Switanek et al examine the dependence of snow depth (SD) on temperature (T), precipitation (P), and elevation (E) in the Austrian and Swiss Alps. By using historical data from weather stations, they build a statistical model (SnowSens) to estimate seasonal SD based on these predictors. The statistical model is trained with data from 1901-1970/71, then evaluated over 1971/72-2021. The model performance is compared with that of the physics-based model SNOWGRID-CL for a subset of weather stations. Finally, the statistical model is used to estimate SD over the entire domain and some conclusions are drawn on future changes of SD at specified elevation bands.

The authors claim that SnowSens is used to “forecast snow depth” (SD), although SD estimates are produced with contemporaneous observed T and P. The model is, as presented, an emulator of SD driven by P and T, and not a forecasting tool. This and other major concerns listed below diminish the significance of this work, and should be clearly addressed before the paper can be considered for publication:

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 valid point about “forecasting.” The authors will change the terminology used in the paper to reflect how the performance of the model is being evaluated.

Major

1. SnowSens is not a forecasting tool. SnowSens forecasts could be produced if SD lagged T and P, or T and P were themselves forecast, which is not the case in this study. The authors call “forecasts” what seem to be out-of-sample estimates of SD used to validate their model. Therefore, the authors should give a more clear explanation of how their model should be applied. Is this statistical model expected to outperform more advanced state-of-the-art physics-based models? Or, is it more a diagnostic and analysis tool? Perhaps the authors should emphasize applications such as that discussed in L343-355 and Fig. 12, with estimations of future SD based on projected T and P.

Thank you for this comment. The authors agree that we need to do a better job in explaining how our model should be applied. 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. 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 or new climatological terrain (as Reviewer 1 referred to it).

We do agree that the model values used for validation in the paper are not exactly “forecasts.” This is because, as the reviewer has pointed out, we use seasonal temperature (T) and precipitation (P) to say something about contemporaneous or concurrent snow depth (SD) anomalies. However, we are very much proposing that the model and/or the sensitivity plots be used to make actual forecasts of SD given projected future ranges of T and P. We have used a perfect prognosis approach to quantify the uncertainty of one part of the modeling chain as it concerns seasonal snow depth. We have asked the question: Given specific known values of average seasonal temperature and precipitation at the various stations in the region, how accurately can we “forecast” the values of SD? We have answered this in the paper, and our associated skill measures tell us this. As with any model, though, there can be uncertainty that is added anywhere along the modeling chain. If observed future values of T and P differ from what were forecast by CMIP6 for example, then this adds to the uncertainty and will ultimately degrade the quality of the forecasts of any snow model that is used. The whole point of the validation of our model is to give users the confidence to apply it in a true forecasting framework. So, consider an example where we know the average CMIP6 seasonal forecasts of T and P over a period of time such as 2031-2080. Given those conditions, with those specific forecasts of T and P, we can make an actual forecast of SD over that same period of time. We will clarify these points in the revised version of the paper. As a comparison, the SNOWGRID-CL model is also being run with concurrent data. The SNOWGRID-CL model also makes “forecasts” or estimates of SD, for a particular day, given T and P from the same day.

To further illustrate our approach and applicability, consider an illustrative example where some climate model can make forecasts out for the next six months. One might want to observe whether that model is capable of capturing, in a forecast framework, the precipitation patterns associated with ENSO. Initially, the model developers might run the model over six month periods using reanalysis data as input in order to observe how well the “known” set of conditions can be used as model input to simulate a “known” set of precipitation observations. If, over many cases, the precipitation patterns do not align well with ENSO patterns, then perhaps that model should not be trusted in new and unknown cases. On the other hand, if the model is found to perform well, then the model can be useful to provide a set of actual precipitation forecasts which can be conditioned on forecasts of ENSO (e.g., the NINO3.4 region will be 1.5°C above average over the next six months). Our approach is similar. We show, over a set of “known” cases, that the model is quite capable of quantifying how different T and P anomalies translate to anomalies of SD. Now, with a skillful model one can make actual forecasts of SD given different a range of future values of T and P. We will make this clearer in the revised version of the paper.

2. The statistical model seems to work best at larger scales (e.g., averages over elevation bands), 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. Can the authors comment on this?

