Reply to Reviewer #2

We thank the reviewer, Dr. Birkeland, for his insightful comments. Below we provide answers to the major points and indicate how we will improve the manuscript.

I selected "Major Revisions" for this paper simply because some of my comments suggest that the paper may benefit from re-analyzing some of the data. However, I do not think that such a re-analysis should be overly difficult, and so my review probably falls somewhere between "major revisions" and "minor revisions".

This research aims "to characterize the avalanche danger levels based on expert field observations of snow instability." This is an important goal and is something that would be beneficial for avalanche forecasters and for a better understanding avalanche dan- ger. The authors utilize an interesting, unique, and high-quality data set. While I believe the authors have produced an interesting paper, I believe it would benefit from some changes and clarifications. I have four major comments:

1. First, when I was initially and quickly reading the paper and looking at the title, I assumed that the "avalanche danger" referred to in the title was the regional avalanche danger rating. However, this is not the case. Rather the "avalanche danger rating" is really the local nowcast provided by the observer. As pointed out in the paper, there is a scale mismatch between a local rating for a particular area and a regional avalanche danger rating. In addition, the authors point out that they are utilizing a local nowcast in their abstract. Despite this, I believe the authors should more clearly define the differences between these two ratings and if they decide to utilize the local nowcast avalanche danger then that should be specified in the title.

We refer to local nowcast and regional forecast as introduced by Techel and Schweizer (2017). The study relates to the avalanche danger as described for an area of about 100 km^2. The avalanche danger level summarizes the avalanche conditions regardless of the type of assessment (regional forecast vs. local nowcast). We will further clarify this point in the revised manuscript. Moreover, as we relate the danger level estimate to observations at the local scale it makes sense to use the local nowcast rather than the regional forecast.

2. Second, along the lines of my first comment, the authors acknowledge that there is circular reasoning in their data since the observers are making snowpack and avalanche observations and are also assigning the local avalanche danger. Undoubtedly the observers are taking their observations into account when they are assigning the local avalanche danger. I believe this is somewhat problematic. In this scenario the authors may actually be testing what snowpack observations the observers happen to associate with a particular local avalanche danger rating rather than the more general question of which snowpack observations are associated with a danger level. I am wondering why the authors didn't simply compare the local observations to the regional avalanche danger assigned for that region for that day? That way there would be independence between the snowpack observer and the assessment of the avalanche danger, and the results would reflect the frequency of making these observations for a given regional danger level rating. I would suggest either utilizing the regional danger ratings or providing a solid rationale for not using them in the paper.

We use the local nowcast since it is clearly a better descriptor of the avalanche conditions (Bakermans et al., 2010; Jamieson et al., 2009; Techel and Schweizer, 2017). The two main reasons are: (1) The local nowcast assessment is done after the regional forecast assessment and is no longer a forecast, hence it does not include forecast errors due to, for instance, errors in the weather forecast. (2) There is a scale mismatch between the local observations we relate the danger level to and the regional forecast. The regional forecast is by definition broader and cannot take into account peculiarities within the region. The regional forecast has to address the highest danger prevailing in the region. Hence, it is well possible that in some subregions the danger is actually lower (Figure R2). This danger assessment (local nowcast) should then be related to the local observations. That's the approach we follow.

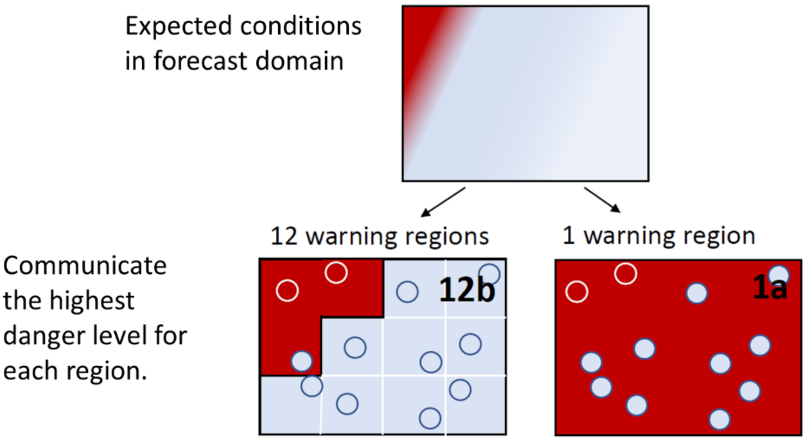
Expected conditions in forecast domain
12 warning regions
1 warning region
Communicate the highest danger level for each region.
12b
1a

