GENERAL COMMENTS

Millan-Arancibia and Lavado-Casimiro describe how they determine regional rainfall thresholds for landslides in Peru at the national scale. They use recently developed methods to objectively assess these thresholds, which makes the study interesting more from a technical than from a scientific perspective for those who aim at implementing early-warning systems. The study is a bit hard to read and seems unorganized in some parts. As a consequence, parts of the methods, results and conclusions were not clear to me. I have few general comments and more specific ones below, which should be addressed before publication in NHESS.

1. There are quite many specifications and clarifications needed in order to make the methods they used unambiguous and reproducible. This also resulted in quite a long list of specific comments below.

2. Some paragraphs seem unnecessary wordy or seem like a random list of unrelated statements, which makes it difficult to follow. For example, L. 177 "TSS is more objective than simple random estimate", it could be explained what makes TSS objective (e.g. balancing TPR and FPR). Some of these arguments are in the text but unorganized and unclear. I think the authors will easily identify such paragraphs themselves when editing. See also comments below.

3. I miss mainly two discussion points. One is the spatial variability of thresholds and the origin of this. Can it be explained with climatology/lithology or is it related to the quality of the data set? See also comments to Figure 7. The second point is related to how calibration/validation is performed, there is almost no discussion about that. I appreciate that this important step is taken and I understand that the dataset is new and short. However, I think it should be stated more clearly that a validation set of one year is quite short and there is a risk of overinterpreting. I suggest to at least discuss other possible validation techniques than splitting years, and flag that as a topic for future research.

4. There are some results and conclusions that are not clear or surprising to me, which should be checked. For example, I would expect Imean-D and E-D thresholds to result in the same performance, but this is not the case here. See comments below.

SPECIFIC COMMENTS

L. 24: Citation needed for the original cause and the different processes leading to saturation

L 27: (e.g. Prenner…)

L. 31: rainfall thresholds

L. 35: time

L. 31: The literature you cite only consider statistical methods. Berti et al. (2020) or Tang et al. (2019) are examples for thresholds based of physically-based modelling. Please also change “physical bases” to “physically-based models”

L. 37: in the way it’s written it makes one think that the difference between the global and national rainfall thresholds is that one is based on antecedent precip and the other on empirical-statistical approaches. Please rephrase. Also, if you use “antecedent”, does it have the same meaning as in L. 29? Antecedent conditions can refer to the conditions prior to the triggering rainfall or prior to the exact time of landslide occurrence. Please specify and use consistently.

L. 45: I think this section is to justify the methods used. Given the uncertainties in the rainfall product that you mention later in the ms one could ask why you’re not using physically-based modelling, which consider the actual mechanisms causing landslides, to back-calculate rainfall thresholds. Hence, I would also mention the challenged accompanied with such models: mainly the many high-quality input data such as soil information that is needed, which is associated with high uncertainties, too.

L. 56: maximum at what scale? Daily, annual?

L. 60: gridded

L. 80: Just out of curiosity. It’s funny enough that the precipitation dataset is named after Peru’s national liquor. Is PISCOpd_Op actually the abbreviation of something?

L. 84: Can you give some information on the number of rain gauges or the average distance? Maybe even add them to the map in Figure 2 if you have such a map.

L. 85: What do you mean by “multipliers that are based on monthly climatology”?

Table 1: I’m not sure this table is so important. To me, only the spatial resolution and the time period is of relevance. But why compare these two datasets if you only use one of them?

L. 92-93: these two sentences can be simplified, now it is confusing. So SLIP covers the period 2018-2020 but you have greater certainty for 2019 and 2020?

L. 101: Figure 3

L. 88-101: I don’t understand how the two landslide databases were combined. They time periods do not overlap and Figure 3 only starts at 2019. If one event was excluded it should be 382 events in total. So which was your study period?

Figure 3: What is this rainfall? One grid cell? Which location are we looking at? And the colour is one rainfall event? Please specify in the caption and add labels a) and b= to subplots.

L. 103: Since you describe the sequence of your methods here, Figure 1 would fit here. And describe the steps in the text and refer to the figure.

L. 116: How can the PISCO report Pr>0 and the station Pr=0 if Pisco is interpolated from the stations?

L. 118: How were rainfall events defined? Are two events independent if they are separated by at least one non-rainy day?

