Switanek et al. provide an analysis of the multivariate dependence of relative snow depth anomalies over the Austrian and Swiss Alps to temperature and precipitation anomalies. Besides showing past trends of relative snow depth trends, they use the estimated sensitivities to predict snow depth and compare it to a degree-day snow model. The multivariate approach is interesting and has a lot of potential for understanding past changes and predicting future changes. However, some major reservations need to be addressed or discussed first. Finally, it is unclear what the paper is mainly about. I tended to follow what was written in the title. But there are also other elements within that need to be linked to the research aims (a lot of trend analysis of relative changes and comparison to a degree-day model).

The paper's structure is somewhat unfamiliar, because it does not follow the standard approach of intro, methods, results, discussion, but instead guides the reader through a research journey with a lot of motivation used, e.g., in the methods description. Personally, I enjoyed reading it. But, a major drawback is that methods are sometimes difficult to find, since they are spread out. Furthermore important elements are missing, the research questions/aims and the discussion. I honestly don't know, if I should recommend a standard paper structure or not, but definitely the missing components need to be added.

The authors would like to thank the reviewer for their time and effort in providing useful feedback concerning our paper. The reviewer has made a comment about the structure of our paper. We would like to be clear that our paper does contain the standard sections mentioned by the reviewer (i.e., intro, methods, results, discussion/conclusion). The last paragraph in our Introduction outlines the primary focus of the paper. The reviewer has questioned: What is the paper mainly about? Our main goal of the paper is to use observational records to show the sensitivity of snow depth to temperature and precipitation anomalies at different elevations. And furthermore, we show that these empirical-statistical relationships are quite robust over longer periods of time, and as a result we can use historical sensitivities to make surprisingly skillful forecasts of "future" snow depth. One could use a physically-based model to investigate these sensitivities, but they might not align with the observational records themselves. Therefore, we use the observational data itself to inform us and produce a data-driven model to better quantify these sensitivities. While that is the main focus of the paper, we do also provide some additional trend analysis in order to provide the specific context for the data we used in our study. The second additional component of the paper is the comparison of the forecasts from our proposed methodology to an existing model, SNOWGRID-CL. The authors find these additions to be strengths, rather than a distraction, from the paper. However, if it is seen as beneficial to the paper to remove anything relating to the observed historical trends, we could proceed in that direction.

Major points

1. I would expect temperature and precipitation to have different effects in the accumulation and ablation phases of the snow cover. But in your model, using seasonal averages, accumulation and ablation are treated together. Did you perform tests for differences in sensitivities between start and end of the snow season?

   This is a good observation of the reviewer. While the authors agree that greater model complexity has the potential to further improve forecasts, that is precisely what we are trying to avoid in this paper. The main objective of the paper is to show how effective a simplified data-driven, empirical-statistical model performs in making forecasts of long-term changes to snow depth. We consider some of these simplifications (e.g., seasonal averages of our predictors, or using a type of localized linear regression model) to be a strength. This allows researchers and other end-users to very easily visualize how different combinations of changes in precipitation and temperature would be projected to translate into changes in snow depth. As we state at L369: "The SnowSens model is not to be seen as a replacement for physically-based models such as the SNOWGRID-CL." We show how large simplifications can still provide very useful and skillful forecasts, most especially concerning long-term trends averaged over elevation bands.

2. One major drawback of your method is the strong need for extrapolation of the sensitivities in "unknown" climatological terrain. In my opinion, the chosen approach using local linear regression produces unrealistic values, especially at the boundaries and beyond the training domain (Fig 5a-d). Moreover, it smoothes out a lot of local effects (Fig 5 comparing the different columns); this might be a reason why SnowSens does not capture interannual variability. I don't know a simple remedy to this, but at least this needs to be discussed.

   Thank you for this comment, and the authors appreciate your skepticism. It is true that we use extrapolation in our methodology. To be clear, the sensitivities, shown in Figure 5, are constructed only using data over the calibration period 1902-1971. This same period is also used to calibrate our SnowSens model. Then, forecasts of snow depth are evaluated over the validation period 1972-2021. Therefore, if the model was systematically producing unrealistic values, then that would adversely affect our skill

measures. We do not find this to be the case. As stated in the paper, we find the trends of the band-averaged forecasts to track very well with observations over the 1972-2021 validation period (see Figure 11).

