RESPONSE TO 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 morefrom 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 tome. I have a few general comments and more specific ones below, which should be addressed before publication in NHESS.

Comment response: Thank you very much for your review, in the new version of the mn we have tried to make it not a bit difficult to read and also not seems unorganized, considering all your comments. Additionally, this document is highly important for the scientific community related to landslides in Peru since this type of work has not been developed in Peru, which, in addition, faces the limited availability of data compared to other countries. Lastly, other investigations also faced similar difficulties (e.g., Kirschbaum et al., 2015; Abraham et al., 2019).

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.

   Comment response: Thanks for the comment. All your comments and the list of specific observations have been taken into account and included in the new version of the mn.

2. Some paragraphs seem unnecessary wordy or seem like a random list of unrelated statements, which makes it difficult to follow. For example, in 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.

   Comment response: Thanks for the observation. All your comments and the list of specific observations have been taken into account and included in the new version of the mn. We have made an exhaustive revision of the mn and we have identified some paragraphs and we have organized them with greater clarity to avoid their difficult reading.

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 at least to discuss other possible validation techniques than splitting years, and flag that as a topic for future research.

   Comment response: Thanks for the observation. We have taken into account your observations and recommendations and have included them in the discussions of the new version of the mn. Regarding the first point of discussion:

   "Regarding the variability of the thresholds, we can explain it mainly to the rainfall climatology in Peru. It can be seen that the magnitudes have a relationship with respect to the spatial distribution of rainfall in Peru, that is, low thresholds related to rainfall of lesser magnitude in the arid zones in the western part of Peru (Pacific region), thresholds

intermediates related to the increase in the magnitude of rainfall in the middle part or mountainous region (Andes region) and the highest thresholds related to wet regions (Amazon region). However, the Andes 1, Andes 3 and Andes 6 regions do not have this relationship, so this discussion is not conclusive and is considered to be related to limited data, so it is suggested that this variability be discussed in future research that include more shallow landslides events data."

Just to comment, that the lithology in Peru is still highly general and we hope in the future to do exercises with lithological data (e.g., soil tests) that we are developing at small basins level.

About the second point, regarding calibration/validation we have added your observation and we have discussed about it, as you can see below:

"The calibration/validation methodology, based on take one year of observations for validation set, which was used in other research works (e.g., Dikshit et al., 2019; Kirschbaum et al., 2015), is quite short and there is the risk of overinterpretation. It is therefore highly recommended for future research to expand the dataset and explore other calibration/validation methods, for example, a random selection of the differentiated data set for the calibration and validation (e.g., 70% for calibration and 30% for validation) (Brunetti et al., 2021; Gariano et al., 2020)."

In addition, in our future research we hope to advance in these limitations in Peru, for example, our perspective is to expand the database, for which we are working with INDECI (entity in charge of the attention of the population when landslides occur) for future studies that include greater data extension.

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.

   Comment response: Thanks for observation. We have taken into account your comment. For better understand, according to the way we have defined the variables for a dataset, Imean, that is affected by D, does not have the same distribution as E. For example, two events with the same E (e.g. E=10), can have different D (e.g. D equal to 2 and 4 days), therefore, the Imean of both resulting events are different (Imean equal to 5 and 2.5 respectively), so the threshold could not be defined as the division of both. A more specific example for a example dataset is shown in the specific comments below.

RESPONSE TO SPECIFIC COMMENTS

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

Comment response: Thanks for the observation. The citation is: Lynn Highland. 2006. Landslide Types and Processes. USGS Fact Sheet 2004–3072. But it was removed for better understand according the general comments.

L. 27: (e.g. Prenner…)

Comment response: Thanks for the observation. It was edited in the new version of the mn.

L. 31: rainfall thresholds

Comment response: Thanks for the observation. It was edited.

L. 35: time

Comment response: Thanks for the observation. It was edited.

L. 31: The literature you cite only considers statistical methods. Berti et al. (2020) and Tang et al. (2019) are examples of thresholds based on physically-based modelling. Please also change "physical bases" to "physically-based models"
Comment response: Thanks for the observation. It was added the citation examples and edited "physical bases" to "physically-based models". Additionally, we have recently instrumented some basins to collect more accurate data for future research, where we could explore 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.

Comment response: Thanks for the observation. The text has been rephrased in order to clarify the main idea, as you can see below.

