The Cryosphere Discussions

The Cryosphere Discuss., referee comment RC1
https://doi.org/10.5194/tc-2022-108-RC1, 2022
© Author(s) 2022. This work is distributed under
the Creative Commons Attribution 4.0 License.

Comment on tc-2022-108

Frank Techel (Referee)

Referee comment on "Combining modelled snowpack stability with machine learning to predict avalanche activity" by Léo Viallon-Galinier et al., The Cryosphere Discuss., https://doi.org/10.5194/tc-2022-108-RC1, 2022

Review "Combining snow physics and machine learning to predict avalanche activity: does it help?" by Viallon-Galinier et al.

The authors present a random-forest algorithm, which predicts the occurrence of natural avalanches running to the valley bottom in the Haute-Maurienne part of the French Alps. The algorithm is trained using a long-term record of avalanche observations, a highly unbalanced data set with 100 times more non-avalanche days compared to avalanche days. From my perspective, the novel - and certainly very challenging aspect of this study, is the prediction of (often single) avalanche events for aspect-elevation segments. The algorithm's predictive performance is characterized by recognizing many of the observed avalanche days, but having a very high false-alarm rate (only 3% of the predicted avalanche days coincided with observed avalanche days). The manuscript is well written, and most sections are easy to follow. Questions, however, arise with regard to the definition of the target variable (Sections 2.1-2.3, 2.5.1, Discussion), the stability indices for dry snow (Sect. 2.4.1), and the way the variable importance is presented and interpreted (Sect. 3.2 and Fig. 4).

Please find below some comments regarding these three points. I hope these comments will be helpful in improving the manuscript.

General comments

(1) Definition of the target variable and subset used for training and testing

- You defined avalanche days (AvD) and non-avalanche days (nAvD) by aspect-elevation-segment (AE segment). For a specific AE segment, an AvD is fulfilled if at least one avalanche running to the valley bottom (below the blue line in Figure 1) was observed, while nAvD are all other days (l 148-149). If possible, please provide an indication regarding the minimal avalanche size that would be typically required to reach this run-out zone in the study area.
- Overall, I think that the description of AvD and nAvD could be improved. Particularly, what is considered a nAvD is not fully clear. Furthermore, as nAvD were 100 times more frequent compared to AvD, it could be valuable to use a more strict definition of nAvD, excluding for instance days when avalanche activity was uncertain (l 96-101). Not doing so, will inevitably reduce the performance statistics, not because the model performs poorly, but because the target variable is uncertain.
- Some avalanche events had uncertain dating (l 96). Please indicate the number of these events.
- You removed avalanche events with an uncertainty on the release date of more than three days from the data set (l 97-98). Were these days and AE segments then treated as nAvD, or removed from the data set?
- In case the uncertainty of the release date was two or three days, you assigned the last day as the date of release (l 98-99). Did you treat the two previous days as nAvD, or were these removed from the data set? On l 146-148 you explain why the time derivatives are required and that avalanches may release when the stability is lowest. This is somewhat different to how you assigned the avalanche release date when this was uncertain.
- You state that the data set provides a "nearly exhaustive screenshot of natural avalanche activity" (l 93). To me, less than 3000 avalanches in 110 paths in 58 years do not seem exhaustive at all. Consider rephrasing this sentence, for instance to "a representative screenshot of avalanche activity of avalanches running to valley floor" or similar.
- There are 110 avalanche paths and 24 AE segments. - If you consider the topographical distribution of potential start zones, are all AE segments equally often represented? For instance, the distribution in Figure 2 shows that there were 100 times more avalanches in the South aspects compared to the North-East aspects. Is this due to more start zones in South aspects or because activity was indeed higher? Providing more information on the distribution of start zones per AE segment would help the reader to understand this relationship. Consider showing the AE distribution of potential start zones in the study area, maybe in a plot similar to Figure 2. If they were distributed rather unequally, please discuss how you considered this in the analysis, and what impact this may have on the results.
- You attempt to predict both dry-snow and wet-snow avalanches with the same algorithm. I suspect that this probably contributes to the poor performance of the algorithm as a dry-snow avalanche can't be correctly predicted by a tree, which learned conditions favorable for a wet-snow avalanche, and vice versa. This should be discussed.
- Does the EPA provide information on the wetness of the avalanche? Please briefly indicate whether it did or not and if it did, why you preferred to develop one rather than two algorithms. It could also be discussed that splitting the data into wet and dry snow conditions using the simulated stratigraphy and learning two separate algorithms may have helped to address the different release mechanisms in a more appropriate manner, which would potentially also cause fewer false alarms.
- Why did you pick 15 Oct until 15 Mar as the winter season? 15 Oct seems rather early, and 15 Mar rather late. Please explain.
- Why did you use a 1 cm threshold as minimal snow depth? (l186) Or did you use 10 cm, as stated later in the manuscript (l 299)? Both values seem rather low snow depth values considering that avalanches must be rather large to reach the run-out zones. Also along this line: how did you treat cases when there was no snow in a lower elevation band, but some snow in the highest elevation band. I suspect that avalanches