Figure R2: Effect of the spatial resolution of a forecast (or the size of forecast region) on danger assessment and quality. In the upper square, showing a forecast domain, the expected avalanche conditions represented by a higher (dark red) and a lower (light blue) danger level (RF) are shown. This situation will be assessed and communicated differently, depending on the size of the warning regions used by the warning service (12 vs. 1 region). The circles represent local danger level estimates (LN) (adapted from Figure 6.2 in Techel, 2020).

Therefore, we are not convinced that it is advantageous to use the regional forecast. In our view, it is imperative to use the verification data, or at least the best possible assessment. This is obviously, by definition, not the forecast. While full independence is certainly desirable, we think it is almost impossible to achieve in the context of avalanche forecasting. Also, the observers are biased by the forecast. While the local nowcast has the disadvantage of introducing a potential bias, which we openly discuss, the regional forecast is obviously incorrect on 1-2 days per week and does not match the scale of observation. We think this obvious inaccuracy and the scale mismatch is more severe than the potential bias introduced by the local nowcast. Therefore, we prefer to relate the observations to the local nowcast.

Alternatively, we have explored an approach where we used the regional forecast (RF) as starting point for the local assessment. The corresponding classification trees are dominated by RF when predicting LN; the local observations rarely show up in the tree. The only exception is in the case when RF = 3 and there are no signs of instability, then the tree predicts LN = 2. In all other branches LN is predicted solely by RF. This is not surprising given the still relatively high agreement rate between RF and LN. However, this approach does not provide any guidance on how to assess the local danger level based on field observations, which is what we actually aim for.

3. Third, this paper utilizes a unique dataset of snow profiles and observations from Switzerland. Approximately 95% of the data are from the region of Davos. Later the authors explain that some of their results, such as the predominance of persistent weak layers, may be because so much of their data are from the Davos area. I believe the paper would benefit from using only those snow profiles and observations from around Davos, rather than a highly unbalanced dataset consisting of almost all profiles from Davos and then 5% from other areas. This would still retain 95% of the data but would remove some of the variability introduced by the other 5% of the data.

We admit that at first glance the selection seems questionable with regard to geographical origin. However, we initially did not care about origin, but selected the profiles purely based on quality. We have now assessed whether those 5 % of profiles that were not recorded in the region of Davos, introduce any bias. As shown below (Figure R3, Table R1) this seems not to be the case. The profile characteristics are very similar. In particular snow depth, a good indicator of snow climate, is not significantly different (median snow depth 120 cm vs. 108 cm; *p* = 0.13). Also, the snowpack structure index, which relates to the importance of faceting, is not different either (1.78 vs. 1.70; *p* = 0.7).