"For example, there is been developed empirical–statistical approach to the estimation of global thresholds (Caine, 1980; Guzzetti et al., 2008; Kirschbaum and Stanley, 2018), and national thresholds (Leonarduzzi et al., 2017; Peruccacci et al., 2017a; Uwihirwe et al., 2020)."

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 considers the actual mechanisms causing landslides, to back-calculate rainfall thresholds. Hence, I would also mention the challenges 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.

Comment response: Thanks for the observation. It was edited, as you can see below.

"This empirical approach is widely applied because its analysis and implementation do not require the constant monitoring of the other physical variables on which other types of most robust models are based (e.g., physically-based models), and this drawback of the robust models is the main advantage of empirical approaches and its applicability over large areas (Rosi et al., 2012). Another advantage for its application is that it is not subject to the challenges accompanied with other models, mainly the many high-quality input data, such as soil information that is needed, which is associated with high uncertainties too."

To comment, we are recently developing studies on a local scale with less uncertainties that we will use to define rainfall thresholds at local scale (Asencios Astorayme, 2020a, b).
https://repositorio.senamhi.gob.pe/handle/20.500.12542/478
https://repositorio.senamhi.gob.pe/handle/20.500.12542/476

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

Comment response: Thanks for the observation. It´s daily scale. It was edited.

L. 60: gridded

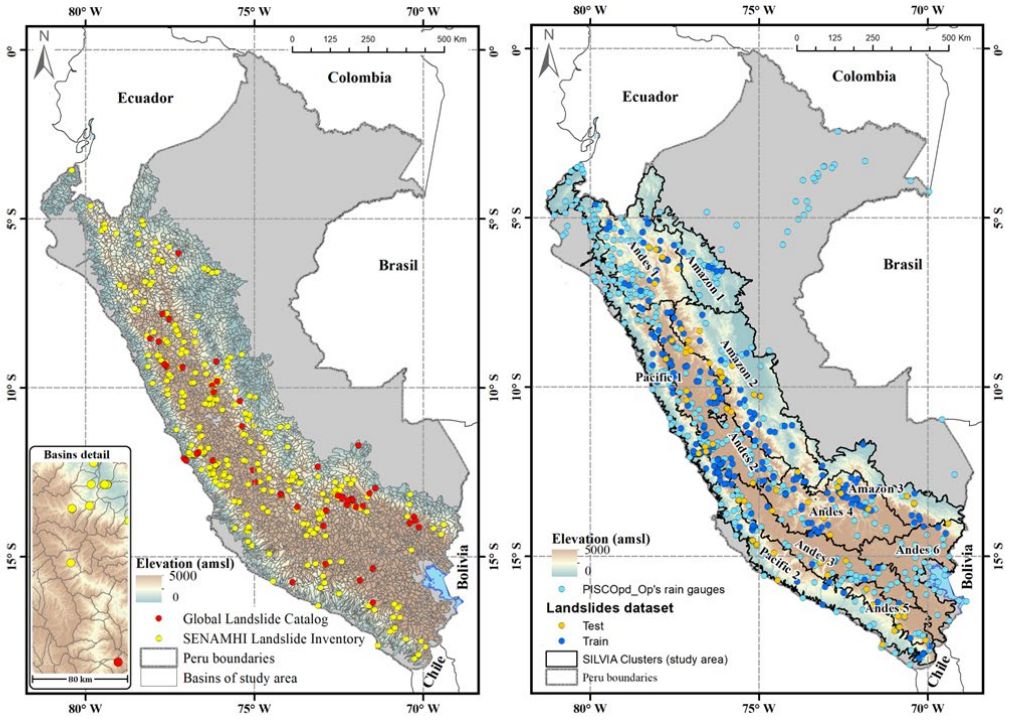
Comment response: Thanks for the observation. It was edited.

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?

Comment response: Thanks for the observation. Yeah, the name helped us a lot as a hydrometeorological service to be able to spread the information in a fun way. The PISCO is derived from: **P**eruvian **I**nterpolated data of the **S**ENAMHI's **C**limatological and Hydrological **O**bservations. PISCO is a base name of different products of SENAMHI, i.e., PISCOpd_Op is derived from **PISCO P**recipitation-**D**aily-**Op**erative Gridded data. It was edited for better understanding, as you can see below.

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.

