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 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 observation threshold 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 rewrote 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. For each day, aspect and elevation band, we classify avalanche*

and non-avalanche situations. A given day on given aspect sector and elevation band is considered as an avalanche situation if at least one avalanche is reported for this day (after filtering of observations and attribution of dates). All other situations are non-avalanche ones.. Moreover, we changed 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 improved the discussion on this limitation in the discussion section:

Observation may not be possible every day (e.g., poor visibility or remote sites), and only avalanches are reported (i.e., no information on the observation that no avalanche occurred). This means that the dataset does not allow to clearly define non-avalanche situations: some situations may be identified as non-avalanche situations, while an avalanche occurred but was not reported. The data also suffers from uncertainty on the dates of avalanches.

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

An uncertainty is associated to each event. We now provide indication on the part of the recorded events that have uncertainty over 1 and 3 days: 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 was 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 rephrased 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 introduced 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). Hence, 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 introduced 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 corrected this error 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 now 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 corrected 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 renamed 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 added 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 now 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 added 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 were 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 added "*This [. . .] step acts as 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 rephrased 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 included 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 included it in the revised version of the paper.

Answer to Karl W. Birkeland (RC2)

Léo Viallon-Galinier Pascal Hagenmuller Nicolas Eckert

General comments

> In this paper the authors present a method using random forests to predict natural avalanches running to the valley bottom in the French Alps. Their methods appear to be solid, and the question they are trying to answer is important. In comparison to previous research, the novelty of their approach is that they make their predictions at the spatial scale of specific elevations and aspects. The paper is generally well-written and clear. I believe this research makes a valuable contribution, but I also feel there are issues that should be addressed prior to publication.

We thank K. W. Birkeland for this detailed and useful review. We answer point by point to the different issues raised below.

> Here are a few of the major issues that I believe should be addressed:

> - It would be helpful for the reader to better understand the spatial characteristics of the starting zones of the approximately 110 avalanche paths in the study area. Looking at Figure 1, it appears that most of the starting zones will have either a NW or a SE aspect. I am not sure about the distribution of the starting zone elevations. A Figure like Figure 2 (which shows the distribution of avalanche events by aspect and elevation) should be created for the avalanche path characteristics. In fact, it would be useful to pair this new Figure with Figure 2 so the reader could assess the effect of the avalanche path characteristics on the number of avalanches in each elevation/aspect zone.
> - Along these same lines and again looking at Figure 1, I assume that the elevations and aspects of the avalanche starting zones are not evenly distributed in the 24 classes (three elevation and eight aspect categories). How does this affect the analyses? I understand that the authors would like to use the 24 elevation/aspect categories used in avalanche forecasts, but I wonder if it is appropriate to use all 24 categories for a dataset that appears to be unbalanced in the distribution of avalanche starting zone characteristics? How is this affecting their 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 introduced this in the discussion: *we here trained 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 sensitivity of*

certain aspects or elevations (eastern crests during easterly returns). Hence, the model may not be transferable directly to other areas without a new calibration.

> - Another issue is the inclusion of both dry and wet snow avalanches in the same analysis. This was also pointed out by the other reviewer. Since we know that the avalanche release mechanisms for these two primary categories of avalanches are quite different, as are the meteorological factors that lead to instability, why are these included in the same analysis? Perhaps this is because both wet snow stability indices and dry snow stability indices are included? Wouldn't it be better to split all the avalanches into "dry" and "wet" categories, and then proceed with the analysis on each of these two subsets of the data?

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 these 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 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 path.

We introduced a paragraph in the discussion section to summarize this.

> - The other reviewer also mentioned another issue I believe needs to be addressed. The dataset does not include all avalanches that occurred, but rather it consists predominantly of avalanches running to the valley floor. I assume these are almost all quite large avalanches. Can you provide a range of the size of the avalanches? Are they all Size 3 (on the Canadian or the U.S. destructive scale) or larger? Or perhaps size 4 or larger? What effect do the authors believe that this bias toward large avalanches has on their results?

