Review of "A user perspective on the avalanche danger scale – Insights from North America", by Morgan, A., Haegeli, P., Finn, H. & Mair, P.

General comments

This paper investigates how different levels of avalanche danger, visualized by the avalanche danger scale, are interpreted and applied by backcountry users. As described by the authors, "snow avalanches are a serious natural hazard […] that can threaten settlements, transportation corridors, critical infrastructure" (row 32-34). In addition, avalanches pose a lethal threat to recreational users of mountainous terrain.

The avalanche danger scale is a color coded 5-level scale that summarizes current avalanche conditions. The scale is ordinal and uses natural language to describe the different levels (1 - low, 2 - moderate, 3 - considerable, 4 - high, 5 - extreme). Although the numbers used (i.e., 1 – 5) suggests a linear scale, the underlying scale is meant to be exponential (i.e., $b \cdot a^{danger\ level}$, where b and a are some constants). Based on Munther's calculations on avalanche activity in Switzerland, the hazard increases two-fold between each level. This suggests that $a = 2$. In addition to the forecasted danger level, bulletins usually also include a main message, a description of the most important avalanche problems, distribution of these, and a snowpack history. The danger scale is commonly placed at the top of the information pyramid.

Many papers that analyze behavior and perception of risk in avalanche terrain include the avalanche danger as a factor. However, these papers usually focus on a few specific danger levels and do not estimate differences in perceived risk at different danger levels. A few papers have investigated if avalanche bulletin readers understand the information in the avalanche bulletin and draw correct conclusions from the information. A general finding is that the danger level alone is not sufficient to make correct terrain choices.

In contrast to previous papers, this paper analyzes how backcountry users perceive the increase in avalanche danger from one level to another. The paper further analyzes if different danger levels affect users' terrain choices differently, and if it is possible to identify different groups, with different perceptions of the danger scale and different behavioral responses. Finally, the authors also test if respondents remember the different avalanche danger levels, their names, and if participants can correctly rank the different levels.

The authors elicit perceived avalanche hazard by asking respondents to provide an interval of risk on a scale from 0 (no avalanche hazard at all) to 100 (widespread large natural avalanches reaching the valley bottom). The danger levels were stacked on top of each other so that respondents could easily compare one danger level to another.

The authors use a categorical scale to evaluate how the respondents use the danger scale to make terrain choices: 1) "I go primarily based on the danger rating", 2) "I go mainly on the danger rating, but I check other bulletin information", 3) "Avalanche problems and forecast details are the basis for my decision, and 4) "The danger rating alone prevents me from going".

The empirical analysis is based on a relatively large sample of backcountry users (N = 3195). The authors use a latent class mixed effects model and conditional inference trees to analyze the data.

Mixed effects models are used to analyze hierarchical data where observations are clustered (and possibly correlated) within groups (e.g., pupils in a school). Latent class mixed models allow researchers to identify latent group structures in a dataset.

In this paper, the authors use latent class modelling to identify different classes (or patterns) in danger scale recall, perceived functional form and use of the avalanche danger scale.

Conditional inference trees, also called unbiased recursive partitioning, is a non-parametric class of decision trees that uses statistical theory to select variables. The method is commonly employed in machine learning. The aim of the model is to identify variables (or rules) that partition the data into groups that are maximally different from each other. The rules are based on the statistical association between the explanatory (or predictor) variables and some outcome.

In this paper, the authors use conditional inference trees to understand how the response pattern in danger scale recall, perception and use questions relate to the background of the respondents. The paper is well-structured and the authors provide a clear and in-depth description of the data and empirical analysis.

Evaluation:

Both professional and recreational users of avalanche terrain rely on avalanche bulletins for information about avalanche hazard. It is therefore important to understand how users understand and use the information presented in the bulletins. Previous research shows that a fair share of backcountry users relies heavily on the danger level when they make decisions on where to go in the backcountry. It is therefore interesting to understand how backcountry users perceive and use the different danger levels of the scale. **In conclusion, the topic is relevant and should fit well within the scope of NHESS.**

The analysis of how people perceive and use the avalanche danger scale is novel and worthy of investigation. I am more skeptical about the value of testing if people remember the names of different danger levels. The latent class mixed effect model has not been previously applied in this field of research. It constitutes an interesting approach to gain a more detailed understanding of the data. The Conditional Inference tree method is an interesting approach to understand the association between background variables and the identified response patterns in the latent class analysis. The method can further identify new, interesting groups in the data. However, I have several concerns with the paper, described in detail below.

