Reply to Referee #1

The authors tackle a very important problem for the snow and avalanche community: how to provide synthetic indicators relevant for avalanche forecasting from potential huge amont of simulated snow data. I am very impressed by the obtained results. Moreover, the paper is very well written and is comprehensive with a deep analysis of the model behavior and a detailed presentation of the data pre-processing which essential for machine learning approaches. It might be sometimes difficult to catch the key results among all the results presented, but with a second read it becomes clear enough. However, the interpretation of the model explanatory variables should be qualified as the used feature importance metric is affected by correlation between the input variables and contains only partial information of the underlying « physics ». Overall, I suggest accepting this very good paper with only minor revision I have listed below.

Dear Pascal Hagenmueller

Thank you for your detailed and constructive comments on our manuscript. Please find below our replies (in blue) describing how we will address your comments in the revised manuscript.

Minor comments:

Abstract : add somewhere that the study domain is mainly around Davos and in Swiss.

We will add the study domain in the abstract of the revised manuscript.

L11 : give number of points in the validation data set

We will provide the number of points in the validation data set (N=121) in the revised abstract.

L14-16 : you provide the accuracy for discriminating the non avalanche / avalanche days. However, if the data is not balanced it is difficult to interpret. Use the same clear sentence as in l390-392.

We agree that only providing the accuracy is not sufficient for an imbalanced data set and will follow your suggestion when revising the abstract.

L44 : the model MEPRA (Giraud, 1992) is one of the first model that tried to combine different metrics of snow instability into a synthetic index. Add historical reference in the text.

Thank you for pointing this out. We will include this reference.

Fig. 1 : « virtual slope simulation » => « simulated snow profiles »

We will change the wording as suggested.

L83 : give reference of the rutschblock score from 1 to 7 or explain its meaning.

We will provide a reference where the test procedures are described.

L89 + 105 : « RB tests failed adjacent to layer of persistent grain types ». I do not understand what is meant here. Do you mean: the weak layer revealed by the RB test is in 64% cases composed of FC, DH or SH ?

In the data set of observed Rutschblock tests, the height where the RB failed was indicated as an interface between two observed layers. The failure layer was thus one of the two layers adjacent to this interface. In 64% of the cases in the DAV data set, one of the two layers adjacent to the failure interface was composed of persistent grain types (facets, depth hoar or surface hoar). We will clarify this in the manuscript.

L90 and throughout the text : « to evaluate the model », it is not clear what is the model here. Indeed, the « basic » model predicts whether a weak layer - slab system is unstable or not. I understand that you applied your model more extensively to simulated snow profiles. But be more specific.

We agree with you that we need to be more specific which model we are talking about. In L90 we refer to the application of the RF model on complete snow profiles. We will be more specific when using the term “model” throughout the manuscript.

Fig. 2 : x-label and plot title appear in the same form which is confusing.

We will increase the font size of the plot title to resolve this confusion.

L214 : « similarity criteria » to be defined. Do you mean criteria 1-5 ?

Yes, with similarity criteria we refer to criteria 1-5. We will add “1-5” in L214, to make this more clear.

L274-277: reword. Not clear to me. The use of the probability is not related to the fact that you want to apply the model to any layer of the profile ???

Thank you for pointing out that the wording is confusing. We will reword the sentence to clarify our intention to obtain information on layers from the complete range of the instability spectrum using the output probability that a layer is classified as unstable.

Fig. 8 : I do not understand the role of Fig.8b as the goal is here to see how the model works on intermediate instability classes.

We included Fig. 8b, which shows the average values of $P_{unstable}$ in the unstable and stable classes of the SWISS data set, as the corresponding classes of the DAV data set (in the upper left and lower right corner of Fig. 8a) were used for the training of the model. In our view, the average values of $P_{unstable}$ for the unstable and stable classes of the SWISS data set contain important additional information, as they are independent from the training of the model and are still higher respectively lower as compared to the average values of $P_{unstable}$ from all the intermediate classes in the DAV data set.

L325-329: I did not understand your point here, could you be clearer to explain your point (L330-331).

In L325-331 we analyze mean values of $P_{unstable}$ and proportions of profiles classified as unstable for different subsets of the data not used for training: First for the two marginal RB stability classes “poor” (i.e. RB class = poor and LN $\in \{1,2,3,4\}$) and “good” (i.e. RB class = good and LN $\in \{1,2,3,4\}$) and then for the two marginal LN classes LN ≥ 3 and LN = 1, merging all RB classes (poor, fair, good). As the decrease of $\langle P_{unstable} \rangle$ and the proportion of profiles classified as unstable was more pronounced from the LN ≥ 3 to the LN=1 subset than from the “poor” to the “good” RB class, we concluded that the simulated stability correlated more strongly with the local danger level estimate (LN) than with the observed stability at a point as assessed with a RB test. We will rewrite the paragraph in the revised manuscript to improve the clarity.

