Reviewer 1:

I would like to thank the authors for replying extensively to all issues that were raised. They added discussion elements to all critical comments, which helps making their study more trustworthy.

We would again like to thank this reviewer for his time and effort in reviewing our paper.

Finally, two more comments:

1. Good choice of color scale; a diverging one makes perfect sense since 100% is equal to normal conditions. Please make sure that the color midpoint (where the scale changes from red to blue) matches the value midpoint (100%, where it changes from negative to positive anomalies). At the moment, it seems that the color midpoint is at ~120% or so.

We have changed these plots again.

2. Figure 13 would need some context with existing literature, and comparison of your sensitivities versus results found in previous literature. 10.5194/tc-11-517-2017 is a good choice, but also 10.1002/joc.4205 or 10.1007/s00382-012-1545-3 and many others.

We have added in a couple of these citations and added this sentence at L390: "These expected changes to snow depth, given different temperature and precipitation scenarios, can be compared to prior studies such as Schmucki et al. (2014) and Marty et al. (2017a).

Reviewer 2:

Second Review for the paper "Snow depth sensitivity to mean temperature, precipitation, and elevation in the Austrian and Swiss Alps" by Switanek et al. While I appreciate the authors' effort to address the issues raised in the previous review, I found some answers unsatisfactory. The following comments still require, in my opinion, the authors' attention.

We would again like to thank this reviewer for his/her time and effort in reviewing our paper.

Major

1. The authors state: "We have changed "forecast" to "predict" throughout the paper. We have also added the paragraph between L278-288. This discusses how

we are making the predictions in our paper, but how we envision forecasts can be made using a range of future projections of temperature and precipitation." I find the explanation in L278-288 cumbersome and should be simplified. As pointed out in the first review, the authors provide estimates of snow depth anomalies based on anomalies of contemporaneous temperature and precipitation. This makes SnowSens a statistical model or emulator, but not a prediction tool. As per the AMS Glossary of Meteorology: https://glossarytest.ametsoc.net/wiki/Predictability, predictability is "the extent to which future states of a system may be predicted based on knowledge of current and past states of the system". I do not deny that the emulator could be driven by future meteorology (obtained from, say, a climate model) to estimate future snow depth, but this is, in my opinion, secondary to this paper and doesn't make SnowSens a forecasting or prediction tool. While the authors should point out those potential applications in the paper, I recommend, for accuracy and clarity, that the authors frame their work as the development of a statistical tool to estimate snow depth from meteorological and elevation data.

We have changed the terminology to "estimates" throughout the paper.

2. Regarding the authors reply to my previous comment: "The statistical model seems to work best at larger scales (...), but it may fail at representing e.g., interannual variability at smaller scales, where processes such as orographic precipitation as well as blowing and sublimation of snow can greatly affect the snowpack." The authors state: L343-344 "... we want to be clear that the SnowSens model still exhibits substantial skill for the year-to-year seasonal predictions at the station scale", and then refer to Fig. 9 to support their claim showing "all seasonal SnowSens predictions versus observations over the validation period using all of the stations in the study domain" pooled in a scatter plot. I contend that Fig. 9 does not support the claim, as it doesn't look at the temporal skill of individual stations, but all times and stations combined. For instance, what would be the result of showing the anomaly correlation coefficient of the (perhaps detrended) snow depth estimates over the validation period at each station individually? I do not suggest to do this, since the emphasis of the paper is on larger scales, but the authors should avoid such statements and, at the very least, recognize in the paper the challenges and limitations of using their tool at station/regional scales, where blowing and sublimation of snow can greatly affect the snowpack, and temperature and precipitation alone may not suffice as predictors (e.g., Sexstone et al 2018 https://doi.org/10.1002/2017WR021172).