The observation network was designed at the end of the 18th century when avalanche sizes were not yet normalized. Hence, the avalanche size 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 evaluated. We thus prefer not to give an indication that would 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 good overview of 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 avalanche activity of avalanches reaching low altitudes.*. Then, we do not expect a bias on the results, as least as long as the model is not applied on other areas. We developed this idea in the discussion.

> - While the authors reference some of the more recent work on predicting avalanches with random forests, I feel like they might want to also reference some early work that attempts to better predict avalanche activity using the statistical techniques available at that time. These older papers had more the more modest goal of trying to predict avalanche days (without elevation/aspect of the starting zones), but

> they were a first step in this direction. This does not have to be a comprehensive review at all, but just a sentence or two with some references would be nice to see. Some older examples exist of researchers using discriminant analysis (examples: Bovis, 1977; Foehn and others, 1977), nearest neighbor techniques (example: Buser, 1983), and binary regression trees (example: Davis and others, 1992). Also, who was the first to use random forests for this type of work? Perhaps one of the authors who you already reference?

Thanks for pointing this lack. We added a paragraph to the introduction to shortly summarize the history of machine learning and avalanches. The pioneering works were performed by Bois and Foehn in the 70s, with linear methods, while the first use of classification trees were by Davis et al in the late 1990s and random forest models were firstly used in the 2010s (e.g. Mitterer et al, 2013).

> - Finally, one thing that perplexes me about this research is why new snowfall is rated so low in importance (Figure 4). This is completely different than prior research, which typically rated snowfall as the most important factor for dry avalanche release. Why do the authors believe this is the case? Is it because the "snow depth and variations" class is capturing this essential information? Or is it because of this information is captured (fully or partly) in some of the stability indices? Or is it the mixing of the dry and wet snow avalanches into one dataset? It might also be related to the fact that the dataset consists of only large avalanches. What do the authors think?

The variables we use contain a lot of redundancy and correlations. The random forest select the variable that allow the best separation into two groups at each step. We observe here that post-processed variables such as snow depth variations or new snow depth on 24, 72, 120h seem to be slightly more relevant than bulk snowfall the given day (or on 3 days), This is not contradictory with previous studies as it does not mean that snowfall do not contain relevant information. It just means that other variables are more relevant for the information related to new snow.

We have several possible ways of explanation. The first is that contrarily to most of the previous studies, we use large-scale modelled meteorological information rather than locally observed meteorology. We know that the Haute-Maurienne massif experience some heterogeneous meteorological conditions, especially during easterly return events, for which modelled meteorological information may not be fully representative. The second one is that most of the meteorological information we consider (precipitations and wind) are identical for all aspects and elevations and temperature are identical for all aspects and highly correlated between elevations whereas we know that the snowpack are generally quite different. The snowpack variables are able to summarize the history of past conditions that have built up the snowpack while meteorological information is not at the correct time scale for this. We added a paragraph in the discussion to discuss this result on meteorological variables: *Meteorological information only was unsufficient. Contrarily to many other studies [e.g. Buser et al., 1989; Mayer et al., 2022], we did not use observed meteorological information but large-scale modelled information [Durand et al., 2009]. Thus, the meteorological information is uncertain and nearly identical for all aspects and elevations while underlying snowpack are generally significantly different. Therefore, we did not expect a good prediction at high spatio-temporal resolution with only meteorological information.*

> Despite the above comments, I believe this is valuable research and is deserving of publication once the authors address or respond to these issues.

> I have also attached an annotated PDF, which includes corrections to some typographical errors, as well as further suggestions and suggested wording changes.

We gathered our answers to the attached comments below.

> I hope the authors find my comments and suggestions useful.

> Karl Birkeland

> Some possible older references (the authors may have other/different older references they wish to cite):
>
> Bovis, M.J. 1977. Statistical forecasting of snow avalanches, San Juan Mountains, Southern Colorado, U.S.A. Journal of Glaciology 18(78), 87-99.
>
> Buser, O. 1983. Avalanche forecast with the method of nearest neighbors: An interactive approach. Cold Regions Science and Technology 8, 155-163.

