Answer to F. Techel comments

Léo Viallon-Galinier Pascal Hagenmuller Nicolas Eckert

> Dear authors
>
> Thank you for responding and addressing the points raised by the reviewers. From my perspective, the manuscript addresses most of the points appropriately.

We thank Frank Techel for his detailed and constructive comments that allow improving the paper. We answer point by point below. The original review is reported in green and associated with the answers in black. Quotations from the paper are in italic font and proposed changes in purple italic font.

Main comments

> I have two points, which I still feel are insufficiently addressed in the Data, Methods, and/or Discussion sections:
>
> 1. The first point refers to the way the data is filtered/labeled, and here, particularly the uncertainty of the release date. The labeling as a function of the release date uncertainty is (l104-106):
>   - 0 days: the release date is labeled avalanche situation
>   - [1, 2, 3] days (24% of the data): the last possible date within this range is labeled as an avalanche situation
>   - 3 days (28%): these cases are removed After removing the cases with uncertainty >3 days, the data set used for analyses, therefore, contains about one-third of the 2500 avalanche situations, where release date uncertainty is [1, 2, 3] days.
>   - Is this +/-[1,2,3] days? - Please include this information.

The number of days refers here to the length of the period on which the avalanche can have occurred. For instance, if an observer reports that an avalanche has occurred between the 21st and 23rd of January in a given path, we consider that the uncertainty of the report is 3 days (<= 3 days) and we arbitrarily consider that the avalanche occurred on the 23rd of January. If the observer reports an avalanche between the 10th and 14th of February, the uncertainty is 4 days (> 3 days) and we do not include the avalanche event in our database. We now detail in the text the definition of the uncertainty as the *length of the period on which the avalanche can have occurred* and provide an example. The paper now reads:

To associate meteorological and snow conditions to each observed avalanche, we remove observations with an uncertainty (length of the period on which the avalanche can have occurred) of more than three days on the release date, from the dataset. When the uncertainty is larger than one day, the last day of the period was defined as the day of the avalanche event. For instance, if an observer reports that an avalanche has occurred between the 21st and 23rd of January in a given path, we consider that the uncertainty of the report is 3 days (<= 3 days) and we arbitrarily consider that the avalanche occurred on the 23rd of January.

> 1. - You describe that the last day of the respective period is labeled an avalanche situation (l104-105). Please provide the respective information on how the other days within this range are labeled after this statement (I guess they are labeled non-avalanche situations as indicated on l166-168). Add a sentence as to why this approach is warranted as the other days within the range of the release date uncertainty also have a reasonably high chance that the avalanche occurred then.

As you noted, the definition of avalanche situations is provided lines 166-168. The other days within the range are labelled depending on the presence of other observations. If no other observations are reported for this day (after filtering and attribution of dates), the day is labelled as a non-avalanche day.

1. • Please discuss that one-third of the release date labels were uncertain and that this will likely impact the classifiers' observed performance. In that regard, could you provide an indication of how many of the 623 missed avalanche situations (Table 4) were missed simply because of this release date uncertainty? Being more clear about these errors would help to interpret whether the model really has a low precision or whether the target variable is just rather noisy (errors in the labels). For instance, Brenner and Gefeller (1997) demonstrate nicely the impact of erroneous labels on observed precision values (referred to as positive predictive value in their paper). This doesn't only apply to the uncertainty in the release date but also on errors in the labels in general.

We agree that the uncertainty related to the avalanche release date may affect the scores presented in this paper. Advanced and dedicated statistical tools that could deal with this uncertainty should be employed but are out of the scope of this paper. We chose to consider events with moderate uncertainty (<= 3 days) as certain with an occurrence date arbitrarily chosen. We may have kept only certain events and non-events (days not covered by any period of avalanche occurrence). However, in this case, there are only very few non-avalanche situations, as some uncertainties can be up to several months.

For the precision value, the *precision is highly influenced by the base rate (proportion of avalanchee situations). Here, avalanche and non-avalanche situations are highly unbalanced* as explained in section 4.1. Considering a balanced evaluation set, we would have a precision of around 76%. We added this in the text to better illustrate the dependency of the precision to class balancing. The uncertainty on the date is a drawback of the dataset that is discussed, but is not the main reason of that explains the scores. We also checked that some variations around the definition of avalanche situations does not change significantly the results.