As stated in the paper and above in this response, “The SnowSens model is not to be seen as a replacement for physically-based models such as the SNOWGRID-CL.” More advanced state-of-the-art physics-based models have their place, and we are not trying to replace them. 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. We consider some of the simplifications that we use (e.g., seasonal averages of our predictors, or using a type of localized linear regression model) to be a strength. We have already pointed out (L341) that the SnowSens model does underestimate the observed interannual variability for any given individual station. The authors will 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 values. 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.

3. L322. Related to the previous comment: to provide a comprehensive assessment of the modeled SD “year-to-year variability”, it would be beneficial to include results of the anomaly correlation coefficient (ACC) of estimated and observed SD. Given the results in Fig. 10 and the comment in L307-308, ACC for the estimated SD at weather stations may be low. If so, the authors should clearly and explicitly address this shortcoming of their method. I would be curious to know whether (and how) the authors plan to overcome this.

We thank the reviewer for this comment. As stated above, our methodology is not designed to better capture the year-to-year variability at individual stations in comparison to a more state-of-the-art physics-based model. Given what we have already said about the applicability of our model, we do not advise placing too much weight on RMSE_SS values or anomaly correlation coefficients at individual stations or point locations. We typically avoided using something like ACC (with the exception of Figure 11) because it is not well suited in evaluating skill over trending time series. Additionally, ACC does not show or reflect whether any biases are present within a model with respect to observed time series. In contrast, a skill score such as RMSE_SS evaluates how well modeled values match the observed values (a close match will better minimize the errors between the model and observations), which includes information related to the observed trend, along with whether bias exists in either the mean or the variance. In Figure 1, which can bee seen below in this document, there are a set of 5 synthetic “observed” values which are all below normal (i.e., solid black line). The values of Model 1 (dashed red line) have a high ACC, but have large error residuals from the observations because it is not capturing the systematic mean change, which could result from an underlying trend. Model 2 (dashed-dotted green line) has the highest ACC, but is dramatically underestimating the observed variability. Model 3 (dotted orange line) has the lowest ACC, but the highest RMSE_SS. This tells us that, on average, the squared differences between observations and Model 3 are closer to one another than between observations and Model 1 or Model 2.

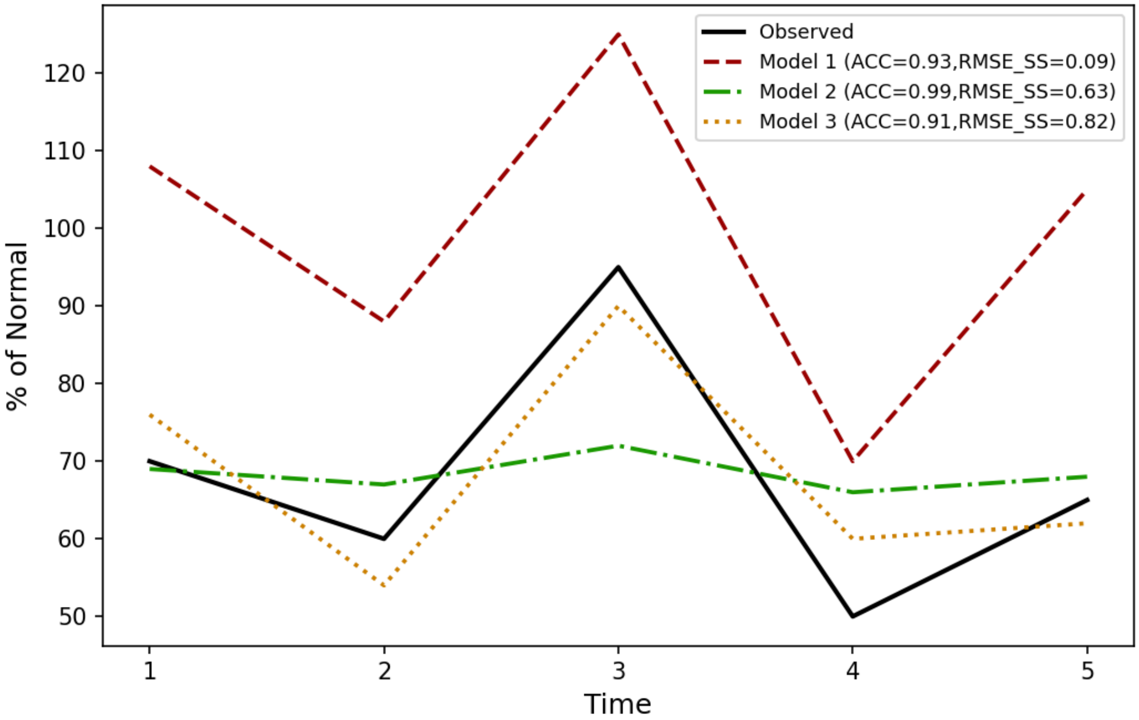

Figure 1: Four synthetic time series are plotted in order to illustrate how effectively different skill metrics evaluate the performance of a model.