Davis, R.E., K. Elder, and E. Bouzaglou. 1992. Applications of classification tree methodology to avalanche data management and forecasting. Proceedings of the 1992 International Snow Science Workshop, Breckenridge, Colorado, 123-133 (available at: https://arc.lib.montana.edu/snow-science/item.php?id=1245).

Foehn, P.M.B. and others. 1977. Evaluation and comparison of statistical and conventional methods of forecasting avalanche hazard. Journal of Glaciology 18(78), 375-387.

Attached comments

We only detail hereafter the main comments of the PDF. We took into account all the detailed suggestion in the attached PDF in the revised version.

> Page 1: I would suggest re-wording the title to make it more direct, while keeping the same meaning. I'm also not sure that "snow physics" is appropriate... perhaps it would be more accurate to state that you are really using modeled stability indices? Another thing that is important to emphasize throughout the paper is that we are talking about predicting large avalanches in this study (that go to the valley floor). Given all this, one suggestion would be: Does combining modeled stability indices with machine learning help with predicting large avalanches?

We propose changed the title to *Does combining modelled snowpack stability with machine learning help with predicting avalanche activity?*. We believe that the method presented here is not specific to large avalanches. Moreover, in the specific case of Haute-Maurienne, even though EPA observation dataset record avalanches that reach the valley floor, due to the specific geography of this area and the large number of observed paths, it is representative of the overall avalanche activity.

> Page 4: Are all the avalanches in the database naturally triggered? Or are there also some artificially triggered avalanches?

Avalanches reports are based on the observation of avalanche deposits. Observers have few information on the origin of the avalanches. The observation network was designed to give an overview of natural avalanche activity. Hence, natural avalanches are natural ones. However, we cannot ensure that no triggered avalanches are present in the dataset.

> Page 11: This is interesting. I am curious why snowfall does not look important in Figure 4, and rainfall also does not look important. It seems that snowfall should be among the most important variables, and the other snowfall variables (snow depth and variations in depth) are important. I will be interested to see how this is explained in the Discussion.

We answered in the general comments and now include further discussion on this point.

> Page 12: This is a surprising finding since new snow is often one of the most important variables for discriminating between avalanche and non-avalanche days.
>
> I find this to be really surprising since others in the past have had some reasonable results (though definitely not perfect) with only looking at meteo variables such as new snow, wind and temperatures.
>
> I am hoping you will discuss this in your discussion section.

We answered in the general comments and now include further discussion on this point.

> Page 14: I understand the other measures, but I think it would be helpful for the reader if you more explicitly explained "threshold" and what a value of that threshold means in the context of this Table. What is a "good" threshold, or is there such a thing? Larger numbers or smaller numbers or ??
>
> I read through some of the explanation in Section 2.6.3, but it was still not clear to me so I went and did my own work to try to better understand it.
>
> Now that I understand it better, perhaps your wording is OK. But, you could have another look and see. I would like it if there was something even in this figure legend that told us how to interpret these threshold values. Clearly they don't line up the same as the AUC values for the different sets of variables.

There is no good or bad threshold here, we reported this value as an information, as it traduces some details of the behavior of the model. We adapted the legend to make it clear that, contrarily to other values, this is not a score, but an additional information.

Answer to Simon Horton (RC3)

Léo Viallon-Galinier Pascal Hagenmuller Nicolas Eckert

General comments

> This study presents a statistical model to predict avalanche and non-avalanche days using a combination of weather data, modelled snowpack properties, and modelled stability indices. The model is developed with 58 years of avalanche observations from a region in France. The study is designed to examine the added value of stability indices in statistical models for avalanche activity. While statistical models have been widely developed and tested in the scientific literature, investigating how recent advances in snowpack modelling and snow mechanics could improve these models is an interesting and worthwhile objective that is well suited for The Cryosphere. My main concern is how some of the methodological choices likely impacted the results and conclusions. I also think the study missed an opportunity to present their spatially distributed results (i.e., by aspect and elevation) which could be of value to avalanche forecasters. Please see my specific comments for suggested revisions to this paper.