Comment response: Thanks for the observation. For the PISCOpd_Op purpose, we use 416 rain gauges and them were added to Fig 1 (before Fig 2).



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

Comment response: Thanks for the comment. These multipliers are the ratio between the value of the monthly background grid at location x (extracted from PISCOp monthly climatology) and the value of the monthly back- ground grid at the gauge location for every gauge (derived from rain gauges) to create a set of multipliers from the gauges to the given grid cell. For more information about genre Interpolation Method is shown in: van Osnabrugge, B., Weerts, A. H., & Uijlenhoet, R. (2017). genRE: A method to extend gridded precipitation climatology data sets

in near real-time for hydrological forecasting purposes. Water Resources Research, 53, 9284–9303. https://doi.org/10.1002/ 2017WR021201.

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

Comment response: In consideration of the observation, we decided to remove the table and show only the relevant information (i.e., spatial resolution and the time resolution).

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

Comment response: Thanks for the observation. The SLIP covers the period 2014-2020, it was corrected, and we have more certainty from 2019-2020 just because we were more data and number of events these last years. It was edited, as you can see below.

SLIP was implemented in January 2019 and has 330 records from the 2014–2020 period. Therefore, there is a greater degree of certainty regarding the number of events recorded in recent years.

L. 101: Figure 3

Comment response: Thanks for the observation. It was edited.

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

Comment response: Thanks for the observation. According previous comment, the period was 2007-2020. The number of events was edited. The figure it's just an extracted period for show how we define an event.

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.

Comment response: Thanks for the observation, we have taken into account your comment and the figure has been modified. It's a daily rainfall data for one basin (from GEOGloWS discretization, fig1) where occurred a landslides event. The purpose of the figure was to show how its defined rainfall events (each color it´s a rainfall event). The figure it's just an extracted period for show how we define an event. It was edited, as you can see below.

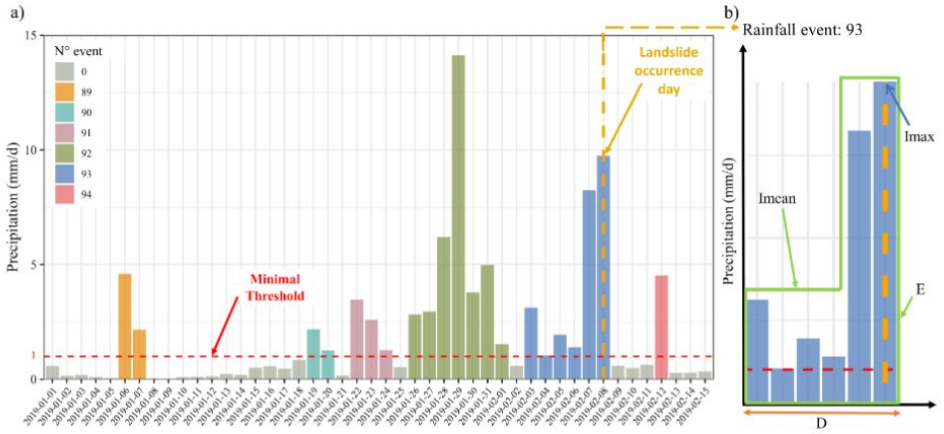


Figure 3. a) Extract from the precipitation time series (rainy period 2019) for an example basin, where the estimated rainfall events are observed (each color is a rainfall event, the lead-colored event 0 is the non-rainy days). b) An example of a rain event associated with the occurrence of a landslide, in this case the rain event No. 93, where the variables analyzed for the estimation of thresholds are shown: the maximum daily intensity I_{max} (mm/day), the accumulated precipitation E (mm), the duration D (day), and the mean daily intensity $I_{mean} = E/D$ (mm/day).

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.

Comment response: Thanks for the observation. It was edited, as you can see on manuscript edited, moreover we put Regionalization subsection before the Rainfall threshold model subsection because we think it help to manuscript better understand.

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

Comment response: Thanks for the observation. The principal reasons for this, is because in the interpolation method it's affected by monthly climatology. Therefore, it is not an exact interpolation, but rather an approximate one, since it tries to represent gridded data at the national scale. Another comment, we are developing another rainfall products that have the purpose of improve the representativeness of rainfall products where there are no terrain data based on novel methodologies with which we think to include them in future research about landslide thresholds. Additionally, the installation of radars and more rain gauges is planned in Peru, which will be assimilated in future rainfall products.

