Reply to Referee #1

Overall, the paper is very interesting and tackles a relevant problem for the snow and avalanche community. The main methodology remains relatively simple and was already applied to different avalanche hazard data but the authors provide a deep analysis of their results to understand their algorithm behavior. In particular, they try to overcome the difficulty that their target variable (the forecasted avalanche danger) is an imperfect ground truth of the avalanche danger. The text is well written and easy to follow. The figures are of high quality. The paper is quite long but a reduction would be at the cost of completeness. My comments mainly concern minor clarifications of the methodology or some statements/findings should be qualified. I have only two major comments that should be adressed before publication.

We thank Pascal Hagenmuller for the positive review of our manuscript and the constructive comments. We will revise the paper following the suggestions. Please find below our replies (in blue).

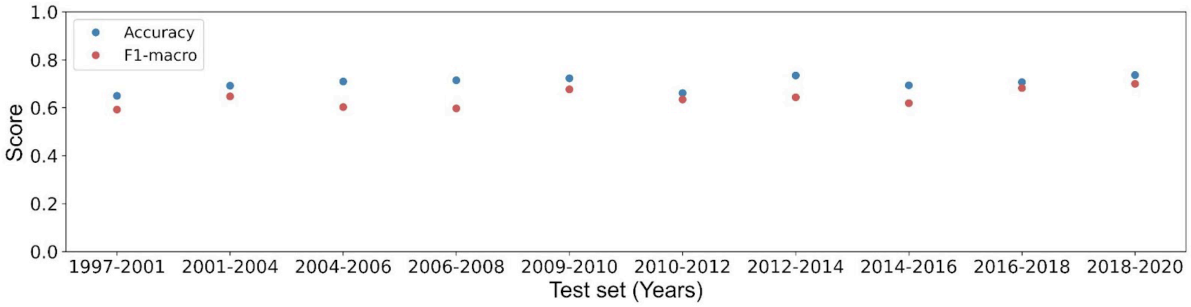
Major comment 1:

In the paper, the algorithm was trained on the winter seasons 1997-1998 to 2017-2018 and evaluated on the latest two winters 2018-2019 and 2019-2020 (line 215-221). The paper findings are thus only based on these two particular years that may exhibit specific avalanche situations. I do not understand why the authors have not repeated their evaluation by extracting any two successive years in their data set and using the rest of the data for training the random forest. Therefore, I am not completely convinced that some of the presented results (some of them based on tiny differences on the evaluation scores) are perfectly robust given the high inter-annual variability of snow conditions.

We would like to thank the reviewer for this comment. The optimization process of the random forest model has been done using a 5-fold cross-validation method (please see Sections 3.4 and 4). To this end, the training data were divided into 5 subsets, in this case containing several winter seasons with an approximate size of 20 % of the training data. For each set of hyperparameters in the random grid search and the grid search, each model was tested 5 times, such that each time, one of the 5 subsets was used as a test set and the other 4 were part of the training set. The F1-score estimate was averaged over these 5 trials for each hyperparameter vector. For instance, Figure 5b shows the box plot of the F1-macro, using the 5-fold cross-validation method, with the variation of the number of features.

The two recent winter seasons were used for a final evaluation of the model's performance. These test sets, for either the D_{tidy} or D_{forecast}, contain enough data samples for each danger level (Figure 3d).

In addition, following this suggestion, we will provide a plot of the accuracy and F1-macro of a random forest model (choosing the optimized hyperparameters and selected final features) evaluated over 10 folds with an approximate data size of 10 % (see Figure 1 below). Each test fold contains two successive winter seasons, and the remaining data are the training set. As the amount of data in the first winter seasons is considerably smaller, these first folds contain more than two winter seasons.