L389-390: could you plot on Fig. 13 the avalanche and non-avalanche days as defined in this paper.

We regret but do not understand what you ask us to do here.

L402: « Figure c » => « Figure 15 c » ?

Thanks for pointing out this typo. We will change to Figure 15c in the revised manuscript.

L425: « they were mostly developed to align complete profiles ». In practice, this is not true as a parameter of the model can be used to align only a sub part of the profile. In particular it is used to relax the assumption that the snow-ground interface must be matched. Besides, it is not a limit of the method since for the manually matching you also look below the weak layer for stratigraphy markers (eg. MF-crust). « these additional parameters are not included in the current available automated methods » It is implemented and shown in Viallon-Galinier et al. (2020). Actually your manual method seems to works fine enough and you do not necessarily need an automatic method. You might see the automated matching method as a further development to reduce the time spent to prepare the data but you do not need to say something wrong about the automated method limits.

Thank you for pointing out that these matching algorithms can be used to align only a sub-part of the profile and the parameters grain size and density are implemented in the model used by Viallon-Galinier et al. (2020). We will rewrite the paragraph and include this reference in the revised manuscript.

Section 5.3 : all your analysis is based on the feature importance as computed by the scipy package. First, here, you do not give any information on the « sign » (> or <) of the important feature. For instance, it is not clear (and there is no info about that) whether it is high or low values of « mean density divide by mean grain size » that promote instability. To be added. Besides, the feature importance are somehow « shared » between correlated variables. For instance, viscous deformation might be correlated to the initiation criteria such as SK38 (stress over strength) which is itself correlated to strength, stress (and so importance shared …). Your comment about the absence of initiation criterion must therefore be qualified. Moreover, your comparison of your model score (6 parameters, training) to the « physical » model with only two parameters and no training is unfair (L. 478).

Thank you for your recommendations on how to improve Section 5.3. To include information on the "sign" of the relationship between the features and the target response, we will show partial dependence plots in the appendix of the revised manuscript. A partial dependence plot shows the effect of a given feature on the output prediction, marginalizing over the values of all other features (Friedmann, 2001).

While training the RF model we aimed at avoiding "shared" feature importance between correlated features by excluding pairs of features that were highly correlated (Pearson's r > 0.8). The correlation between viscous deformation rate and the skier stability index SK38 in our training data set was rather low (Pearson's r =- 0.19). Even when removing all features with correlation coefficients with SK38 exceeding 0.5, SK38 still appears at the lower end of the feature importance ranking.

We agree that the comparison of our trained model with the untrained threshold-based model using only the critical crack length and the initiation criterion as input features is somewhat unfair. In the revised manuscript, we will make the limitations of this comparison more explicit. We will also note that when training a decision tree of depth two on the DAV data set, the five-fold cross-validated accuracy is lower when using the critical crack length and the failure initiation criterion as compared to using only the critical crack length. This clearly indicates that for our data set the strength-over-stress initiation criteria have a very limited information content.

Fig. 13 and 14 and Section 4.2.5: the results at the regional scale are very interesting but never discussed in the paper. In particular, the model apparently failed (?) to detect clearly the big avalanche events (high AAI) at the regional scale. Add a discussion on the inherent difficulty to predict high AAI from only slab stability indices (size, spatial distribution, natural release).

We agree that there are some discrepancies between the predictions of our RF classifier and the observed regional avalanche activity and that we did not mention these results in the discussion. Our main goal with these two figures was to show the potential applicability of our RF classifier for avalanche forecasting and the overall promising results. As we only used simulations from one field site for this comparison, there can be a number of reasons why these discrepancies occur, including a lack of information on spatial snow distribution and on potential avalanche size as well as incomplete or biased avalanche data. As these are well-known problems when using avalanche observations for validation, we will briefly discuss the results shown in these figures in section 5.5, but we do not want to discuss these potential error sources in great length.

References:

Friedman, J.H.: Greedy function approximation: A gradient boosting machine, The Annals of Statistics, 29(5), 1189-1232, https://doi.org/10.1214/aos/1013203451, 2001.

Giraud, G.: MEPRA: an expert system for avalanche risk forecasting, Proceedings ISSW 1992. International Snow Science Workshop, Breckenridge, Colorado, U.S.A., 4-8 October 1992, pp. 97-106, 1993.

Viallon-Galinier, L., Hagenmuller, P., Lafaysse, M.: Forcing and evaluating detailed snow cover models with stratigraphy observations. Cold Regions Science and Technology 180, 103163, https://doi.org/10.1016/j.coldregions.2020.103163, 2020.