This comment makes the claim that "Figure 9 doesn't look at the temporal skill of the individual stations." While Figure 9 doesn't explicitly show the temporal skill of the individual stations, by themselves station by station, Figures 11a and 11b does exactly this. The reviewer keeps bringing up anomaly correlations as one way that we could show this skill station by station. It seems that the reviewer did not find the time to read our response carefully, or understand what we have said, regarding anomaly correlations. We explained in our response to this reviewer why we have chosen to use a skill metric such as the root mean

squared error skill score, and why that, in our case, it is a superior metric to use. This is because, critically, the RMSE skill score accounts for any mean bias that may exist between the modeled and observed values. One can have a time series of modeled values at a particular station that are highly correlated with observations, but the mean of the modeled distribution could be half of what we find in the observed distribution, for example. In cases like that, an anomaly correlation will not suffice in telling us whether or not we have model skill. See Figure 1, at the bottom of this document, for a real example from our study. Figure 1 plots the modeled and observed time series of the station where our model skill is the worst. It has a RMSE skill score of -0.51, while at the same time the anomaly correlation is quite high, at 0.61. Our SnowSens model is capturing the observed variability fairly well in this case, but it has overestimated the expected average decrease in snow depth with respect to what actually occurred. As a result, our skill score in this case was worse than climatology.

Next, we are confused as to why the reviewer believes that our Figure 9 “does not support the claim that the SnowSens model still exhibits substantial skill for the year-to-year seasonal predictions at the station scale.” Yes, that figure does pool our entire set of modeled values and plots them against observations. Though, to be clear, each scatter point is a modeled/observed value corresponding to an individual season for an individual station. The scatter points in Figure 9 are comprised of all of the individual time series (modeled and observed) for all of the stations in the validation period. Yes, it is possible in Figure 9 that a station for which the SnowSens model is not skillful can be obscured by the cloud of all of the other stations for which we do have skill. Though, if enough modeled values at individual stations were not skillful in Figure 9, then we would not observe the level of skill that we do. Figure 9 gives us a view of the average skill of our model in predicting season-station anomalies, as measured across our entire study domain and for our validation period 1972-2021. Figure 9 could be thought of as a regionally averaged measure of our interannual skill at the station level. As mentioned above, Figure 9 does not explicitly look at the temporal skill station by station, but Figures 11a and 11b show precisely that. Those are the skills across time in the validation period for every station that we use in our study. Those are more local measures of skill. It seems the reviewer did not see that we have already shown the temporal skill for all of the individual stations. In the paper, we already applied a statistical significance test for the regional skill provided in Figure 9 (L149 in the paper states that we find that this regionally averaged skill is statistically significant with a p-value much less than 0.01, because we generated 10,000 bootstrapped samples and none of them were even close to the skill we observe with our model). We used bootstrapping again to test the statistical significance of the forecast skill at each individual station (Figs. 11a and 11b). We find that there are only a total of 11 stations which either perform worse than climatology or worse than randomly simulated time series (using a p-value < 0.01). Therefore, the SnowSens modeled time series of snow depths from more than 95% of the stations in our study are found to exhibit positive skill and be statistically significantly skillful. Figures 11a and 11b from the paper already

show the temporal skill of the individual stations, and more than 95% of those stations contain statistically significant positive skill.

We have added some relevant statistics to the paper. See L372-376.

3. Regarding the question about extrapolation of T , P , and SD in (old) Fig. 5e,f,g,h to create the maps in (old) Fig. 5i,j,k,l beyond the range of observed values. In their response, the authors refer to L353-363 and Fig. 10 in the new MS. I am not convinced that this figure shows the "effectiveness of the SnowSens model in its ability to predict in new climatological terrain". Here again the authors pool several stations (and years?) in a scatter plot to support their claim. The authors state L362-363 "while we are extrapolating to "unknown" climatological terrain, we find the model is quite capable of performing well in that new terrain." I'm not convinced that the dispersed cloud of points in Fig. 10 justifies such a strong claim. I think the authors should, at the very least, mention that extrapolation from the trained ranges should be taken with caution.