4. Based on Fig. 9, SnowSens tends to underestimate SD more than SnowGrid-CL does, particularly for high SD. This suggests that SnowSens may not work well at estimating high snow accumulations and more generally in cases of

extreme snowfalls. Can the authors comment on whether/how their model could/would handle extreme events?

This point is also concerning year-to-year variability. See our comments above to the last two points. The SnowSens model is not making any attempt to forecast extremes, and that is outside the scope of our current study. We are proposing to use the SnowSens model to make forecasts of expected changes to band-averaged seasonal SD given future projections of T and P. As stated above, we will make this clearer in our revised version of the paper.

5. L309-316. Are the values reported in Table 2 for bias corrected SNOWGRID-CL? Please clarify. If not, please provide the bias corrected values as well.

Thank you for this comment. We will make this clearer in the revised version of the paper. The first skill metrics listed for SNOWGRID-CL (which is the first row in Table 2, and it is labeled: SNOWGRID-CL RMSE SS (HS)) are for the absolute or raw modeled values. This tells us that the absolute values of the SNOWGRID-CL model are not particularly skillful, i.e., not much closer to observations than climatology. This is due to the fact that the SNOWGRID-CL model contains substantial mean biases. After correcting for these systematic mean biases, the skill can be seen to improve (i.e., this is the third row in Table 2, which is labeled: SNOWGRID-CL RMSE SS (HS^*)).

6. How sensitive is the statistical model to the bin size discussed in L194-L204. Is it robust to changing bin sizes?

While the performance of the SnowSens model can change as a function of bin sizes, we do not observe the performance of the model to be overly sensitive to what we consider reasonable choices of bin sizes, given the ranges of the data and the sample sizes. When using the bin sizes from the paper (window with sizes, 0.8degC T and 40% P), the RMSE_SS across all modeled versus observed values is 0.26 (row 4 in Table 2) and RMSE_SS is 0.19 for its ability to capture the SD changes from one period of time to another for all of the stations (row 7 in Table 2). If we use a window size of 1.0degC T and 50% P, then these skill scores are 0.26 and 0.20, respectively. And, if we use a window size of 0.6degC T and 30% P, then these skill scores are 0.26 and 0.17, respectively. While we present a working and skillful model, a user may deem it appropriate in their application to use a different bin size than what we have put forward in the paper.

7. Following on the previous comment, have the authors considered quantifying the uncertainty of their statistical model?

We thank the reviewer for this question. While we appreciate the motivation to quantify all of the uncertainty associated with our model, that is beyond the scope of our study. We have presented a methodology with which we have evaluated the performance of modeled values with respect to observed values. The skill scores that we have listed in the paper do provide a quantification relating to how “certain,” or how much variability, there exists in the observations about the modeled values. However, future work can focus on better quantifying the uncertainty of the model as a function of different elevations, T anomalies, and P anomalies.

8. I may have missed it, but how the authors extrapolate T, P, and SD in Fig. 5e,f,g,h to create the maps in Fig. 5i,j,k,l beyond the range of observed values? For example, in Fig. 5i for the 0-500 m band, how is the map created for temperature anomalies greater than 3° C or precipitation more than twice the normal values? It seems unlikely that the model will perform well out of the observed range.

Here, we include some of our responses to Reviewer 1 (see above), who also raised a couple of questions about extrapolating to new values. We are planning to integrate much of the following content into the revised version of the paper.

It is true that we use extrapolation in our model. 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 2, seen below in this response to the reviewer, we have plotted the cases which fulfilled these criteria. Figure 2a 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). The average of the forecasts and observations over these cases are the same; they are both 33% of normal. Figure 2b 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.

And here is another example of extrapolating to large temperature anomalies. If you look closely at Figure 5i (from the paper), 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 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.

