Response to referee comment by Clàudia Abancó

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

Specific comments:

Title: It may be a bit misleading. It does not deal with the limitations of the thresholds but more with the uncertainties on their definition? I would suggest reconsidering it...

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".

L65: I would suggest adding a few references of studies using different MIT, as it is said they range from 10 min to 6 h but no references are given (although they appear later, in 3.3., but I would add them here too)

We will add references here, including the ones in section 3.3.

L82: Although the two methods that are going to be compared are mentioned in the abstract, I would list them here too

We will do that.

L80-85: I miss here stating as an objective (maybe as a secondary one) the analysis of the performance with local vs. Regional dataset, which is stated in the abstract.

Thank you for pointing this out. The local vs. regional analysis should be mentioned here. We will add that we investigate how the uncertainties associated with the methods used for a local data set with local rain gauges change when the same methods are applied to a regional data set with gridded rainfall information.

L89-98: Cite Figure 1 in this paragraph

We will do that.

L106: Rain gaugeS in plural? If there's more than one, why in the Figure only 1 is shown? Why only data from 1 is used?

There are two other gauges in the basin which are not suited for this study. One has been moved during the study period and the other is sheltered by trees. Furthermore, the used rain gauge is the one closest to the triggering area. We will add this explanation.

L111-113: If the geophones and depth sensors have been removed, which sensors (less maintenance) have been installed?

We created a bit of confusion here. Badoux et al. (2009) describe the alarm system and some of these sensors have been replaced. However, these are not the same sensors used for the debris-flow monitoring conducted by the research institute WSL, which this paper focuses on. Here, only information on the time of debris-flow occurrence was used as it was detected at the force plate or earlier by geophones upstream. We will add explanations on these sensors and indicate them in Figure 1.

L115: By including the citation of Badoux in line 113 I think you could delete it in Line 115, it is clear for the Reader that details can be found there

We will delete this last sentence as it is redundant.

L122: Any reference where snowmelt has been observed (even if not as sole trigger)?

We will add that although snowmelt in many places adds considerable amounts of liquid water to the debris (Mosterbauer et al., 2018), it has never been observed to be the sole trigger in Illgraben.

Figure 1: Why only force plate is indicated? I would suggest adding the other sensors (the new ones replacing the geophones)?

We will add the locations of the relevant sensors and label them (see also response to L111-113).

L135: 5 mm? This sounds like a very low number...

It does sound low but the Illgraben ID thresholds are also lower than in other places. Debris flows are either triggered in individual gullies or in the main channel. When the latter is the case, water drained from multiple gullies with low infiltration rates concentrates in the channel. This possibly explains the low thresholds.

L136: Does the local rain gauge not have a thermometer? Also, I would suggest moving Temperature, lighning strikes and other parameters to another paragraph, as I understand these are all variables for the Machine learning, but not actually for the main ID thresholds comparison? I was a bit confused reading about rainfall and changing to temperature abruptly, as it's the first time you mention the temperature variable

There is a thermometer at the local rain gauge but because the sensors are not properly shielded, the measurements are unreliable.
We will make two paragraphs to better separate variables only used for machine learning.

L160: Reference of TSS?

We will add references.

L169: Include Area Under Curve as clarification of what AUC stands for

We will do this.

L173: I understand that you used the same criteria for both trigg and non trigg rainfalls?

Yes, discretizing the time series into rainfall events was done before separating into trig. and non-trig., so we used the same criteria. We will specify this.

L175: Delete ? before Deganutti

LaTeX couldn't find the reference due to a typo. We will correct that.

L180 and Figure 2: I think this is more results than methods?

We decided to add it to the method part because it was a necessary step in the process of defining rainfall events and in the result section we wanted to focus on the results addressing our research questions. As mentioned in the text (L. 178), we followed Bel et al. (2017) for this purpose, who thoroughly discusses ID threshold sensitivity to MIT.

Figure 2 (b): Sensitivity? This word may be confused by SE? Could it be called "Analysis of ID-threshold parameters with changing MIT"?

Good point. We will follow your suggestion.

L184: I would not say that β stabilizes, but reduced the increasing tren dat MIT 3h... (in Fig 2b)

We will adapt the sentence accordingly.

L197: Actually, if it was snowing the data from the rain gauge would not be valid, right? As it is not heated... Have you considered this? If so, maybe you could mention here.

Yes, this is an additional reason not to trust the rainfall measurements when it was cold. We will add it here. However, most debris flows occur in summer when solid precipitation is rare even for the higher parts of the basin. Therefore, there are only very few data points where this is the case.