running almost to the valley bottom are probably rather unlikely in these situations (-> nAvD), even if conditions in the start zone would favor avalanche release.

(2) Presentation and interpretation of variable importance (Sect. 3.2 and Fig. 4)

- Fig. 4 shows the variable importance, aggregated (summed) by groups of variables. This is a rather unusual way of presenting variable importance and makes the interpretation of the plot rather difficult. For instance, snow depth and variations (SDV) and dry snow stability indices (DSSI) have the same cumulative Gini importance (about 0.18), but the first contains 7 variables, the latter 30. This means that on average each SDV variable has a higher importance (0.18/7 = 0.025) compared to a single DSSI variable (0.18/30 = 0.006). This only becomes clear from the plot when making these calculations. This is also somewhat indicated in the text (l 259-260).
- To me, it was not intuitive, which of the 7 variables belong to snow depth and variations (SDV). I was able to figure this out after going back to Table 1. Maybe you could somewhere describe this more clearly in Table 1 and/or Figure 4? For the other variable groups, this was clear.
- Did the depth of the weak layers, described in Table 1, not play a role in the RF models? It seems to be missing in Figure 4.

(3) Variable definition (Sect. 2.4.1)

You selected the five weakest layers in each profile (l133-136). Please explain why you used five layers and not just the weakest one. Furthermore, I wonder whether the stability of the five weakest layers isn't highly correlated? What would happen if you train the RF only with the weakest layer? Please elaborate more on how you selected the five weak layers if the local minima for Sn, Sa, Sr, + two crack propagation indices were in five different layers, and how if they all indicated the same weak layer.

Technical comments

- l 60: consider rephrasing this sentence as *machine learning approaches evaluation* is somewhat awkward to read
- l 63: consider replacing of *of interest* with *suitable*, or similar
- l 72: in this study could probably be deleted
- l 77: consider removing *largely*
- l 87: consider adding *was* before extensively
- Figure 1: please show the runout area more clearly, for instance by shading it
- l 97-98: consider rephrasing the second part of this sentence (*from the data set* at the end of the sentence)
- l 144: typo *Considering* --> *considering*
- l 146-148: somewhat awkward to read, consider splitting or rephrasing this sentence

- l 180: consider rephrasing the beginning of this sentence to *We use two classes* or similar
- l 186: You mention that the first selection criteria causes undersampling. What impact did the second selection criteria have?
- l 207: typo *probabilityy* --> *probability*
- l 215: Consider changing *truly* to *correctly*, or similar
- l 243: typo *closed* --> *close*
- l 250: add *day* after *avalanche*
- l 298: what does *leading to strong results* mean. A recall of 3% is not really *strong*. Consider rephrasing.
- Discussion: It would be rather nice to see an exemplary time series of the model predictions for one winter season for all 24 AE segments, together with the corresponding observed avalanche activity. This may help the reader to get a better impression on the correlation between avalanche activity and model predictions.
- l 351-353: this statement is correct, but maybe more importantly, this lowers the observed performance of the classifier as AvD predictions may be counted as a false alarm when in fact there was a (smaller) avalanche