Reply to reviewer comments by Simon Horton

Frank Techel

Correspondence to: Frank Techel (techel@slf.ch)

Dear Simon

thank you very much for your very detailed and helpful review of our manuscript. We greatly appreciate the time and effort you put into this review.

Please find below our reply (in blue) to your comments *(in italics)*.

General comments

The paper addresses the subjective nature of avalanche hazard ratings by exploring a large dataset of hazard ratings and field observations. Since hazard ratings are defined by subjective terms, this study presents a unique approach to quantifying some of these terms. This could improve the consistency of hazard assessments and risk communication, which is an important contribution. The scientific methods applied in the paper are rigorous, and the conclusions drawn from the results are appropriate and interesting.

Thank you for this positive feedback to our manuscript.

My main comment is the key results of the study could be more clearly communicated and emphasized. I found there were some additional data and results that distracted from the key messages, and as a result I needed to read certain sections twice to properly understand the relevance. I think this could be easily addressed by restructuring and shortening some sections. My other main comment is I think more discussion about the application of the results (e.g. improved forecasting methods, danger rating definitions) would make the contributions clearer to the reader.

Specific comments

Relevance of additional data sets: The key conclusions of the study appear to come from the Swiss Rutschblock test and avalanche data, however while reading the results there are numerous references to patterns between the Swiss/Norwegian and RB/ECT data. I found this distracting from the main research question about the contributing factors to danger ratings. I would consider restructuring some of the sections so the main research question is addressed first, and then perhaps distinct sub-sections discussing how the core results differ between SWI/NOR and between RB/ECT results. There are also quite a few Appendix figures with these additional data sets which disrupts the flow while reading the results. Thank you for pointing this out. We agree, the structure of the manuscript should be improved to make it easier for the reader to follow the line of argumentation. We intend to restructure the Results section (Sect. 4) in a way that we first present the results using the Swiss data only, showing relevant figures and tables in the main part of the manuscript. Results relating to the methodology (bootstrap-sampling and classification approach), which we consider equally important to highlight strengths and limitations of the approach taken, we will move to a separate subsection. Other results, like the comparison RB to ECT, or comparing Swiss to Norwegian data, will be moved to a separate subsection with additional figures in the Appendix. In the Discussion section, however, we will keep the subsections where we discuss the limitations of the data used and the methodology, as we consider these relevant to understand the limitations of the study. Readers who are primarily interested in the findings of the study relating to the research questions, should be able to ignore these subsections easily.

Applications/next steps: I found the discussion focused too much on the limitations of the study rather than focusing on an in-depth discussion of how the results relate to the main research question. I would suggest shortening the limitations and adding more discussion about how these results could be applied to improve danger ratings. Linkages to existing hazard assessment frameworks (e.g. ADAM, CMAH) are discussed, but I would have found it interesting to read more about how the results could improve or unify the existing frameworks. For example, the stability matrix in Fig. 6 has many parallels with the likelihood matrix from the CMAH (Statham et al. (2018), Fig. 2), the Bavarian matrix, and ADAM. While these are discussed, I think the strong quantitative data in this study are well positioned to make an informed critique on existing methods and more suggestions for future directions.
Applying stability distributions: While the bootstrap sampling method is appropriate to derive stability distributions and define classes, it is somewhat theoretical and not clear how the derived classes could be applied in practice. Can these distributions help us understand more about stability conditions for a given hazard level? For example, from Fig. 3 could we assume that roughly 17% of slopes are unstable at 3-Considerable and 38% of slopes are unstable at stable at 4-High? I understand the challenges with making that inference, however, I wonder if the numeric nature of the study could help give some additional meaning to terms like few, several, and many. And although the theoretical meaning of these terms are clearly defined, how can forecasters actually assess whether the frequency of unstable locations is few/several/many? - Concerning these two points (application-next steps / applying stability distributions): With this study, it is our intention to provide the avalanche forecasting community with results regarding the contributing factors of avalanches grounded in data and robust methodologies. Primarily, we hope these findings will stimulate the discussion in the avalanche forecaster community when revising key terms and their definitions (a current project in the working group of the European Avalanche Warning Services) and the Avalanche Danger Scale. - We intend to take up some of the quantitative key results again, like the proportion of very poor tests at 4-High or the quantitative ranges describing the four frequency classes, and put these in relation with other studies (e.g. Thumlert et al., 2020)

Methods: The beginning of each sub-section could use a bit more context about how that step is relevant to exploring the link between danger ratings and a contributing factor. - We will add more context at the beginning of each section.

Figures: The figures are clear, legible, and support the main messages of the study well.

Thank you for pointing out typos and other improvements in your technical comments. We will address them when revising the manuscript and provide a detailed point-by-point reply when submitting a revised version of the manuscript. Please find our feedback regarding the more important points below.