We thank S. Horton for his detailed and constructive review. We answer point by point to all his comments hereafter.

Specific comments

> - Manuscript structure: The paper was well structured with complete and logical flow of information. The graphics were also clean and easy to interpret.
> - Sampling of days to include in the study: I question some of the choices made about filtering the data set and how that impacted the results. A few things stand out as dramatically impacting the set of avalanche days and non-avalanche days that were analyzed:
>   - Why was the period restricted to Oct 15 to Mar 15? Doesn't this remove a large portion of large wet avalanches from the study? What is the purpose of including wet snow stability indices when many of the wet snow avalanche days have been removed? Do you have any information about wet versus dry avalanche activity in the EPA data set? Similarly, I question how meaningful including days in October and November are for predicting full path avalanches.

The period is from 15 Oct to 15 **May**. We corrected this error in the revised manuscript. The choice of dates inevitably contains arbitrariness. Our goal was 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, 2021], for instance. We now specify these reasons in the text.

> - – Second, the threshold of 1 cm (or 10 cm in other parts of the manuscript?) seems very low considering the avalanche observation data only considered avalanches reaching the bottom of avalanche paths. I think a larger threshold would be much more appropriate. Choosing a threshold depth for avalanches grounded in literature or deriving one form your data set would be more appropriate (e.g., calculate the distribution of snow depths on avalanche days and chose a low percentile as a cut-off). I assume this would be on the order of 100 cm and would remove many of the non-avalanche days from the study.

Sorry for the inconsistency. We corrected 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. We summarized this in the text.

> - – I suspect plotting the avalanche activity by day of year and snow depth would reveal informative patterns about when discriminating avalanche and non-avalanche days is actually important to avalanche forecasters. A model informing the likelihood of large natural avalanches in mid-winter and late-winter is likely much more helpful than a model informing whether the snowpack depth has reached the threshold for avalanches.

We work on 58 years and 24 aspect-elevation bands. It would be impossible to visually interpret something from this large amount of data. Moreover the central question of this paper is on the interest of physically-based stability indicators to summarize information from snow cover models before applying machine learning methods. We thus propose, according to the proposition of an other reviewer, to provide an example of one representative output for one aspect and elevation. This will not provide a systematic evaluation of the interest for forecasters, that is done by the overall scores but allow for a qualitative interpretation. For instance, this will highlight that the model do not only inform on the fact that a sufficient snowpack depth is reached.

> - – By removing more of the uninteresting non-avalanche days, the dataset would be more balanced. This would likely diminish the obvious impacts of snow depth on the resulting models and put more weight on the stability indices, which would better suit the objective of the study.

We do not currently have elements to justify that the weight of snow depth is linked to unbalancing. All previous studies show a high importance of snow depth, new snow depth or overall precipitation, despite they use different levels of unbalancing (e.g. Davis et al, 1999; Hendrikx et al., 2014; Scwheizer et al., 2009; Sielenou et al., 2021). The new snow depth is also the first criterion for most practitioners and forecasters for natural avalanche activity.

> - Weak layer selection: The choice of always selecting 5 weak layers seems unusual and was not adequately justified. What is the benefit to this method over choosing a threshold value to identify weak layers? Could there be adverse effects to having many extra layers in the analysis that are potentially stable and uninteresting? For example, wouldn't this diminish the importance of the stability indices compared to a dataset that only included layers that met some type of threshold stability criteria?