2. You include Aspect-Elevation-(AE)-segments (Figure 2) that hardly ever have any avalanche observations. For instance, A=NE, E=1800 m, has one avalanche recorded in 58 years x 210 days. The base rate of avalanche situations is <0.0001. In the same aspect, E=3000 m had no avalanche recorded, and E=2400 m had six avalanches. It is, therefore (basically) impossible to make correct positive predictions. - Why are there so few avalanches in this aspect? Is the number of start zones per AE segment much lower than for other AE segments (please provide this information somewhere, maybe in a plot similar to Figure 2)? Or are these avalanche paths particularly difficult to observe, and, hence, observations less reliable? Or are there other factors, which cause these differences? –> Please discuss the implications of including these aspects and how this will impact the observed classifier performance.

In this work, the AE segments are not associated with avalanche paths but with the starting zone of each avalanche event. In particular, avalanche paths could cover large start zones, with different elevations and aspects. Hence, it is not possible to provide a repartition of starting zones. We nevertheless keep this idea for further work, in which we can group avalanche paths to define areas of interest and then ensure a number of observed avalanche paths per modelled segment. Moreover, the Haute-Maurienne topography is specific as the main avalanche activity comes from easterly returns. Hence, the eastern avalanche paths are expected to have a higher avalanche activity, especially for dry snow avalanches. The two sides of the valley also have different aspects, uses, which finally leads to differences in the avalanche activity. This leads to different avalanche activity in the different AE sectors. We do not believe this is a problem by itself. We try here to predict the avalanche activity of Haute-Maurienne. Training on this specific dataset allows us to capture all these local effects that affect the Haute-Maurienne avalanche activity. In a second time, we will work on the extension of this method. If we want to train the method in one area and apply to another, we have to get rid of these local specificities but this is not the goal of this study. We discuss it in section 4.3: *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 (e.g., during easterly returns). Hence, the model may not be transferable directly to other areas without a new calibration.*

Is is very tricky to compare the performances between aspects and orientation. As you pointed out, the base rate is different. However, we checked that the scores were not directly correlated to the number of observed avalanches. For instance, at 2400 m, the true positive rate is 80.7% and the false positive rate is 31.5% for aspect SE (207 observed avalanches) and respectively 79.4% and 30.7% for NW aspect (504 observed avalanches). At 3000 m, the true positive rate is 75.5% and the false positive rate is 34.4% whereas there are only 45 observed avalanche situations. The main goal of the paper is to evaluate the respective role of different input variables on the prediction performance. However, we ensure that that the model is able to perform an analysis based on the data provided and the result is not only linked to unbalance (at least when dealing with true positive rate and false positive rate, the

precision being, by construction, highly related to the balance between avalanche and non-avalanche situations).

> Given these two points (1, 2), I don't agree with the statement on l.444 "more representative scores compared to other studies". While I fully agree that you used a "robust and conservative" (l320) LOYO method to analyze the model performance, potential sources of error in the avalanche situation labels leave quite a few questions open. I am aware that it is in the nature of the observational data used that these inevitably have a fair amount of uncertainty in the labels (i.e. wrong release date, uncertainty with regard to observations of 'none'). While you make the reader aware of this (Sect. 2.2), I suggest taking up this point when discussing the results, particularly when discussing precision values as this is most impacted by the low base rate and such errors.

We removed *more representative* and replaced it with *robust and conservative* as proposed.

Further remarks

> Figure 3: please describe in the caption (or change the x-axis labels) whether the year-month labels indicate the beginning or the end of the month. Please scale the x-axes for the same interval [0,0.2] as in (b). I am not sure about the journal's requirements, but you may have to move the y-axis labels to the left. Consider adding a dashed horizontal line indicating the base rate of avalanche situations in the data set (about 1%). This would allow to easily see that the predicted probability, despite being rather low in absolute terms, is actually much higher than the base rate in some cases.

We adapted the legend to precise the x axis and reminded the base rate, which is also defined in the methods section. We adapted the figure to have a similar y-axis.

> l315: Consider rephrasing this sentence as precision strongly depends on the base rate (of avalanche days; see also Brenner and Gefeller, 1997).

We split the sentence to improve the clarity of the message.

> l336: replace "in many countries" with "in France". Not every warning service uses the three elevation bands 1800 m, 2400 m, and 3000 m.

We reworded the sentence to avoid this confusion. "This spatial resolution enables to capture the spatial distribution of the expected avalanche activity in one region. This latter information is crucial to evaluate and describe the avalanche danger at regional scale (Morin et al., 2019)."

