Reply to Referee #2

General comments

The paper presents the development of a machine-learning model capable of assessing the avalanche danger level based on input data from automatic weather stations and a snowpack model in the Swiss Alps. The models are trained using a large data set of forecasted danger levels and a filtered subset of "re-assessed" danger levels from local nowcasts.

Compared to previous studies the presented paper uses a much larger and well-refined data set. The trained machine-learning models achieve performances comparable to human forecasters throughout the region of the Swiss Alps. Previous studies did either have either poorer performance or were more limited in their spatial extend.

The topic is of scientific interest and value for avalanche researchers, forecasting services and stakeholders. The topic is within the scope of NHESS. The authors present their study in a clear manner. The manuscript is well written and structured. The abstract provides a good summary of the goals, methods and conclusions of the presented study.

Tables and figures are of high quality and readability contributing to the good overall impression of the paper. The language is precise and understandable. The paper is long. However, it combines the field of avalanche forecasting and machine- learning using the Random Forest algorithm and needs to (and does) explain both concepts to the reader potentially being unfamiliar with one or both of them. I therefore only have minor suggestion on how to shorten it - see specific comments.

It is not clear from this paper how you apply or intend to apply the model in a forecasting setting since it is trained and run on input data measured and modeled at an automatic weather station. I also miss a discussion on the how to apply the models in an operational setting and the expected benefits in supporting the human avalanche forecaster - see specific comments.

We thank Karsten Müller for his positive evaluation of our manuscript and the constructive comments. We will revise the paper following the suggestions. Please find below our replies (in blue).

Specific comments

l-171 Your models are trained on station data. That means they require a measurement and a subsequent SNOWPACK model output to be applied. Thus, RF#1 and RF#2 as described in this paper only provide a hindcast or nowcast.

Yes, we agree. We will clarify in the new version of the manuscript that the models presently provide a nowcast.

In order to be used operational your models need be run with input data from weather prediction models and the corresponding output from SNOWPACK at the location of IMIS stations. As far as I can see this is not addressed in your paper. Please add or reference information on how this is or could be done. I expect that the transition from the spatial resolution of the weather

model to the station site (especially in mountainous terrain) poses some scaling issues which might have an effect on performance/accuracy. This should be addressed in the discussion e.g. in connection to section 6.3.

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.

l-207 Why do you only filter by elevation and not by aspect? I assume you do not filter by aspect because most (all used?) IMIS stations are on a flat field and thus cannot be assigned an aspect. Please add a short explanation.

The data sets for training and testing the models were computed with SNOWPACK simulations for the "flat" field. We have also tested the models with input data from SNOWPACK simulations computed for virtual slopes with different aspects (N, E, S, W). The results obtained for the different aspect are, however, outside of the scope of this paper. We will clarify in the revised manuscript that SNOWPACK simulations were computed for level terrain.

l-216 It seems legitimate to use the most recent winter seasons as test data. However, it should be ensured and stated that these do not exhibit any special avalanche conditions not or barely seen during previous winters - have you considered/tested a random draw from all data with an equal amount from each month as an alternative? If yes, what was the effect on model accuracy.

We discarded random train/test to avoid correlated data from closer stations and days to be in the training and test sets at the same time. The random forest model was optimized by 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. The two recent winter seasons were used for a final evaluation of the model's performance. This test sets, for either the D_{tidy} or D_{forecast}, contain enough data samples for each danger level (Figure 3d).

Please refer to the reply to Referee #1 where we address this issue in more detail, please also see the Figure we provide there.

l-275 "Note that this last step..." - what do you mean by this sentence? It is not clear to me to which "last step" you refer and what the effect on model performance is. Could you clarify?

We refer to the step of feature selection by Recursive Feature Elimination (RFE) as described in that paragraph. In the sentence we provide an explanation why we used RFE rather than relying on internal feature ranking. We clarify this issue in the revised manuscript.

l-355 While the section "Exemplary case studies" is useful for the reader in order to get an overview over potential model outcomes in relation to published avalanche forecasts, it is not

necessary for the understanding of the paper. Considering that the paper is already very long, I suggest to move this section and Fig.8 to the Appendix or provide it as supplementary material.

