Answer to Frank Techel (RC1)

Léo Viallon-Galinier Pascal Hagenmuller Nicolas Eckert

> 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.

We thank Frank Techel for this detailed review that will improve gloabally the manuscript. We provide below a point by point answer to all his comments.

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.

The observation network was designed at the end of the 19th century when avalanche sizes were not yet normalized. Hence, the european avalanche size scale is not explicitly used. The minimal size is indirectly defined for each avalanche path by the position of the observation threshold that should be crossed by avalanches to be recorded by the observer. Empirically, we can imagine that no size 1 avalanche are recorded. Avalanches of size 2 may be recorded, especially if an accident is related to this avalanche or if the avalanche reached high altitude infrastructure and most of the recorded avalanches may be of size 3 or more. However, this was never explicitely evaluated. We thus prefer not to give an indication that will not be properly supported. Moreover, due to the high number of avalanche paths in Haute-Maurienne and the steepness of the slopes, we believe that EPA provides a excellent overview of local natural avalanche activity, as we point out in the description of the dataset: *Besides, the steep topography of Haute-Maurienne reduces the effect of the threshold of observation as most of the avalanches reach the valley floor, providing a representative screenshot of the overall avalanche activity in the 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.

We will rewrite paragraph 2.5.1 to make clearer the difference between avalanche and non-avalanche situations: *Avalanche activity is based on EPA records in the selected area. Days were classified into two categories: avalanche*

and non-avalanche situations. For a given day, aspect sector and elevation band, it is considered as an avalanche situation if at least one avalanche is reported for this day, aspect and elevation band and a non-avalanche situation if no avalanche observation were reported.. Moreover, we change the name from avalanche/non-avalanche day to avalanche/non-avalanche situation as a situation is defined for a given day and AE sector.

It is not obvious how to chose non avalanche days to remove. When the uncertainty is provided, it concerns the date of avalanche days. It is not possible to remove all days in the uncertainty ranges as these uncertainties can be important, up to several months for most remote sites that are not observed systematically. Moreover, other uncertainties exists but are not reported. When no information is reported, we assume that no avalanches occurred. However, it can simply signify that the visibility was too low to observe anything or the observer was not available a given day. In these cases, no specific information is reported to discriminate between non-avalanche days and uncertain situation. This is one of the major drawbacks of this dataset. We currently work on alternative datasets to overcome these limitations, but no other dataset allow for a constant observation method on such a long period. We will improve the discussion on this limitation in the discussion section.

> - Some avalanche events had uncertain dating (l 96). Please indicate the number of these events.

An uncertainty is associated to each event. We will provide indication on the part of the recorded events that have uncertainty over 1 and 3 days in the revised version: it concerns respectively 28.6 and 23.6% of the dataset on the considered area.

> - 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?

Observations are not considered when uncertainty is more than three days. The 23.6% of observations that have an uncertainty over 3 days, most of them concerns remote site that are not visited regularly and 16.5% of observations thus have an uncertainty of one month or more. It is not possible to remove complete months from the analysis. For each day and AE segment, the status is then determined depending on whether an avalanche is reported or not in the remaining data. This remark will be taken into account in the rewriting of paragraph 2.5.1.

> - 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.

Previous days are assigned depending on other observations. It is considered an avalanche day if an observation related to the considered day and AE segment is reported and non avalanche day otherwise. Information on previous day may help the model to deal with the uncertainty on dates but we fully agree that the inclusion of derivatives is to discriminate the most critical moment is something unrelated.

> - 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.

We will rephrase the sentence to *provide a representative screenshot of the overall avalanche activity in the area.* The word *exhaustive* is indeed incorrect, the dataset do not report exhaustively avalanches, but we believe that due to the specific context of Haute-Maurienne (steep slopes, a lot of avalanche paths recorded), the dataset is an excellent indicator of the overall avalanche activity.

> - 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.

Avalanches are observed only if they are located in pre-defined avalanche paths and reach the observation line. However, information reported contain the elevation of departure of the avalanche and the aspect of the area where it started. This may not be directly linked to avalanche path information as a path may be globally south-oriented but sides may look south-east or south-west, or even East or West for avalanche paths with large departure zones. We use the information related to the precise recorded avalanche as soon as it is available (most of the time), that is why we presented the data for the observed avalanches rather than for the avalanche paths.

Moreover, as the valley turn, we have globally north and south facing paths in Lanslevillard and more South-East and West facing path in Bessans for instance. However, even though a wide variety of aspects is represented in terms of avalanche paths, this does not ensure that all paths are equivalent. Some may have more forest than other, or different vegetation influencing susceptibility to avalanches, some may be steeper than others, etc. We cannot ensure an equal representation of all possible aspects and elevations with similar conditions.