> l375: add "or rain" after "with solar radiation" (rain is a strong driver of snowpack wetting and avalanche release, i.e. Conway and Raymond, 1993)

Added.

> I hope these comments are helpful.

> Kind regards,

> Frank Techel

References

> Brenner and Gefeller, 1997

> Conway and Raymond, 1993

Answer to Editor comments

Léo Viallon-Galinier Pascal Hagenmuller Nicolas Eckert

Main comments

> The paper describes a study combining weather data, modelled snow data and modeled snowpack stability data with observations of avalanches employing the random forest method to forecast avalanche activity (avalanche/non-avalanche days) for 24 for elevation and aspect segments.
>
> The manuscript is easy to follow; the research objectives are relevant and doubtless challenging. The manuscript is a valuable contribution to the field of numerical avalanche forecasting.
>
> As reviewers have pointed out – and I share their concerns – it is questionable whether the avalanche dataset selected is fully suitable to reach the goals of the study. With less than 3000 avalanches in 58 years observed in 110 avalanche paths, about every second winter an avalanche is observed in a specific path. Whether this frequency is suitable to predict daily avalanche activity in 24 aspect and elevation segments remains to be shown.
>
> This said I do neither oppose the use of the dataset and nor question the study overall, I simply invite the authors to reflect whether the question put in the title can be adequately answered and how well the results can be generalized.
>
> On the title, for instance, my recommendation is not to word it as a question, or at least put the question in a way that it can be answered. For instance, to answer the question you would need to consider combinations other than stated in the title. However, you only explore one method. I suspect you primarily wonder about the value of modeled stability information (compared to, for instance, weather and snow data only), in particular in view of your goal to forecast for 24 aspect and elevation segments. By the way, stability information, mostly not modelled, has been used in several previous studies, even back in some early work on numerical forecasting in the 1990ies – and was shown to be important.

We thank Jurg Schweizer for both edition and useful comments that help us improve the paper. We answer point by point below. The original review is reported in green and associated with the answers in black. Quotations from the paper are in italic font and proposed changes in purple italic font.

We fully agree that the selected dataset can contain bias and is a specific representation of the avalanche activity. Before publishing the paper, we ensure that the main results are not largely affected by the main drawbacks we identify in the discussion phase (threshold effect, Haute-Maurienne specific climate, etc.). We also discuss the advantages and limits of the dataset in section 4.3. In particular, we underline that the trained model is specific to Haute-Maurienne due to the dataset used (line 424 and following), even though we believe the method can easily be transferred to other areas.

We adapted the title not to word it as a question.

Specific comments

> Below you will find some more specific comments:

> - Lines 13-14: I think it would be valuable to show how the model would perform with a simple target variable (not considering 24 aspect and elevation segments). Moreover, I think “cutting-edge” should not be related to data here.

We removed *cutting-edge*.

- Lines 20-21: I recommend rewording. Long-term and short-term are somewhat oddly used here. I suppose you refer to hazard assessment in the context of hazard mapping vs. avalanche forecasting. Hazard mapping and avalanche forecasting are both mitigation measures having a long-term and a short-term effect, respectively.

We now use *mapping* and *forecasting*.

- Line 24: I am not sure whether you focus on forecasting or prediction.

As forecasting has a connotation of prediction on future (defined in the first paragraph), and we work on data of the past, we then use the term prediction in the rest of the paper, except when dealing with practical use. Moreover, the term prediction is a common term for the machine learning community.

- Line 34: Please cite the corresponding peer-reviewed publication rather than a magazine article (Schweizer and Jamieson, 2007).

Replaced.

- Line 53: Some early models that used stability information were described by Schweizer et al.

(1994) and Schweizer and Föhn (1996).

- Line 58: Schirmer et al. (2009) used output variables from a numerical snow cover model as input for statistical forecasting.
- Line 65: It has been shown that the type of data (input variables) can at least be as important as the method applied. For instance, regardless of the method, with meteorological data as input only, the performance is always limited (e.g., Schirmer et al., 2009)

Thanks for pointing out these references. We added two sentences in the introduction to acknowledge these contributions:

The first machine learning models [Navarre et al., 1987; Buser, 1989] mainly rely on meteorological observations, simple snow observations and avalanche records. The use of modelled snow information was therefore developed to complement or replace observations [e.g. Schirmer et al., 2009; Sielenou et al., 2021] and expert analyses were introduced to provide appropriate variables [Schweizer and Fohn, 1996].

