This NHESS Discussions paper provides a detailed and objective investigation of numerous factors related to development of rainfall thresholds for landslide forecasting. It relies on a database of landslide occurrence across Switzerland and four alternative configurations of rainfall data in daily and hourly resolutions. Beyond the issue of data temporal resolution, the authors investigate the effect of uncertainty in landslide timing, sparseness of rain gage data, duration of records, normalization of rainfall thresholds for different regions, and the role of antecedent rainfall in threshold performance.

Overall it is a very relevant topic and a very nice contribution. In fact, it provides quite a few surprising and constructive insights that can inform future considerations for landslide threshold development, so I wonder if the title could be rephrased to reflect the various novel contributions of the work, not just the limitations? Ultimately, the paper should definitely be published in NHESS with some quite minor revisions. In particular, the investigation of antecedent conditions was not entirely clear to me, so the description of the methods and results could be improved. Otherwise, numerous edits would enhance the clarity of other aspects of the study, which I have outlined by line number below.

We thank Ben Mirus for the review and the constructive comments. We address here all the points raised in the review.

1: When it comes to landslides, I have started to prefer "forecast" over "predict" since it implies less specificity on location and/or timing. Also, for a concise abstract one could delete phrases such as "In this paper" as it's not needed.

We agree

2: You are not quite providing a comprehensive evaluation of "landslide prediction performance," since that can take many forms, but rather specifically of "rainfall threshold performance."

We agree

15: Avoid ending on a negative note. Perhaps rephrase to state that is it worth the additional effort to build antecedent rainfall into threshold curves?

Rephrased to: "Finally, we demonstrate that there is predictive skill in antecedent rain as a proxy of soil wetness state, despite the large heterogeneity of the study domain, although additional effort is required to build this into rainfall threshold curves."

21: In a new paper we provide further updates and review of reports on economic losses in the U.S. as well as analysis of over 300,000 landslides (Mirus et al., Landslides, 2020, DOI: 10.1007/s10346-020-01424-4).

We will add this reference

50: Specify that you focus on "different temporal resolution of data." Even though this does also relate to the negative consequence of lower density and duration of rainfall measurements.

We agree

60-61: Might be worth clarifying that these studies have in fact demonstrated the utility of including antecedent conditions, but at a relatively narrow scale comparted to the effort you explore here. However, as you know, Wicki et al. (Landslides, 2020, DOI 10.1007/s10346-020-01400-y) have already *evaluated soil moisture at the regional scale for landslide warning. Also, probably our other paper from 2018 is more appropriate for citating here related to comparing antecedent rainfall and soil state (Mirus et al., Landslides, 2018, DOI: 10.1007/s10346-018-0995-z).*

We were planning on citing Wicki et al. in the revised manuscript, unfortunately it wasn't yet published when we first submitted. We also added a reference to Mirus et al., 2018. A sentence has been added mentioning the scale of previous studies as suggested: "… and the inclusion of antecedent rainfall which provides additional information on soil state prior to landsliding, typically studied at local scales (Glade et al., 2000; Godt et al., 2006; Mirus et al., 2018a, b)".

62: It's not clear what a realistic comparison means, so it might be more accurate to state "… an extensive, objective comparison between real rainfall data at hourly and daily resolutions for…"

Rephrased to: (a) to provide and extensive comparison between hourly and daily rainfall data for the definition of rainfall thresholds, considering several practical consequences of choosing a higher temporal resolution

67: What is "TSS"? Should introduce all acronyms before using and also repeat definitions in figure captions and tables for clarity.

We will implement this

80-83: This is a bit unclear and maybe includes several typos or confusing phrasing, so I had to re-read a few times:
- Rainfall not raninfall
We will modify this
- Clarify that you used two different hourly gridded data, not two-hourly gridded rainfall. Just avoid that source of confusion.
We will implement this
- Initially it was unclear how hourly data could be derived from RDI, so I thought it was a typo until later reading the disaggregation methods.
Suggested revision: "We used two different hourly datasets that were derived by disaggregating the RDI such that the daily sums match that of the corresponding RDI cell at the same 1 x 1 km resolution."

We agree

87: Is it possible to give a range of distances to explain what you mean by "quite sparse"?

It is visible in Figure 1. We added a reference to the figure, as well as an estimate of the average density (ca. 1 rain-gauge per 900 km^2)

89: It may be unclear to some readers what the fourth record is. You have only described the daily RDI and two hourly records RHIR and RHIG (derived using the RDI and RHG). Consider listing out all four record names here to avoid confusion.

