Response to referee comment by Ben Mirus

We thank Ben Mirus for providing a constructive review and his interest in our study. Here we respond to the comments and address how we will implement the changes in a revised manuscript.

General comments

There are quite a few minor details that are missing in the abstract, most of which could be gleaned eventually from reading the entire paper, but should be included in the abstract for completeness.

We will adapt the abstract according to the answers to the specific comments below.

…I could not find information on how storms and ID are defined for the regional datasets, nor did I find a clear explanation that multiple durations for antecedent rainfall conditions were explored in the RF model.

We will add these specifications according to the answers to the specific comments below.

…some further details on the rationale behind the scenarios selected for different optimization strategies would be helpful.

We will add details on the rationale behind the scenarios according to the answers to the specific comments below.

Specific comments

I found the title quite critical, when actually the paper is not only about limitations, but rather a comprehensive evaluation of multiple approaches to debris-flow forecasting. Perhaps the title could be revised to more fairly represent the important contributions of this work.

We agree that the title is a bit unrewarding. We will think of a new title and for now propose "Evaluating methods for debris-flow prediction based on rainfall data in an Alpine catchment".

L3. I'm not sure I agree that there are no standardized procedures. I think it's more reasonable to state that there are multiple competing methods that have not been objectively and thoroughly compared at multiple scales.

We will adapt the sentence accordingly.

L6. Consider stating "record duration" since you are talking about time, not a distance (length).

That makes sense. We will make adaptations accordingly.

L12. Regional landslide dataset with local rainfall input or with a regional rainfall database? This is a critical detail that needs to be clarified in the abstract even if it can be determined later in the paper.

The regional landslide database is combined with a regional rainfall database, i.e. with a gridded interpolated daily rainfall from stations. We will specify this.

L13. If these implications are important, is it important enough to list them in the abstract? Also, state here whether the RF model was tested for just local or also for regional?

We left the reader in the dark here. We will add that the major implication is that the appropriate method depends on the available data set.
The RF model was only tested for the local case. We will add this information to the text.

L15. I found this "30-min maximum accumulated rainfall" a bit confusing as it isn't really standard terminology. Is this the greatest accumulated depth of rainfall observed within a given 30-minute period of a storm? If so, wouldn't that be basically equivalent to the peak 30-minute rainfall intensity (I-30)?

Yes, it is the peak 30-min rainfall intensity. We will change the term here and in the rest of the manuscript.

L17. Increase in predictive performance over which other threshold optimization approach/approaches?

RF yields a slight increase over the ID-threshold (LR&TSS). We will adapt the sentence.

L41. Again, it's not that there are none, but that a few established procedures are in use and that those approaches have not been compared objectively and thoroughly.

We will adapt the sentence accordingly.

L82. Could also mention that you evaluate these differences for both a local vs. regional landslide inventory.

Thanks, we will add it.

Figure 1. Legend should explain what the blue shaded channel and also the X marks the Illhorn peak. It wouldn't hurt to put the elevation of the Illhorn and the force plate or catchment outlet to provide easy reference for the steepness of the basin.

We will add these specifications.

L143-146. Consider briefly explaining the gridded daily rainfall product, including the spatial resolution and how it is collected/calculated, as well as what rainfall value was used for the threshold evaluations (i.e., did you use rainfall values from the nearest grid cell, or some grid-cell averaging, or ...?). This is important context for evaluating the ID thresholds at the regional scale vs. local scale.

The rainfall dataset for the regional analysis consists of a 1km x 1km gridded product of daily precipitation sums obtained by interpolation of rain gauges (ca. 420), accounting for local climatology and precipitation-topography relationships. We will add these specifications.

L173-202. I guess you didn't explain the regional data here in Table 1. Perhaps that's not necessary, but you do need to define your MIT for the daily/regional data analysis. How are multi-day storms determined?