The goal of the Random Forest technique is to optimize thresholds to decide whether a weak layer should be considered as prone to avalanche or not, by considering the values of its different stability indices as well as its depth in the snowpack. In this study, the goal is to extract the potentially relevant information from the snow cover model that may not be available to statistical tools otherwise (as nobody will try to put all the output of a snow cover model as an input of statistical model) and then let the statistical method define, from observations what is relevant or not. We therefore do not want to introduce expert thresholds on the different stability indices. Moreover, we need a constant input parameter number, even though we are not able to find a relevant weak layer in the snowpack. We thus selected 5 weak layers as we have five stability indices that may point out a weak layer. It may lead to redundancy of the information in some cases. We already have a lot of redundancy (correlation) in our input variables. We added a sentence to justify this choice: *This approach allow for identifying the five weakest layers, with five complementary ways of estimating the weakness (five stability indices), and have the advantage to provide a constant number of variables for further statistical analysis.*

> - Classification scores and model performance: I wonder how my previous comments impact the resulting classification scores. The precision seems very low, despite the explanation provided. I was also surprised to see the low performance of the meteo subset, as I would expect weather factors to be significantly better at predicting natural avalanche activity than a random model. Especially when considering large natural avalanches, common forecasting experience and past studies have found simple weather indices like 72 hour accumulated precipitation and air temperature to be strong influences. This has me question the representativeness of the dataset/variables and the overall soundness of the results. Can you justify the low performance of the meteo subset in this model?

In this study we explore a large time period and detail the results by aspect and elevation bands. Meteorological variables are highly correlated between elevation and aspect sectors (except incoming radiations that are not in the "Meteo" variable set here). However, in early season or in spring, lower elevations are no longer prone to avalanche while higher elevation may experience higher activity. Weak layer may also form differently depending on aspect. Therefore, we expected meteorological variables to be insufficient to describe the avalanche activity in aspect-elevation sectors. We introduced a sentence in the discussion to precise this results.

- No presentation of results by aspect and elevation: While I understand the decision to aggregate the results from different aspect and elevations to see the overall importance of input variables, I think presenting some of the aspect and elevation patterns would be of great interest as well. First, the question of how well the model can predict the location of avalanche activity would be valuable to forecasters. Second, it's not clear whether the imbalance in the amount of avalanche days by terrain class shown in Fig. 2 impacted the results (e.g., how does the model performance compare on south aspects where there were many avalanche days versus NE aspects where there were few avalanche days).

On the first question of how well the model predicts the location (in a semi-distributed model where location means altitude and elevation) of avalanches, we answer the question as we provide scores for the identification of avalanche activity by classes of elevation and aspect even though only the overall results are presented.

On the imbalance of the avalanche activity we present, it depends on the goal of the study. 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. However, this means that the model may not be directly transferable to other areas. We introduced 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). Hence, the model may not be transferable directly to other areas without a new calibration.*

Moreover, raw results by band of altitude and orientation are difficult to interpret because each aspect and elevation have different number of avalanche observations involved. In a sector where only one avalanche was observed, the score can be either 0 or 1 but this does not carry any information. We nevertheless checked that there is no obvious over- or under-performance for sectors with significantly more or less observations.

- Writing style: I found parts of the manuscript difficult to read, with poor flow between sentences and phrases interrupted by citations. I had to read some paragraphs twice to fully understand the meaning and would appreciate additional editing to improve the readability.

We carefully reread the manuscript to improve the overall flow.

Technical comments

We thank the reviewer for the detailed comments. We answer to comments that were not directly taken into account as proposed or ask for more explanations.

- Title: Is "snow physics" the best way to describe the dataset in this study? It has a broad range of interpretations and when first reading the manuscript I wouldn't have automatically assumed the main data was model-generated stability indices.

With the suggestion of all reviewers, we changed the title to *Does combining modelled snowpack stability with machine learning help with predicting avalanche activity?*

- Lines 11-12: The terms "recall" and "precision" are rather technical for the abstract and would probably have more impact if replaced with plain language descriptions (e.g., predicted X% of days when avalanches were observed), especially considering there are many synonyms for contingency table statistics and some readers may not be familiar with these specific ones.

We edited the abstract with your suggestion.