- Lines 66-67: Suggest rewording to better reflect the aim.

We reworded the sentence by adding the idea that we compare the impact of using stability indices as predictors on the model performance.

- Lines 67-68: Suggest rewording so that the three different types of input data become more obvious.

We reworded the description of the three types of data.

- Line 79: In your replies to the reviewers' comments, you point out many uncertainties related to the dataset. Hence, I wonder whether you can state that the avalanche observations are particularly reliable.

We removed the sentence.

- Lines 97-98: As you describe above, EPA includes primarily large natural avalanches in selected path. It is questionable whether these data describe the overall avalanche activity in the area. There is much more potential avalanche terrain in the area and almost always when there are large or very large avalanches, there will be many small and medium- sized avalanches as well. On the other hand, I suspect, there are also numerous days when no very large natural avalanches occurred, but still many small to medium-sized natural avalanches at higher elevations. Hence, what is observed and recorded in EPA cannot provide a complete picture of natural avalanche activity. All these additional smaller natural avalanches are very likely also relevant for operational avalanche forecasting. Hence, in conclusion, I suggest you reword and specify the statement.

We removed the end of the sentence to keep only the idea of the reduction of the importance of the observation threshold in the specific case of Haute-Maurienne: *Besides, the steep topography of Haute-Maurienne reduces the*

effect of the observation threshold as most avalanches flow far downslope, close to the valley floor.

> - Line 133: I recommend you add "for a given depth" or "a given layer interface"

We added *for a given layer interface*, as suggested.

> - Line 146&147: I recommend replacing humid by wet or moist.

We prefer the generic term *humid* as it does not directly refer to the liquid water content scale from Fierz et al. (2009), especially as we vary the threshold to define a layer as humid. Wet or moist would each one be inappropriate with some thresholds.

> - Line 156: As far as I remember, Reuter et al. (2022) have recently applied a time-dependent index to describe natural failure initiation.

As far as we know, it is a derivation of Conway and Wilbour (1999) index. We added the reference as an illustration of further use of the work of Conway and Wilbour.

> - Line 160: It is confusing to the reader that changes in snow depth are associated with stability indices. Hence, I suggest regrouping the input variables.

The change in snow depth is now grouped in the derivative group, which is obviously more consistent. We adapted the results accordingly. This change highlights the importance of the time-derivatives, which we now discuss.

> - Line 166: I agree that the model resolution you select is more demanding. However, the question remains unanswered whether this additional sophistication represents an added value given the overall rather poor performance of the model from an operational point of view.

In this paper, we decided to work with the presented geometry (AE segments). The goal of the paper is not the evaluate this choice. However, we studied this point (not shown) and we were able to prove that this specific geometry provides an added value compared to a similar algorithm applied at the massif scale.

> - Lines 174-175: I suggest using more descriptive variable category names, for instance, weather, snow and stability. Certainly, "Bulk" does not seem appropriate to me. Also, in Table 1, you refer to derivates, which seem to include snow depth changes (though you only refer to dry snow indices). Please clarify.

We reworded the "Bulk" group as "Simple snow".

> - Lines 219-221: I suggest you also mention that the recall is, commonly in the forecasting context, called the probability of detection (POD) or hit rate. This would ease comparison with previous work.

We added this term.

> - Lines 221-222: I suggest you also mention that the false positive rate is called the false alarm rate (FAR).

We added this term.

> - Line 236: Random classification is usually associated with the diagonal in the ROC diagram.

Edited.

> - Line 254: elevation

Corrected.

> - Lines 257-258: The example you provide is probably one of the most prominent avalanche winters in your dataset. There were three, well-known major avalanche cycles in late January and February in 1999. It would certainly be interesting to know the performance in such a "catastrophic" winter. In addition, are these model results obtained with the model trained with all input variables?

We now precise that all the variables were used for this figure.

> - Lines 262-263: Please adapt the caption in Figure 4a to the description provided here.

Adapted

- Line 270: It may be worth mentioning that sensitivity (75.3%) and specificity (100-23.6%) are similarly good, which is nice, but that due to the unbalanced dataset the precision is low (3.3%).

We introduced the specificity and added the comment you suggested.

- Lines 278-279: The text here, in the revised manuscript, is unchanged, but Figure 5 (previously Figure 4) changed, i.e. the importance scores are different now. Are all changes in Figure 5 reflected in the text?