And yes, we have already pointed out (L341) that the SnowSens model does underestimate the observed interannual variability for any given individual station. Perhaps the authors can do a better job stressing in our revised version of the paper the most appropriate application of our proposed methodology. In our revision, we would more strongly recommend that a user of our methodology should not place too much weight on the forecasts for any one station or any one point location, but rather should focus more on band-averaged forecasts or using some other type of aggregation (see the next paragraph). For the paper, we wanted to be transparent about how the skill of the SnowSens model compares to something like the SNOWGRID-CL model. Therefore, we initially show the interannual skill at the station level.

Here would be a good place to discuss the extrapolation that we use in our model. Later, the reviewer has this comment when discussing L210: “Personally, I would not trust the values far beyond (>1degC, 50% prec) what one sees in Fig 5e-h.” In Figure 1, seen below in this response to the reviewer, we have plotted the cases which fulfilled these criteria. Figure 1a shows the 95 cases where the average seasonal temperature in the validation period was greater than 1.0degC and less than 50% of normal precipitation. One can see that there is not perfect agreement between the individual forecasts and observations. That would be true for any snow model. Though, the error of the SnowSens forecasts are less than half of the climatological forecasts (indicated by RMSE_SS>0.50). When the temperature anomaly is greater than 1.0decC and the precipitation anomaly is less than 50% of normal, the averaged forecasts and averaged observations are both identical, they are both 33% of normal snow depth (see the orange square in Figure 1a). Figure 1b increases the sample size by using a threshold of less than 75% of normal precipitation. This gives us 988 cases. Again, the average forecast error is less than half of climatological forecasts. The average of the forecasts and observations over these cases are 42% and 41%, respectively. So, while we are extrapolating to “unknown” climatological terrain, we find the model is quite capable of performing well in that new terrain, especially when aggregating over a number of cases.