- Line 20 "Human infrastructure" is an unusual term and could probably be described better.
- Line 19-23: These first few sentences are examples where the position of citations interrupts the readability.
- Line 42: The phrase "delimitation lines around avalanche-prone conditions" is verbose and could be more concise and clear.
- Lines 50-52: Nice context and motivation for this study!
- Line 52: I question whether adding mechanical stability indices would "reduce the complexity of statistical

> tools". These tend to be relatively complex variables dependent upon many other parametrized variables, and in my view are more complex than a simple model based on variables like snow depth and air temperature. I suggest removing "reduced complexity" and directly stating what is meant by complexity (i.e., models with fewer variables and interactions).

We rephrased this sentence and add "*reduced complexity compared to a model that directly uses the snow model output*".

> - Lines 62-63: This important sentence stating the objective of the study should be written to be more clear and specific. I had to read this multiple times and was still unclear on the big picture aim of the study.

We reformulated this sentence as follows: *On this basis, this paper aims to determine whether combining machine learning on avalanche data and mechanical stability analysis of snow profiles helps predict avalanche activity.*

> - Line 72: remove "an" from "study an area"
> - Line 75: What is meant by a "series of events" being reliable? Is this refereeing to reliable observations of the events?

Yes, we updated the text to make it clearer.

> - Line 80: Please justify this date range. As mentioned above, the early part of this range likely contains many uninteresting non-avalanche days and the late part of this range omits large spring avalanches. This date range criteria could be dramatically influencing the results and their interpretation.

The period is from 15 Oct to 15 **May**. We corrected this error in our original version in the revised manuscript. The choice of dates inevitably contains arbitrariness. We would like 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 now specify these reasons in the text.

> - Line 88: Can you comment on the typical size of these avalanches that reach the run-out threshold (e.g., using the EAWS scale https://www.avalanches.org/standards/avalanche-size/). This would help readers better understand the type of avalanches this model predicts. Also, are all these avalanches natural or are any of the paths modified or controlled with explosives (because snowpack would impact the representativeness of the snowpack model)?

The observation network was designed at the end of the 18th century when avalanche sizes were not yet normalized. Hence, the avalanche size 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 evaluated. We thus prefer not to give an indication that would 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 good overview of 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 avalanche activity of avalanches reaching low altitudes.*

> - Line 98: Please describe how avalanche date uncertainty is defined? Do observers estimate a range of dates?

Yes, uncertainty is estimated by observers based on their previous visit. We added a precision on that in the text.

> - Line 116: Was the entire study area treated as a single massif in SAFRAN or was SAFRAN run for each municipality? If a single massif, why is it meaningful to show the three municipalities in Fig. 1?

We work on a SAFRAN massif from the simulation point of view. Observations used are those of the three selected districts. The three districts are plotted to delimit the overall studied area on the map and help the reader to locate the area of interest.

- Line 131: I think a bit more detail about these indices could be included in this section rather than referring to another paper. Providing equations and/or describing some of the key snowpack outputs used to calculate strength and stress would be valuable. Also, the only reference for Viallon-Galinier (2021) in the reference list is https://doi.org/10.1016/j.coldregions.2020.103163, but I think these citations are intended to refer to https://doi.org/10.1016/j.coldregions.2022.103596 which is not listed.

Sorry for the error on the citation, we corrected it in the revised version.

We included a direct reference for each stability indices. We also explain the main principles of each stability indicator.

- Line 133: The choice to select five weak layers from every profile is not adequately justified. Also see my specific comment about how this may impact the results. Also, when defining the local minimum is one layer identified for each separate indices or is there some type of weighted average? If the former case, are there situations where a layer may be duplicated because it is the minimum for multiple indices?

See the main comment answer.

- Sect 2.4.3: I really like the addition of these time derivatives and think it is an interesting part of the study!
- Sect 2.5.1: With such a rich observation dataset I wonder why the simplest binary metric for avalanche activity was chosen. I would expect between the large set of avalanche observations and the types of stability indices included in the models you could try to predict more advanced indicators such as weighted avalanche activity indices, percentage of paths in an aspect-elevation sector that released, etc. The chosen indicator is fine, but perhaps the choice could be justified a bit more.

