

## Reply to referee #2 (Samuel V. Verplanck)

We thank the reviewer for the positive, constructive review and helpful feedback. We reply below point by point in [blue](#).

The authors present a detailed study of snow stability metrics during a winter season in Davos, Switzerland that had multiple persistent weak layers and cycles of avalanche activity. The data was centered at the Steintälli study site above Davos, Switzerland which included stability tests, image analysis of propagation saw tests, and snow-micropenetrometer observations. Using data from the study site's automated weather station, they ran the snow cover model SNOWPACK to generate simulated stratigraphy and modelled stability metrics. In their analysis they include information in the broader region via the avalanche activity index (AAI) and the daily public avalanche danger forecast.

They provide detailed comparisons of weak layer identification, weak layer properties (shear strength, density, fracture toughness), and stability metrics. Six stability metrics are included in the analysis.

They found general agreement between all six stability metrics which coincided with periods of higher AAI and avalanche danger. They provided some nuanced comparisons such as the ML approach, P\_unstable, and combined skier stability index and critical cut length (pAp sk38 intersection with r\_c) performing slightly better than the traditional skier stability index (RTA-sk38).

In the introduction, they provide a particularly well-written overview of the history, various attempts, and state of the science for different approaches to quantify snow stability. Their methods are generally clear, organized, well-referenced, and justified. Their conclusions follow well from the thorough results and discussion sections. The richness of the dataset is particularly valuable.

I recommend publication, after consideration of the following comments and technical notes:

- The authors state the dataset from Steintälli 2015/2016 provides a valuable benchmark for the evaluation of stability metrics and that the “wealth of data makes it an attractive case study” (lines 476-477). Thus, the title of the article may be too broad for the article's content, as it could, perhaps, be considered more of a case study. Consider adjusting the title accordingly.

[We agree that the work presented refers to a single winter and region, and can be considered a case study. Still, we prefer to keep the title as is.](#)

- The very high AAI on 9 February 2016 that only overlapped with one or two (I can't quite tell) stability metrics jumped out in Figure 4 but was only briefly mentioned in results and discussed some in lines 406-411. Consider drawing more attention to this result. I find it remarkable that the highest AAI day, by far, was missed by most stability metrics.

We agree that the high avalanche activity on 9 February 2016 is remarkable, more so since it was missed by all metrics – and also in the public forecast. For that reason, we discuss it in detail in lines 406-411. We state the reason (mainly snow drift) and why the models cannot account for this.

- Lines 107-108: Why is the propagation saw test not considered a stability test?

For the propagation saw test, the weak layer needs to be identified first, using a stability test such as the CT, ECT, or RB. Moreover, a true stability test should provide information on failure initiation *and* crack propagation.

- Figure 1: Text in upper left and grain form symbols are not legible when printed.

Thank you for pointing this out. We will improve Figure 1 in the revised manuscript.

- Lines 209-210: Are you basing this statement off higher number of taps for CT's and ECT's? If so, there is first a decrease, then, a subtle increase but the trend seems rather weak and based off N=1 for each test day.

The statement is based on the general trend between late January and early March. Taps increase from about 11 to 23/29.

- Figure 2: A bit confusing to see the same data plotted twice for shear strength.

Showing the shear strength in both panels should ease comparison with the other parameters. Since the scales on the y-axis differ, comparison would not be straightforward otherwise.

- Figure 4: This figure synthesizes the major results from the analysis and could have improved clarity. For example, the stability metrics themselves could be labels for the vertical axis in (c). There is not a clear difference in labeling between the skier stability index in (a) and (b). There are borders in the shaded regions in (a) but not (b).

Thank you for these suggestions. We will improve panels (a) and (b) accordingly. We considered various options for additional labelling in panel (c), but in the end, we discarded them all. Figure 4 is a busy figure, indeed, and can only be understood in combination with the caption.

- Table 5: Consider restating the PC, POD, and PSS acronym definitions in the caption.

Thank you for the suggestion. We will do so in the revised manuscript.

- Line 334: I can't find a definition of the acronym PON.

Thanks, correct; we will define it in the revised version.