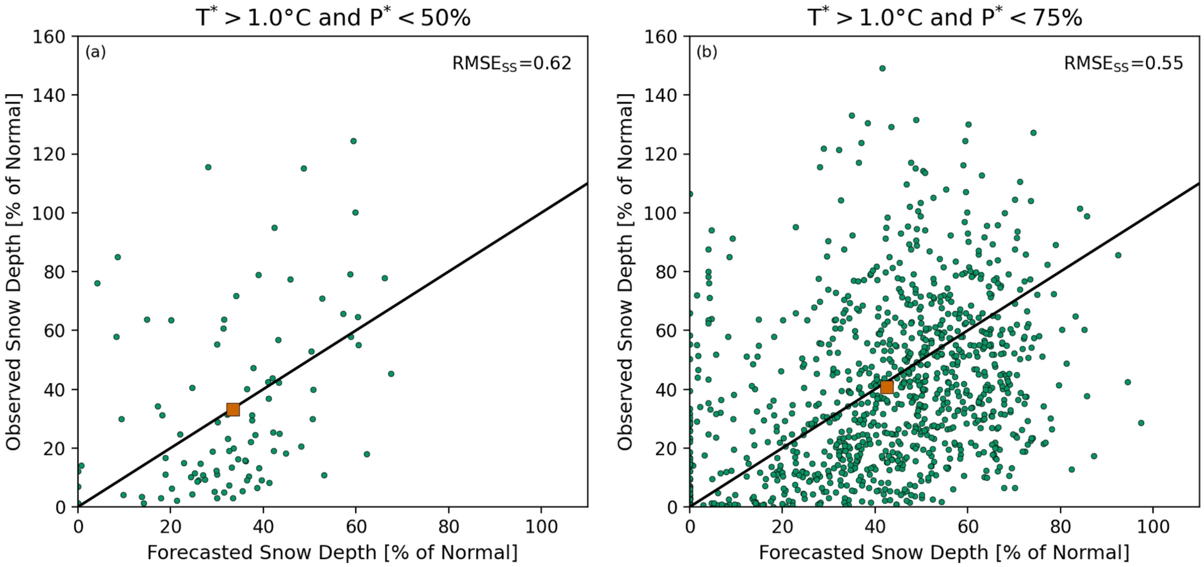

Figure 2: Figure 2a shows the 95 cases where the average seasonal temperature in the 1972-2021 validation period was greater than 1.0degC and less than 50% of normal precipitation. Figure 2b increases the sample size by using a threshold of less than 75% of normal precipitation. This gives us 988 cases. The larger squares are the “forecasted” or simulated and observed averages over these cases. The skill scores, for these two different criteria, are shown in the top right of the subplots.

9. Table 3 states that the result are statistically significant at $p< 0.05$. What statistical test is used to establish this?

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

10. L249-259. In describing Figs. 6 and 7, the authors make good points regarding the nature of SD and how those fitted straight lines could be misleading. Another point is that the sample size may be different each year (e.g., there may have been considerably less stations at the start of the recording period compared to present time as is clearly the case for Fig. 6d, making the trend largely uncertain). Can the authors comment on this and provide a measure of uncertainty associated to these straight lines?

Thank you for this comment. We have already accounted for the fact that each year can have a different number of measurements. This is done by fitting a least-squares regression line to all of the scatter points in Figures 6 and 7 in the paper (independently for each elevation band), instead of a least-squares fit to the regionally-averaged SD. Yes, if a particular year only had five measurements and another year had 100, then getting regional averages would apply equal weighting to the both of these years even though they have dramatically different data coverage. In Figure 3, below in this document, we have provided a simplified example using synthetic data. In this example there are two individual data points for times 1 and 2, while there are five individual data points for times 3 and 4. One can fit a regression line to the individual blue points, and this is shown as the blue line. This is what we have done in the paper. In contrast, one could average the values at each time and then fit the regression to those time-averaged values (shown as the red line, this could be thought of as the average of all of the available stations for a given season). Because we are not using a very large number of points in this example, the blue and red lines are not all that different. We are only illustrating that they are different, and it is important to know which data is used to fit your regression.

The reviewers comment on the uncertainty of the trend is a good one. In the revised version of the paper, we will be sure to add confidence intervals to the plotted trend lines.