The different groups were slightly adapted but the results did not change. We carefully checked the whole paragraph.

- Line 280: In Figure 5 it is indicated that there are 25 variables from the dry snow stability group whereas in the text you refer to 30.

This was corrected.

- Line 294: Here you denote snow depth and its derivatives as "bulk", in contrast to the variable categories described in Table 1. Please clarify and/or improve Table 1.

Corrected. It concerns snow depth and new snow depth.

- Lines 313-314: What means "e.g. 2021"? By the way, both studies are published now. Please update the references (Mayer et al., 2022; Pérez-Guillén et al., 2022).

Updated and corrected.

- Line 336. I am not aware of any country who explicitly specifies the avalanche danger in 24 aspect and elevation segments.

We reworded the sentence to avoid this confusion. "This spatial resolution enables to capture the spatial distribution of the expected avalanche activity in one region. This latter information is crucial to evaluate and describe the avalanche danger at regional scale (Morin et al., 2019)."

- Lines 347ff: It seems odd to me not to consider snow depth and in particular new snow depth as meteorological variables. In any case, you cannot conclude that meteorology is irrelevant and that this contrasts with other studies that probably all considered new snow depth. Moreover, it is doubtful that the alleged difference stems from modelled vs. measured new snow depth. Most readers would probably agree that new snow depth (precipitation) is a meteorological driver of avalanche danger. Hence, it seems that your conclusion depends on your choice of categories and cannot be generalized.

This sentence supports the statement *Meteorological information only was insufficient*. Meteorological variables include snowfall, rainfall, temperature and wind speed and direction only (Table 1) and do not contain any snow depth for instance. Moreover, we do not say that meteorology is irrelevant in the general case, we state that it is insufficient to predict avalanche activity at high spatio-temporal resolution. We fully agree that it may be sufficient at larger scales and most studies also include other information, such as bulk snowpack information. We now insist in the first pointed sentence: *Meteorological information only was insufficient to predict avalanche activity with our method*.

- Line 353: Isn't new snow depth the significant variable rather than snow depth?

We added *snow depth and new snow depth*.

- Lines 559-360: I am not sure how this last sentence refers to the importance of new snow depth.

We are not sure to have understood the comment. For sure, evaluating new snow depth (e.g., in kg/m2) does not require a snow model but we show that adding variables (e.g. snow depth variations) derived from the full stratigraphy simulated by a snow model improves the prediction score.

- Lines 371-373: As far as I remember van Herwijnen et al. (2016) and van Herwijnen et al. (2018) showed that dry-snow avalanche activity was correlated with snowfall on times scales of 2 or more days, while for wet-snow avalanches the correlation was shorter and with energy input. Hence this seems to be in contrast with what you describe. However, I agree with your statement in the following sentence.

It is an interesting comment and more work is required to investigate this point. However, a comparison of the results of the two studies are not straightforward as we present results on time-derivatives of stability indices whereas van Herwijnen et al. uses correlations on snowfall and energy-related variables. The snowpack model in-between also accumulates information on mass and energy balance on longer periods and this finally influences the stability indices. Moreover, the results given by Figure 5 have to be handled with care, as variables are highly correlated. For these reasons, we prefer to only generally check that the overall time scales are coherent with the underlying processes and do not go further in the interpretation. In the future, we will work on the reduction of the number of input variables. This will save computational time but also allow more interpretation of the variables importances, especially if we are able to select a set of variables that are not too much correlated. We keep in mind this comment to compare with the results of van Herwijnen et al. in further research steps.

- Line 383: Previously you referred to 2779 events, not 2518. Please clarify.

We corrected the sentence for the 2779 events as it refers to the number of observed avalanches, whereas 2518 corresponds to the number of avalanche situations once combined by aspects and elevation sectors (avalanche situations).

- Lines 384-285: Please refer to my previous comment on "the representative screenshot of the overall avalanche activity". In addition, I am not sure I understand why in Haute Maurienne the scarcity of reported avalanche events is not a problem.

The scarcity of observation is a challenge. We used different techniques to balance the dataset for the machine learning method. We already have nearly two orders of magnitude between the number of avalanche and non-avalanche situations. But in some lower massifs with large forested areas, the ratio could be dramatically lower. In this case, we do not expect the balancing mechanism we use to be sufficient to provide useful results. We add *as our balancing methods may become insufficient* to explain the difference with Haute-Maurienne.

- Line 412: I am not sure I understand what you mean here with "the interest of physics".

