Author response

Discussion of '*A quantitative module of avalanche hazard—comparing forecaster assessments of storm and persistent slab avalanche problems with information derived from distributed snowpack simulations*'

Table of contents

Line numbers in parentheses given after Author Responses refer to the track changes document.

1 Responses to referee comment #1 (Zachary Miller)

1.1 Overview

1.1.0.1 Referee Comment:

The manuscript titled "A quantitative module of avalanche hazard—comparing forecaster assessments of storm and persistent slab avalanche problems with information derived from distributed snowpack simulations" sets out to improve avalanche forecasting quality through the development of an additional toolset and analyzes the effectiveness of this toolset in this paper. The authors leverage recent scaling developments in the utilization of the SNOWPACK model to produce spatially distributed snow cover outputs over ten winter seasons for the Glacier National Park, Canada. They then post-process this data to produce numerical predictions of the characteristics of storm snow and persistent avalanche problems and compare those results against the timeseries of daily human hazard assessments. Their comparison is extensive and thorough, both evaluating broad trends and specific event day-to-day evolutions of avalanche problems. They describe their methods clearly despite the relative complexity required in post processing and inter-comparison of their datasets. The results and discussion clearly establish where their work fits within the current sphere of research and the contribution their methods offer to the snow and avalanche science community.

I do not have any major issues with this manuscript and feel that it is of very high quality. The largest question raised is whether or not they considered comparing their two hazard assessments (simulated and human) against observed avalanche records? I realize that the additional effort is probably outside the scope of the current project and that there are known limitations to observational records but also believe that their specific domain - Glacier National Park, Canada - has potentially one of the most complete and thorough records available due to the high level of professional avalanche activity in the terrain in and around the park. These records, and additionally utilizing the Avalanche Hazard Index (Schaerer, 1989), could provide a relative "truth" in the avalanche hazard characteristics being compared. I would

appreciate a response to my question but do not feel that this further analysis is required for publication given the robustness of the work presented.

1.1.0.2 Author Response:

Thank you very much for your evaluation and the supportive assessment, we highly appreciate it!

We appreciate your comment about considering observed avalanche activity as additional validation data set. Glacier National Park, Canada, indeed has valuable data on avalanche activity that could provide interesting insights for validating the simulations. Our main reasons for not including such comparison in our current manuscript are as follows. We focused on one validation data set (human assessments) and aimed at extracting as much useful information as possible from the comparison to the simulated data set, exploring the challenge from several perspectives. This made the methodology already quite complex and including yet another different data set would risk making the story line more confusing. Moreover, the comparison against avalanche activity comes with several challenges that would have blown up the current manuscript. First, human observed avalanche activity comes with known caveats. For example, "no observed avalanches'' does not always mean that there were no avalanches (e.g., bad weather and visibility, limited terrain available for observations), and the timing of observations might not coincide with the release of avalanches or the peak modeled instability. Automated detections of avalanche activity in Glacier National Park are limited to the highway corridor, which hosts an artificially managed snowpack. In other words, it is only suited to validate storm snow instabilities, but not so much persistent problems. Then, there remains the challenge of translating artificial triggering of storm slabs with explosives to the modeled storm slab characteristics. Is the currently used instability model p_unstable (Mayer et al, 2022) suited, or would we need to prefer the avalanche day predictor by Mayer et al (2023)? How to include process-based indices? Overall, we agree and see it as highly important to carry out these comparisons against observed avalanches. A dedicated study might be best suited for this endeavor and ongoing conversations at the international working group AvaCollabra could help in designing a research approach in the most meaningful way. We included a statement in our Conclusion section that highlights the opportunity for a validation against observed avalanche activity (see our responses to the specific comments further below).

Thank you also for the specific technical corrections and comments. We incorporated them into the revised manuscript and added clarifications where required. Our detailed revisions are documented in the specific comments below.

1.2 Specific technical corrections and comments

1.2.1 Comments on language

1.2.1.1 Referee Comment:

Line 100 – use of word "over" is confusing within the discussion of the ordinal likelihood of avalanches scale, simply remove to clarify

Line 106 – use of word "over" is confusing within the discussion of the ordinal North American Public Avalanche Danger Scale, simply remove to clarify

1.2.1.2 Author Response:

Thank you, done!

