An extensive set of snow depth measurements from the European Alps is evaluated to explore snow climatology regions and temporal trends. Raw data is collected from various sources and subsequently harmonized, quality controlled and temporal gaps were filled. Based on the snow depth time series, five regions with different snow climates are derived using Principal Component Analysis and k-means clustering. Linear long-term trends and short-term trend variability are computed from the snow depth time series for the five derived regions. Finally, snow depth observations are compared with gridded air temperature and precipitation data, which is either reanalysis based or inferred from spatially interpolating observational data. The authors find decreasing trends of snow depth for the majority of the stations and substantial differences in trends between regions.

This study addressed a very relevant topic: the evolution of snow depth in the European Alps during the last half century. The authors quantify these changes based on an extensive compilation of in-situ time series of snow depth from different regions. Additionally, the authors provide the unified dataset as an online resource, which can be very useful for further applications. The manuscript is well written and applied data and methods are sufficiently explained. My comments concern therefore primarily smaller ambiguities and (potential) errors.

We thank you for the detailed review of our manuscript, the appreciation of our work, and the constructive comments, which we shall address in a revised version.

General/major comments

• Some sentences are extremely long and should be shortened to increase readability.

We shortened sentences as much as possible throughout the manuscript.

• Selection of different time periods

I'm confused about the different time periods that were used for this study: The PCA and k-means clustering was performed for 1981 – 2010, long-term trends were computed for 1971 – 2019 and shortterm trends for 1961 – 2019. Could you elaborate in more detail why you use (these) three periods?

We chose different periods because of the different nature of the analysis. The PCA and k-means aimed at having the largest spatial density, so we chose the 30-year period with the largest number of stations. The long-term trend analysis aimed at a tradeoff between coverage of stations and as long as possible period. The moving-window short-term trend analysis will be removed in the revision (see responses to Ross Brown's review comments). We shall add an explanation also in the manuscript and provide more information in the data overview sections on these two periods and station subsets.

• Which data (raw, gap-filled) was used for which analysis/plot?

In the manuscript, I was sometimes a bit confused which data was used for which application. E.g. in section 2.4 you use the raw data without any gap filling, right? But for the subsequent analyses you always use the gap-filled data (as explained in appendix section A.3)? This point should be more explicitly stated in my opinion.

Thanks for pointing out this ambiguity. Actually, we used the gap filled data for all analyses. We now explicitly state this in all related method sections (overview as well as statistical analyses).

• Gap filling method

I struggled to follow the explanation of the gap filling method – particularly from line 641 to 654:

- I do not understand what "crossing a calendar day window with a year window" means.
- Shouldn't the "window of 31 days" be "window of gap length + 30 days"?
- "mean of the daily values" -> does this refer to the daily climatology?

- “and the weights were based on the vertical distance between candidate and reference station.” àwhy are horizonal distances not considered?

- Finally, I wondered if reconstructed values have a smaller temporal variability (on smaller scales) because you apply climatological values in your method. But this is not the case, right? Because you only compute the ratios from climatological values (daily means)?

We rewrote the description of the gap filling in an algorithmic way and also provide an additional explanatory figure. We hope that this helps to clarify our procedure. Regarding your specific comments:

- The horizontal distance was only considered for selecting candidate stations and not for the weighting, because we wanted to have univariate weights and not multivariate. In that case, we found the vertical distance to be more important than the horizontal distance.
- The temporal variability of the reconstructed series should not be affected. The climatological values are only used to calculate the ratios. The reconstructed values is then based on the daily value(s) of the neighbouring series. Consequently, the daily variability in the reconstructed series stems from the daily variability in the reference series’. The only loss of variability could occur because the reconstructed value is an average of up to 5 values from up to 5 reference stations.

Point comments

L150-151: I don’t understand this sentence – could you rephrase it?

Done.

L157-158: this sentence is a bit odd: why do three different climate forcing zones create four main climate regions?

Yes, it is counterintuitive. However, the three forcings combined with the topographic effects result in gradients along the North-South and East-West directions. And these two gradients, if intersected, result in four regions. We modified the wording accordingly.

L179-189: I struggled to read this sentence (because it is so long). To increase readability, the providers from Italy could e.g. be listed with bullet points.

As suggested, we restructured the data providers of Italy as bullet points.

L195-196: “with a few expectations of monthly/seasonal data form the HZB and SMI.” -> was the monthly/seasonal data also used in this study?

Actually, no. We removed this part of the sentence, to avoid misunderstandings.

