Reply to Reviewer #2

We thank reviewer #2 for the thorough and detailed review of our manuscript, which we greatly appreciate. In the following, we address each of the points raised. Black text indicates the reviewer's comments. The blue text shows our responses to the comments.

The authors use simulated snowprofiles from SNOWPACK around AWS as both avalanche day predictor and avalanche size predictor for natural dry avalanches. They validate these indicators using a long-term dataset of re-analyzed danger levels and a short-term avalanche observation dataset. They further show how their predictors improve on simple avalanche day indicators by adding information about stratigraphy and stability.

The study is very well written with good and clear figures and tables. Additional information is found in the appendix. I find the study interesting for a large part of the avalanche community both scientific and practical.

My main critique is that used datasets and methods were developed in other studies. It is very time-consuming to read all these associated studies to gain a better insight into the methods. This is especially true for the binary avalanche day classification.

The discussion does a good job at listing a handful of limitations of this study, however, falls short in explaining what impact these limitations have. I am pointing towards some of these limitations in my comments below.

20: please reconsider this sentence. You write that erroneous forecasts may cause costs as false alarms may lead to economic loss. Isn't that the same thing?

Forecasts may be erroneous in two ways, false alarms and misses. Both may cause costs, which we explain with this statement.

23: I think accurate forecasting of natural avalanches in space will never be possible. You write about forecasting the location of avalanches also in the first sentence in the abstract which is a bit misleading.

We doubt that accurate forecasting of natural avalanches in space will never be possible, as physics-based models and computational power will likely further improve in the future. Moreover, accurately forecasting natural avalanches in space does not necessarily refer to the scale of a slope but can also refer to regional-scale activity. We will remove the word "location" in the abstract.

Figure 1: great figure that is very helpful in understanding which datasets serve which purposes.

Thank you, we are glad that this figure is helpful.

97: it seems like you are referring to Table 3 before you refer to Table 2.

We will remove the first reference to Table 3, as we want it to remain close to the Results section. Table 2 will then be referred before referring to Table 3.

130: what is the rational behind selecting the deepest WL as Pcrit? If I understand your method correctly, the DH layer at around 50 cm in Figure 3 should be Pcrit? Thank you for clarifying this!

This is probably a misunderstanding. We defined P_{crit} as the maximum value of $P_{unstable}$ over the respective profile. The $P_{unstable}$ – value of the deepest weak layer indicated by the instability model was termed $P_{deep.}$ Figure 3 should clarify this. The sentence "In case of ties, we selected the layer deepest in the snowpack" means that if the maximum of $P_{unstable}$ is not unique, i.e. the maximum of $P_{unstable}$ is assigned to more than one layer across the profile, then the values for z_{crit} and gt_{crit} (depth and grain type of the weak layer) are taken from the deepest of those layers. Only very few cases had non-unique values of $P_{crit.}$

148: I am not familiar with sn38. Could you explain very briefly where and how it is used?

We will add the definition of the natural stability index sn38, which for each snow layer describes the ratio of shear strength to the shear stress exerted by the overlying slab on a 38° slope.

150: What is the rational behind assuming that new snow height is not aspect dependent? You describe in the sentence below that there is a snow redistribution routine in SNOWPACK.

New snow depths used in avalanche forecasting are usually given by representative flat field measurements in line with meteorological precipitation measurements. In accordance with this, the height of new snow provided by SNOWPACK is the same for the flat field and the slope simulations. The new snow amounts hn1d and hn3d considered in our study are therefore indeed independent of aspect.

In addition, we also considered the thickness of precipitation particle layers as a further parameter in our analysis. For this parameter, which should capture the amount of recently fallen including snow transport by wind, we employed the snow redistribution module of SNOWPACK.

Section 3.1.1: I had to obviously read the Hendrick et al. (accepted) paper to better grasp your definition of avalanche days. I do not think that it is particularly favorable to the readability and comprehensibility of this paper, that one must read up on the methods in another paper. I, however, do think that the algorithm is clever.

We refer to Hendrick et al. (2023) as this is the publication where we first developed this definition of an avalanche day definition. However, we included all essential information in our manuscript making it not necessary to read the paper by Hendrick et al. (2023).

There are a couple of things that I was wondering about, that do not necessarily have to be answered in a revised manuscript:

- How do the different gap check requirements compare to the size of the forecasting regions (I know that you have dynamic regions) as well as to the typical size one of your observers can cover to do avalanche observations?

  This comparison is one of the few points, which we did not repeat from the description given in Hendrick et al. (2023): 250 km^2 is approximately the size of the spatial units used in the forecast in Switzerland, while 5000 km^2 is close to the average size of the seven snow-climatological regions in the Alps.

- We often say that avalanches are rare events and given that you have had a median of two avalanches per aspect and elevation in an area of 250 km2 around an AWS, I wonder if this is true?

In our data set, avalanche days were indeed comparably rare (we had ten times more nAvD than AvD). We do not know the number of avalanche release areas in the aspect and elevation band surrounding a station, but presumably a median of two avalanches represents only a fraction of the potential release areas, and thus of potential avalanches, in this area. Hence, avalanche activity may still be considered a comparably rare event, even on avalanche days.

209: I might have missed it, but what is “thr”? Threshold?

We abbreviate the best-splitting threshold for binary classification using thr. This abbreviation will be clearly introduced.

Section 3.1.2: this section was very hard for me to comprehend and only after reading the results from 4.2 onwards, it became more clear what you were doing.

We will revisit this section and try to make it clearer.