Major comment 2:

The input meteorological and snow data is not forecasted but derived from measurements at AWS. This is somehow expressed in section 2.1 but it appears clearly to me only when it is discussed at the end of the paper (line 536-540): the predicted avalanche danger is a nowcast and not a forecast. I think this should more clearly stated in the abstract and in the methodology as the reader can easily be mixed by « the prediction of the nowcast of the forecast ». Besides, the authors mention in the abstract (line 18-19) that a prototype was used during one winter by the Swiss avalanche warning service. However, there is no more mention of this in the paper (except the same statement in the conclusion). This is not the main scope of the paper but it is legitimate to ask how the nowcast was used/accepted by the warning service.

We will clarify in the Abstract and Section 2.1 that the SNOWPACK simulations, and hence the model predictions, rely on measurements from AWS, and are therefore a nowcast prediction.
Moreover, we intend to include a short paragraph with an outlook on the setup used during operational testing of the model during the winter 2021-2022.

Minor comments:

L.3-4 « based on their experience ». Not only. I guess the forecasters also follow some general guidelines as for instance, picking the right level in the EAWS bavarian matrix.

Yes, forecasters do follow EAWS guidelines and definitions, as for instance the European Avalanche Danger Scale. However, for the final assignment of a danger level a forecaster will strongly rely on his or her experience.

L.13 « the accuracy ». This term should be defined in the abstract or replaced by plain text, e.g. « the danger level was correctly predicted in the 72% of all cases ». Besides, the danger scale data is highly unbalanced, therefore accuracy might not be the best indicator of the algorithm performance (as explained and shown later in the paper). For instance, I can reach an accuracy of 60% by predicting always predicting 3 in Belledonne (France).

We will reconsider using the term in the Abstract.

L.14 « better than previously developed methods ». Remove. I think this is a bit slippery to compare to previous methods as the data, the evaluation strategy, etc. may be different.

We agree and will remove this part of the sentence.

L.16-17 « the accuracy of the current experienced-based Swiss avalanche forecasts ». I would say « agreement » instead of accuracy as we cannot certainly consider the local nowcast as a perfect ground truth too.

We agree. In this context, we will exchange the term 'accuracy' with 'agreement between forecast and nowcast assessments'.

L.23 « predicting stability in time and space ». Generally, the avalanche size is supposed to be also a characteristic of the avalanche danger.

That's correct if we talk about the definition of the danger levels. The sentence simply refers to a description of avalanche forecasting that was coined by McClung (2000).

L.28 « expert judgement » and general guidelines.

The final decision is in fact an expert judgement – in contrast to, for instance, a decision by an algorithm. Of course, the expert will consider all kind of rules, guidelines etc. for the decision.

L.47 « the only solution is to use avalanche detection systems ». No it is not the only solution, it is « another » solution. One may also take into account the uncertainty in the human based observation.

We will reword the sentence so that it becomes clear that we refer to obtaining complete avalanche catalogues.

L.68 « intrinsically noisy ». Could you please develop/explain this statement or give some references.

Thank you, we will explain the meaning of noise. Essentially, where there is judgment, there is noise (see Kahneman et al., 2021).

L.68 « danger level is the most relevant component for communicating the avalanche hazard ». Replace by « an important component ». Indeed, depending on the target public (e.g. mountain guides), the information pyramid of the avalanche bulletin might be different (e.g. avalanche problems on top).

Please note the information pyramid used in public avalanche forecasts is the same regardless of the target audience (as shown in the EAWS recommendations; EAWS, 2021). Of course, the other elements according to the information pyramid are relevant as well, but less so overall.

L.69 « dry-snow conditions ». It might be not clear to every reader how you define dry- snow conditions. Here, I expected that you set a threshold on liquid water content. That is not the case. As far as I have understood there is always an avalanche danger level for dry snow conditions in the avalanche bulletin but sometimes there is also a wet avalanche danger scale when it is higher than the dry one. Is that correct? Please explain it somewhere in the introduction.