To make my discussion here meaningful, I first need to point out an error in the text, the authors say that class 7 in figure 5 is concave, and that class 1 and 6 are concave. This is wrong. The function in figure 5a is convex according to the commonly accepted mathematical definition. The functions in figure 5 e and f are concave. Below, I refer to convex as a function that has a slope that increases at higher danger levels, and a concave function as a function where the slope diminishes with the danger level.

A. Elicitation of perceived hazard at different danger levels.

Ambiguity of question asked

In the survey, the respondents are asked to: "For each of the five danger rating levels, use the two grey sliders to indicate the severity range of the associated avalanche conditions on a scale from 0 (no avalanche hazard at all) to 100 (widespread, large natural avalanches reaching valley bottoms)." It is not clear if the respondents are asked about their experiences of different avalanche hazard at different danger levels or if they are asked about what hazard each level *should* represent. This may seem like a detail, but it may be important for the results. The reason is that the forecasted avalanche danger may be wrong, e.g., a forecasted level 3 that is really a level 4. If the question is understood as "the spread in danger during a forecasted level 3", then we should expect to see overlaps between different danger levels (here, level 3 and 4). If the question is instead meant to capture the conceptual meaning of the danger levels, we should not expect such overlaps. The problem is that different respondents may have interpreted the question differently.

Bias

The respondents were asked to choose an upper and lower bound for the hazard at each danger level. I am sympathetic to this approach, because it is probably a lot easier to do this, than to provide an estimate of the exact level. However, this approach also causes problems. In the graphical analysis, the midpoint is used to identify the functional form. However, this presumes that the respondent thinks that midpoint represents the most common level of hazard at a given avalanche danger. This is a strong assumption if the respondents rely on their experiences of the forecasted danger level (i.e., when the hazard can be under- or overestimated). This is most evident for danger level 1 and 5 (a forecasted level 5 can be a real level 4 but never a level 6. Level 1 is similarly censored from below). However, it is also possible that there is bias in how levels 2 - 4 are forecasted. The problem is that we do not know what this bias is, and the approach used to elicit the functional form does not allow us to identify it. This problem could have been avoided if the respondents had been asked to identify a spread around a chosen point on the scale.

Construct validity

The aim of the question is to identify the perceived functional form of danger scale. Given that the danger scale is exponential with a base 2, this means choosing intervals around the points $b \cdot 2$ (low), $b \cdot 4$ (moderate), $b \cdot 8$ (considerable), $b \cdot 16$ (high), and $b \cdot 36$ (extreme). A participant who knows that the hazard doubles on each step thus first needs to understand that s/he should choose $b \cdot 2^{danger\ level}$ to find a point estimate on the scale, and then place the spread around this point. This is a challenging task. Since the endpoint of the scale represents a level 5 situation, this means that 100 should be included in the spread around level 5. Similarly, the endpoint for level 1 should include zero. This may lead to a different spread around these levels. Since the scale is exponential, the spread will increase with higher danger levels, unless the respondent is expected to leave gaps in the scale (suggesting that some hazard levels are not forecasted). In conclusion, my fear is that the complexity of the task makes it more into a test of math and graph skills than a test of avalanche hazard perception.

It may be that I am making a big fuss about something that is not a real problem. It is possible that people with expert knowledge of the avalanche danger scale would be able to rate the different danger levels in a way that would produce a convex function. To know if this is the case, the scale needs to be validated. As of now, we do not know what a "correct" answer looks like, or if it is at all possible to answer in a "correct" way. Unfortunately, the problem with the elicitation method severely limits the value of the analyses of the responses.

Recommendation: ask a panel of avalanche forecasters to answer the same question as in the survey. Use this as the "ground truth" and compare it to the responses used in the current dataset. In the expert survey, I would also ask the experts to provide a point estimate to get an idea about if the midpoint can be used as a point estimate for the functional form. To check that the ambiguity of the question (spread in actual hazard during e.g., forecasted level 2 versus which hazard level a level two should represent) does not affect responses, I also recommend to re-run the survey question on a small sample of backcountry users where you clarify what it is that you are after.

B. Use of the avalanche danger scale

As mentioned by the authors, the "use" question was relatively crude. It only asks to what degree the decision to go into the backcountry depends on the danger level, and not if the danger level affects terrain choices in the backcountry. It may also be mentioned that category 1 encompasses category 4 (both imply that the respondent rely mainly on the danger scale).