1.2.1.3 Referee Comment:

Line 119 – "layers that were exposed to the snow surface" is confusing, perhaps adjust simply to "layers that were the exposed snow surface"

1.2.1.4 Author Response:

Thank's for the heads up. We changed the phrase to *"…layers that were exposed to processes happening at the snow surface before the snowfall…"*. (L119)

1.2.2 Date tags for crusts

1.2.2.1 Referee Comment:

Line 120 – "represent rain events that form a crust" implies that rain is the only way surface crusts form and the date tags probably include additional crust formation events. Perhaps something like: "represent crust formation events at the snow surface such as rain or insolation-driven melting."

1.2.2.2 Author Response:

You are right, crusts can form due to different processes. In this case, our date tags only partition layers into precipitation-driven groups of layers. This means, that insolation-driven crusts would be grouped with other layers based on the timing of their burial. We added a footnote, *"Please note that other processes leading to crust formation, such as insolation, are not captured by the date tags as presented in this study. In this case, the present approach groups crust layers with other layers based on their burial date"* (L122). This decision has no effect on the results presented herein.

1.2.3 Confused values for p_unstable

1.2.3.1 Referee Comment:

Line 180 — It seems as though the likelihood of avalanches should be represented by the layer with the highest p_unstable value rather than the lowest?

1.2.3.2 Author Response:

You are completely right. We meant the *highest* value, not the lowest. We corrected this and also changed the confused "≤" signs in this section. (L177, L179, L182, L184)

1.2.4 Figure 5

1.2.4.1 Referee Comment:

Figure 5g — Color and interquartile range of the air temperature line is the same as p_unstable in plots 5c-5f and, therefore, is slightly confusing. Consider changing it to dashed or a different color to differentiate it.

Figure 5h – Color of median high of new snow is the same as storm slab instabilities in plots 5d and 5f, and, therefore, is slightly confusing. Consider changing the color to differentiate it.

1.2.5 Author Response:

Thank you for these suggestions. In the revised manuscript, we use different colors for these panels.

1.2.6 Figure 6

1.2.6.1 Referee Comment:

Figure 6 — It appears, but is not explicitly described, that assessed values are solid and modeled values are hatched. Consider mentioning this in the figure description or adding a legend.

1.2.6.2 Author Response:

We added a sentence to the caption that clarifies this misunderstanding. We did not add a legend, since it would basically look identical to panels (b) and (d).

1.2.7 Comments on language

1.2.7.1 Referee Comment:

Line 299 – "increased strongly" doesn't make sense in reference to the simulated depth of the weak layer, perhaps remove "strongly" or change wording to "increased substantially"

1.2.7.2 Author Response:

Agreed! We changed it to *"increased substantially"*. (L309)

1.2.7.3 Referee Comment:

Line 304 – "short and moderate peaks of modeled instability" is confusing since "moderate" is not defined within the spectrum of modeled instability

1.2.7.4 Author Response:

We rephrased the sentence to *"...several short and mid-level peaks of modeled instability around the threshold of 0.77..."*. (L314)

1.2.7.5 Referee Comment:

Line 331 – "as" is meant to be "was"

1.2.7.6 Author Response:

Fixed!

1.2.7.7 Referee Comment:

Figure 8c & 8d – "Turn-on" and "Turn-off" are confusing category names and I am interpreting it them as if the avalanche problem was assessed ("Yes"/"No") or added/removed?

1.2.7.8 Author Response:

We changed the labels to *"Added"* and *"Removed"*. The revised text also uses these terms.

1.2.7.9 Referee Comment:

Line 359 – "is unstable" should be "are unstable"

1.2.7.10 Author Response:

Thanks, fixed!

1.2.8 Hazard ratings for storm and persistent problems

1.2.8.1 Referee Comment:

Line 361-363 – You mention that deeper splits show a generally higher hazard rating associated with storm snow problems than persistent problems. Is this due to the relative frequency of lower hazard ratings associated persistent problems (aka "spicy moderate" or existing for weeks) vs. the short-term spiking hazard commonly found with storm snow problems (quick to rise and quick to fall)? If so, I believe it is worth clarifying the relative temporal effects of both types of problems because your results discussion seems to say that the model simply pairs lower hazard with persistent problems.