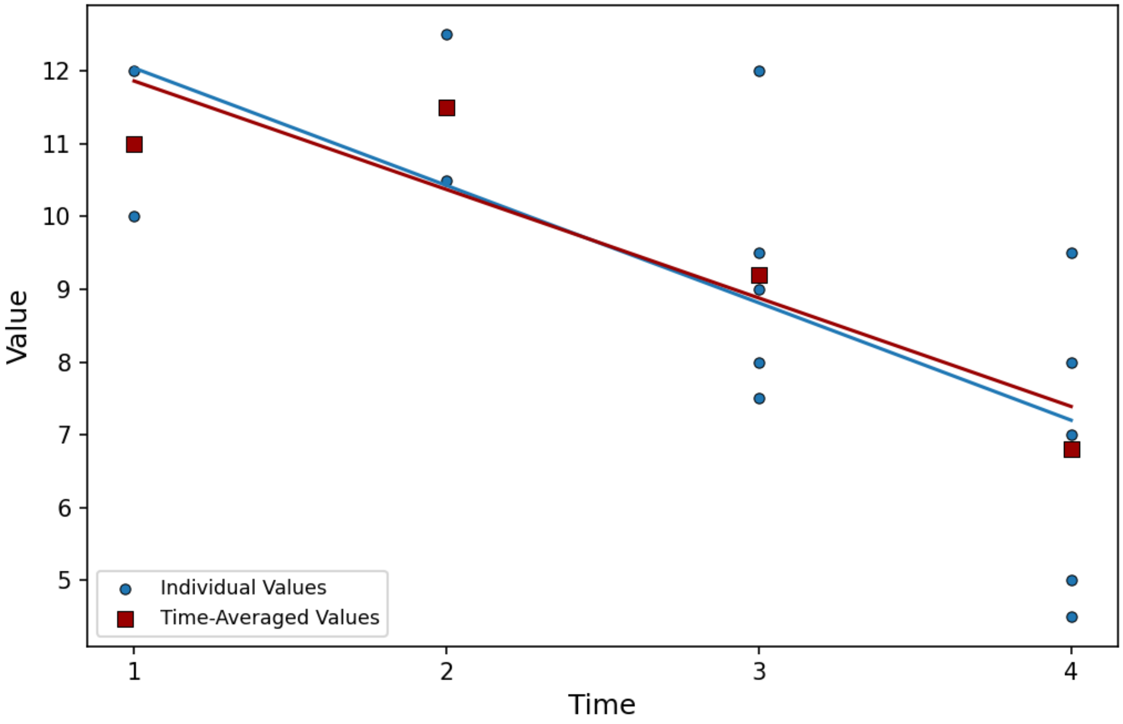

Figure 3: Synthetic data which has two individual data points (blue circles) for times 1 and 2, and five individual data points (also blue circles) for times 3 and 4. The red squares are averages of the blue circles at each time step. The blue

regression line is fitted to the individual blue points, while the red regression line is fitted to the red points.

Minor

1. L1 Delete “incredibly” and “climatic and”

We can remove “incredibly.” However, we see snow depth as something that not only results from a particular climatic regime, but it also influences the climate. Consider the influence of albedo on the energy budget that is present on the ground or in the atmosphere, and how that changes as a function of SD.

2. L32-39. altitude → elevation

Thank you. We can be more consistent with the wording.

3. L47. “However, these studies suffer … strong dependence of snow depth on elevation”. Please clarify.

We are primarily referring to some prior work conflating the absolute changes in snow depth and the fact that higher elevations climatologically have greater depths of snow to begin with. At L36, we write: “However, this altitudinal dependence of the snow depth trends is conflated to some degree with the fact that stations at higher elevations also typically receive more snow.”

4. L50. “This allows us to remove the influence of elevation …”. Please clarify. “Remove” from what? The dependencies established in this study are strongly affected by elevation.

We remove the influence of elevation on the absolute values of SD. Even after removing this influence, there is still an influence of elevation in a relative sense. We will clarify this point in the revised version of the paper.

5. L68-71. Please clarify what homogenization means in this context and why one or the other choice is not expected to change the results.

Homogenization in this context is referring to the possible shifting up or down different parts of the snow depth time series at certain stations. This can happen if they were found to experience a noticeable break (or shift) in the time series due to something like moving the station up or down in elevation. We state at L69, “Through personal communication, the authors of a recent homogenization study in the Alps (i.e., Resch et al. (2022)) have indicated that there are not any systematic changes in snow depth one way or the other as a result of the homogenization procedure (Marcolini et al., 2019; Buchmann et al., 2022).” For that reason, we do not expect the results in our paper to differ all that much from some potential future study that chooses to use homogenized data.