We will clarify the meaning. Essentially, dry-snow conditions mean that dry-snow slab avalanches are the most prominent danger. When other avalanche types become prevalent, those are specifically addressed and communicated, and when wet-snow or glide-snow avalanches dominate, the danger level refers to these avalanche types only.

L.98 and elsewhere « 1700 CET » check with the editor how you should write time in this journal. « 17:00 CET »?

We will refer to local time in the revised manuscript.

Figure 2. It appears that there can be more than one station per forecast region. How do you deal with that?

The meteorological and snowpack data from each station is an individual data sample to train and test the random forest model. The daily forecast of the region is used to label each data sample. If more than one station is in the same forecast region above the elevation indicated in the bulletin, they are assigned the same danger level label.

L.120 « the reliability, which is the trust ... as 0.9». I do not understand the number. Provide precise definition.

We will clarify in the revised manuscript what we mean when referring to reliability. The reliability of an individual danger level estimate is the scaling factor required to obtain the agreement rate of pairs of local nowcast estimates between several observers within the same warning region. For a more detailed definition, please refer to Techel (2020, Section 4.1.2, p. 35). The reliability is thus the congruence between the assessment provided by two individuals (Jacob et al., 1987), or, in other words, the factor describing the repeatability for obtaining the same danger level assessment within the same (small) warning region (Techel, 2020, p. 34). -

L.147 and 148 « accuracy ». Replace by « agreement ».

We will reword as suggested.

L.163 « level was corrected ». You mean corrected during the morning update? Clarify.

We will clarify this in the revised manuscript: "…was corrected for the purpose of this study. The danger level was not corrected during the morning forecast update, but for the purpose of this study to obtain a data set of danger level labels which corresponded best with conditions.

L.168 « High: (0.3%) » incorrect parenthesis

Thanks for spotting this.

Section 4.1. Which hyper parameters did you optimize? Number of trees, depth of the trees? And what are their final values?

We computed a random grid search and a grid search with a variable number of trees, features to consider at every split, depth of the tree, the minimum number of samples required to split a node, the minimum number of samples required at each leaf node and the maximum number of samples for each tree.

The specific hyperparameters of RF1 are:

- Number of trees =1000
- Maximum depth = 40
- Maximum features = 'log2',
- Minimum samples leaf = 6
- Minimum samples to split =12

The specific hyperparameters of RF2 are:

- Number of trees =1000
- Maximum depth = 50
- Maximum features = ' auto',
- Minimum samples leaf = 5
- Minimum samples to split =10

We will include the final hyperparameter settings in the Appendix.

L.257. Explain with plain text how the feature importance is computed by scikit-learn.

We will include an explanation of the feature importance computation.

L.273-274. Why did you chose 30 features since you already reached the performance plateau for 20 features?

Because with 30 features the models reached the highest scores.

L.295 « This results highlights the impact of using better-balanced training detain RF#2 and less noisy labels ». I am not convinced by this statement. Indeed, you have already indirectly balanced your data set by weighting the different classes by 1 / frequency.

We have not balanced the training dataset. We tested some balancing techniques using oversampling, undersampling, SMOTE methods, but we did not achieve an improvement of the performance. The weights are used to penalize misclassification for each class in a different way.

L. 308-314. I am wondering if the observed bias is not linked to how you weight the different classes. Do you use the same weight for both D and D_tidy even they do not contain the same frequency of danger level? Please clarify how it is done.

We used the same strategy for each data set, but independently. This means that class weights are obtained for each separate training set. Indeed, using wrong proportions could lead to biases if class separability also changes drastically depending on the learning set. Therefore, the model trained with D_{tidy} has different weights than the model trained with D_{forecast}.

L.317 « The performance of both models improved when tested against the best possible test data ». Misleading statement (for RF2) to be changed. Indeed, you explain correctly that the RF perform at best on the set of data they were partially trained on, no link with data quality for RF2.

We agree and will modify this statement.