1.2.8.2 Author Response:

Unfortunately, we can only observe this relationship, but not infer *why* this is. However, given other publications who have made similar observations, we can discuss this topic in detail. We did not make any changes to the discussion of the revised manuscript, because (former) L450–460 speak to your question (L363–473). We did, however, add a statement to the Results section (L372f) where this topic first comes up and refer readers to the Discussion.

1.2.9 Typo sentence from earlier manuscript version

1.2.9.1 Referee Comment:

Line 380-381 – Delete sentence since it seems you are submitting the manuscript for publication in a peer-reviewed journal and I hope you've taken those additional research steps.

1.2.9.2 Author Response:

Thanks for spotting this relict of a prior version of the manuscript. All of the associated research steps had been integrated in the meantime. (L391f)

1.2.10 Kudos

1.2.10.1 Referee Comment:

Line 423-425 – Very concise distillation of the effectiveness of your model – nice job.

1.2.10.2 Author Response:

Thank you very much!

1.2.11 Language around removal of persistent problems

1.2.11.1 Referee Comment:

Line 444-445 – How do forecaster's removal of reported persistent problems appear arbitrary? That is a big statement to make given the multitude of factors forecasters must balance to make that call.

1.2.11.2 Author Response:

Thanks for this comment and giving us a heads up that our choice of words was not optimal. We certainly value the expertise and judgment of forecasters that lead to their decisions and acknowledge that their decisions are based on a multitude of factors, often hidden in messy or sparse data sets. With the preceeding sentence leading up to this topic and a change of wording in the given sentence, the revised statement should now be clear and non-offensive: (L468)

> One of the situations where simulated avalanche problem characteristics could be particularly helpful for forecasters is the removal of persistent slab avalanche problems. Both @Horton2020b and @Hordowick2022 found that forecasters struggle with the decision to remove persistent slab avalanche problems. We also found evidence of this issue in our analysis where the timing of forecasters' removal of reported persistent problems was often much later and mostly uncorrelated to the simulated avalanche problem characteristics. Hence, in these difficult to assess situations, the simulations might provide valuable information about the instability of the relevant weak layers.

1.2.12 Avalanche observations

1.2.12.1 Referee Comment:

Line 545 – Understanding the truth of avalanche hazard is infinitely complex. Has your team considered comparing your results with observed avalanche activity to quantify the accuracy of avalanche depths/size and loosely distribution (despite the inherent bias in physically observed avalanche records)? I wonder if this could help clarify a relative truth especially when the simulated and reported hazards differ?

1.2.12.2 Author Response:

Great comment! We responded to this idea in detail in the Overview (Section 1.1) at the beginning of this author response. We also added a statement to the Conclusion section: *"Furthermore, future research should leverage existing datasets on avalanche activity, such as remote detections and manual observations, to better identify times when model simulations align with actual conditions."* (L565f).

2 Responses to referee comment #2 (Veronika Hatvan)

2.1 Overview

2.1.0.1 Referee Comment:

The manuscript titled "A quantitative module of avalanche hazard—comparing forecaster assessments of storm and persistent slab avalanche problems with information derived from distributed snowpack simulations" presents a well-executed study aimed at enhancing avalanche forecasting by integrating numerical snowpack simulations with human judgment. The study introduces a spatial approach to extracting the characteristics of storm and persistent slab avalanche problems from distributed snowpack simulations, tracing individual snowpack layers over time and space. This approach enables the calculation of spatial distributions of avalanche problem characteristics and presents the data in familiar hazard chart formats aligned with the (CMAH).

The study leverages snow cover data spanning ten winter seasons from Glacier National Park, Canada, to examine the agreement between snow cover simulations and human assessments for persistent and storm slab avalanche problems. The comparison is thorough, addressing both seasonal trends and day-to-day evaluations. The authors clearly describe their methods, which are up-to-date and well-suited to the study's objectives. This work aligns well with recent advancements aimed at integrating snowpack modelling more closely with operational forecasting workflows.