L. 131: events

L. 134: I think that E-D and Imean-D should result in the same thresholds, only that b(E-D) = b(Imean-D)+1. That's what I get when substituting Imean with E/D. So there is no point in comparing both thresholds. This said, I'm surprised that by the number is table 3. Either I'm misunderstanding something or something went wrong here. Please clarify.

L. 135: a and b are scale and shape parameters, but in the log-log space they become intersection and slope of the linear threshold

Figure 4: These box plots are nice but it's not clear from the text why you show them. Is it to show that the two can be separated well? Considering the methods you use, it would be nice to see some AUC curves instead which would also help you in explaining the methods

L. 146: Max precip at what time scale? And what is the motivation for using this for regionalization and one of the other indices?

L. 158: Please consider rewriting or reorganizing this section. The information to certain steps are spread across the entire section, for example, how the dataset was split into calibration and validation data sets.

L. 179-182: This will be confusing for many readers. You have two definitions for TSS, and two for sensitivity. Please be consistent and avoid introducing alternative definitions if they actually mean the same. Also, the TSS itself doesn't seek to maximize TPR and 1-FPR, but you do so by choosing a threshold that maximizes TSS.

L. 185: Please be more specific. It's not clear what you did using ROC, TPR, FPR. Which is the "most widely used technique"? Did you choose some variables with large AUC and dropped the others. If so, what was the threshold AUC. Or did you define thresholds by maximizing TSS? There are many possible

L. 192: It's not clear to me how exactly the validation was performed. Was the performance of the validation data set calculated for the thresholds determined with the calibration data set or was a new threshold determined for the validation data set to see if the performance is similar?

L. 196: The values of 0.4 and 0.7 seem somewhat random. Could you elaborate a bit on the meaning of these values? Are these values commonly used or why is this classification needed?

L. 205-214: I'm surprised that Imean-D and E-D don't have the same performances. See comment L. 134.

Table 2: "D (days)". Is this the full data set, calibration or validation? How many events per region? The same for Table 3.

L. 270: do you mean first in Peru? Please specify.

L. 277: Table 3

L. 280: Yes, landslide detection is sacrificed but false alarms are reduced. There are various scores one could chose depending on if you want to give more weight to the detection or false alarms. But you chose TSS because it's a good balance between the two.

L. 283: What is a high-impact stream?

L. 284: what do you mean by constant landslide occurrence?

L. 284: Imax-D-D?

L. 285: di xyou mean entire event?

L. 286: is background condition scenario the antecedent condition scenario?

L. 286-290: I can't follow. If you're the validation results are better than the calibration, then maybe your validation set is too small. I don't see how you can conclude the importance of antecedent conditions from this. Also, the sentence "in the validation stage…showed growth in calibration performance" is confusing

L. 296: The absence of extreme events does not imply poorer threshold performance. An option would be to do calibration/validation on more data splits.

L. 298: "the number of landslides was lower than in other years" but the only reliable year you can compare with is 2019, right?

L. 313: Again, you mean first in Peru, right? Please specify.

L. 315: Well, you cannot compute empirical-statistical thresholds without landslide observations so this is not really an advantage. An advantage is that you have used datasets available at the national scale to objectively determine and compare rainfall thresholds.

L. 318: it is still not entirely clear to me what process we are talking about. Here you say shallow landslide and earlier you mention streams and debris flow. Is it a mix of processes? Please add some information on this in the dataset description and clearly define what collection of processes you refer to when using "landslide" throughout the ms.

L. 324: More interesting would be why the performances can be so different. Can you say something about that?

L. 329: high sensitivity to what?

Figure 7: Is there a reason for showing sensitivity/specificity? Wouldn't it be easier to interpret if you would just colour according to TSS?
This figure is very interesting and shows high spatial variability in the thresholds. Can you say something about this variability? E.g. is the threshold higher in wet regions? See e.g. Leonarduzzi et al. (2017) Figure 7 or Marc et al. (2019).

Marc, O., Gosset, M., Saito, H., Uchida, T., Malet, J.P., 2019. Spatial Patterns of Storm-Induced Landslides and Their Relation to Rainfall Anomaly Maps. Geophys. Res. Lett. 167–177. https://doi.org/10.1029/2019GL083173