Natural Hazards and Earth System Sciences Discussions

Nat. Hazards Earth Syst. Sci. Discuss., referee comment RC1
https://doi.org/10.5194/nhess-2021-341-RC1, 2022


# Comment on nhess-2021-341

Pascal Hagenmuller (Referee)

Referee comment on "Data-driven automated predictions of the avalanche danger level for dry-snow conditions in Switzerland" by Cristina Pérez-Guillén et al., Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2021-341-RC1, 2022

Overall comment :

The paper tackles an interesting problem of providing decision-aid tools to avalanche forecasters based on modern simulation tools and large quantity of data. The authors used random forests to reproduce the regional avalanche danger (human forecast) on a four level scale in dry conditions from meteorological data measured at automatic weather stations and the corresponding simulated snow conditions. They evaluated their algorithm on two winter seasons (2018-2020) and showed that the model is able to predict the danger level chosen by the avalanche forecasters with an accuracy of about 75%. The avalanche danger is not directly measurable and the forecasted avalanche danger cannot be considered as a perfect ground truth. This limits a lot the capacity of this approach. However, to assess the quality of this accuracy, the authors elaborated interesting evaluation strategies based on different data sources: the nowcast of the local danger and on a subset containing verified regional danger data.

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.

- 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 then based on tiny differences on the evaluation scores) are perfectly robust given the high inter-annual variability of snow conditions.
- 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.

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.
- 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).
- 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.
- 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.
- L.23 « predicting stability in time and space ». Generally, the avalanche size is supposed to be also a characteristic of the avalanche danger.
- L.28 « expert judgement » and general guidelines.
- L.47 « the only solution is to used 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.
- L.68 « intrinsically noisy ». Could you please develop/explain this statement or give some references.
- 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).

- 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.
- L.98 and elsewhere « 1700 CET » check with the editor how you should write time in this journal. « 17:00 CET » ?
- Figure 2. It appears that there can be more than one station per forecast region. How do you deal with that?
- L.120 « the reliability, which is the trust … as 0.9». I do not understand the number. Provide precise definition.
- L.147 and 148 « accuracy ». Replace by « agreement ».
- L.163 « level was corrected ». You mean corrected during the morning update ? Clarify.
- L.168 « High: (0.3%) » incorrect parenthesis
- Section 4.1. Which hyper parameters did you optimize ? Number of trees, depth of the trees ? And what are their final values ?
- L.257. Explain with plain text how the feature importance is computed by scikit-learn.
- L.273-274. Why did you chose 30 features since you already reached the performance plateau for 20 features ?
- 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.
- 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_tiny even they do not contain the same frequency of danger level? Please clarify how it is done.
- 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.
- 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.
- 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.
- Figure 10. Recall on the figure or in the legend the « sense » of Delta. E.g. Delta_ elevation = station elevation - bulletin elevation limit.
- Table 3. Add the distribution of increasing, equal and decreasing danger level for each level.
- L420-430. Add the unit « m » when giving numbers for Delta_elevation.
- L.455; « intrisically noisier ». Again give justification when you state that earlier in the text.
- L.475 « REF2 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).
- 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).
- L.527. « phenomenon » . Avalanche danger is not a phenomenon.
- Section 6.5. Clarify if the described studies apply also only to dry snow conditions.
- Conclusion. Mention the fact that for the moment it is only a nowcast tool.

Pascal Hagenmuller