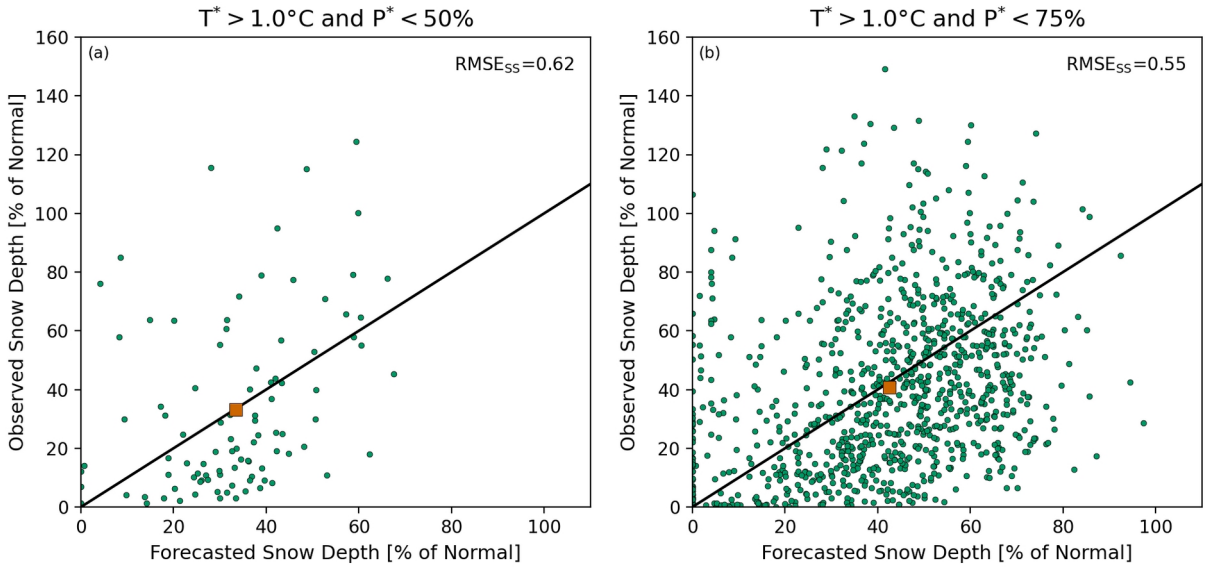

Figure 1: Figure 1a shows the forecasted and observed anomalous snow depths for the 95 cases where the average seasonal temperature in the 1972-2021 validation period was greater than 1.0degC and less than 50% of normal precipitation. Figure 1b increases the sample size by using a threshold of less than 75% of normal precipitation. This gives us 988 cases. The larger orange squares are the forecasted and observed averages over these cases. The skill scores, for these two different criteria, are shown in the top right of the subplots.

3. I understand the choice of elevation bands, but in a changing climate context, I could also imagine a lot of potential for statistical methods to learn across elevation, at least what concerns temperature, given its strong dependence with elevation. However, this probably requires going away from anomalies to absolute temperature and snow depth values. Did you test the multivariate dependency also for “raw”, ie., absolute values of temp, precip and HS? Would it work? Also without subdividing by elevation?

   We would like to thank the reviewer for bringing up this point. In the Conclusions, we state at L374: “If these sensitivities continue to remain persistent into the future, then this modeling approach can be expected to yield skillful forecasts for the next 50 years.” We only used data from 1902-1971 to forecast snow depths for the period 1972-2021. These forecasts were shown to be skillful. As a result, one can logically conclude that the sensitivities over the last 120 years have been reasonably stationary. Given this information, we then propose that these methods could produce skillful forecasts over the next 50 years. In contrast, we are not proposing that the historical sensitivities be applied for the next 200, or 500, or 1000 years. A

user should periodically update the sensitivities, in addition to testing their effectiveness in a cross-validated framework, prior to making another long-range forecast. For example, people in the year 2050 should not solely rely on data from 1902-1971 or 1902-2021 to say something about the future of snow depths. They can, and should, also incorporate data over the more recent period 2022-2049.

The reviewer asked about if “raw” or absolute values can effectively be used. In our study, we found that constructing the SnowSens model using absolute values across either, 1) elevation bands or, 2) all of the stations, produced forecasts that performed substantially worse than the normalized version of the model (the forecasts from the absolute model also performed worse than climatology). We can show why normalization is a critically important step when using our proposed methodological approach. First, take a look at Figures 2a-2c here in this response to the reviewer. The bar plots show the distribution of values for absolute temperature, precipitation, and snow depth for the Austrian and Swiss stations between 500-1000 meters. The average station height of the Austrian stations used is 745m, while it is 742m for the Swiss stations. So, they are not much different in elevation between the two regions. However, one can observe that the Swiss stations are generally warmer and wetter than their Austrian counterparts. At the same time, the Swiss stations have lower seasonal averages of snow depth. Let’s take a further subset of these Austrian and Swiss data points over this 500-1000m elevation band. Those observed data points of the subsets of data are shown as the scatter plots in Figures 2d-2e. A Student’s t-test shows that the means (for temperature, precipitation, and snow depth) of the subset of Austrian data points (Figure 2d) are all statistically significantly different than the subset of Swiss data points (Figure 2e). Looking closely, we find that while this subset of historical observations in Austria has a greater absolute temperature and less absolute precipitation than the Swiss subset, the Austrian stations have significantly more absolute snow depth than the Swiss stations. As we decrease temperature and increase precipitation, we should expect snow depth to increase. However, this is exactly the opposite of what the absolute data is telling us. By simply using the absolute data alone here, we get the wrong signal. This is an example of a regional or spatial climatological difference that we can address through normalization. After normalizing the data, we can better leverage information across a larger region.