The MIT for daily rainfall is 1 day and multi-day storms consist of a sequence of wet days. This definition of events overestimates event duration and underestimates mean intensity compared to hourly data, but compensates for this by longer and more dense records (see Leonarduzzi and Molnar, 2020). We will add these specifications.

L192. Initially, I assumed that this 3-90d antecedent conditions meant the cumulative rainfall total measured between 3 and 90 days prior to the storm event. While there needs to be some explanation of why 3 days was selected as a cutoff (why not 2 or 1 day?), there also needs to be a clearer explanation that multiple potential durations of antecedent rainfall were considered. This only becomes apparent in Figure 7 and the associated analysis of the RF results and variable importance.

Your assumption is correct. We cut off at 3 days to reduce the number of variables and it didn't make any difference in the results. However, this certainly requires some clarification. We will adapt the paragraph accordingly.

L208. Yes, and thus does not consider the rate of false alarms.

We will add this clarification.

L204-216. These paragraphs could benefit from an explanation of the shape parameters in terms of how they influence ID threshold shape/position, and then subsequently the rational for why the two contrasting optimization approaches (LR-TSS vs, TSS-TSS) were selected. It might not be clear to all readers the significance of these choices.

We will add the ID equation so the reader can directly see where scale (α) and shape (β) parameters come into play. We'll add an explanation saying that in the log-log space α controls the intercept and β the slope of the threshold line. We will also stress that the two approaches mainly differ in the way β is obtained: TSS&TSS considers both triggering and non-triggering events while LR&TSS only considers triggering rainfall events. This is a fundamental difference.

L223. Consider clarifying the "... original complete (or 17 year) record..."

We'll add this clarification.

L248. By "classical" ID thresholds, you mean those optimized with ROC statistics (LR-TSS, TSS-TSS)?

Yes, we will clarify this.

Figure 3. Difficult to see what the minimum number of debris flows are in each month, but it looks like they're all zero. If so, consider just stating in the figure caption. (b) also, see previous comment about maximum 30-min accumulation. Isn't this just more or less equivalent to the peak I-30 (i.e. 7.2mm/h)?

We will improve the visibility of the number of debris flows and/or clarify it in the caption. Yes, it is peak I-30 and we will change the term (see also answer above).

L256-257. If seasonal snowmelt is a relevant control on rainfall triggering, then the antecedent precipitation variable ought to somehow account for this, but I suspect it cannot.

We do not have direct evidence that snowmelt is a relevant control on debris-flow triggering. Seasonality in snowmelt is of course present, but snow cover cumulates over a long period of time, so it is not straightforward to account for this as antecedent precipitation. We hoped that the thresholds in Fig. 4 would clarify this but the seasonal thresholds were not conclusive. We could only confidently account for antecedent rainfall. Accounting for antecedent wetness, including infiltration from snowmelt, would require hydrological modelling since we don't have any measured data on snow or soil moisture.

L260-261. Again, these non-conforming observations might also be related to the fairly coarse consideration of antecedent rainfall.

Thanks, we will adjust the sentence accordingly.

L266-268. As a discussion point, it could be interesting to compare this range in parameter variation to the ranges of typical ID thresholds reported in the literature, say for example the difference between values for Caine vs. Guzzetti et al. ID thresholds. I have not done this comparison myself, but it could be worth looking at.

Bel et al. (2017, Fig. 12) did a thorough comparison of ID thresholds for debris-flows torrents. We will refer to this study here.

L372. Even though 20% seems low, this is actually pretty good performance overall for an ID threshold relative to others developed worldwide, so that just further highlights the multitude of complex interactions that lead to debris-flow triggering and justify the need to explore more data-rich approaches like the RF you propose.

Thanks, we will add this observation to the discussion.

References

Leonarduzzi, E. and Molnar, P.: Deriving rainfall thresholds for landsliding at the regional scale: Daily and hourly resolutions, normalisation, and antecedent rainfall, Natural Hazards and Earth System Sciences, 20, 2905–2919, https://doi.org/10.5194/nhess-20-2905-2020, 2020.