The applied methods and developed approaches are of high quality and contribute significantly to the goal of further integrating snow cover modelling results into avalanche forecasting workflows. The comparison between modelling results and human assessments provides a valuable foundation and insights for future applications.

2.1.0.2 Author Response:

Thank you very much for reviewing our manuscript and providing very supportive feedback! We highly appreciate it!

We incorporated your suggestions in the revised manuscript as explained by the detailed responses below.

We will include your suggestions into the revised manuscript, particularly use terminology more consistently and clearly (e.g., p_unstable), and make Figures 5 and 6 more accessible. We will also include an introductory paragraph that aims to reconcile the two different terms frequency distribution and spatial distribution, which are being used in Europe and North America, respectively.

2.2 Specific technical corrections and comments

2.2.1 Consistent language: model predictor p_unstable

2.2.1.1 Referee Comment:

Line 174: I noticed a small typo in the phrase 'characteristics avalanche problem type'; I assume it should be ' characteristic avalanche problem type'? Additionally, I concur with the comment by Frank Techel that, for clarity, it would be beneficial to consistently use the term p_unstable when referring to the model predictor. This distinction will help avoid (my) confusion, when differentiating it from human assessments (e.g., Line 174 and other locations).

2.2.1.2 Author Response:

We fixed the typo and adjusted the language around human reported "likelihood of avalanches" versus model predictor p_unstable to be more consistent throughout all sections. Thank you for this comment!

2.2.2 Layer depth

2.2.2.1 Referee Comment:

Line 175 – 178 & 180 - 181: To me, it is unclear how the depth of the identified layer differs from the depth of the deepest unstable layer. To my understanding, these two are the same. Consider revising for more clarity, otherwise I would be interested in a reply to clarify this distinction.

2.2.2.2 Author Response:

We clarified this in the revised manuscript: *"… and the release depth corresponds to either the depth of the identified layer (i.e., stable case) or the deepest unstable layer (i.e., if one or more layers are classified as unstable, with* $p_{unstable} \geq 0.77$*)"* (L178f).

2.2.3 Typo

2.2.3.1 Referee Comment:

Line 178 & 181: I assume this is a typo, and p_unstable ≤ 0.77 should instead read p_unstable ≥ 0.77.

2.2.3.2 Author Response:

Correct. Thank you, we fixed this.

2.2.4 Terminlogy spatial vs frequency distribution

2.2.4.1 Referee Comment:

Line 186: I do not understand the term 'spatial distribution' as it relates to Figure 4, as I don't see any spatial component represented in the figure. Consider revising this term for greater clarity.

2.2.4.2 Author Response:

We added a footnote to the relevant mention of the term 'spatial distribution' to clarify potential confusion. A more detailed author response is given to the same comment from Frank Techel further below.

The footnote reads: *"Please note that the term 'spatial distribution' is used slightly differently in North America and Europe. European forecasters primarily use the term 'frequency distribution' and only refer to spatial distribution when they know the exact locations in the terrain (e.g., on a map view). Since this paper builds on the CMAH, we use the term spatial distribution throughout this manuscript."* (L170)

2.2.5 Figure 5

2.2.5.1 Referee Comment:

Figure 5: For clarity, it would be helpful to choose different colours for air temperature and HN24, as the current colours are very similar to those used for p_unstable and avalanche problems in the upper panels. Additionally, I assume that air temperature and HN24 are based on modelled data rather than measurements? It might be beneficial to add a small comment on this for clarity.

2.2.5.2 Author Response:

We changed the colors and added a clarification to the caption that the variables are derved from model runs.

2.2.6 Figure 6

2.2.6.1 Referee Comment:

Figure 6: It took multiple views to understand that the hashed bars represent modelled data and the full-colour bars represent assessed data. To improve clarity, consider explicitly stating this in the figure caption for easier understanding."

2.2.6.2 Author Response:

Thanks for the heads up. We fixed this (see similar comment from Referee #1).

2.2.7 Typo sentence from earlier manuscript version

2.2.7.1 Referee Comment:

Line 380: Remove sentence – you are already submitting to a peer-reviewed journal.

2.2.7.2 Author Response:

Thank you, we removed the sentence.