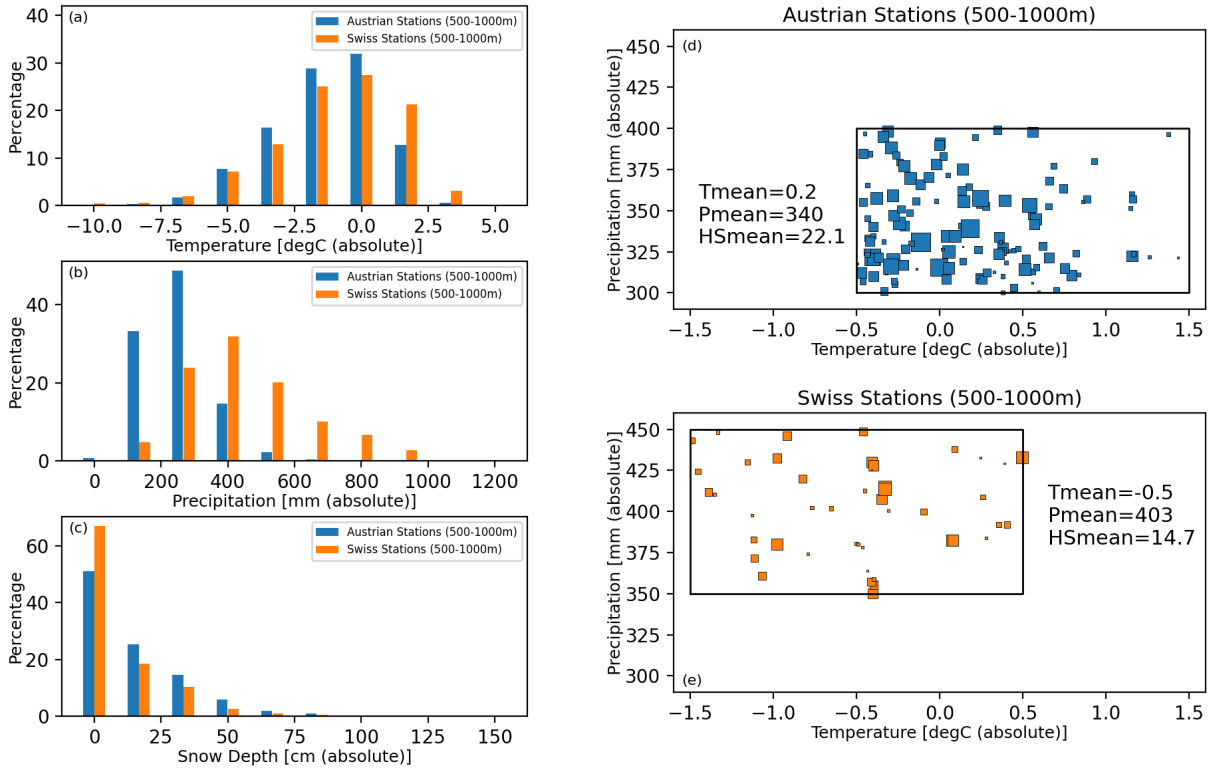

Figure 2: Figures 2a-2c are bar plots that show the distribution of absolute temperature, precipitation, and snow depth for the Austrian and Swiss stations between 500-1000 meters over the historical period 1902-1971. The average station height of the Austrian stations is 745m, while it is 742m for the Swiss stations. The percentages of the blue and orange bars in each subplot (2a-2c) will sum to 100%. The bar plots are comprised of 1,755 observed data points for Austria and 558 data points for Switzerland. A subset of these Austrian and Swiss data points are shown as the scatter plots in Figures 2d and 2e, respectively. The size of the squares reflect the values of absolute snow depth. So the larger the snow depth, the larger the square. The subset of Austrian data points have a greater absolute mean temperature, less absolute mean precipitation, and greater mean absolute snow depth than the Swiss data points.

Minor points:

- L1: What climatic cycles do you? Maybe better rephrase, since climatic cycle can mean something like the Milankovitch cycles.

  Yes, we can make this clearer. Though, Milankovitch cycles operate between tens of thousands and hundreds of thousands of years. Since our paper focuses on a time horizon on the order of ~50 years, it would be unlikely that a reader would be confusing these two.

- L40: Might be worth mentioning doi:10.1002/joc.8002 who also attempted something similar for snowfall

  We can do that.

- L49-54: this belongs into methods. Please provide here a more conceptual statement how you go beyond the state-of-the-art and what your research questions, aims, or hypotheses (choose one) are.