L196-199: I’m confused by this sentence: automatic measurements are used both from France and the Aosta Valley, right? Then I would change “only for France or” to “only for France and”. Anyway, I think the sentence should be rephrased to improve comprehensibility.

The sentence was split and rephrased.

L205: “(see also Fig. 2b)” should be “(see also Fig. 2c)”, right?

True. Thanks for spotting the error.

L245: why is the data for Austria only available until 2016?

Because of the processing and quality checking performed by the Austrian Hydrographical Service, it takes some time until the data are published online (we added this information also in the revised manuscript). In our case, when we accessed the data (early 2020), only data until 2016 were available. Since a few months, the records have been updated to include 2017, however, we cannot manage to re-analyze this update in the revision and, furthermore, results are not expected to change significantly including a one year extension of the Austrian series.

L253: "The frequency by elevation (50 m bins were used to calculate to proportion)…" -> I don't understand this part.

We plotted so-called frequency polygons, which are basically histograms shown as lines and not bars. This makes it easier to compare different distributions. But, as for histograms, it is necessary to specify the bin width. We tried to improve the caption to make it more understandable.

L257-258: why is this criterion applied? Wouldn't e.g. a threshold of 50% make more sense?

Yes, a different threshold would also make sense. However, then, also the meaning of the figure would be altered. We chose this simple threshold to show availability of stations per year. Another option would be to show stations with the threshold you proposed, but then it also depends on which season the 50% apply (Dec-Feb, Nov-May, or Oct-Sep). Yet another option would have been to show only the stations used in the analysis. We decided to stick to our simple threshold, and tried to clarify the intent.

L273: did you consider to use the "elbow method" to find the optimal number of clusters?

We shall look at the elbow method for the revision. We also compared it to the average silhouette coefficients, which we used initially. The elbow method identifies 4 clusters as optimal if we apply the k-means on the scaled observations, and 5 clusters if we apply it on the PC matrix. The silhouette analysis identifies 4 clusters as optimal, if applied on the scaled observations. If applied on the PC matrix, then 2 clusters are optimal, followed closely by 5 clusters. We checked maps of the clustering results for all combinations (see figure below; we also shall add it in the supplementary material), and, to be honest, all choices of clusters make "sense". The different number of clusters and whether observations or PC results are the input, all highlight different aspects of the snow depths and their hierarchy (e.g. as seen by increasing the number of clusters). These are the elevation, North-South gradient, and East-West gradient. We finally decided to leave our analysis as it was with 5 clusters, because they agree best with our knowledge of the climatic and topographical drivers of snow depth in the Alps. This information on this process shall also be added in the revised methods section.