We will list the names as suggested

175: Consider adding Thomas et al. (WRR, 2019, DOI: 10.1029/2019WR025577P) here as well regarding investigations into satellite measurements for landslide warning.

We will add the suggested reference

179-180: Since this is the opposite of what is normally done, I think a slightly more detailed explanation is needed. I was not able to fully grasp the methods or interpretation of the results in Figure 7.

We always work with rainfall events as defined at the beginning of section 2.2. Therefore, by “duration” we refer to the actual duration of the events defined accordingly (given the number of dry hours to separate individual events). All of these events were observed as triggering (if a landslide happened during or right after them) or non-triggering otherwise. According to the optimum ED threshold, we can separate them also into predicted triggering (above the curve) or non-triggering. The intersections of these prediction/observation gives us the 4 groups: false alarms, true predictions, misses, and true negatives. If the antecedent rainfall is the parameter explaining the “ED failures”, we would expect that misses were associated with high antecedent rainfall, and false alarms with very low antecedent rainfall. We investigate this by averaging within each of the 4 groups the antecedent rainfall for each event duration. We decided to do it separately for each event duration (all events of duration 1 day, all events of duration 2 days etc.), because we suspected there could be differences also relative to the duration (which could also be a proxy of storm type).
We expanded the description in the paper accordingly, to better introduce and explain the methodology.

239: Consider listing “his/her/their”, using only the pronoun “their,” or more simply revising to “overconfidence in the threshold predictions.”

We will fix this

300: As you note, the 3-15 day approach of Chleborad and others is indeed specific to the Seattle area. It would be possible to evaluate what time-scales are most appropriate for distinguishing between rainfall linked to the “trigger” versus the “cause” as outlined by Bogaard and Greco (2018). We re-evaluated the appropriate timescales for ID and cumulative rainfall thresholds (Scheevel, Baum, Mirus, and Smith, 2017; doi: 10.3133/ofr20171039) as well as rainfall-saturation thresholds (Mirus, Morphew, and Smith, 2018, already cited herein). Thus, it is possible that other timescales are better for Switzerland. Can you clarify which timescales you tested, how, and why those times were selected? Again, the methodology for considering antecedent conditions was not totally clear to me.

We didn’t focus on the triggering vs antecedent rainfall threshold. We simply followed the exact approach of the Seattle area, supported by the fact that the optimum (TSS maximization) threshold for event duration was of 3 days. We decided not to try to improve and optimize this, but rather chose to follow a different approach to verify if indeed the antecedent wetness signal was visible even over such a scale. Hopefully the explanation in response to the comment of lines 179-180 improves the understanding of the approach and methodology.

308: Confusing. Revise to clarify that they were wrongly predicted as triggering, but no known landslides occurred due to low antecedent rainfall.

Changed to: At the same time some false alarms (non-triggering events above ED curve) were wrongly predicted as triggering, but no landslide was observed due to the very low antecedent rainfall

340-348: This paragraph has a number of typos and grammatical errors relative to the rest of the paper, so perhaps these were overlooked in the authors final editing?

We revised the paragraph as follows: "At the hourly resolution also the richness of the landslide database is affected, as not only the date but also the timing of the landslide must be known. Staley et al. (2013) addressed this issue and showed the overestimation of thresholds when considering peak rainstorm instead of triggering intensity. This is common practice, when the actual timing of the landslides is unknown. It generally leads to overestimation of the triggering events' maximum intensity, but potentially also other triggering events' parameters. Here, the optimum threshold does not seem to change much, especially when the threshold is obtained maximizing TSS. This is true if at least the landslide date is known. Constraining the timing of landslides on the actual date seems a better choice whenever possible. Allowing a larger window (48h centered on the actual timing) leads to bigger threshold changes, both if maximising TSS or following the frequentist approach. Nevertheless, in both cases, the performances are overestimated if the peak intensity is used to time the landslide, giving the user overconfidence in the threshold values themselves."

343: Careful to clarify that the threshold / triggering intensity is over-estimated, not the measured max. rainfall itself being overestimated.

Changed to: This generally leads to overestimation of the triggering events' maximum intensity, but potentially also other triggering events' parameters.

