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

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 will introduce this in the discussion: *we here train the model with the Haute-Maurienne data. Some climatological or terrain features may lead to a predicted avalanche activity specific to the Haute-Maurienne area, especially with a higher senitivity of certain aspects or elevations (eastern crests during easternly returns). The model may not be transferable directly to other areas without a new calibration.*

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 will carefully reread the manuscript to improve the overall flow.

Technical comments

We thank the reviewer for the detailed comments. We answer to comments that will not be 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 will change 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 will edit 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 propose to rephrase this sentence and add "*reduced complexity compared to a model that would directly use the snow cover 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 propose to reformulate this sentence as follows: *On this basis, the aim of this paper is to determine if the combination using machine learning techniques of observed avalanches data and mechanical stability analysis of snow profiles helps for predicting 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 will update 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 will correct 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 will 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 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 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 will add 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 will correct it in the revised version.

We will include 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 will carefully check 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 will introduce 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). 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 will correct 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 will 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 will 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 will be 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 will adapt as suggested, as it will improve 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 will rephrase 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.