However, as we point out in the description of the dataset, we believe that due to the high number of observed avalanche paths and the steepness of the slopes, the recorded EPA avalanche activity is a good proxy of overall avalanche activity of the Haute-Maurienne valley. Hence, when the goal is to predict the avalanche activity of Haute-Maurienne, the use of a realistic avalanche activity, including unbalance between elevation and aspects seem relevant. This means that the model may not be directly transferable to other areas. We will introduce this in the discussion: *we here train the model with the Haute-Maurienne data. Some climatological or terrain features may lead to a predicted avalanche activity specific to the Haute-Maurienne area, especially with a higher senitivity of certain aspects or elevations (eastern crests during easternly returns). The model may not be transferable directly to other areas without a new calibration.*

> - 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.

The EPA being an observation of avalanche deposits, with remote observations from valleys, it provides few information on processes in starting zone. In particular, although the deposit is often described as mainly dry or mainly humid, the wetness of the snowpack in the starting zone is not reported. It is therefore difficult to classify between dry and wet avalanches based on information reported in the dataset. More generally, classification between dry and wet avalanches is not always obvious, especially during the progressive wetting of the snowpack (from top to bottom) during spring or when dry snow falls over a wet snowpack.

By removing litigious situations and using snow cover modelling, it remains possible to define two subsets of wet and dry snow. However, we do not agree on the need of splitting a priori the two types of avalanche processes. We use the tree-based RF model. If the wetness of the snowpack is a critical factor to identify the situation, it should be selected during the optimization process as one of the top split in the tree directly by the model, especially as we provide relevant indicators to identify if the snowpack is rather dry or wet, such as the height of wet snow or the mean liquid water content. Then, the two branches will analyze different characteristics depending on the situation (dry or wet). The model should then be able to deal with different situations. In the dry snow this is also required as we have to identify situations where a persistent weak layer is involved from situations where only the new snow have to be considered, for instance. We thus do not think that a split between dry and wet situations would help the classification, even though we know that it is a common approach in the avalanche community.

We nevertheless tested to focus on the wet snow situations, as it is closer to the analysis done by forecasters. We extracted the situations for which the snowpack is mainly wet from the whole dataset (both for non avalanche days and avalanche days), based on snow cover modelling. The performance on the resulting model, focused on wet snow was not better than the full model. We then do not pursue in this direction.

We will introduce a paragraph in the discussion section to summarize this.

> - 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.

The period is from 15 Oct to 15 **May**. We will correct this error in our original version in the revised manuscript. The choice of dates inevitably contains arbitrariness. We wanted to include a large variety of situations. In France, the avalanche bulletin is produced from early November to early June, which is coherent with the selected date range. Our choice is also coherent with the choice of other studies, such as [Sielenou et al., 2021], for instance. We will specify these reasons in the text.

> - 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.

Sorry for the inconsistency. We will correct the value line 186. We use a threshold of 10 cm to remove days with no or few snow on ground. The threshold is inevitably arbitrary. We would like to keep all situations that could lead to avalanches and therefore select a conservative value. It therefore allows for a significant undersampling, especially on low elevation bands. The statistical algorithm then have the role of selecting optimal values to separate between avalanche and non avalanche situations. Hence, this threshold is chosen to be conservative and not optimal in any way.

For the way we compute avalanche and non avalanche day, it is important to notice that there is no relation between the three elevation bands and eight aspect sectors we consider. It provides 24 situations composed of meteorological and snow conditions as well as avalanche observations each day. We propose to rename avalanche day and non avalanche day to avalanche and non-avalanche situation and better explain this specific approach in the material and method parts to limit misunderstandings.

(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).

Considering individual importance is misleading because the variables we use have important redundancy. The importance is therefore shared between several variables containing redundant information. We propose this visualisation because it is a common approach and allow to test easily selections of variables. However, if the results indicate main trends, precise values have to be handled with care as we point out lines 235 to 238. We propose to add a reminder that absolute values have to be treated cautiously in the result section.

> - 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.

We will precise it both in Table 1 and legend of Figure 4.

> - 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.

Depth has an importance, we forgot this group in Figure 4. We will add this data in the revised version.

(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.

We have five ways of identifying a weak layer through the five dry stability indices we selected. That is why we selected five weak layers (see line 133). In some situations, the five weakest layers may be highly correlated, if no identical. In some other situations, we know they are different. Beyond this question on the weak layers, a lot of our variables are highly correlated.

Technical comments

Thanks for these detailed comments. We do not answer to all technical comments that will be taken into account as proposed.

> - 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

We tried both representation and find that a shading does not allow for a better interpretation of the figure.

> - 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?

We will add "*which corresponds to an oversampling of the minority class*" at the end of the paragraph.

> - 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.

We propose to rephrase to *leading to trustworthy evaluation results*.

> - 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.

We think that 24 different AE segments would not bring relevant information while overloading the paper. The main question of the paper is the interest of stability indices in combination with machine learning algorithms. We hence propose to include an illustrative example on one year, chosen to be representative of the results of the model.

> - 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.

We agree with this remark and will include it in the revised version of the paper.