356: This may be a good point to mention that whereas hourly records capture some indication of rainfall intensity, daily rainfall totals tend to represent a metric somewhere in between the intensity of the storm and its cumulative depth. It might provide another explanation for why the daily still performs reasonably well, since not all landslides are triggered by brief, high-intensity events.

We agree with your reasoning. In fact, we believe that hourly intensity can be considered a real intensity, close to the physical process, while daily rainfall is more representative of a weather system (rather than a storm). This is also one of the reasons why in some climates with longer storms triggering landslides, daily thresholds work.

357: Consider specifying which two methods.

We will specify the methods

364: The only method? I'm sure that some alternative approaches could be proposed by others. Consider avoiding such an absolute statement.

We're definitely not suggesting the frequentist approach is the only method. By saying "a method like the frequentist approach would be the only option" we mean any methodology that only considers triggering events, such as the frequentist. We will rephrase it to avoid confusion.

365: For brevity: "Lastly, we demonstrate the benefits of normalizing the rainfall thresholds using high quantiles of rainfall intensities, quantiles of event properties, MAP, or RDN. These are all particularly useful when using daily data, but we suggest MAP as it is general and a widely available climatological variable."

We will change it as suggested

371: Still not clear what you mean by “realistic comparison.” Suggest: “… of providing a rigorous and objective comparison between…”

By realistic we mean that it not only considers hourly or daily rainfall, but also takes into account the implications of choosing hourly rainfall (i.e. shorter rainfall records, lower quality rainfall records, less landslide events available) mentioned several times in the manuscript. Rephrased to: providing a comparison between hourly and daily rainfall resolutions, which considers data limitations associated with choosing a higher temporal resolution.

372: “…unknown landslide timing, and more sparse rain gage networks…”

We will change it as suggested

376: Suggest: “…more appropriate for forecasting landslides since it better captures triggering intensities, several other aspects…”

We will change it as suggested

380: Suggest: “…daily data are not far behind, potentially since it does tend to capture cumulative storm totals that may also be relevant for landslide triggering.” [?]

As mentioned in the comment relative to lines 356, we agree that rainfall intensities and events at the daily and hourly temporal resolution have different meaning. As mentioned above, we believe daily rainfall is rather representative of weather systems and cumulative storm properties which last in most cases less than 1 day.

383-385: This is true and useful, but unlike your other conclusions, it is nothing new. Consider clarifying that your results further underscore/reinforce previous findings about the importance of non-triggering events.

It is definitely not a new conclusion, but due to the large number of studies in which this is still not done (without explicitly mentioning why it was not possible to include non-triggering) we still believe it’s an important message. We will improve the conclusion accordingly: “Whenever continuous rainfall records are available together with a landslide inventory, our work underscores the importance of including non-triggering events in the definition of optimal rainfall thresholds, not only because false alarms are an essential factor in warning systems, but also to increase the robustness of the threshold estimation.”

392: “…these would increase...”

We will change it as suggested

Figure 1. Define NASS and mdi. Increase legend size for RDI in Swiss map.

We will increase the size of the legend and remove “NASS” from the figure, since it’s a not needed acronym.

Table 1. I think you know the dates of all the landslides during the various time periods, no? If so, please revise “known timing” should really be “known date and time” to avoid any confusion. Same for the subsequent figures.

We will revise as suggested

Figure 2. Define all acronyms used in figure and caption.

We will add all missing acronyms explanations

Figure 4. It’s a bit surprising that the duration in the upper plots (hours) have the same axis numbers (10ˆ0 and 10ˆ1) as the duration does in the lower plots (days).