  We respectfully disagree with the reviewer here. It is standard practice to provide a brief outline of what is done in the Introduction of the paper. We do provide one research aim. See our initial response above. The trend analysis provides relevant context for our work, while the forecast comparison to an existing model provides a necessary level of legitimacy of our newly proposed methodology.

- L99: so y_clim,i should not be a time series but a fixed value for every station, right? Maybe state it explicitly. Also your RMSEs are then the average over all stations?

  y_clim at station, i, can be thought of as a single value or a time series array where all values are the same. We will make this more clear. And yes, as the equations 1 and 2 indicate, the averages are performed over all of the stations.

- Sec 3 is more than just methods, it contains a lot of background information and motivation

  See above. It is primarily trend analysis that is the additional component of the paper. We found this to be a valuable addition to provide the necessary context with the specific data that we are using. However, if it would improve the paper to remove this content, we can consider doing this.

- L126: Not sure I agree that Nov-Mar performance should equal to Nov-May. See also Major point 1.

  We respectfully disagree with the reviewer here. One can easily compute Nov-Mar and Nov-May seasonal averages of snow depth. We have done that, and their similarity can be observed in Figure 2b. This does not mean that the April-May average cannot also have its own variability, it is just that the April-May contributions to Nov-May average snow depth are obscured, to a large extent, by the larger contributions from Nov-Mar. Also, keep in mind that we are showing and comparing the similarity of the normalized quantities, and not their absolute quantities. The normalized quantities Nov-Mar (normalized with respect to Nov-Mar, station-by-station) are very strongly related to Nov-May (normalized with respect to Nov-May, station-by-station). Put another way, when the Nov-Mar average

snow depth, for a particular station, was about 20% above average (or 120% of normal), then we can expect that the Nov-May will also be quite close to 20% above average. We compute a mean absolute error between the two normalized seasonal averages of 3.0%. So, on average, a Nov-May percentage anomaly will vary about 3.0% above or below a Nov-Mar percentage anomaly. The Nov-Mar anomalies explain 99% of the variance of the Nov-May anomalies.

- Sec 3.2. is unclear. Please describe better how you performed the interpolation. Eg, “function of the inverse distance”? “adjusted to match”? Also not clear if your interpolation takes into account the effect of elevation? The five nearest stations might not be equally representative in that regard.

  We say at L142 that we use inverse distance weighting. It is true that elevation can influence the absolute values of these meteorological quantities. However, since we use normalized temperature and precipitation anomalies, it doesn’t particularly matter to us or our model what the absolute values of these predictors are. That said, if one were to produce and use “better” predictor data along with our methodology, this should only improve our model performance.

- Related: Why did you not use LAPrec or the gridded HISTALP to extract this information? They use homogenized input, but at least for LAPrec, the spatialization is much more complex and takes topography well into account.

  We made a choice of the data to use for our study and to construct our sensitivity maps. While beyond the scope of our paper, it could be useful for a future study to compare the influence that different data sets have on the results.

- L150ff: Seems like research questions to me, not methods.

  We are providing local context, in this section, for the methods that are being presented.

- L156: which correlation coefficient (Pearson, Spearman)?

  Good point. We use Pearson correlation. We will make that clearer in the revised version of the paper.

- Fig 4: Please do not use rainbow scales, since the changing colors introduce artificial visual breaks. Use a continuous scale such as viridis, scico (https://www.fabiocrameri.ch/colourmaps/), or similar. Moreover, figure looks quite overplotted, maybe it could help to sub-divide by

elevation bins? Ok, I see this comes as Fig5. So maybe in Fig4 you could focus on a few single stations instead or omit?

Thank you for the good suggestions. We can think about how to improve the visibility of these figures.

- L209: how did you define “nearest quartile” in 2d?

  Thank you for pointing this out. We will have to make it clearer what we have done there. We use a Euclidean distance measure which essentially equates the distances of a 10% precipitation anomaly with a 0.2decC temperature anomaly. So, a data point that had the coordinates of (0.4decC warmer, 0% of normal precip) with respect to a point of interest, and another data point with coordinates (0.0decC, 20% of normal precip), would be treated as the same distance. We did not find the model to be overly sensitive to providing more or less weight to the temperature or precipitation axes.