3 Responses to community comment #1 (Frank Techel)

3.1 Overview

3.1.0.1 Referee Comment:

3.1.0.2 Author Response:

Thank you for your interest in this manuscript and for providing feedback to improve it!

We included your suggestions in the revised manuscript as detailed below.

3.2 Specific technical corrections and comments

3.2.1 Consistent language: model predictor p_unstable

3.2.1.1 Referee Comment:

While explaining the link between the distribution of p_unstable and the likelihood of avalanches (in CMAH) makes sense (Sect. 3.1), consider referring to p_unstable rather than the likelihood of avalanches when referring to p_unstable (e.g., L174). This would ease understanding when referring to model predictions and when to human assessments.

3.2.1.2 Author Response:

Thank you for this comment! We think that the changes made to the revised manuscript (see similar comment from Referee #2) will make the manuscript more accessible and understandible.

3.2.1.3 Referee Comment:

Along that same line, you use expected p_unstable for the first time on L211. I presume it is meant to be introduced on L191(?) though it is referred to as the expected likelihood of avalanches. Only later I noticed that in Figure 4a, the expected p_unstable values are shown but this is nowhere mentioned (or I missed it). From Fig. 4a, I took that the expected p_unstable is the mean of all the p_unstable-values in the plot. On L189 you refer to the likelihood of avalanches (but I presume this is again p_unstable) for which you derive various percentiles. Consider indicating that the 50th percentile is what you call expected p_unstable.

3.2.1.4 Author Response:

With the changes made to the terminology around "likelihood of avalanches" and "p_unstable", this issue is basically resolved, too. The expected p_unstable (and expected depth) is now properly introduced in former L191, and we added explicitely that it represents the 50th percentile. (L194–196)

3.2.1.5 Referee Comment:

L164: you say that you used the threshold p_unstable >= 0.77 to define layers with poor stability, as proposed by Mayer et al. (2022). But afterwards, you seem to analyze exclusively p_unstable; at least in all the figures p_unstable is shown. - Why did you primarily explore p_unstable and not the proportion p_unstable >= 0.77? I would expect that this explains why your distribution of 2 (moderate) was wider compared to Switzerland, while the distribution

of 3 (considerable) was wider in Switzerland than in your data. (L404-406). You also say something along that line. — Out of curiosity, while analyzing, did you plot Figures 7a, c, e, g and Figure 8g and h using the proportion unstable rather than the expected p_unstable?

3.2.1.6 Author Response:

Our analysis focuses primarily on the examnation of the distribution of p_unstable, where the proportion of unstable grid points (with p_unstable >= 0.77) is only one piece of multiple. The threshold of 0.77 is also used to compute the expected depth, which uses only unstable layers as subset.

As explained in our Reply on CC1, we re-plotted the relevant figures with the proportion unstable (as suggested in the comment) and the 90th percentile of p_unstable (which was shown to be an important variable during our CTree anaylsis, Fig. 10). While the re-plot of Fig. 8 did not add value, the re-plot of Fig. 7 indeed shows an interesting perspective with added nuance. Since this figure is partly redundant to the results of our CTree analysis, we placed the additional figure in the Appendix (Fig. A1) and added a paragraph to the Results section (L271–276).

> In addition to the expected p_{unstable} (i.e., 50th percentile), we explored other variants that describe the distribution of p_{unstable}, such as the 90th percentile of p_{unstable} or the proportion of unstable grid points (with $p_{\text{unstable}} \geq 0.77$) (Fig.~**??**). While the general patterns remain similar, this comparison illustrates that not one single descriptor best discriminates between the different danger ratings but that a combination might be required (also see CTree analysis further below). Notably, the 90th percentile of p_{unstable} best discriminates the danger rating *Low* from the others. While it is located close to the threshold of 0.77 for *Low*, it is substantially higher for all other danger ratings.

3.2.2 Typo

3.2.2.1 Referee Comment:

L178: I assume this is just a typo, it should probably read >= rather than <=

3.2.2.2 Author Comment:

Yes, thank you. Fixed.