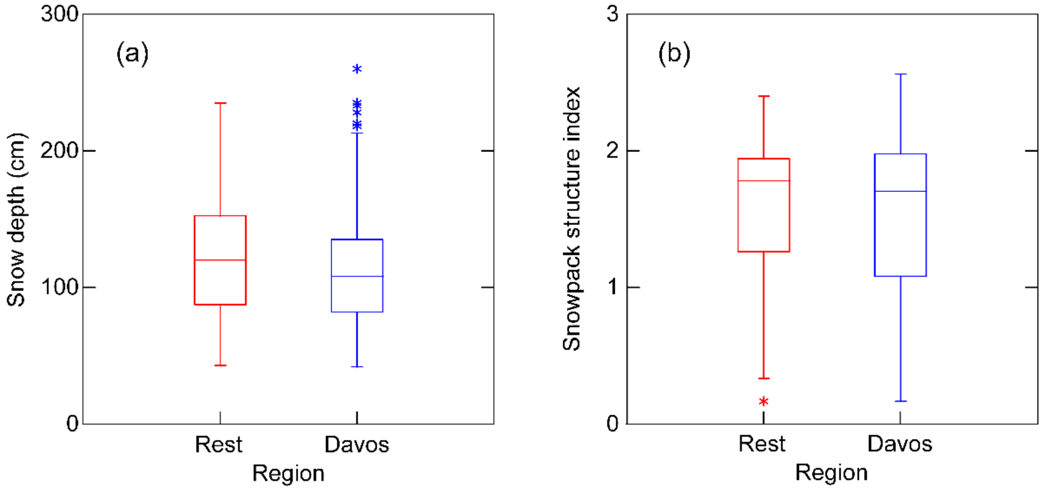

Figure R3: Comparison of selected profile characteristics from the region of Davos (*N* = 561) and from other regions (Rest, *N* = 28): (a) Snow depth and (b) snowpack structure index. For median values and *p*-value see Table R1.

At first glance, these results may seem surprising. However, these are fully in line with the observation that in our study many snowpack characteristics were similar to those reported in previous studies that included profiles that originated from regions with clearly different snow climates such as the Columbia mountains of western Canada. For instance, in the study by Schweizer and Jamieson (2003) profiles from the Swiss Alps and the Columbia Mountains were jointly analysed. While there were some differences between the profiles from the different snow climates, the characteristics of instability were largely the same in both sets of profiles. This suggests that characteristics of instability may well be similar in different snow climates. Of course, their frequency of occurrence will likely be different.

Table R1: Comparison of selected profile characteristics from the region of Davos (*N* = 561) and the other regions (*N* = 28). Median is shown (mode in the case of FL grain type) and level of significance based on non-parametric Mann-Whitney U-test.

Parameter	**Rest**	**Davos**	***p***
Elevation (m a.s.l.)	2488	2470	0.367
Slope angle (°)	34	33	0.301
Snow depth (cm)	120	108	0.131
Slab thickness (cm)	41	38	0.052
FL grain size, avg (mm)	1	1	0.643
FL grain size, max (mm)	1.5	1.5	0.846
FL grain type	Facets	Facets	0.872
RB score	4-5	4	0.737
5-class stability	Fair	Fair	0.761
$\text{SNPK}_{\text{index}}$	1.78	1.70	0.702

4. My fourth comment relates back to my third. The authors spend considerable effort (and content in their paper) characterizing the snow profiles. This provides some interesting results that I believe should be retained, but it is outside the primary stated goal of the paper. As stated above, I believe this analysis would have more meaning if the data were restricted to just the 95% of data from the region of Davos. Then the snowpack characterization part of the paper can provide a characterization of the Davos area snowpack rather than "mostly" the Davos area snowpack with 5% of the profiles and observations from other areas. In addition, this part of the paper should be better highlighted in the abstract and perhaps the title as well since nearly as much attention is paid to this snowpack characterization as is paid to the relationship of the snow profiles and observations to the danger levels. I think that this "characterization of a snowpack" in a region is quite valuable and will set a baseline for future work which could compare this characterization against the characterization of the snowpack in other regions or other countries.

We thank the reviewer for his favorable assessment of the part on snowpack characterization. As shown above, the variability introduced by those 5 % of profiles is not really significant. Hence our conclusion that our results reflect the snow climate of Davos and might be of limited value for other regions is actually not fully supported and seems too restrictive. Given the insignificant bias introduced by those 5 % of profiles, we see little advantage in re-doing the analyses.

While I do have some substantial comments that I believe the authors should address, I do think that this is important work and that it should be published following revisions.
In addition to the above comments, I have some more minor comments and suggested typographical corrections:

Line 29: Delete "at times"
Line 40: Delete "were" and replace with "have been"
Line 68: Replace "inexistent" with "nonexistent"

Many thanks; we will make these changes as suggested.

Line 84: Can you provide more specifics about how you defined "an experienced observer"?