Section 5.3. Reading this section raised a question on the methodology. The training is done on all station together (any station.day adds a line in the data set) or is there a RF per station ? Clarify in the methods and maybe discuss these two approaches.

We will clarify when revising the manuscript. In fact, the models were trained with all the stations together. The amount of data per station varies widely as not all the stations from the IMIS network were installed at the same time or were operative in the same period of time. In addition, lower elevation stations were more often filtered due to the elevation filter used.

L.404-406. The impact of a slight distribution difference of the danger level on the overall accuracy might be quantified and I doubt that it is the reason for the geographical differences.

We have quantified the differences of the danger level forecast distribution and a discussion of the possible explanation of the geographical differences in Section 6.4.

Figure 10. Recall on the figure or in the legend the « sense » of Delta. E.g. Delta_ elevation = station elevation - bulletin elevation limit.

We will modify the Figure as suggested.

Table 3. Add the distribution of increasing, equal and decreasing danger level for each level.

We will add this information in the revised manuscript.

L420-430. Add the unit « m » when giving numbers for Delta_elevation.

We change as suggested.

L.455; « intrisically noisier ». Again give justification when you state that earlier in the text.

We will provide an explanation in the Introduction section.

L.475 « RF2 performs better on D_tidy ». Not the point here and not a justification of what is stated just before. RF2 performs better on D_tidy compared to RF1 because it is trained on D_tidy (the test subset).

Yes, we agree and will change accordingly.

L.485 « cost sensitive learning ». I am wondering whether this is somehow not equivalent to duplicating the minority classes and the following statement « reflecting the positive impact of balancing the training ratio » seems over-stated (no proof).

Cost-sensitive learning means to apply a heavier penalty on misclassifying the minority class. For this, we set the class weight as an inverse of the class frequency in the training dataset, focusing on the minority classes. This is a different technique than duplicating minority classes in the training set. Duplicating instances could be a viable technique for specific classifiers and models. For random forests specifically, duplication would have the effect of balancing the probabilities when uniformly sampling the training bags for each tree. This is equivalent to penalizing classes in the cost function, with weights proportional to the frequency of each class. Rather than duplicating, one could also sample data points according to the inverse of observed frequencies, so that no exact duplicates are present. In our analyses, this led to no benefit, and observing more data points and possibly using a large ensemble of trees is always more beneficial.

L.527. « phenomenon » . Avalanche danger is not a phenomenon.

We will reword as suggested.

Section 6.5. Clarify if the described studies apply also only to dry snow conditions.

All three studies mentioned focused as well on dry-snow conditions. We will clarify this.

Conclusion. Mention the fact that for the moment it is only a nowcast tool.

We will emphasize that we present results for the nowcast mode of the model. We have already tested the model in forecast mode, and it performed equally well (as presented by Perez et al., 2021). We will include a short outlook paragraph on the potential for running the model in forecast mode in the Discussion section.

References:

EAWS, 2021; https://www.avalanches.org/downloads/#informationpyramid

SLF, 2021; https://www.slf.ch/en/avalanche-bulletin-and-snow-situation/about-the-avalanche-bulletin/interpretation-guide.html

Hutter, V., Techel, F., and Purves, R. S.: How is avalanche danger described in textual descriptions in avalanche forecasts in Switzerland? Consistency between forecasters and avalanche danger, Nat. Hazards Earth Syst. Sci., 21, 3879–3897, https://doi.org/10.5194/nhess-21-3879-2021, 2021.

Kahneman, D., Sibony, O., and Sunstein, C. R.: Noise - A flaw in human judgment, Hachette Book Group, New York, U.S.A., 454 pp., 2021.

Pérez-Guillén, C., Techel, F., Hendrick, M., Volpi, M., van Herwijnen, A., and Schweizer, J.: Operational test of automatic danger level predictions in Switzerland. Colorado Snow and Avalanche Workshop, 14-15 October, 2021.