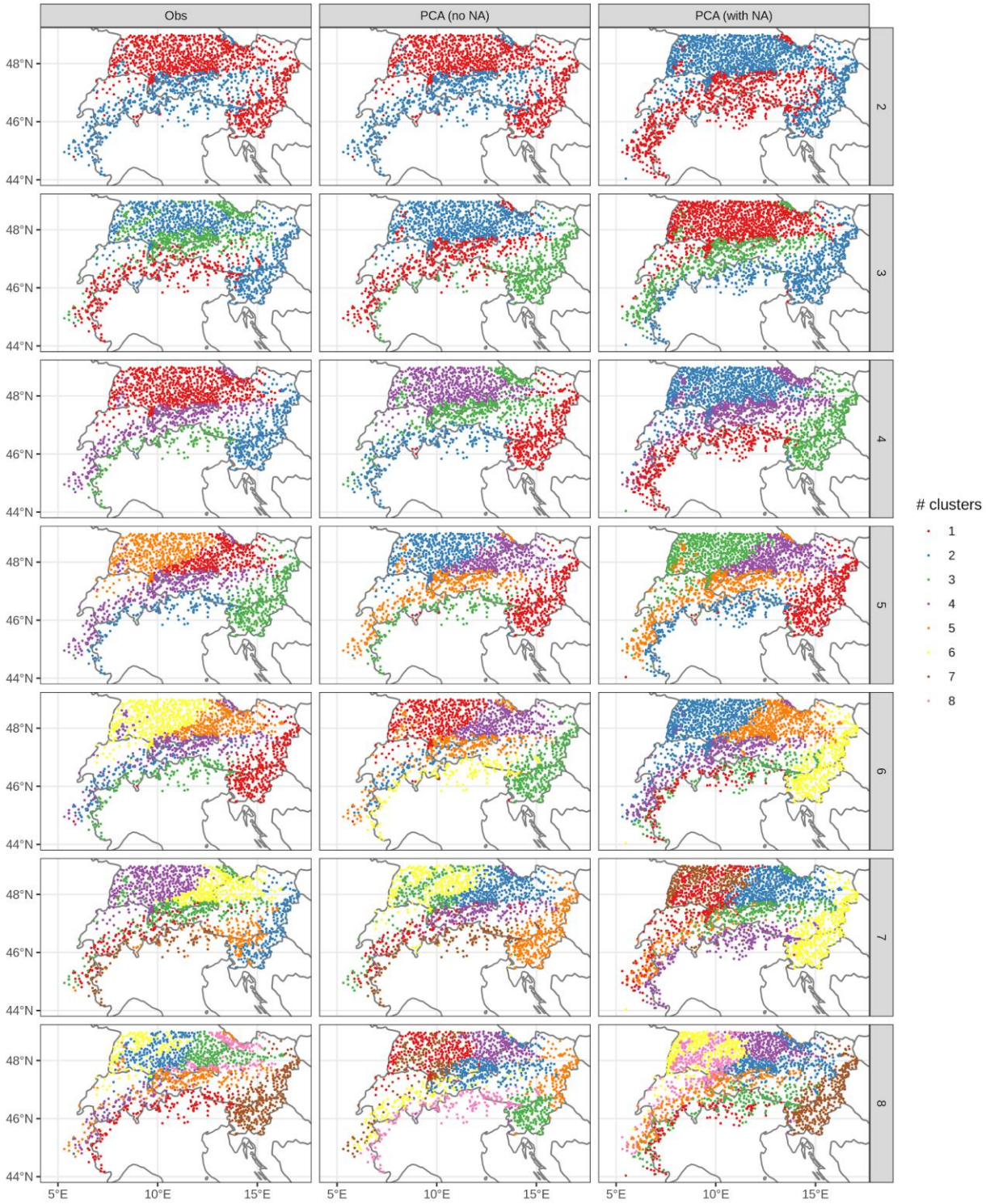

Figure R1: Results of k-means clustering. Rows show the number of clusters (order is arbitrary, so colors might not match within a row). Columns identify the input matrix for the clustering algorithm: For Obs scaled daily observations of snow depth were used and for PCA the PCA matrix. The two PCA columns stand for the standard PCA that does not allow missing values (no NA), which corresponds to the same station set as in Obs, and the second PCA column is for the modified PCA algorithm that allows missing values (with NA) and has a higher station coverage.

L273-274: what do you mean by "as well as clustering directly on the daily observations."?

We applied the k-means clustering on the PCA matrix with the principal components as input, and, as comparison, we also applied the k-means clustering directly on a (scaled) matrix of daily snow depth observation. We tried to make this clearer in the manuscript.

L287-288: remove "using assessed"

Done.

L337: "and is highly correlated to elevation up to 1000-1500 m, and mostly constant above." I don't understand this part.

What we meant is that the PC2 is related to elevations up to 1000-1500m. We removed the word "correlated" and rewrote the sentences to make it clearer that PC1 and PC2 are both elevation driven, but PC1 mostly >1000m and PC2 mostly <1000m..

L348: the clustering was performed on the 5-dimensional PCA loadings, right?

Yes. We rewrote the sentence to make it clearer.

L367-368: please provide a reference here for the HISTALP subregions

Done.

L370: what you do mean by "estimated data-driven"?

We meant that the clustering was performed automatically using snow depth data, and no manual re-assignment or modification was done on the clustering results. We modified the sentence in the manuscript accordingly.

L377: what do you mean by "unique stations"?

Unique in the sense that they have no similar neighbours in any cluster. We now added this information in the manuscript.

L386: I would not write "matches" here because there are 4 vs. 5 regions. I would rather write that the obtained regions are similar.

Rewrote according to your suggestion.

L404: does this statement refer to a specific month? Or the entire winter?

To the whole winter. Clarified it also in the manuscript.

L408-409: "mean North & high Alpine" is odd. Do you mean: "While in December, the mean negative trend was stronger in North & high Alpine"?

Thanks for pointing this out. We rewrote the sentence.

L422: "Points with lines indicate" -> "Lines indicate" (or are there points with no lines?)

No, you are right, all points have lines, even though they might be hardly visible. We rewrote the legend.

L467-468: "Moreover, we assume that most of this seasonal imbalance is because there is no or no significant snow cover in that month" I don't understand this part

We meant the seasonal imbalance of station observations, since some low elevation stations do not record outside the winter season. We rewrote the sentence accordingly.

L493: I guess "~100m" should be "~100 mm"

Yes, thank you. We also changed it to the ≈ sign.

L515-516: "because this implies less chances that precipitation falls at the "right" time." I don't understand this part

We replaced "right time" with "concurrent with low temperatures".