Thanks, we will consider this suggestion. For the time being, we think it makes sense to keep this section in the main part of the manuscript. These cases are illustrative examples of the model output and show the overall performance in different situations. In addition, it helps to interpret the videos provided as supplementary material.

l-328 What is the "daily averaged accuracy"? Is it the average of the predictions from RF#1 and RF#2 or is it the average of the results from all stations within a forecasting region with regard to Dforecast for that region?

We will remove "averaged" as it was computed the daily accuracy per model.

l-405 The last two sentences in this section should be revised. I understand it such that performance was lower because the danger levels (1 and 3) - that have highest prediction performance - are less common in these regions. However, I had to read it several times to understand what you mean.

We will clarify these statements in the revised manuscript.

It would also be interesting to know if you could identify common traits for stations/sites that had a high accuracy (e.g. >0.8): specific elevations, typical snow or weather conditions?

Thank you for this interesting suggestion. However, the paper is already long and we prefer to provide more insight into site-specific performance in a future publication.

l-540 see comment for l-171

We will include more details about the setup in forecast mode.

l-573 Your features include several stability indices and information on weak layers. Does that mean the provided stability information from SNOWPACK is not good enough to detect/predict persistent weak layers or the stability related to them?

As mentioned, it is presently not fully clear where the regional differences stem from. In fact, our model does include stability information, but a concurrent study has recently shown that for stability prediction actually better predictor variables exist than, for instance, the skier stability index provided by SNOWPACK (Mayer et al., 2021).

l-591 Could you discuss the intended operational application of the models and their main benefits to the human forecaster in more depth. I could imagine that the models would be useful in deciding when to increase or decrease the danger level and to assess the spatial or temporal extend of a given danger level.

As mentioned above, we will include a short outlook paragraph on the setup used during operational testing of the model during the winter 2021-2022

l-602 It would also be interesting to know in the discussion what your expectations on model performance are. I would argue that your results are as best as it can get. You state that a human forecaster has an average accuracy of 76%. You use the assessment by the human forecaster as your labels. Thus, the model inherits human mistakes and biases. For RF#2 these biases are somewhat corrected for or at least replaced by biases or mistakes in human assessed nowcasts.

We agree. As we measured the model's performance using a noisy target variable (forecast danger level, "tidy" danger level), we cannot expect obtaining a model accuracy that is higher than the (unknown) errors in the forecast. Even a perfect model would show (seemingly) mediocre performance if compared with a ground truth which is very noisy. We explicitly address this issue in Section 5.2 and also illustrate it with the exemplary case studies.

l-603 It is not clear from your paper that your model "predicts" avalanche danger. I read it that your model can be used to validate or quality control a published forecast once data has been measured at an IMIS station.

We agree. However, we have already used the model in forecast mode with good success. We will shortly describe this in a new outlook paragraph as mentioned above.

l-610 see comment for l-573

Please, see reply to your comment above.

Technical comments

l-141 "...which jointly account for more than 75% of the cases." Change to "which jointly account for 77% of the cases."

We will change as suggested.

Fig.3 ideally the y-axis of the DL proportion [%] plot for Dforecast would have the same maximum value - currently these are 50% and 40%.

We will modify the Figure as suggested.

l-311 "...the two models...", missing "s" l-318 remove one "particularly"

l-422 spelling "Eq. 1"

Thank you for spotting these typos.

l-463 Split this sentence in two.

We will reword the sentence.

l-474 "...only the 10%..." - remove "the"

We will change as suggested.

l-581 Change to "..., predicting high probabilities for both danger levels."

We will change as suggested.

l-587 remove one "the" and the end of the line

We will change as suggested.

References

Mayer, S., van Herwijnen, A., and Schweizer, J.: A random forest model to assess snow instability from simulated snow stratigraphy, EGU General Assembly 2021, online, EGU21-12259, https://doi.org/10.5194/egusphere-egu21-12259, 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.