Reply to Referee #2

In "A random forest model to assess snow instability from simulated snow stratigraphy" an ensemble machine learning approach is used to classify instability in profiles from the SNOWPACK model. I enjoyed reviewing this manuscript and recommend that it be accepted subject to minor revisions based on the quality of the work and its importance to advance the field of artificial intelligence in avalanche research. I have a few thoughts for the authors to consider while preparing their final submission.

Dear Edward Bair

Thank you for your constructive comments on our manuscript. Below we describe (in blue) how we will address your comments when revising the manuscript.

1) Why are rutschblocks still being used as the test of choice? For example, Schweizer and Jamieson (2010) report unweighted average accuracies of ECTs as 0.81 - 0.95. For the rutschblock, the range is 0.67 - 0.88 when score or release type is used. Using results from a more accurate stability test might improve the performance of the random forest model used here.

We agree that the ECT is widely used among practitioners. However, for research purposes the rutschblock test is well suited and well validated data sets exist. Moreover, the rutschblock differentiates between stability classes more clearly than the ECT. As recently shown, for the very poor and poor stability classes the RB test results correlate more strongly with instability than the ECT results (Techel et al., 2020a, 2020b). For instance, Techel et al. (2020a) compared the correlation between RB and ECT and slope stability using a common data set and the same base rate of slopes rated as unstable. In contrast, the review of stability tests by Schweizer and Jamieson (2010) compared many different studies, which either explored the RB or the ECT and rarely ECT and RB in the same study. These studies used different approaches in terms of defining what stable/unstable means and how the data was selected, and hence, the base rates of unstable profiles in the data set differed. As shown by Techel et al. (2020a, Sect. 5.5) for RB and ECT, and by Brenner and Gfeller (1997) from a theoretical perspective, these definitions and the base rate have a strong impact on the resulting performance statistics. Thus, Techel et al. argued that comparisons should primarily be made when exploring stability tests using the same approach and a common data set.

2) At 27 pages with 15 figures and 2 tables, excluding the 2 appendices, the article is too long. The Cryosphere is unusually vague in article size limits, but it is expected to fit with 12 journal pages. In any case, the article's length dilutes its important findings, which show that random forests can be used to classify profiles based on stability with high accuracy. Perhaps some of the details regarding hyperparameters and explanation of the widely-used random forest model could be omitted or moved to an appendix.

We will carefully revise the manuscript, make the language more concise, and consider moving certain sections to an appendix to shorten the manuscript.

3) The finding that viscous deformation is the most important predictor is only briefly discussed. This finding deserves further discussion as it highlights how profiles alone are inadequate to classify instability.

Loading rate is one of the most important avalanche predictors, stated in Atwater and Koziol (1953) and before. The viscous deformation parameter appears to be an indirect measure of this.

Thanks for this comment. We will discuss the importance of loading rates and how these might be linked to the viscous deformation rate in more detail in the revised version.

Minor comments from the annotated PDF

L54 This is not a huge gap in the literature since SK38 and r_c have been separetely validated. Could you provide more motivation for why evaluating both metrics together is vital?

It is correct that both indices have been validated separately. However, to predict snow instability information on both failure initiation (SK38) and crack propagation (rc), and combined threshold values are required (e.g. Reuter et al. 2015). Such an approach has not yet been investigated for simulated snow profiles.

L62 citation?

We will move the citation for the RF classification (Breiman, 2001a) from L220 to L62.

L63 I suggest deleting both instances of "rather"

We prefer to keep the terms "rather", as it is not possible to unambiguously define what stable and unstable snowpack conditions are. Obviously, "unstable" profiles were observed on slopes that did not avalanche.

L82-84 Hasn't the science moved past rutschblocks? Why are they still being used over ECTs ? For example, Schweizer and Jaimieson (2010) report unweighted average accuracies of ECTs as 0.81 - 0.95. For the rutschblock, the range is 0.67 - 0.88, when score or release type is used.

Please see answer to question 1).

L201-203 Or maybe that's because the whole concept of a weak layer that always be pointed to as the culprit in an avalanche is too simplistic. It's great to have an identifiable weak layer for studies like this, but sometimes (for example in storm slab avalanches) there is not an easily definable weak layer.

We agree with you that identifying the weakest layer can be challenging, however, we adhere to the fact that the existence of a weak layer is a prerequisite for the formation of a slab avalanche (Schweizer et al., 2003). While it is clear that for storm snow instabilities it can sometimes be difficult to identify a weak layer in a manual snow profile, this does not mean that there is no weak layer. The fact that we cannot easily identify one in our snow pits just highlights the inherent difficulties in obtaining good data from such manual measurements. However, as in our dataset most weak layers were persistent weak layers, this is not a big issue.

Moreover, in our study, we aimed at identifying the simulated layer most similar to the observed rutschblock failure layer, rather than identifying a weak layer from scratch. In L201-203 we describe that this matching was not always unambiguous, but this was rather due to differences between simulated and observed snow profiles.

L220 put this citation at the first mention of RF on l 62. Since RF is already defined there, "Random Forest" need not be spelled out here.

We will follow your suggestion in the revised manuscript.

L222 delete accounting

We will replace the wording by: ..., this model can account for complex mutual dependencies ...

L412 linked

We will replace “allowed linking” with “linked” in the revised manuscript.

L526 detection of

We will replace “detecting” with “the detection of” in the revised manuscript.

References

Brenner, H. and Gefeller, O.: Variations of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence, Stat. Med., 16, 981–991, https://doi.org/10.1002/(SICI)1097-0258(19970515)16:9<981::AID-SIM510>3.0.CO;2-N, 1997

Schweizer, J., Jamieson, J. B., and Schneebeli, M.: Snow avalanche formation, Rev. Geophys., 41, 1016, https://doi.org/10.1029/2002RG000123, 2003.

Schweizer, J. and Jamieson, B.: Snowpack tests for assessing snow-slope stability, Ann. Glaciol., 51, 187–194, https://doi.org/10.3189/172756410791386652, 2010.

Techel, F., Winkler, K., Walcher, M., van Herwijnen, A., and Schweizer, J.: On snow stability interpretation of extended column test results, Nat. Hazards Earth Syst. Sci., 20, 1941–1953, https://doi.org/10.5194/nhess-20-1941-2020, 2020a.

Techel, F., Müller, K., and Schweizer, J.: On the importance of snowpack stability, the frequency distribution of snowpack stability, and avalanche size in assessing the avalanche danger level, The Cryosphere, 14, 3503–3521, https://doi.org/10.5194/tc-14-3503-2020, 2020b.