Natural Hazards and Earth System Sciences Discussions

Nat. Hazards Earth Syst. Sci. Discuss., referee comment RC1
https://doi.org/10.5194/nhess-2021-135-RC1, 2021

Comment on nhess-2021-135

Ben Mirus (Referee)

Referee comment on "Evaluating methods for debris-flow prediction based on rainfall in an Alpine catchment" by Jacob Hirschberg et al., Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2021-135-RC1, 2021

This paper presents a very important and thorough investigation of sources and impacts of uncertainty in different methods for determining rainfall intensity-duration (ID) thresholds for debris flow forecasting. It is a particularly useful and interesting contribution that can have immediate implications for researchers interested in developing landslide warning thresholds with the ID approach. The paper compares different optimizations of ID threshold including the true skill statistic, a regression analysis, and a random forest (RF) machine learning model. These methods are rigorous and rely on a solid local-scale dataset from the well-characterized Illgraben catchment in Switzerland, as well as the Swiss national landslide inventory database and a daily rainfall product. Using bootstrap resampling of their dataset the authors calculate robust uncertainty estimates and evaluate the value of information and observational record duration for defining robust ID thresholds. Their RF model does not introduce compelling improvements in threshold performance, but the framework provides useful insights in the value of information for improving landslide forecasting capacity.

Overall, the paper is very well written. 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. However, 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. Beyond these missing details, some further details on the rationale behind the scenarios selected for different optimization strategies would be helpful. I have outlined a number of specific suggestions below related to these and other minor concerns, so these comments should be addressed prior to final publication in NHESS.

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.

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.

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

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.

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?

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)?

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

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.

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

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.

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.

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?

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.

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

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.

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

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

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)?

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.

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

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.

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.