3.2.3 Terminlogy spatial vs frequency distribution

3.2.3.1 Referee Comment:

L186: I don't understand how the point cloud in Figure 4 can provide a spatial distribution. I am aware that spatial distribution is the term used for the number of triggering locations in CMAH. In this context, I found it rather confusing as the distribution in the plot doesn't have a spatial component. - Consider changing to something like "the number of potential triggering points within a region can be gauged from the distribution of p_unstable". At least to me, p_unstable provides primarily an indication of potential instability considering a Rutschblock test. Of the unstable locations, only a small fraction will be sufficiently unstable to result in human-triggered (or natural) avalanches (Mayer et al., 2023).

3.2.3.2 Author Comment:

Thanks for the suggestion of how to rephrase the sentence to satisfy both North American and European terminology with respect to spatial versus frequency distribution. We value your paraphrase and included it in the revised manuscript. (L188f)

Regarding the more general conversation around the terminology spatial vs frequency distribution, we see two key aspects: accessibility and accuracy of the terms.

The term "spatial distribution" was selected in the CMAH to prioritize accessibility. This term can be relatively easily understood by the avalanche community because it emerged from avalanche practitioners and academics alike. While "frequency distribution" might be the more precise term (more on that in the next paragraph), it can be more challenging to interpret in its English translation. Although the German origin "Häufigkeitsverteilung" appears both precise and accessible, "frequency distribution" does not represent a satisfiable translation. Imagine german-speaking backcountry skiers were asked about the "Frequenzverteilung" or "Häufigkeitsverteilung" of snow instability, which term would be better understood?

When it comes to accuracy of the term, spatial distribution and frequency distribution are very closely connected. We know that the snowpack and its properties vary with space, and the accepted term for this is spatial variability. This spatial variability of snow is the reason why avalanche forecasters are interested in the number of trigger spots. These trigger spots are distributed across space, i.e. "spatial distribution", but a forecaster does not necessarily need to know (and does not precisely know) where exactly they are; this is when the "spatial distribution" looses its space component and becomes the "frequency distribution". Frequency distribution therefore seems more precise, but not necessarily more accurate. Particularly in the context of simulations we actually do know the exact distribution of the points across space (every point on the chart in Fig.~4 can be placed on the map). Does discussing frequency distribution on a chart versus spatial distribution on a map add value, when both visualizations actually refer to the same underlying variable?

Ultimately, a term should be selected based on both these perspectives. (Are there additional ones)? It would be desirable to reconcile the diverging terminology in North America and Europe, and we hope that this discussion can provide a small piece to that effort.

3.2.4 Expected depth

3.2.4.1 Referee Comment:

L192: How do you proceed if none of the profiles fulfilled the p_unstable >= 0.77 - criteria for deriving the expected depth when no such layers were present?

3.2.4.2 Author Response:

When none of the grid points were unstable, we did not compute the expected depth and assigned NaN values to those results. We added the footnote *"Note that an expected depth is therefore only computed when there is at least one unstable grid point."*. (Footnote 5, below L202)

3.2.5 Correlation with danger level

3.2.5.1 Referee Comment:

You mentioned twice that p_unstable correlated more strongly with the danger level than the likelihood terms (e.g., L311-312). This is interesting. - Is this maybe due to likelihood estimates being less reliably estimated by forecasters as compared to the reliability of danger level estimates? Or is this maybe linked to the fact that p_unstable is a mix of Rutschblock stability and danger level (p_unstable actually correlated more with danger level than RB stability, see Mayer et al., 2022 [p.4601])?

3.2.5.2 Author Response:

Thank you for this comment! Our results show more consistency between human–model for the higher-level variables. We observed this pattern not only for the danger rating, but also for the presence or absence of a problem, for example. Therefore, I do think that the more intricately-to-assess characteristics, such as likelihood, are less reliably and less consistently estimated by the forecasters. However, I am grateful for your hint that the local danger level indeed plays a role in the training process of p_unstable by filtering for stable and unstable profiles that were observed during low and high danger days, respectively. We added a sentence to the revised Discussion to include this subtle but important detail: *"Although $p_{unstable}$ was only trained with snowpack characteristics, the local danger rating was used to filter only for stable and unstable profiles at Low and High hazard days, respectively, which may create a subtle correlation."* (L481f)