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.

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.
- **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.
- **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.

- **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.

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…"
- Table 1: Why is snow depth change in stability?
- Line 196: Is there a word missing in "goal is not to substitute to the machine learning algorithm"? The sentence is unclear.
- 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.
- 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.

Initial review

Editor

Dear Jurg Scweizer,

I found the study interesting however have several concerns about some of the methods and results that I recommended be addressed in a revision.

Best, Simon Horton

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.

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.
  - 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.
  - 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.
  - 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.
- **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?
- **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?
- **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).
- **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.

Technical comments

- 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.

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

- 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.
- 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.
- 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?
- Line 185: A 1 cm threshold seems very small for full path avalanches.
- 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?
- Fig. 3: Please specify the range of uncertainty in the caption (i.e., 20th to 80th percentile).
- 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.
- 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.
- 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.
- 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.
- Line 300: I disagree that the obvious non-avalanche days have been removed (see Specific comments).
- 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.