Section 3.3: Is it a problem that there is a mismatch between the number of observed avalanches per aspect and elevation within 250 km2 of an AWS in the training dataset (N=2) and the number of observed avalanches regardless of aspect and elevation within 1000 and 5000 km2 which is a minimum of 1? In my mind you are getting more avalanche days due to a less stringent threshold in your validation dataset than in your training dataset.

It is correct that the avalanche-day definition is slightly less stringent compared to the definition used for training and testing the P(AvD)-models. To make it clearer where the avalanche day definition differs from Section 3.1.1., we will adjust the paragraph, where we describe the definition for the validation data set.

255: Do I understand you correctly that you are using SNOWPACK simulations forced by Weissfluhjoch data for the entire validation dataset?

It is correct that the for the validation part concerning the avalanche activity data from the region of Davos (data set AV3), SNOWPACK simulations were forced with Weissfluhjoch data.

260: What is the rationale behind removing simulated snow depth < 30 cm?

We removed data points that had a simulated snow depth of less than 30 cm, as avalanches are very unlikely to release on such shallow snowpacks. We thereby mainly reduced the number of data points labeled as non-avalanche days. We consider the days with sufficiently high snow depths as more relevant.

263: Do you mean that avalanche days were in general associated with new snow, both 24 and 72 hours?

Yes, this is correct, we will add “72 hours” in the sentence you are referring to.

284: 12 or 13 cm?

Thank you for pointing out this error: It should read 12 cm, as is shown in Figure 4 and in Table A2. We will adjust accordingly.

297: Interesting observation about persistent weak layers needing less new snow for natural triggering. I thought that there was not much difference between the strength of persistent and non-persistent weak layers during and immediately after snow fall. However, non-persistent forms sinter quicker than persistent ones (Alec's lab study from 2013). And I believe that Ben Reuter and others showed in 2018 that non-persistent forms were initially as weak as persistent forms, however, gaining in strength quicker.

In this paragraph, we compared the optimal threshold for the thickness of precipitation particle layers z_{pp} to differentiate between AvD and nAvDs for different subsets. We found that the threshold for z_{pp} was higher for the case when the critical weak layer determined by the instability model consisted of precipitation particles compared to the case when it consisted of persistent grain types. From this, however, we cannot directly deduce that persistent weak layers need less overloading than non-persistent weak layers, because the variable z_{pp} does not describe the depth of the weak layer (i.e. the non-persistent weak layer could be anywhere within the precipitation particle layers contributing to z_{pp}).

309: What is the physical explanation for taking the mean of both models?

There is no clear physical explanation for taking the mean of both models. The two models may be inaccurate in different situations, and by averaging the models, one model's strength can balance out the other model's weakness to some extent.

317: A median failure depth of 30 cm for size 1 avalanches is surprisingly high in my mind. Where were you surprised about that result? I must confess that I am positively surprised that simulated weak layer depth is such a good predictor of avalanche size. I thought that discerning avalanche size is much more complex.

A median failure layer depth of 30 cm may seem high, considering all size 1 avalanches. Presumably, there is a bias towards reporting "larger" size 1 in our data set. Interestingly, Bellaire and Jamieson (2013), in an ISSW proceedings paper pointed out by Reviewer 1, also observed a median failure layer depth of 30 cm for size 1 avalanches. We will refer to this publication when revising the manuscript.

Figure 7: median values are not always readable.

We will increase the font size for these median values.

Figure 8: the numbers indicating respective portions above and below threshold are not always readable.

We will increase the font size for these proportions.

395: With the target variable including either a lot or no avalanche activity, what would you expect your results to be if medium avalanche activity days were accounted for? How does the time stamp of 12:00 LT for your model simulations influence the results? Do you foresee some problems with regards to when observers record avalanche activity during the day? I am also convinced that "medium avalanche activity" might characterize many avalanche days with considerable avalanche danger (ref Fig 8).

Our approach to train the model using a binary target variable (widespread vs. no avalanche activity) was driven by the demand for high-quality labeling of the training data. We would also like to emphasize that the criterion defining the AvD does not require "a lot" of avalanche activity within

the 250 km^2 surrounding of the AWS, but solely an AAI of at least 0.01, which corresponds to a single small avalanche. With the criterion for increasing avalanche activity within the larger surroundings (1000 km^2, 5000 km^2), we, however, ensured that activity was widespread and thereby presumably reduced the number of data points where single observers allocated observed avalanches to a wrong date.

Defining a third label with "medium avalanche activity" is difficult. Which avalanche activity index would this label refer to? And which P(AvD)-value would we assign to medium avalanche activity? With our approach, we decided to go the other way around, conduct the training with only two labels, but evaluate on a data set covering a much wider spread of activity. The evaluation on the danger level set then provided plausible results, regarding the wide range of P(AvD)-values on days with considerable avalanche danger (level 3), which contains days with "medium avalanche activity" as you mentioned.

The timestamp of 12:00 LT for the model simulations influences the results in so far, that for example the critical new snow amount may be under-/overestimated if avalanches occurred several hours after/before this timestamp. To overcome this uncertainty more accurate data with exact avalanche release times would be necessary.

469: …or get rid of the avalanche danger levels altogether (just my personal opinion and somewhat confirmed by Figure 10)

500: pretty interesting!

References

Bellaire, S. and Jamieson, B. "On estimating avalanche danger from simulated snow profiles." In Proceedings of the International Snow Science Workshop, Grenoble–Chamonix Mont-Blanc, pp. 7-11. 2013. https://arc.lib.montana.edu/snow-science/item.php?id=1740