We changed to *The impact of using physically-based indices of snow stability as predictors of avalanche activity instead of simpler variables*.

- Lines 419-420: Please reword.

We split the sentence. Hope this is clearer: *These alternative statistical methods could be further compared to our random forest approach. It may provide improvements in the prediction scores or strengthen our results on the effectiveness of combining snow physics and machine learning for predicting avalanche activity.*

- Line 429: In my understanding "avalanche prediction" means to predict the exact time and location of a single avalanche event. I think so far, we rather forecast avalanche activity.

The wording was confusing, we added the precision *avalanche activity prediction*.

- Lines 434-435: I suggest rewording the statement: the combination proves to be valid.

We reworded as *proves to be useful for avalanche activity prediction*.

- Line 435: I guess new snow depth was a significant variable, not snow depth.

Corrected.

- Line 436: As far as I remember, Jamieson and co-workers have shown that stability indices, though derived from measurements (not modeled), were valuable inputs for forecasting (e.g., Zeidler and Jamieson, 2004).

We added the reference.

- Line 439-440: While I agree that snow cover models and thereof derived stability information is valuable, I recall that your model is not that great in identifying avalanche prone situations. The precision is low. Please consider rewording.

We reworded the sentence, which now reads: *Our results also underline the interest of physically-based snow cover models and stability indices for identifying avalanche-prone conditions.*

- Line 443: I suggest deleting “cutting-edge”. The random forests method has become a rather standard tool over the course of the last decade. In addition, I am not sure you can call the data cutting edge.
- Line 446: Similarly, not convinced the stability indices are that cutting-edge.

We removed both *cutting-edge*.

References

Mayer, S., van Herwijnen, A., Techel, F. and Schweizer, J., 2022. A random forest model to assess snow instability from simulated snow stratigraphy. The Cryosphere 16(11): 4593-4615.

Pérez-Guillén, C., Techel, F., Hendrick, M., Volpi, M., van Herwijnen, A., Olevski, T., Obozinski, G., Pérez-Cruz, F. and Schweizer, J., 2022. Data-driven automated predictions of the avalanche danger level for dry-snow conditions in Switzerland. Nat. Hazards Earth Syst. Sci., 22(6): 2031-2056.

Reuter, B., Viallon-Galinier, L., Horton, S., van Herwijnen, A., Mayer, S., Hagenmuller, P. and Morin, S., 2022. Characterizing snow instability with avalanche problem types derived from snow cover simulations. Cold Reg. Sci. Technol., 194: 103462.

Schirmer, M., Lehning, M. and Schweizer, J., 2009. Statistical forecasting of regional avalanche danger using simulated snow cover data. J. Glaciol., 55(193): 761-768. Schweizer, J. and Föhn, P.M.B., 1996. Avalanche forecasting - an expert system approach. J. Glaciol., 42(141): 318-332.

Schweizer, J. and Jamieson, J.B., 2007. A threshold sum approach to stability evaluation of manual snow profiles. Cold Reg. Sci. Technol., 47(1-2): 50-59.

Schweizer, M., Föhn, P.M.B., Schweizer, J. and Ultsch, A., 1994. A hybrid expert system for avalanche forecasting. In: W. Schertler, B. Schmid, A.M. Tjoa and H. Werthner (Editors), Information and Communications Technologies in Tourism, Innsbruck, Austria, 12-14 January 1994. Springer Verlag Wien, New York, pp. 148-153.

van Herwijnen, A., Heck, M., Richter, B., Sovilla, B. and Techel, F., 2018. When do avalanches release: investigating time scales in avalanche formation. In: J.-T. Fischer et al. (Editors), Proceedings ISSW 2018. International Snow Science Workshop, Innsbruck, Austria, 7-12 October 2018, pp. 1030-1034.

van Herwijnen, A., Heck, M. and Schweizer, J., 2016. Forecasting snow avalanches by using avalanche activity data obtained through seismic monitoring. Cold Reg. Sci. Technol., 132: 68-80.

Zeidler, A. and Jamieson, J.B., 2004. A nearest-neighbour model for forecasting skier-triggered dry- slab avalanches on persistent weak layers in the Columbia Mountains, Canada. Ann. Glaciol., 38: 166-172.