C. Latent class modelling

The authors use a latent class mixed effects model. This seems like an adequate approach, as it can identify interesting response patterns in the data. However, as noted by the authors, the model did not converge in the sense that the model fit kept improving at a higher number of latent classes. Instead, the authors stopped the process when the new latent classes did not provide meaningful information. I do not doubt that the research group took great care in analyzing the data and determining the number of latent classes. However, I still fear that they read more into the data than what is actually there. It seems reasonable to divide the responses into convex, linear, and concave functions, with narrow and wide spreads. However, although the classes are significantly different from each other, they are all relatively linear from a practical perspective (eyeballing the graphs, the step size in mid-points between different levels appears to be about 15-20 for all functions). Based on the AIC and BIC, all seven classes should be included in the model, but so should the discarded classes. This makes me wonder how much we gain from including so many classes and how much we should read into the differences.

Recommendations:

1. Perception of the avalanche danger scale:
    a. It would help if we knew why the kept classes are more meaningful than the discarded classes. I therefore recommend that a set of discarded classes are presented in an online appendix. These classes should consist of the first e.g., 3-6 classes that improves fit but that the authors deemed meaningless. I would also like to see a short discussion of why the discarded classes have less meaning than the included classes.
    b. Since the differences in functional form are relatively small, it would help if the authors discussed the practical implications of the differences. It would also be interesting to know which analytical functional form the different graphs represent (i.e., how far from an exponential scale are they?).

c. If the practical implications of the differences are important, I think that you should put more emphasis on the fact that there was no clear pattern among people who perceived the danger scale as convex with narrow spreads.

D. Conditional inference trees

Conditional inference trees offer a novel and interesting approach to understand the association between background variables and response patterns. In this paper, the trees are used to understand groupings according to 1) the share of correct responses on danger level rank question, 2) latent class combinations with respect to the functional form of the danger scale, and 3) latent class combinations with respect to use of the danger scale.

The conditional inference tree is very helpful to understand the association between background variables and the ability to correctly rank the different danger levels. The authors make an admirable effort to explain the results. However, the final nodes of the inference trees for perception of avalanche hazard and use are very complex and difficult to understand. The reason is that the nodes represent different combinations of latent classes. In the case of *use* of the danger scale, the latent classes further represent different combinations of user types. The problem is made worse because the different latent classes are ordered in a way that doesn't help the reader. I understand that the statistical software labels the classes and that the authors use these class names to enable replication. However, it makes it cumbersome to read the text. I had to scroll up and down to recall what the different classes represent, and make notes, to have a chance to interpret the figures.

Recommendations:

I think that it would be beneficial if the authors renamed the classes so that the follow a logical order and given names that describes the class (e.g., for perception: class 1 – convex, class 2 - linear, class 3 – concave. For user group; class 1 – rely mainly on danger level at low levels, class 2 – rely mainly on danger level at high levels, class 3 – rely mainly on other information in avalanche bulletin).

Minor points:

1. I am not sure why it is interesting to analyze if people recall the names of the different danger levels. If kept in the paper, I would like to see a clearer discussion about what we learn from the results.
2. Section 3.3. I think that it would help to first describe the different classes (give some intuition for them) and then describe how many participants fall into each class. I suggest to organize the classes into Convex, Linear, and Concave.
3. Figure 5. I would like to see a caption and figure order that is consistent and easy to remember. I found the figure very confusing, because the caption mixed "a" to "f" with the class labels, and the class labels did not follow any logical order (to me as a reader).
4. Section 3.4. I would have preferred to read a description of the order of the different classes before the presentation of the list, as this would have helped me understand the order. If possible, order the classes based on how "correct" the strategy is. If that is not possible, give the classes a name that describes the group. This will help when the reader is trying to understand figure 8.

5. Discussion: I would like to see a discussion of why we should expect to see a relationship between a specific functional form of the danger scale and a certain use of the scale. It is not clear to me if one of the use strategies is more correct than another and if a e.g., convex functional form would imply more or less reliance on the scale relative to other information in the bulletin.
6. Discussion: the authors say that the results indirectly validate the Bulletin user typology. I would like to argue that this is incorrect. Both the user typology and the question about use of the avalanche scale asks how a person uses the avalanche bulletin, i.e., they are two ways to ask the same thing. The results therefore suggest that we can measure self-assessed use with two methods. To validate if the bulletin user typology represents actual differences in use, we would have to look at actual use. I would like to see a more nuanced discussion about this.
7. Discussion: The discussion about risk perception seems misplaced. The questions in the survey ask about the objective risk for widespread very large avalanches and not about ability to mitigate the hazard. Although the discussion is fascinating, I don't think that it belongs in this paper.
8. All datafiles are available on osf. This is great! However, I would also like to see that the questionnaire (in full) was publicly available. Information about the length (average completion time) should be provided in the text, together with information about where in the survey the questions were placed. The latter is important since the task may have been perceived as relatively challenging for the respondents.