Dear Pascal Haegeli,

We would like to thank you for your detailed review and constructive comments. We greatly appreciate your help in improving our manuscript. We have modified the paper following your suggestions. The changes are incorporated in the revised version of the manuscript. A point-by-point response to all the comments is provided below (in blue). In addition, we provide a marked-up manuscript version showing the changes made. We hope that the manuscript is now suitable for publication in Natural Hazards and Earth System Sciences.

I look forward to hearing from you.

Sincerely,

Cristina Pérez Guillén

(on behalf of all authors)

Reply to Editor's comments:

1. Overall, the writing is quite complicated with lots of long and convoluted sentences. I think that simplifying the language would make the manuscript easier to read and more accessible to readers. While I highlighted a few grammatical errors in the manuscript, I will leave the copy-editing to the editorial team at Copernicus. However, I recommend having a native English speaker proofread the manuscript before submission.

We agree and tried to simplify some parts of the manuscript. We changed the text following most of your suggestions.

2. I have some suggestions about how to better describe the source and nature of the inherent noise in avalanche danger rating datasets. The attached PDF for details. I believe that a more detailed discussion of this challenge in the introduction would be useful and eliminate the need to explain this several times throughout the manuscript.

Thanks for your suggestions. We added this description in the Introduction (Lines 70-73) and moved a paragraph of the Discussion to Section 3.1.2, where we introduce the compilation of the "tidy" data set (Lines 160-165).

3. I think it would be useful to explain the reliability of the danger rating assessments that was derived by Techel (2020) in simpler terms (L-130). The current description is hard to understand for readers not directly familiar with Frank's work.

We agree and now provide a simplified explanation (Lines 131-132).

4. One of the reviewers raised some questions about the splitting of the data set and the use of the last two winter seasons (2018/19 and 2019/20) for the validation of the model. While you reiterate your explanation from the original version of the manuscript in your response to the reviewer's comment and provided additional evidence about how the selection of the validation winters does not make a difference, you did not actually expand your explanation in the manuscript. I suspect that many NHESS readers will have similar thoughts as the reviewer, and I feel that it would be prudent to proactively address these potential questions. Hence, I suggest that you expand Section 3.4 slightly to better explain your choices and maybe provide a brief characterization of the winters that were used for validation. I am not questioning your choices, but I think that explaining them better will prevent questions from future readers.

We agree and extended the explanation about the 5-fold cross-validation method in Sections 3.4 (Lines 245-253) and 4.2 (Lines 293-297).

5. I feel similarly about your choice to select 30 features even though the performance does not seem to substantially improve after 20 features (L-295). Explicitly explaining your reasons will prevent future readers from having the same questions.

Yes, we opted to develop a model with 30 features as it has the highest scores. A simplified model would have some benefits if some data were missing, and we will consider this when

developing future models. Unlike other machine learning methods, random forest models are practically insensitive when adding less important features.

In addition, we provide below a plot comparing the scores (accuracy and F1-macro) of two random forest models trained with 20 and 30 features. We evaluated them over each fold and the final test set, showing that with 30 features the performance is overall higher than with 20 features.