Answer to Simon Horton review

Léo Viallon-Galinier Pascal Hagenmuller Nicolas Eckert

General comments

> I appreciate the thoughtful revisions to this manuscript which have clarified many of my initial concerns. Many of the methodological choices are now clearer and better put the results into context. In particular, the explanation of how of a correctly predicting avalanche activity in specific elevation and aspect terrain classes is more difficult that predicting activity at a regional scale, is much clearer. With these rationales clearer I still have some concerns about how the variable groupings impact the results and a few suggestions to strengthen some arguments. Thanks, it was an enjoyable and interesting manuscript to read.

We thank the reviewer for his positive feedback and hope that the following adjustments match the reviewer suggestions and improve the paper. We answer point by point below. The original review is reported in green and associated with the answers in black. Quotations from the paper are in italic font and proposed changes in purple italic font.

Specific comments

> - **Variable groupings**. I find the limited choice of meteorological variables misleading, especially when this group performs similar to a random classifier. First, not all the meteorological inputs appear to be included, especially ones that would specifically improve the prediction at the spatial resolution investigated. For example, solar radiation and wind direction should be primary drivers of snowpack variability across different aspects and precipitation, temperature and wind speed should be driving variability across elevations. The resulting snowpack properties on different terrain classes are ultimately a product of the meteorological inputs, and by not including these in the analysis it is not a representative comparison of a model with and without stability indices. With these groupings some of the predictive skill being attributed to stability indices does not truly reflect the added value of a snowpack model. Similarly, including snow depth change in the stability group may also inflate the importance of stability indices. First, based on many previous studies this is expected to be one of the most important predictors of avalanche activity. Second, it would make more sense to consider this a meteorological variable or, specifically in this study, a derivative of a bulk variable. While I understand the argument that the current structure shows how stability indices aggregate the information in a way that explains the dataset well, the goal of the study suggested by the title is to ask "does it help?", which I think requires a precisely thought out structure that compares the ability to predict avalanche-situations with and without stability information.

We changed the title according to J. Schweizer proposition.

On the input variable, we selected a set of variable that is not uncommon but necessarily arbitrary. In particular, solar radiations are not included as even though it is quite important for the snowpack, it is not commonly measured nor used directly as a meteorological variable. However, wind speed and direction are included.

We fully agree that the snowpack is the result of the evolution under the meteorological conditions, but we here consider that this synthesis is the goal of the snow cover model. We then do not re-do the job of the snow cover model by introducing additional variables. The same way, we assume that stability indices are a good summary of the complex stratigraphy represented by the snow cover model. It is also possible to consider that all information is produced by the snow cover model and use all output variable as predictors. The same way, it is possible to use only the input variables of the snow cover model (meteorological variables) as once again, all the final information comes from this early data. The goal of this study is to compare three common levels of information (meteorological data, bulk snowpack information, stability indices). This separation is somehow arbitrary but not fully original as it

roughly corresponds to the main classes of McClung and Schaerer, 1993.

> - **Variable importance**. The way the results of the Gini importance are presented in Fig. 5 make it difficult to follow the discussion in lines 277-286, because individual variables are discussed in the text but only the group values are shown in the figure. For example, although snow depth is reported to be the most important, its value is not reported in the text and the group of dry snow stability indices collectively have more importance in the figure. I think these results could be presented in a clearer and more consistent way. In some ways this relates to the previous comment about some arbitrary groupings. Perhaps presenting the Gini importance values for every variable in an appendix would help, perhaps sorted by importance within each group.

We do not believe that Gini importance for individual variable have any sense. As explained in section 2.6.4 and reminded in the presentation of the results, the analysis of such variable importance is only mathematically justified when variables are independent. This is absolutely not the case here. Therefore, we keep Figure 5 only because this method is very commonly used with random forests [e.g. Sielenou et al., 2021; Mayer et al., 2022] and allow for a first easy overview of the variable importance but the values have to be handled with care as there is a lot of correlation between variables and even between the groups of variables. For instance, two variables perfectly correlated will have roughly the same Gini importance with a value corresponding to half of the Gini importance of one of the two variables if the other variable was removed from the dataset.

To improve readability, we added the part of the sentence that mentioned the detail on variables inside a group.

> - **Oversample low snow depth days**. Initially from a forecasting perspective the 10 cm threshold to remove non-avalanche situations seems overly conservative. My concern is the importance of snow height would dominate over any influence of stability indices, especially in a dataset dominated by full path avalanches. After reviewing the results in more detail, I am now satisfied with this concern by seeing that the model with stability indices alone, without snow depth, perform at a similar level and thus must capture the important influence of snow depth some way. I'm wondering if this specific point should be emphasized more to highlight that the impact of threshold snow depth can be captured within the set of stability variables.