Technical comments

- *p1 line 7: Although less precise, saying "frequency of unstable locations" may be simpler to understand when reading the abstract only*
- *p1 line 13-14: Consider adding "simulated stability distributions" (the snowpack distribution isn't simulated)*
- *p2 line7: Preferable to use consistent terminology from the list of key factors, i.e. "probability of avalanche release" instead of "release (or triggering) probability"*
- *p2 line 13: Similar to above, starting the paragraph by repeating the term "frequency and location of triggering spots" would make it clearer the paragraph ties back to the list of key factors*
- *p2 line 23: missing citation*
- *p2 line 24: According to the CMAH spatial distribution also considers spatial density. Statham et al. (2018) "Spatial distribution considers the spatial density and distribution of an avalanche problem and the ease of finding evidence to support or refute its presence."*
- *Table 1: The "data from" column heading isn't clear if the data is from just a single season or all seasons up to 2018/19 (as explained in footnote). Consider a more precise heading or list season ranges in the table (e.g. 2002-2019)*
- *p4 line 2-4: These two sentences aren't necessary, as they are discussed below.*
- *p6 line 4-5: Please be consistent with order of reporting SWI and NOR data, in this sentence NOR is described first.*
- *p6 line 8: It would be helpful to start this section by explicitly explaining the purpose of this step is to relate the snowpack test data to one of the explanatory factors in the study (i.e. probability of avalanche release)* - We will add some explanation in this regard.
- *p6 lines 19-26: This is an example of how the addition of ECT data confuses the reader and distracts from the main point.*

- *p7 line 2: It would be helpful to start this section by explicitly explaining purpose of this step to relate the snowpack test data to one of the explanatory factors in the study (i.e. frequency of triggering spots)*

- *Sect 3.2: This explanation of the bootstrapping method (and the accompanying Fig. 2) are very clear, concise, and effective!*

- *p7 line 15: What effect does an equal number of samples for each rating have considering there are likely a higher proportion of days with ratings of 2 and 3. The sample of 10,000 will likely have a skewed number of unstable tests from high danger days. Does this impact the interpretation of the results?* - An equal number of samples for each danger level is important, when the danger level for each combination is sought. For instance, if only 1% of the samples would have been 4-High, the danger matrix in Figure 6 would essentially never show a 4-High, as 3-Considerable would dominate these cells due to their larger weight. The definition of the class thresholds changes little, as the median proportion of very poor tests VP_{med}) drives them. When using a typical distribution of danger levels forecast in Switzerland instead (1-Low to 4-High, 18%, 43%, 36%, 2%, respectively), the variable which defines the class intervals VP_{med} is the same with 0.08. – We will add an explanation in that respect.

- *p9 line 28: Slightly confusing, perhaps add "… distribution of observed data for all days at a given danger level represent…"*

- *p10 line 1: Consider different verb than "complemented"*

- *Sect. 4.1.2: This section has many references to appendix figures, which disrupts the flow because the reader is compelled to flip back and forth to the appendix. The confusion could be reduced by introducing Fig. 4 earlier, which clearly shows the most relevant results, then followed by more discussion about the sensitivities to sample size, etc that reference the appendix figures.* - As pointed out before, we intend to restructure this section to make it easier for the reader to follow the line of argumentation.

- *p12 line 10: Are these proportions discussed later? They seem meaningful for interpreting stability test results (e.g. even dangerous days have relatively few sites with very poor stability).* - No, we did not discuss them later. They are the central tendency values for the four classes. - We will take them up at a later stage to highlight the quantitative meaning of the classes.

- *p14 lines 2-9: This is an example of where the comparison between countries seems like a secondary discussion point compared to reporting the main patterns between avalanche size and danger.* - These will be moved to a separate subsection.

- *p15 line 9: In this list the percentages reported in brackets could be misinterpreted as proportion of locations with very poor stability. Perhaps the first reported percentage could explain what the percentage means, e.g. "(53% of sample)".*

- *Fig. 6-8: Good use of figures with a consistent layout showing the lookup table and the supporting data. The idea that Fig 7 and 8 have the exact same matrix structure as Fig 6 wasn't fully clear on the first read, so could perhaps be explained more explicitly in the text.*

- *p20 line 17: "while observations of natural or artificial..."*

- *p20 line 27: Captured "slope stability" or "regional danger"?*

- *Sect 5.3.1: Another consideration when comparing with existing methods is the CMAH assesses the frequency of trigger spots for each avalanche problem rather than snowpack as a whole as done in the EAWS matrix. This may make it easier to answer questions about the frequency of unstable locations for a specific problem type but could make it more difficult when combining avalanche problems into an overall danger rating. Just an additional thing to consider when discussing how we can better assess the spatial frequency of instabilities.* - We agree that assessing the spatial frequency of instabilities in an actual situation is certainly not an easy task. Focussing on a specific avalanche problem may indeed allow to make a process-based guess. Still, as you state, the final compilation into one danger level is not straightforward – unless you opt for the worst combination across the various problems.

- *p24 lines 5-9: An updated citation with more comprehensive analysis is Clark (2019), where the influence of many factors on danger ratings are explored (size, likelihood, problem type, region, vegetation band, etc.). The importance of "likelihood" in Clark (2019) still agrees with the main findings in this study.*

References

Clark, T.: Exploring the link between the Conceptual Model of Avalanche Hazard and the North American Public Avalanche Danger Scale, Master's thesis, Simon Fraser University, 115 p., 2019.

Statham, G., Haegeli, P., Greene, E., Birkeland, K., Israelson, C., Tremper, B., Stethem, C., McMahon, B., White, B., and Kelly, J.: A conceptual model of avalanche hazard, Natural Hazards, 90, 663 – 691, doi:10.1007/s11069-017-3070-5, 2018.

Thumlert, S., Statham, G., and Jamieson, B.: The likelihood scale in avalanche forecasting, The Avalanche Review, 38, 31–33, 2020.