Most of the time, only few avalanches are observed in each aspect and elevation sector. Hence, the selected dataset and space partition do not allow to easily overcome the binary classification. However, we propose to add more discussion on this point as the combination with other data sources may allow to remove this limitation.

- Line 160: Be careful with using the term "the model" throughout the paper when both the physical snowpack model and statistical model are part of the study.

We carefully checked all occurrences to remove ambiguity when it exists.

- Table 1: I appreciate this concise summary of model inputs. Minor corrections are the depth of dry snow weak layers is listed in consecutive rows, units are provided in different columns, and column 2 is missing a title.
- Lines 180-190: Are there also concerns about the imbalance in the aspect-elevation data? For example, based on Fig. 1 and 2 I assume the number of start zones per sector are variable, so is it reasonable to have an equal number of data points for NE and S aspects in the analysis?

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.

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. However, this means that the model may not be directly transferable to other areas. We introduced this in the discussion: *we here train the model with the Haute-Maurienne data. Some climatological or terrain features may bias the avalanche*

activity towards certain aspects or elevations (eastern crests during easternly returns). Hence, the model may not be transferable directly to other areas without a new calibration.

> - Line 185: A 1 cm threshold seems very small for full path avalanches.

The threshold is 10 cm. We corrected it in the revised version. This threshold inevitably contains some arbitrariness. However, the optimization of the Random forest should be able to select the most relevant one, we just remove days for which we are completely sure there is no interest of studying avalanche activity to prevent overwhelming of the optimization algorithm with uninteresting situations.

> - Sect. 2.6: I like the LOYO validation approach used in this study and it is well described here. One minor comment is why was the 20 to 80th percentiles chosen when 25-75, 10-90 or 5-95 percentile ranges are more common?

We have 58 years. Therefore, the choice of a classical 5-95 percentile would require additional assumptions on the statistical repartition of the errors for the computation. The 20-80 percentile does not seem completely unusual and suited to our dataset.

> - Fig. 3: Please specify the range of uncertainty in the caption (i.e., 20th to 80th percentile).

We now specify the uncertainty range in the caption.

> - Line 260: Here and in Fig. 4 a new way of grouping the variables is introduced which differs from Table 2. I can track how these counts arise, but it could be clearer.

We now clarify this.

> - Line 257: What is meant by new snow variations? This sounds like change in snow depth, which is not a variable listed in Table 1. Also, I would consider separating the snow depth from variations in Fig. 4 to see how much of the predictive power was simply due to snow depth reaching the threshold for avalanches versus how much was due to detecting snow depth changes over shorter time intervals.

The sentence was rewritten as *new snow amounts or snow depth variations*.

> - Fig 4: I suggest sorting the rows by WSSI and DSSI rather than time step to more clearly show the impact of different step sizes.

We adapted as suggested, as it improves the readability of the figure.

> - Line 294: Please the describe the context referred to in Rubin et al. (2012), I am curious how such low precision has been justified in other studies rather than highlighting some type of issue with how the study was designed.

The value of the precision is highly linked to the unbalancing of the dataset. We have more than 233 000 non-avalanche situations while only 2608 avalanche situations. If we artificially reduce the number of non-avalanche situations in the test dataset (e.g. by randomly drawing observations), we will increase the precision value while not changing anything to the capabilities of the model. We believe that we provide values that are close to real use case of such model as avalanche situations are far from majority of situations.

> - Line 300: I disagree that the obvious non-avalanche days have been removed (see Specific comments).

We rephrased to say that we removed some of the most obvious non-avalanche days.

> - Lines 310-318: While I understand how the model is build with aspect-elevation specific inputs, I think presenting some of the terrain specific results would be a highly interesting part of the study.

We plan to publish in-depth evaluation of the interest of this AE approach in a different paper. We focus in this paper on the question of the interest of stability indices (and derivatives) in combination with machine learning. We thus think that specific situations will not improve the readability and clarity of this paper.