Experienced observers are those who have done dozens of high-quality profiles. They are almost exclusively (96 %) professional forecasters or researchers with extensive experience in field work. The authors recorded 63 % of all profiles.

Line 97: After reading this paragraph, I am still not certain how the failure layer and adjacent layers are defined. I understand how the authors come up with the failure interface, but how do they necessarily define the failure layer? Was this done manually by the authors? And, with the adjacent layer, was this typically just the layer adjacent to the failure interface that was not the failure layer? I think this might be less confusing if the authors talked about the failure interface, and then the layer above that interface and the layer below that interface. They could then also quantify how often the "failure layer" is below or above the layer interface.

We follow the approach that was introduced by Schweizer and Jamieson (2003). In all profiles, the failure interface was reported. While in most cases the failure or weak layer is obvious, we considered the softer of the two layers as the failure layer (FL) and the layer across the failure interface as the adjacent layer. If there was no difference in hardness, we selected the lower layer as the failure layer, and the layer above the failure interface as the adjacent layer. The failure layer was in 53 % below the failure interface, and in the remaining 47 % above.

Line 115 – It seems like assigning 3+ to Considerable is arbitrary. Why not apply High here since 2+ is also assigned to Considerable?

We assumed that + and - refer to a somewhat higher and lower danger level, respectively. Hence, we assigned 2-, 2, and 2+ to *2–Moderate*, and 3-, 3, and 3+ to *3–Considerable*. Occasionally, intermediate values indicated with "2 to 3", those we assigned the next higher danger level, therefore, *3–Considerable* in this example. Overall, intermediate values were provided in only 14 % of the cases.

Line 199-200 and line 525 – I don't think that a decrease in the whole block release of RBs with increasing RB number is necessarily related to a decrease in crack propagation propensity. This may have more to do with the increased damage to the slab caused by harder jumping on the RB that causes partial block releases. Given the complexities involved, I don't think the authors can draw such a definitive conclusion from their data. While I don't have a good dataset to either confirm or refute this conclusion, I have anecdotally seen seasons where PSTs consistently propagate to end – indicating the propensity for crack propagation – for a long time after other tests have indicated that failure initiation is far less likely.

We agree with the reviewer's anecdotical evidence of full propagation in PST tests. However, we do have datasets showing that shear strength as well as specific fracture energy of the weak layer increase with time. This suggests that initiation as well as propagation become less likely with time – in line with our suggestion. Our interpretation is simply a suggestion, clearly not a firm conclusion, and may hopefully trigger some further research.

Line 454: Delete "and" and replace with "or"

We will change as suggested.

Line 528: I assume the "avalanche danger level" in the sentence on this line is the local avalanche danger? As stated in one of my primary comments, it would be helpful to make sure to careful differentiate between the local avalanche danger rating and the "avalanche danger rating". I always assume this latter term is associated with the regional avalanche danger rating.

We will clarify and refer to the nowcast in the revised manuscript as suggested. As mentioned above, we assume that the danger rating describes an avalanche situation regardless of whether this assessment is made in the office, in the field, beforehand or in hindsight.

Line 529: Along the lines of my comment above, the authors state that the local avalanche danger rating agrees with 70% of the regional danger ratings. It would be interesting to more thoroughly compare how those two ratings at the two different scales differ.

Techel and Schweizer (2017) provided a detailed comparison between regional forecast and local nowcast. This topic, essentially verification of forecast, is beyond the scope of the present study. Nevertheless, Figure R4 provides the agreement by danger level. The agreement was 95 %, 66 %, 71 % and 33 % for the danger levels *1–Low*, *2–Moderate*, *3–Considerable* and *4–High*, respectively.

