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.