L532-533: I'm not sure if I understand this sentence correctly. Do you mean homogenization is not so important because such a large number of stations is used?

Partly. Homogenization is important. But given the extent of our dataset it was impossible to apply a common framework to all the data. We made this clearer in the manuscript.

L589: how do you define a network? A country?

By data source i.e. data provider. Clarified it in the manuscript.

L612: how is this surrounding band defined in terms of horizontal distance?

Horizontal distance was not considered here. We added this in the manuscript.

L673-674: I don't' understand this sentence.

We assume you refer to the explanation why our relative MAE is not a "true" MAE. In the standard way, the relative MAE is defined as $\frac{1}{n}\sum_{i=1}^{n} \left|\frac{y_i - x_i}{x_i}\right|$, while our modification is $\frac{1}{n}\sum_{i=1}^{n} |y_i - x_i| / |\underline{x}|$, where $\underline{x}$ is the average of all x_i. We added the formulas also in the manuscript to make it clearer.

L686: what is meant by "ablation scheme of the different stations"?

The local climatic and topographic characteristics that influence ablation. We modified the sentence accordingly.

L701-702: "gap filling snow depth series using simulations of the Crocus snow model for the French Alps" -> I'm a bit confused by this part. Does it state that gap filling was performed by running the Crocus snow model with meteorological forcing?

Yes. Actually, the Crocus simulations were performed independently of the gap-filling used for this study, but we found it interesting to compare the two approaches, since they were both available at the same sites. We added some more information on this in the manuscript.

L721: what is meant by "original observations"? Available observations?

The observations available before gap filling. Modified the sentence to make it clearer.

Figure 1b: this panel is hard to read. Could you enlarge it? To increase readability, station density could be plotted instead (i.e. the number of stations per a certain area).

We splitted the figure in sub panels (for available and used), and put the station density underneath the points for a 0.5*0.25 deg grid. We still show points, because we think they are important.

Figure 2b: how was the polygon for the DEM generated? With a convex hull?

No, a manual outlining along the stations, because a convex hull would include most of the Po valley in Italy, for which we do not have any stations.. We clarified this in the caption.

Table 2: there are typos in the first row (e.g. "(0,1000] m").

We modified the elevation intervals to say “Elevation: (0,1000] m” etc.

Table 3: the spacing between the columns should be improved (-> it is currently confusing that the columns “DJF #” and “MAM mean (min, max)” are so closed together)

We aligned all numeric columns to the right, so the spacing should be better now.

Figure A1: I guess a subset of stations was used to produce this figure, right?

Yes, it is explained in the text, and we added this information to the figure caption.

Figure A2: which statistical quantities (percentiles, outliers, etc.) do the points, lines and box edges represent?

We added this information in the figure caption.

Figure B4: how is the numerical quantity “silhouette width” computed?

With silhouette width we mean the silhouette value or coefficients, sometimes also called width. We modified the figure and provide the formulas for how the silhouette is calculated in the methods section.

Figure C4: For completeness, the table for MESCAN-SURFEX (March to May) should also be shown

Added.

Stylistic comments and typos

L120: “1960–1990. (Lejeune et al., 2019).” -> “1960–1990 (Lejeune et al., 2019).”

Modified.

L144: “while Section 4 concludes.” -> “while Section 4 convers conclusions.”

Modified to “Section 4 provides conclusions”.

L147: “with their typical arc-shaped” -> “with their arc-shaped”

Modified.

L231-232: change to “Station numbers are shown for fresh snow (HN) and snow depth (HS) time series.”

Modified.

L279-280: I would remove the line break here.

Thanks, but since we removed the moving window analysis in the revision, this comment resolved itself.

L312: “significantly to the” -> “significantly from the”

Modified.

L348: “There were” -> “This yielded”

Modified.

L353-354: “South of the main ridge, there were two regions:” -> “Two regions emerge south of the main ridge:”

Modified.

L362: “as has in the north” -> “analogue to the north”

Modified.

L376: “the station in common, and the same common stations” -> “the stations, and the same stations”

Modified.

L400: I would rather use present tense here (and in the following lines).

We noticed our inconsistent use of present and past tense in the results. We mostly prefer past tense for results, and adapt the complete results section accordingly.

L469: remove “supposed to be”

Modified.

L561: there seems to be a space in the word “scientific”

Yes, it looks like this in the PDF of the paper, but in the Word version everything is fine.

L661: “were useful” -> “are justified”

Modified.

L664: “has not been yet used” -> “has not yet been used

Modified.