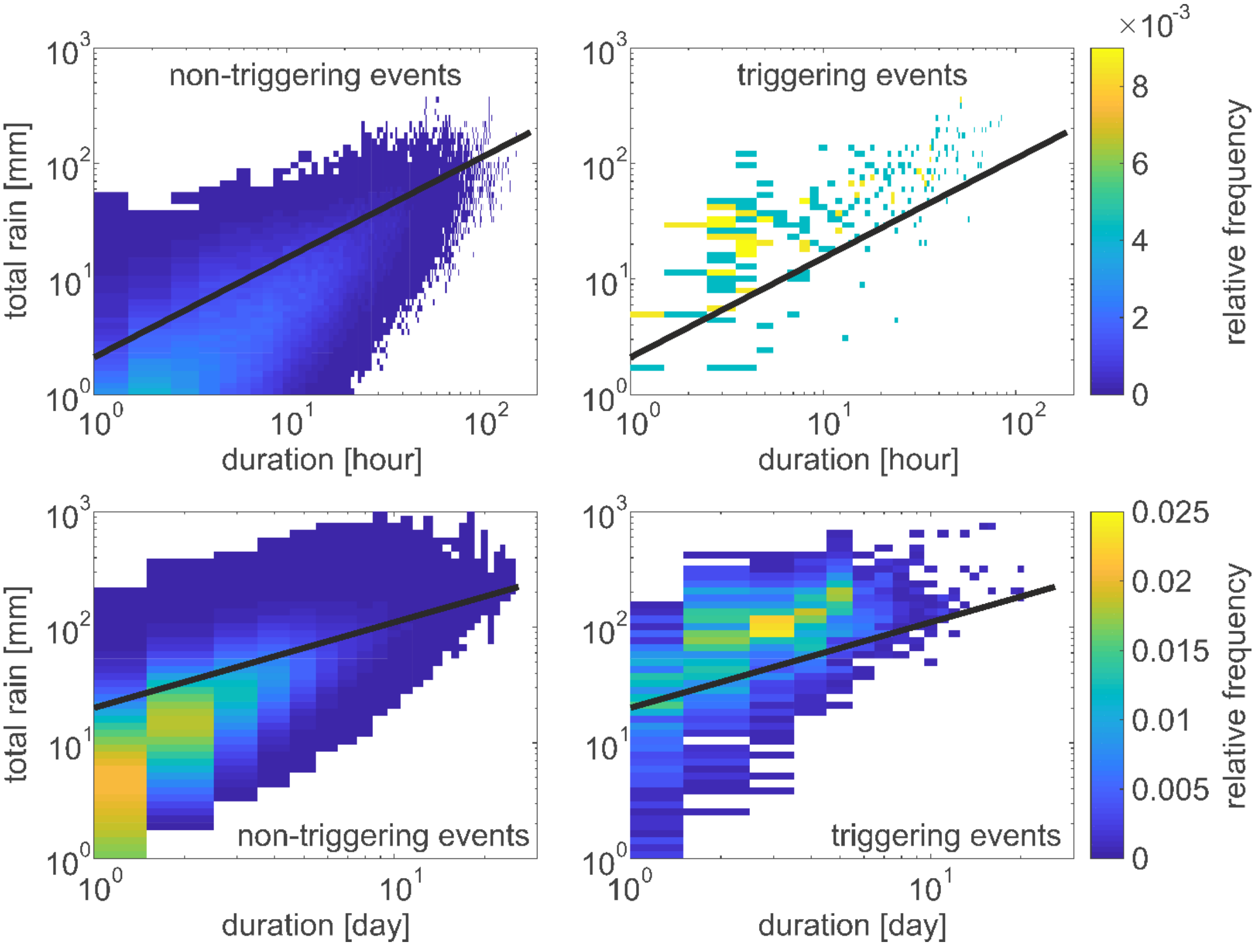
Thanks for this comment. We realized that the figure in the manuscript is an older version where ID thresholds (instead of ED) are used and the axis for the hourly data is indeed cut. We will update it with the correct version (see Review-Figure 1)

Figure 6: Define acronyms like MAP and MDP. I’m not sure I understand the x-axis label and there are no numbers. Please clarify.

The Mean Daily Precipitation (MDP) as the name says is simply the average daily precipitation (equal to Mean Annual Precipitation / 365). We chose to use the MDP because it’s quantitatively comparable to the other precipitation estimates reported in Figure 6, while the MAP would require a secondary axis and be more difficult to directly compare.

Figure 7. Consider labelling (a) and (b) or clarifying that lower plot is Mean Antecedent rainfall (MAR) for 30d. Again, define all acronyms used in caption or legend. These results are a bit confusing and I’m not sure the methods or results are explained clearly enough. What does the “duration” refer to? Duration of the triggering storm event?

Yes, as mentioned here above, “duration” always refers to events’ duration (triggering or non), with events defined given the interarrival time (as described in section 2.2). We added the definition of T (triggering) and NT (non-triggering).



Review-Figure 1 Total rainfall - duration (ED) plots with color scale representing the relative frequency of non triggering (left) and triggering (right) events. The lines represent the best power law curve thresholds obtained maximising True Skill Statistic, above with hourly (RHIR) and below with daily (RDI) rainfall data.