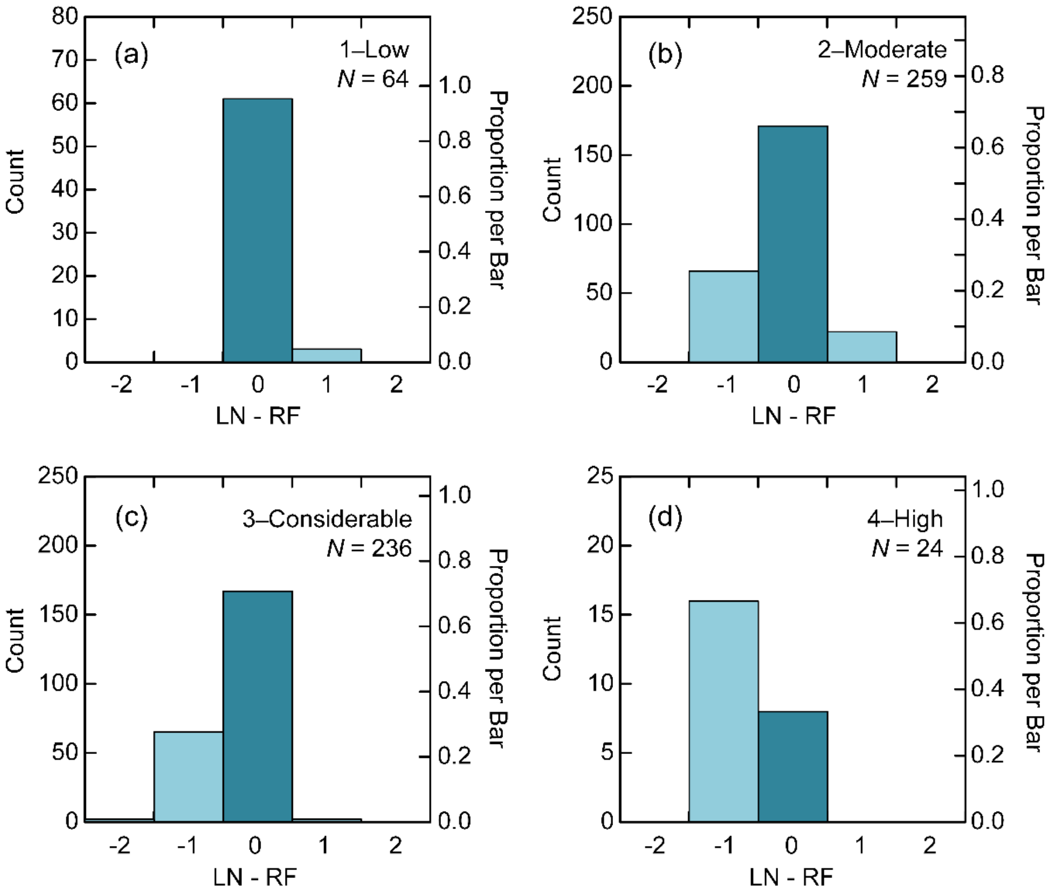

Figure R4: Deviation between regional forecast and local nowcast for each of the four danger levels (a-d) from *1–Low to 4–High* (*N* = 583).

References

Bakermans, L., Jamieson, B., Schweizer, J., and Haegeli, P.: Using stability tests and regional avalanche danger to estimate the local avalanche danger, Ann. Glaciol., 51, 176-186, https://doi.org/10.3189/172756410791386616, 2010.

Jamieson, B., Haegeli, P., and Schweizer, J.: Field observations for estimating the local avalanche danger in the Columbia Mountains of Canada, Cold Reg. Sci. Technol., 58, 84-91, https://doi.org/10.1016/j.coldregions.2009.03.005, 2009.

Schweizer, J., and Jamieson, J. B.: Snowpack properties for snow profile analysis, Cold Reg. Sci. Technol., 37, 233-241, https://doi.org/10.1016/S0165-232X(03)00067-3, 2003.

Techel, F., and Schweizer, J.: On using local avalanche danger level estimates for regional forecast verification, Cold Reg. Sci. Technol., 144, 52-62, https://doi.org/10.1016/j.coldregions.2017.07.012, 2017.

Techel, F.: On consistency and quality in public avalanche forecasting - a data-driven approach to forecast verification and to refining definitions of avalanche danger, Ph.D., Faculty of Science, University of Zurich, Zurich, Switzerland, 236 pp., 2020.