- L210: Why did you not use the actual values for your localized linear regression instead of the bins? In that way, you can maximize the information better, and also include information beyond empty bins (< 50 values). Moreover, in statistics, extrapolating beyond the range of training data is controversial. Personally, I would not trust the values far beyond (>1degC, 50% prec) what one sees in Fig 5e-h. Finally, since you want to get 2d-surfaces, GAMs (generalized additive models) seem like a prime tool to be used (with a 2d tensor product smooth); it would not require to bin your data, and would also work in 3d with elevation as third predictor.

  See above our answer to major point 2. While the forecasts are not perfect, the authors find that the model performs quite well in the climatological region that you propose. With respect to GAMs and a tensor product: As we have said above, our current aim is to show how something quite simple can still perform quite skillfully. However, as we have also pointed out, increasing model complexity has the potential to further improve upon our proposed methods.

- L242 Please explain, why the bias correction is needed.

  Without bias correction, the SNOWGRID-CL model (which is the one we compare ours against) performs about as well as climatology, and substantially worse than the SnowSens model. This is due to the mean bias of the SNOWGRID-CL model (see Table 2 in the paper). For example, SNOWGRID-CL might track the interannual variability fairly well for a station, but its forecast average might be twice as large as the observed average. Calculating the error on the uncorrected forecasts will show that

the model is not skillful, while the skill of the SNOWGRID-CL model dramatically improves with bias correction.

- Sec 4.1. Why this? Not related to the main paper goal, I guess? Also there are some methodological concerns, and missing descriptions: related to data coverage, usage of linear regression for multiple stations (not recommended, because of their correlation, better to a regional/elevation series first), why the arbitrary split in two periods given the know non-linearity of change (papers by Marty and co.).

  At L251, we discuss a couple of caveats to the calculation of the trends. We do not necessarily agree that the beginning and middle of last century are two completely arbitrary starting points. The authors can consider moving the text and figures related to the trends earlier in the paper, so that they are not as prominently displayed in the Results section.

- L307: What test did you use to assess this significance of skill?

  We used bootstrapping to test for statistical significance. We will be sure to add that into the revised version of the paper.

- Fig10 a) and b) scales do not match but should? a) has -0.4 to 0.4 and b) has -0.2 to 0.6

  There is one station that was cut off from Figure 10b that corresponds to the red station in Figure 10a. We did this simply to improve the visibility of Figure 10b. We will add this information to the revised version of the paper.

- L341: Does this also hold for the single series? Would be interesting to see some single stations time series and not only regional averages.

  We need to make more clear where and when our proposed model is most appropriate. For transparency, we compare the year-to-year forecasts, at the station level, to those of SNOWGRID-CL for the Austrian region. However, we propose a user exercise caution in interpreting the forecasts of any one station or point. See above. Rather, we recommend interpreting the results over band-averages or other climatologically aggregated regions.

- L350: Very interesting application of your method. However, 3.2degC is beyond your training range for that elevation range, so the accuracy is highly questionable. Especially, since your numbers are very different compared to previous studies (a comparison with existing literature would be very useful, there are a lot of studies using regional climate models, or snow models forced with climate models).

If you look closely at Figure 5i, it is around temperatures above 3.5degC and below normal precipitation that the model predicts zero precipitation for the elevation band 0-500 meters. While it is true that these criteria are beyond the training range of the data, we find that the model actually performs quite well in these cases in the validation period. There are 32 instances that fulfill these criteria in the period 1972-2021. As indicated by Figure 5i, the predicted values for these 32 cases is always 0% of normal. The observed values for these 32 cases range between 0%-25% of normal, with a mean of 8% of normal. This translates to an RMSE_SS is equal to 0.89, which means that the error associated with the model is nine times less than climatology. So, while a number of the observed values in these cases are not exactly zero, they are quite close to it.

- Discussion of results missing.

  Thank you for this point. We will see where we could expand on our discussion.