L208: This last sentence of the paragraph ("Lately, confusion matrix...") is actually a bit confusing to me. You have not used frequentist method, right? You used linear squares and LR&TSS and TSS&TSS methods if I have understood right. Therefore, the sentence is confusing as it seems that you have calculated the confusion matrix and ROC for the frequentist method...

In this paragraph we reflect on how ID threshold parameters are determined in literature. We will clarify that determining ID threshold parameters using confusion matrix and ROC, on which we rely on in this paper, is an alternative to the frequentist method.

L220: This is also confusing. A record of length 5 years includes 5 annual samples that can include repetition of the same year? Why is the procedure repeated 100 time for each record length? Please clarify

We agree that the sentence is a bit confusing. The procedure is resampling with replacing and is frequently used for uncertainty assessment. Without replacing, we wouldn't be able to estimate uncertainty bounds because for 17 years there would only be one possible combination. Thus, for each record duration (T), we repeat the sampling with replacement 100 times to have enough combinations of years (i.e. for each T there are 100 samples consisting of T years). The stable median and uncertainty bounds confirm that 100 repetitions were sufficient. We will clarify this and add a reference to the resampling method.

L280: I think it would be good to see the total rainfall amounts at some point in a Figure, as it is stated here that long duration need more rainfall (logic, but still nice to see)

We will add a panel to Fig. 3 with total rainfall amounts.

L297: Higher antecedent rainfall amount may lead to higher degree of pore saturation along the entire channel bed, but also, in some cases the antecedent rainfall would mostly contribute to the generation of lateral flow and increase of water table (e.g.: M.N. Papa, V. Medina, F. Ciervo, A. Bateman 2013, Derivation of critical rainfall thresholds for shallow landslides as a tool for debris flow early warning Systems). This could also correlate with the fact that magnitudes are bigger, but I would say that the correlation between the antecedent rainfall and the magnitude it is a tricky point and needs careful evaluation...

Thank you for pointing this out. We will add it to the discussion. However, in Illgraben in Hirschberg et al. (2019) we found that 14 days antecedent rainfall best correlated with the magnitudes, and we also tested shorter durations. Our thinking is that in such a steep gully, the rainfall contributing to lateral flow certainly happens at shorter time scales which is closer to few hours rather than 14 days. This gives us confidence that antecedent rainfall influences the saturation in the channel bed, and therefore the debris-flow magnitudes.

L305: TSS&TSS thresholds are lower for short durations (<4.5 h) and higher for long durations- after this I would add (Figure 5e)

We will add it.

L311: However, the biases decrease to _30% already after 6 years or _25 triggering events- For both? Or only for β? I can't see it that clearly in alpha?

For both. We will match the y-axes in order to make it clearer.

L335: Also, the source of rainfall data is different, right? If I am not wrong the work of Leonarduzzi et al. it was not based only in rain gauge data. Therefore, apart from climàtic, topographic and lihologic uncertainties it may be also from the type of rainfall data?

Yes, you are right. We will add a better explanation hereof the work in which different rainfall data sets and resolutions were tested (Leonarduzzi and Molnar, 2020). We will also add a better description of the regional dataset (as suggested by RC1). 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.

Figure 7:

- *The grey dots are very difficult to see, specially over blue, red and green bars. Change colour of bars or make dots bigger*
- *I find this figure particularly dense and a bit difficult to follow. Some ideas on how it could be made a bit easier to read:*
  - *a) I understand that RF_ID+1 is based in one sigle predictor (the one with best performance). Why not indicating which one instead of leaving the reader to interpret?*
  - *b) Same with RF_ID+var and 4 predictors*
  - *c) Maybe then it would not be necessary to include all the single predictors in the same figure. Either include them in a separate figure or as supplementary material?*
  - *d) If you think it is relevant to keep the same format, I would suggest indicating the selected predictors for each RF model in some way...*

We will follow your suggestion about the dots and the labels. Half of the figure is filled by predictors without predictive performance, so we will visually better differentiate between models and single predictors with predictive performance and the ones without. This should make the figure more accessible. We think it’s relevant that all used predictors remain in the figure to directly see the best predictors without having to look for the irrelevant ones in the text.

References

Hirschberg, J., McArdell, B. W., Badoux, A., and Molnar, P.: Analysis of rainfall and runoff for debris flows at the Illgraben catchment, Switzerland, in: Debris-Flow Hazards Mitigation: Mechanics, Monitoring, Modeling, and Assessment - Proceedings of the 7th International Conference on Debris-Flow Hazards Mitigation, pp. 693–700, 2019.

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.