For this point, the reviewer says, "I am not convinced that Figure 10 shows the 'effectiveness of the SnowSens model in its ability to predict in new climatological terrain.'" The reviewer seems focused on the "dispersed cloud of points in Figure 10." However, it also matters where the cloud is situated. Are the values primarily above or below 100% of normal? Our null hypothesis throughout the paper is that a future value can either always be 100% of normal (i.e., climatology), or could be drawn from the distribution of values over the calibration period for the station of interest. Bootstrapping can again be used to assess our statistical significance of these modeled values in Figure 10. In Figure 10, from the paper, the correlation from the 95 points in Fig. 10a is 0.40, and it has a RMSE skill score of 0.62. Similarly, the correlation from the 988 points in Fig. 10b is 0.34, and it has a RMSE skill score of 0.55. Using bootstrapping with the data from the calibration period, we obtain a correlation and RMSE skill score (at $p<0.01$) of 0.24 and -0.01, respectively, in the case of the 95 points of Fig. 10a. Likewise, we obtain a correlation and RMSE skill score (at $p<0.01$) of 0.08 and -0.30, respectively, in the case of the 988 points of Fig. 10b. The RMSE skill scores through randomly generated simulations are not even remotely close to the skill that we find through our model. This is because the SnowSens model is capturing both some of the observed variability, and also the mean shift. Notice where most all of the points are situated in Figs. 10a and 10b. They are almost all below the 100% of normal for both the modeled and observed values. Even if we focus on anomaly correlation alone, which the reviewer refers to as a "dispersed cloud," then the correlations are still statistically significant ($p<0.01$). Perhaps the reviewer has an issue with us using the word "well" as in, "we find the model is quite capable of performing well in that new terrain." We acknowledge that a word such as that has a level of subjectivity. Maybe we can consider changing that phrase to something such as, "we find the model is quite capable of performing skillfully in that new terrain," because that is not debatable. See L365.

We have added some relevant statistics to the paper. See L360-365.

Minor

1. What “annual climatic cycles”? And, consider deleting “annual” as the statement applies to any time scale, e.g., consider instead “snow depth is an important component of the hydrological cycle and the climate”.

We have changed this sentence at L1 to, “Snow depth plays an important role in the seasonal climatic and hydrological cycles of alpine regions.” This is then better paired with our next sentence which is also referring to seasonal values.

2. L158: “precipitaion → precipitation”. This was mentioned in the previous report (old L134) and wasn’t corrected.

We have corrected this.

3. L208-L218 Define the symbols T x,t , P x,t , and HS x,t . This was pointed out in the previous report (old 165-174) and it wasn’t made clear in the MS.

We have now defined these. See L211-219.

4. Caption to Figure 6 “anomlies → anomalies”. Also in L219. Please check text throughout.

We have corrected this.

5. For the answer to item #22 in the previous review referring to comparisons over periods of different lengths. The authors state: “We chose to use a longer prior period (the 40-year period) in order to increase the robustness of the measured changes…”. What robustness? For consistency, it is preferable to compare the statistics over two periods of equal length, particularly when the data have an underlying trend and no statistical significance is provided. If the periods need to be different, please justify.

We now use two 30-year periods. See Appendix A and Figure A2.

6. L42-45 The authors state “There have been several prior studies that have linked changes in snow depth, at different elevations, across the Alps to changes in air temperature and precipitation”, and then provide several references. Although the references are pertinent, note that they are not all specific to the Alps.

We have just removed “Alps” from that sentence. See L42.