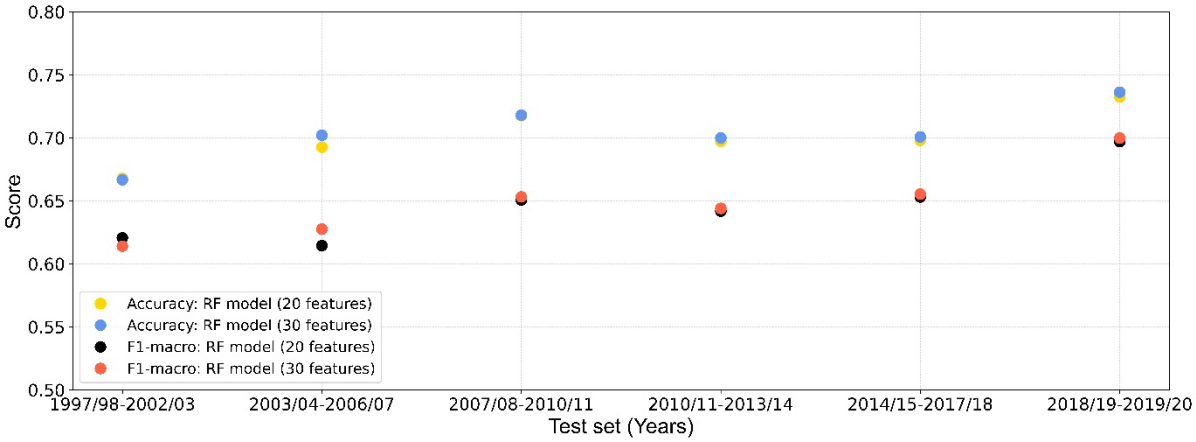

Figure 1. Accuracy and F1-macro scores of the evaluation of two random forest models trained with the optimized hyperparameters and two different sets of features, 20 and 30 features, and used as a validation set one of the five folds (Section 3.4) and the final test set.

6. The current structure of your discussion section is very much focused on the technical aspect of model development, and much of the text seems to mainly summarize information you presented in the result section. While you describe a few important practical insights (e.g., L-534: the model performs as well as the forecasters; L-588: Model performance might be different for different avalanche situations, etc.), they seem to get lost in the technical details. However, in my opinion, these practical insights are the most important for the future operational use of these models and their future development. Hence, I wonder whether organizing your discussion more around practical questions like a) How does the model performance compare to human forecasters? b) What are the situations when the model performs poorly? c) What are the implications of the comparison between RF #1 and RF #2 for future model development in Switzerland and potentially other areas?, … instead of the technical model development aspects (e.g., Training data size and class distribution, Quality of labels, etc.) would be more informative for the NHESS readership and the avalanche community, which both consist of researchers and practitioners. See the annotations in the attached PDF for more details.

We agree and modified the Discussion section following most of your suggestions. We believe that the discussion of technical details of model development is clearly necessary and useful in this first publication. A future study will focus on the practical application and the results observed of operational testing.

We have reduced some parts of the discussion following your advice (Sections 6.1, 6.2 and 6.4). We now include a new sub-section (Section 6.6) where we address the comparison

between RF #1 and RF #2 and the future implications of the use of the models. We do not extend our discussion about the situations when the model performs poorly as it is not the scope of this paper and we have not quantitatively evaluated model performance for different avalanche problems.

Reply to specific comments:

- Line 153: "ground truth data labeling". Could you not just say "for model development and validation"?

We believe that the use of "data labelling" is appropriate when developing a supervised classification machine learning model; it is common and widely used in the literature. Each set of input features should be assigned to a class label in the process of model development.

- Line 178: I am not sure what the purpose of this addition is as it makes it sound that danger levels were corrected specifically for this study, which I do not think is the case.

In fact, some of the danger levels were specifically corrected for this study when compiling the "tidy" data set.

- Line 449. Please explain why you are changing from "accuracy" to "agreement rate". To me, it seems that you are still evaluating the model by comparing it to D_forecast, which is what you did in the previous sections when you used the term "accuracy". You also use the term "accuracy" in the text that describes the content of the table. If you are doing something distinctly differently, please explain in detail and justify the use of the different term.

We agree, in this case 'accuracy' is the right term.

- Line 595: You seem to be presenting new results here, which does not seem appropriate for the discussion section.

In this part of the discussion, we do not intend to present new results, we discuss the results for the regions of Davos and St. Moritz in more detail and suggest possible reasons for the observed decrease in performance in those regions.

- Line 652. While I agree with your statement, the lower danger rating at lower elevations actually does not directly derive from the lower accuracy. This makes me wonder whether

your dataset should have taken the rule "Danger rating outside of core zone one level lower" into account. Is this something to consider for future research?

We tested models labelling the stations outside the core zone with one danger level lower but the performance was lower.