

Reply to Referee #1 (James Floyer)

We thank the reviewer for the positive, constructive review and helpful feedback. We reply below point by point in blue.

General Comments:

The paper compares several snow stability metrics derived from snowpack modeling against a rich dataset of field measurements collected at the Steintälli study site during the winter 2015-16. The stability metrics may be classed as process-based (skier stability index SK38, critical cut length r_c) or data-driven (machine learning snow instability metric P_{unstable}). The paper discusses the challenges of determining which weak layer in simulated stratigraphy output is the most critical and therefore the most likely to cause avalanche activity. The authors consider three approaches to selecting most critical weak layer: maximum P_{unstable} , relative threshold analysis (RTA), and persistent avalanche problem (pAp) approach. In total, six methods that perform the dual tasks of selecting the critical weak layer and quantifying its stability were evaluated, along with the non-modelled, three-day new snow (HN3d) benchmark.

The field measurement dataset is unusually detailed, comprising shear frame measurements, stability tests (PSTs, ECTs, CTs), detailed snow profiles, automated weather station data, and snow micro-penetrometer profiles. This, combined with how the season evolved to develop several persistent weak layers and several significant avalanche cycles, makes it an extremely valuable dataset, well suited to the task of comparing modeled stability metrics.

The manuscript is well written and the analysis is sound. The authors discuss limitations (training data/validation data overlap, and scale mismatch between point observations and regional danger ratings) and the results generally support the interpretations and conclusions. I recommend publication after the following comments are addressed.

Specific Comments:

The overall accuracies of the stability metrics reported in Table 5 suggest a range between 53% and 73% (PC of 0.53 to 0.73). Can you explain why the overall accuracy range suggested in the abstract is stated as 60-70%?

Thank you for spotting this inconsistency. We will change this to 53-68%, i.e., not including the benchmark parameter HN3d.

Several of the metrics have performance skill scores that are very similar, yet no formal confidence intervals are reported. Despite this, there are several places in the manuscript where the relative ranking of the metrics is discussed, admittedly with

cautious language: e.g. L. 495 “ P_{unstable} performed slightly better than the traditional skier stability index RTA-sk38”. How confident are you in their ranking? How confident would you be that the relative ranking would change if applied to a different area, or a different set of conditions in the same area?

We will provide confidence intervals for the accuracy (PC). In fact, given the low number of cases ($N = 74$), the confidence intervals are large and hence overlapping. Therefore, the ranking is formulated cautiously and may change if applied to a different dataset. However, we suppose that again the differences in performance would be rather minor, as was the case when we changed the threshold for the AAI.

Further to the similarity of some of the metrics, the authors explicitly mention (L. 418) that the strong negative correlation between P_{unstable} and $pAp-r_c$ is expected as r_c is one of the 6 features considered to estimate P_{unstable} . Does this erode the statement in the conclusion that a multi-model approach may be advantageous for operational forecasting? Are there additional stability metrics available or being developed that offer a truly independent assessment from existing approaches?

We think the multi-model approach is still advantageous, in particular, since we suggest that the temporal evolution is the feature that is probably the most useful for operational forecasting. Moreover, different ratings between the metrics, which are not uncommon, may always be worth a closer look.

There is one additional metric that was introduced by Gaume and Reuter (2017) that we did not consider since it has never been used operationally. However, it has recently been further developed, now called SK_{38}^{CCL} , and seems to perform well (Borsotti et al., 2026).

The authors are honest about the overlap between the training dataset and the validation dataset (L. 453), where they state that the training dataset for P_{unstable} contained 11 profiles from Davos from 2015/16. Ideally the impact should be characterized by reporting whether excluding the 11 overlapping profiles from the training dataset changes the results. This feels especially pertinent in light of the discussion above, that relatively small differences in performance skill scores are being discussed and characterized. If such reanalysis is not possible, it would be better to highlight this limitation in a stronger way, perhaps by repeating it in the conclusion.

The metric for P_{unstable} was developed using a large dataset (742 profiles) covering many different winter seasons. Given the minor overlap, we think this is acceptable; though, we agree not ideal.

What is the rationale for correcting the avalanche danger level (L. 190)? If the uncorrected danger level was what was available to the public when the regional forecast was issued, isn't that perhaps a more meaningful comparison, as it recognizes the limitations of human forecasting. As presented, the danger level has been altered to improve correlation with the AAI. Both are presented in Figure 4, ostensibly to offer comparisons to two different measures of avalanche danger. However, these are no longer independent, as one has been corrected against the other. I note that the caption describing the avalanche danger level ("Avalanche danger level for the region of Davos") does not indicate that a corrected version of the danger level has been used. At a minimum, I would like to see the caption of Figure 4 updated to indicate that the corrected avalanche danger level is used.

As described, the adaptation of the danger level does not affect the binary classification. The rationale for the adaptation is that for any model development or validation, a verified forecast is more useful. We will update the caption in Figure 4.

PON (L. 334) is not defined (presumably the probability of non-events).

Thanks, correct; we will define it in the revised version.

Figure 4: There is a small orange spot on Pane (c) in mid-March that does not have a corresponding orange shaded region on the lower half of Pane (a). I suggest adding a small shaded region on the P_{unstable} region of Pane (a).

Thank you for spotting this inconsistency. We will correct this in the revised manuscript.

Technical Corrections:

1. L. 146: "Fig.1" should read "Fig. 1" (with space)
2. L. 196: "Fig, 1" should read "Fig. 1" (period, not comma)
3. L. 225, Figure 2, x-axis label: "Winter 2025/16" should read "Winter 2015/16"
4. L. 252: "PWL's" should read "PWLs". Same comment for L. 400
5. L. 262: "10 March 2015" should read "10 March 2016"
6. L. 315: "Tab. 5" is referenced in two consecutive sentences. Delete the second instance.
7. L. 399: "4.2.1 Stability metrics" should read "4.2.2 Stability metrics" (incorrect section numbering)

Thank you for these suggested corrections, which we will all implement.

References

- Borsotti, J., Mariani, A., Villa, G., Metzger, M., Mayer, S., Schweizer, J., Haegeli, P., and Monti, F.: Assessing skier-induced snowpack instability from simulated snow stratigraphy by combining failure initiation and crack propagation, Nat. Hazards Earth Syst. Sci., in preparation, 2026.
- Gaume, J., and Reuter, B.: Assessing snow instability in skier-triggered snow slab avalanches by combining failure initiation and crack propagation, Cold Reg. Sci. Technol., 144, 6-15, <https://doi.org/10.1016/j.coldregions.2017.05.011>, 2017.