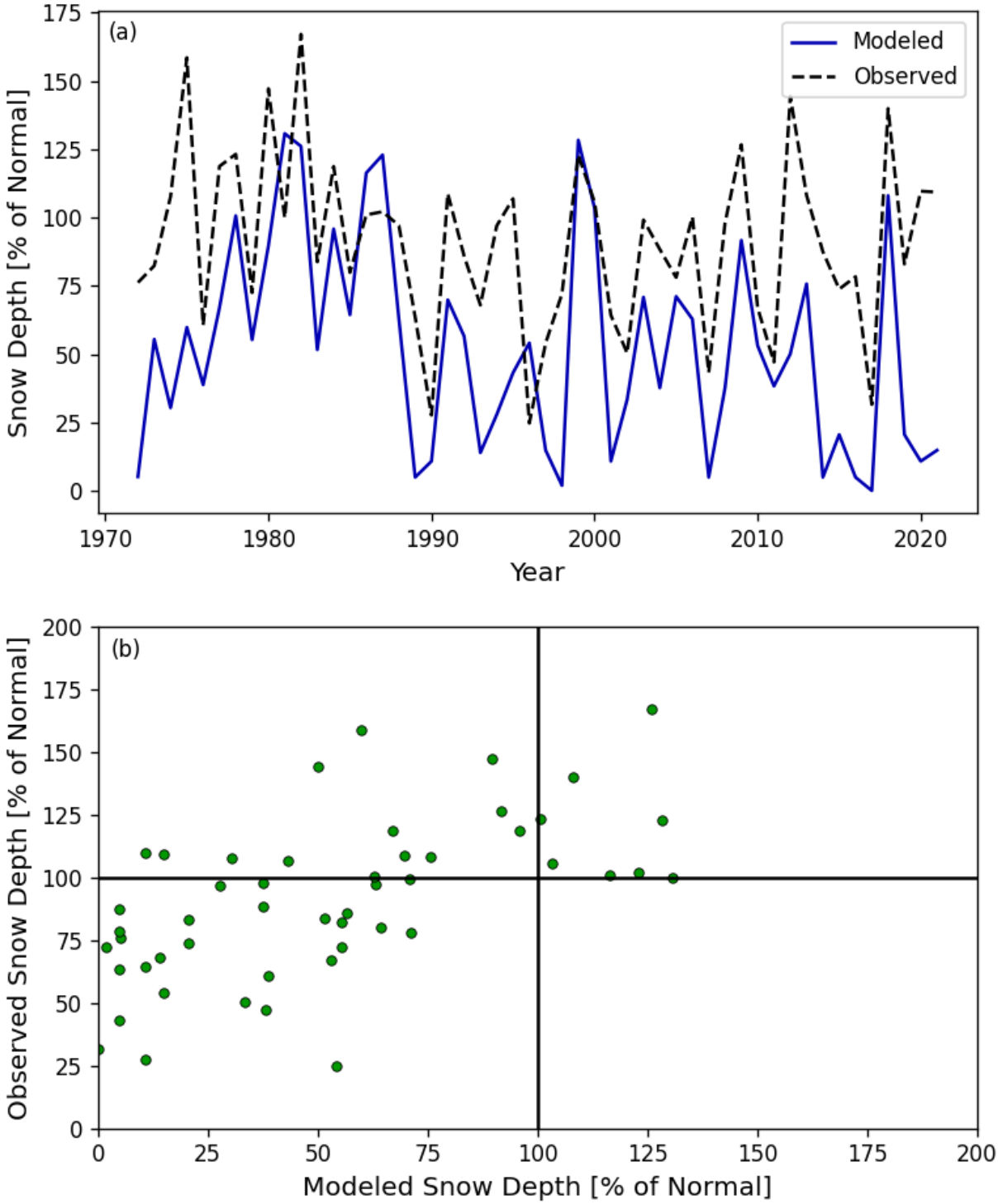

Figure 1: (a) plots the modeled and observed time series, in the validation period, for one station from our study. This individual station has the worst model performance that we found with a RMSE skill score of -0.51 (see Figures 11a and 11b from paper). (b) plots these two time series from (a) as a scatter plot. The anomaly correlation between the two time series is 0.61.