6. L96. Specify what those time series are? Seasonal averages at various years?

That is a good point. We will make this more clear that we are using time series of seasonal averages.

7. L121. Delete “in a given month at a given station”. Unless I’ve misunderstood the statement, it refers to the snow depth coverage of the 291 stations for all the Januaries during 1901-2020.

We will remove that text.

8. L125-127. If so, why not simply use the November-April or the November-May season as in previous studies?

This is explained in the paper at L116, “During warmer months, and especially with stations at lower elevations, an observable amount of precipitation will not always translate to a measured snow depth. This would result in trying to fit a predictor time series (i.e., precipitation), which does vary, with a predictand time series that does not (i.e., snow depth). Therefore, we would like to minimize the number of cases where there is zero measured snow depth.” And since many more stations can have zero recorded snow depth in the months of April and May, we chose not to include those months in our model fit. That way, we can have consistency in the lengths of our seasons for both the predictors, T and P, and our predictand, SD.

9. L134. precipitaion → precipitation

We will fix that.

10. L137. “homogenized stations”? It seems the authors provide a method to homogenize the data, but precipitation and mean temperature are taken over all “available” stations?

We use only temperature and precipitation station data that has already been homogenized by the data provider. We do not apply any homogenization ourselves in the paper.

11. L141-144. This is not clear. In particular, how is the first of the “two time series” computed? Is the second time series an actual time series or an average value over the training period? And, how is the “first time series” adjusted? Do you mean it is super-imposed to the average temperature computed in (2)?

We will rewrite these sentences in an effort to make it more clear to the reader.

12. L154. Delete “the similarly”

We will fix the wording there.

13. Figure 3 shows correlations between SD and T or P, and their dependence with elevation. Given that T and P are not independent variables, perhaps it would be more illustrative to show partial correlations e.g., between SD and T while controlling for P, and between SD and P while controlling for T. In a way, those partial correlations are related to the partial derivatives over the surface shown in Fig. 5.

While our model is not very complex, we do want to layer some of the methodological concepts, piece by piece. We first transform the data into anomalies with respect to average T, P, or SD. Then, we show, like others have before, the correlations between SD and T along with SD and P. This initially presents "forecasting" SD as a one-dimensional problem as either a function of T or P. As you say, using either T or P, independently, only gives partial information about what we can expect with SD. That is why we then introduce Figure 4 (from the paper), and show the problem as two-dimensional. Then, in Figure 5 (from the paper), we break up the data into the elevation bands, and now we present SD forecasting as a three-dimensional problem.

14. L165-169. Unlike P and SD, Eq. 4 shows T "anomalies" relative to the climatology over the training period. These anomalies are not normalized. Why are they called "normalized" temperatures? If there is a need to refer to "normalized" T, P and SD with one term, then perhaps use "reduced", or simply normalize the temperature anomalies with a relevant scaling factor common across stations and years.

In order to avoid confusion, the authors plan to change our terminology in the revised version of the paper. We will present these values simply as anomalies, either as percentages or degrees C from normal.

15. L165-174 Define T x,t , P x,t and HS x,t . In particular, is P x,t the accumulated or averaged precipitation over November-March at station x and year t?

Yes, Px,t is the accumulated precipitation over November-March at station x and year t. At L137, we state, "First, we obtain November-March sums of precipitation and averages of mean temperatures at all of homogenized stations over the years 1901/02-2020/21." So, we use sums or accumulations of precipitation. Though the results would not change if one were to instead use average monthly precipitation. This would simply scale the precipitation accumulations by a common factor. We will clarify this in the revised version of the paper.

16. L177-178. The larger squares are hard to see in the figure. And, what "black lines"?

The larger squares are made using black lines. We will work to make this figure easier to see and interpret.

17. L179-182 “One can observe… two-dimensional plane (not shown)… in the lower-right”. This is not clear. What 2D planes?

Consider the example that we have outlined. We have two predictors, T and P, which are being used to fit some model that can be used to “forecast” SD. With a multiple linear regression fit, then the value of SD depends on both T and P. When one predictor is used, the fit is a line. When two predictors are used, the fit is a plane or a surface. That is what we are referring to. We will improve our description of this in the revised version of the paper.

18. L176-190 This paragraph seems to be a motivation to include an SD dependence not only on T and P, but also on elevation. If so, the explanation could be simplified and made clearer, and previous work explicitly addressing this could be cited, e.g., Moran-Tejeda 2013 [doi:10.1002/grl.50463], Sospedra-Alfonso et al 2015 [doi:10.1002/2015GL063898], Scalzitti et al 2016 [doi:10.1002/2016GL068798].