We believe that we have an even more large conclusion when we state that *The introduction of stability indices and time-derivatives could help identify avalanche-prone situations with machine learning models. This group of variables also gathers a great deal of information as it nearly replaces the information from other variables.* (Section 4.2).

> - Impact of aspect-elevation resolution on performance. Can you explicitly state why the aspect- elevation resolution leads to lower performance due to more precise resolution? I assume it's because it is harder to predict the correct aspect and elevation of an avalanche than predict an avalanche anywhere in a region. Providing a direct explanation of this in the discussion would strengthen the argument that the some of the low performance metrics are justified.

Exactly, the difficulty comes from our evaluation at the aspect-elevation scale, which means that for a given avalanche situation, we have to predict both the avalanche or non-avalanche situation but also in the correct aspect and elevation band. We edited the corresponding paragraph to explicitly explain why aspect-elevation resolution is more demanding with an example:

Our model predicts the probability that at least one avalanche occurs on a given day within a spatial unit corresponding to one elevation band (centred at 1800, 2400 and 3000m) and one aspect (among 8 aspects). This spatial resolution enables to capture the spatial distribution of the expected avalanche activity in one region. This latter information is crucial to evaluate and describe the avalanche danger at regional scale [Morin et al., 2019]. This prediction goal is more demanding than a prediction at larger scales, as generally used in previous studies. Indeed, prediction at aspect-elevation resolution implies to correctly predict the avalanche activity for each aspect and elevation band and not globally at a larger scale. For instance, if one avalanche occurs one day, it implies to identify that we have one avalanche situation but also in which aspect and elevation sector to be considered a success. An avalanche predicted in an other elevation or aspect will be considered as one false negative (in the elevation-aspect it really occurred) and one false positive (in the elevation-aspect it was predicted). It inevitably leads to lower performances for similar models but provides more precise information about the spatial distribution of the avalanche hazard [Statham et al, 2018].

Technical comments

- Line 97-98: Can you clarify what is meant by "observation threshold"? Perhaps this could be clarified by being more specific in the line 94 with "run-out reached a certain threshold distance" and then in line 97-98 sentence make it clear "although the observation data only includes avalanches that reach the threshold distance, the dataset provides an overall representative indication of avalanche activity in the area because the steep topography of the Haute-Maurienne causes most avalanches to reach this distance. . . "

We now explain that it is an *run out threshold (defined for each avalanche path)*.

- Table 1: Why is snow depth change in stability?

We clarified the grouping by associating the change in snow depth in the derivative group and adapted the results accordingly.

- Line 196: Is there a word missing in "goal is not to substitute to the machine learning algorithm"? The sentence is unclear.

We reworded the sentence to make it clearer: *Note that we chose this conservative threshold to remove very obvious non-avalanche situations from the dataset (no snow in the starting zone means no avalanche). We do not expect this threshold to be optimal as this is the goal of the training phase of the machine learning algorithm.*

- Line 278-279: Why are the Gini importance values not reported for these most importance variables, but are reported for subsequent groups? The reporting of these results is inconsistent and difficult to interpret from Fig. 5.

Please refer to the main comment. We use groups that are a little bit more independent than individual variables. We do not believe that Gini importance for individual variable have any sense. As explained in section 2.6.4 and reminded in the presentation of the results, the analysis of such variable importance is only mathematically justified when variables are independent. This is absolutely not the case here. Therefore, we keep Figure 5 only because this method is very commonly used with random forests [e.g. Sielenou et al., 2021; Mayer et al., 2022] and allow for a first easy overview of the variable importance but the values have to be handled with care as there is a lot of correlation between variables and even between the groups of variables.

- Sect. 4.2: It would be interesting to report which specific stability indices had the highest relative importance. Why are these not presented? This could be interesting for future reach into stability indices.

We fully agree with this comment. However, we do not think that a table like Figure 5 could answer the question as some stability indices are highly correlated. Moreover, we think that this goes beyond the main goal of this paper. However, we plan to pursue this work by determining a minimal set of input variables. In this process, we will have to identify the most relevant stability indices. The second method will be necessary. However, it is not realistic to test all the variable combinations with this method. We thus plan to combine our knowledge of the different processes represented by stability indices as well as previous studies to provide a first selection before doing evaluations like in Figure 6.