSavg), or blue (both). The shape indicates which method detected the break and the size is the corresponding break magnitude correction.

- In the legend I suggest using a color other than grey for the circle, square and rectangle (for acmant, climatol and homer) or another color than grey for dHS1, whichever is easiest.
  Colours removed from the shape symbols (acmant, climatol, homer) to better distinguish between methods and variables.
- In addition, there is a “NA” in the legend that looks out of place.
  Yes, done.

Please see the attached pdf for the rest of my mostly technical and minor comments.

Comments from the PDF

L7: 45 breaks in how many of the 184 series? (i.e. please add some information about how many of the 184 series were classified as inhomogenous)
Sentence rephrased to: 45 valid break points in 41 of the 184 series investigated.

L10: done

L20: done

L83: done

L120: done

L125: Should there be a "plusminus"-sign in front of 2?
Yes, rectified

Table 1:
Done

173: In this paragraph, could you also please mention the complementary use of the two variables HSavg and dHS1 and what you want to evaluate by using the complementary approach?
Sentence added: The main focus is on breaks in HSavg, however, the opportunity to use dHS1 alongside as a complementary break point detection approach is discussed in Section 4.3.

192: Please rephrase/clarify this sentence. I read this and thought it was not consistent with what Fig. 3A shows (where most Acmant break corrections are between 10-19 %, but where this is not the case for Homer and Climatol). I was confused by the last part of the sentence ("10-19% for all three methods").
Sentence rephrased to: 10-19% (ACMANT and Climatol) and 20-29% for HOMER.

Figure 3: delete text
Done

L210: done

L213: done

Figure 8:
See comments above

Table 3: Please briefly explain the "code" and "combination" column in the table text.
Changed code to Name and added combination information in the caption. Rephrased to:
Table 3: List of valid break points concordantly identified in 184 investigated Swiss snow time series. Stations are ordered according to break year
and series with break points identified in both HSavg and dHS1 are marked bold. Combination code refers to ACMANT (A), Climatol (C), and HOMER (H).

L261: Instead of the dash, should it be plusminus or ~? Also, add space beween 2 years.
Yes, rectified

L317: changed to:
the six break points identified in both HSavg and dHS1 reveals

L331: rephrased sentence to:
[..] based on a combination of reasons.

L386: 25 % of what? Please clarify.
Rephrased in relation to the total number of series investigated:
We identified 45 valid break points in 41 of 184 (22%) series investigated using a complementary approach [...]

Dear referee

Thank you for your comments and suggestions. Please find our replies below in blue.

The paper is well written. The presentation is clear and well structured. It has been a pleasure to read it.

The paper presents a study assessing the performance of three different applications for identifying homogeneity breaks in snow depth time series. The analysis and results are well described and discussed. My main concern, which is not very large, is related to the comparison of the applications. There are obviously several subjective choices to take in applying both Homer and Climatol, and less in Acmant. One issue the authors mention, and which I recommend to pursue more, is the role of the reference network and how they are established, and how that influences the results. I have a feeling that this is as well an important issue as the break detection algorithm itself. In order to have a fair, consistent comparison of the break detection capabilities I would therefore challenge the authors to apply the three methods with the same reference networks. Would it be possible to apply them on all three combinations of networks?

We agree that selecting/building suitable reference series is as important as the mere detection of change points itself. However, the network-selection process is an intrinsic component of ACMANT, Climatol, and HOMER and cannot be separated.

In ACMANT and Climatol the options for the user specifying the reference (sub) networks and their selection basis are limited. In addition, Climatol uses a geographical distance criterion as the default to select localised reference networks, whereas ACMANT uses a pre-set correlation threshold of 0.4. While it is true that HOMER does allow a geographical distance selection option there is no easy way to match these to the Climatol selections, as HOMER does not consider the vertical distances.
We discuss the different correlation thresholds we applied in HOMER (0.8) and ACMANT (0.4) and the possible implications of this in Section 5.1.

As it is not possible to exclude the network-building from these methods and simple run the break point detections alone, we choose the settings to be as realistic as possible for each method. However, as we are looking for a practical solution and comparison, (and accepting the network-building as an intrinsic component of the method) running the three methods with the same input data is as close to fairness as feasible.

Minor comments:

Line 27: Typo: contemporay → contemporary.
Done

Figure 7: Difficult to read. Use a lighter shade for the terrain, and make the black dots smaller. Consider dark-grey contour lines (rivers, borders).
Changed the contrast to increase legibility