We thank the reviewer in providing us with some relevant citations. We will look if we can make the explanation clearer.

19. L207. valus → values

We will change that.

20. L232 and L234. Consider deleting “real-valued” and use only “absolute” value, or “full” value.

Thank you. We can consider that suggestion.

21. L241. This is confusing. How are HS MOD 1962−1971 ,x,t and HS OBS 1962−1971 ,x,t in Eq. 8 defined? Do they depend on t? And, is the numerator in Eq. 8 missing an ∗?

Thank you for bringing this to our attention. We will work to make this clearer.

22. L271. The comparison is for the last 30-year averages relative to averages over a 40-year period. Why not 30 years for consistency? And, are the dots in the figure averages at all available stations? Sampling errors seem to impact more lower than higher elevations.

We chose to use a longer prior period (the 40-year period) in order to increase the robustness of the measured changes. We could use the most recent 30-year period compared to the 30-year period before, though this will be more subject to sampling variability than using a longer prior period. Though, if it is seen to make more sense to use 30 years for both periods, we can proceed in that direction.

Yes, the dots are the percentage anomalies at all available stations. And no, these are not sampling errors that we observe at lower elevations. This is instead an indication that the snow depth at lower elevations exhibits both greater variability and skewness than the snow depth at higher elevations. It also reflects the zero-bounded nature of something like snow depth or precipitation. The average seasonal snow depth for a station below 500 meters might be something quite close to zero, like 2cm. Some season might be 1cm (or 50% of normal), while another season might be 10cm (or 500% of normal). One can observe the same phenomenon for summer precipitation across the state of California. The average daily precipitation amounts are very small, where the daily averages in July are typically less than 1mm, for example. Then, in the rare events where something like 10mm of rain falls at a station, that would be categorized as an event that is greater than 1000% of normal.

23. L293. As mentioned above, I wouldn’t call this “forecast skill”, as these are not actual forecasts. Perhaps refer to it as a measure of model “accuracy” or “performance”?

We have discussed this above.

24. L341-342. In the panels of Fig. 11, the authors give the correlation coefficients computed for the elevation bands and validation period. These correlations are largely driven by the decreasing trend (particularly at lower elevations). Could the authors add the correlations for the detrended time series?

We respectfully disagree with the reviewer. The trends are not responsible for the level of correlations that we report in Figure 11. The correlations that we reported in Figure 11 between the modeled and observed band averages in the validation period are 0.89, 0.84, 0.81, and 0.75, respectively for the four elevation bands. If we detrend both the modeled and observed time series for each of the bands, then we obtain correlations of 0.89, 0.83, 0.79, and 0.74. They are very close to the correlations which contain trends. We can write this in the revised version of the paper.

25. L367-368 That SnowSens can“skillfully forecast year-to-year variability of snow depth” seems an overstatement, particularly when ACC at the level station were not provided or discussed.

We respectfully disagree with the reviewer. We do not consider that particular text to be an overstatement. The SnowSens model performs statistically significantly better than climatology, which is the baseline measure of skill. The ACC between the SnowSens modeled anomalies versus the observed anomalies is 0.615. This is the result of taking the ACC of all of the modeled vs observed pairings, or the yellow points, in Figure 9 from the paper. We have simplified the viewing of these SnowSens modeled vs observed values, and that is provided as Figure 4 in this document (see below). The number of modeled/observed pairings in the validation period is 11,064. Given that the stations are not independent of one another, the effective sample size would be smaller than 11,064. While the effective sample size might be on the order of three times smaller, we can go to an extreme to clarify our point. Let's instead reduce our sample size by a factor of 100. An ACC of 0.615, with a sample size of 110, has p-value of <1e-13. So, this can be interpreted to mean that there is less than 1 in a trillion chance, on average, that one could achieve the level of skill exhibited by our model, using randomly generated "forecasts." We also produced 10,000 randomly generated simulations, and the maximum ACC value we found in those 10,000 simulations was 0.037. We would argue that it is not in any way an overstatement to say that the SnowSens model has skill when evaluating the performance of year-to-year values at the station level.

As we just pointed out above, Figure 9 shows all of the pairings of modeled values vs observations for the validation period across the four elevation bands. Figure 10 and Table 2 also provide the skill at the station level. We discussed above why we used RMSE_SS instead of ACC.

