Review for the paper “Snow depth sensitivity to mean temperature, precipitation, and elevation in the Austrian and Swiss Alps” by Switanek et al

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:

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.

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?

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.

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?

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.

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

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

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.

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

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?

Minor

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

2. L32-39. altitude → elevation

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

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.

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.

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

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.

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

9. L134. precipitaion → precipitation

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?

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)?

12. L154. Delete “the similarly”

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.

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.

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

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

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

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`].

19. L207. valus → values

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

21. L241. This is confusing. How are $\overline{\mathbf{HS}}_{\mathrm{MOD}_{1962-1971},x,t}$ and $\overline{\mathbf{HS}}_{\mathrm{OBS}_{1962-1971},x,t}$ in Eq. 8 defined? Do they depend on t? And, is the numerator in Eq. 8 missing an $*$?

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.

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"?

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?

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

26. L378 Delete "of the world"