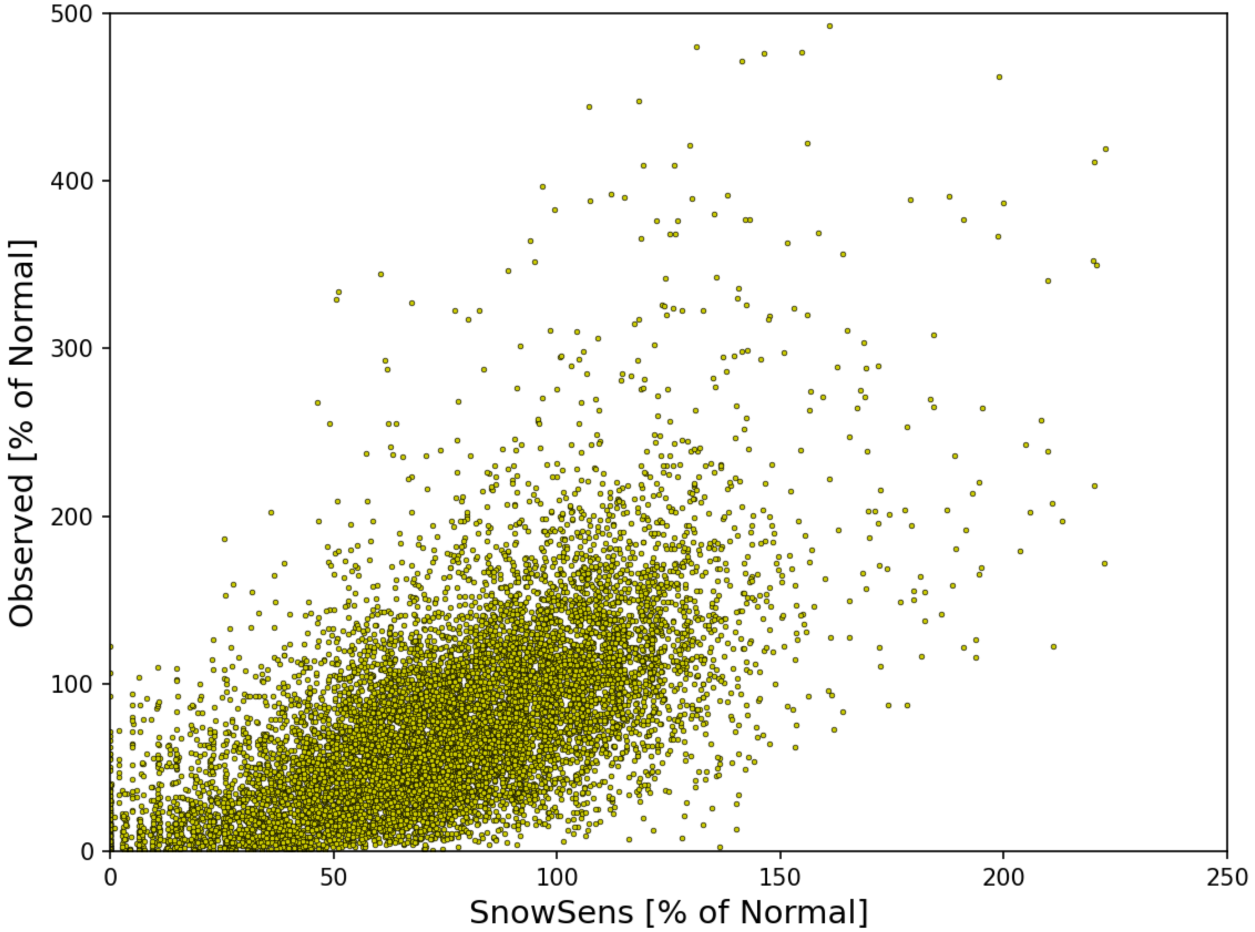

Figure 4: All of the modeled values and observed values in the validation period at the station level. There are 11,064 observed cases. If we had 100% data coverage for the observations over the last 50 seasons, then there would be 14,550 cases (50*291, where we have 291 stations in our study). The Pearson

correlation coefficient between the modeled and observed values in this figure is 0.615. If one computed the correlation, station-by-station, then the average correlation across the stations is 0.66, with a minimum of 0.29 and a maximum of 0.87.

26. L378 Delete “of the world”

We can do that.