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

Comment response: Thanks for the observation. L 109: "For this work, we define an independent rainfall event as a series of consecutive rainy days where it has rained above a minimum rainfall threshold (Figure 3)".

L. 131: events

Comment response: Thanks for the observation. It was edited.

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 by the numbers in table 3. Either I'm misunderstanding something or something went wrong here. Please clarify.

Comment response: Thanks for observation. We have taken into account your comment. For better understand, according to the way we have defined the variables for a dataset, Imean, that is affected by D, does not have the same distribution as E. For example, two events with the same E (e.g. E=10), can have different D (e.g. D equal to 2 and 4 days), therefore, the Imean of both resulting events are different (Imean equal to 5 and 2.5 respectively), so the threshold could not be defined as the division of both. The Fig. X1 shows what is mentioned for an example dataset, where it is observed that E and Imean have different density distributions and therefore their predictive potentials also change (i.e., the thresholds do not have the same Imean relationship =E/D).

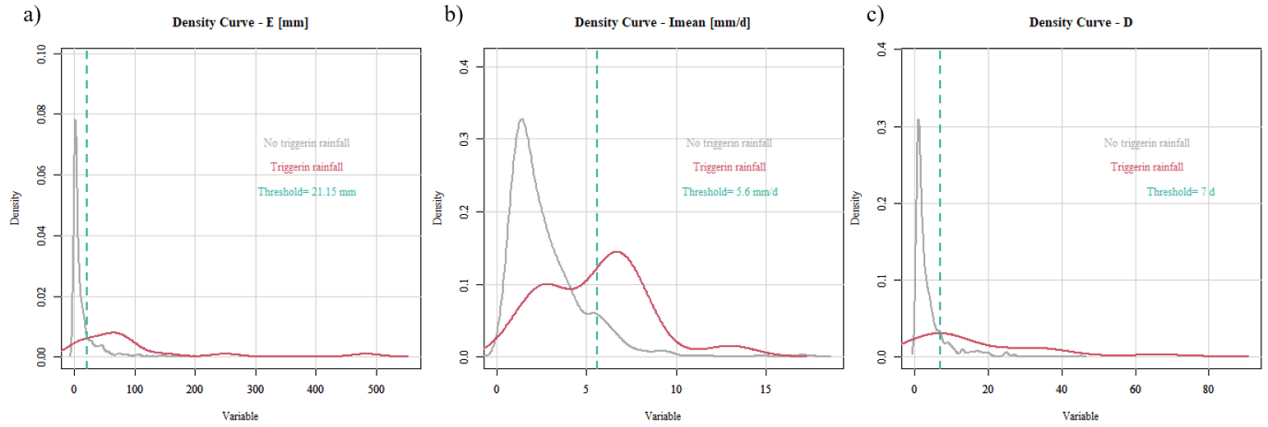


Fig. X1: Density plot of the variables E (a), Imean (b) and D (c) for the same data set, where it is observed that the distributions of the variables E and Imean are different.

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

Comment response: Thanks for the observation. It was edited as you can see below:

"… a and b are the scale and shape parameters of the curve (while for log-log space a is the intersection parameter and b denotes the 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 benice to see some AUC curves instead which would also help you in explaining the methods

Comment response: Thanks a lot for the observation. Indeed, the purpose is to show the ability to separated variables, before determining a threshold, and how it change for each region. The AUC would not help much at first, additionally, this way of showing the potential of the variables has been used in other publications (Martinović et al., 2018; Leonarduzzi et al., 2017). (Martinović et al., 2018; Leonarduzzi et al., 2017). We have taken into account your observation and the text was edited, as you can see:

"Figure 4. Boxplot of triggering (yellow) and no triggering (blue) total cumulative rainfall E for the eleven regions stablished in this study for Peru. The boxplot graphs include outliers and show the potential predictive for the E variable to separate the rainfall events that trigger/no trigger shallow landslides. Also, the plot shows the region variability of the rainfall events that trigger shallow landslides."

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

Comment response: Thanks for the observation. It is daily scale (it was edited). We use this Max Dayli Precip regionalization for Peru in addition to the covariates of relief (altitude) and climatology (average precipitation), mainly because we associate these maximum daily rainfall events with rainfall triggered landslides. The altitude to maintain an orographic similarity, since, in Peru, and in general in South America, the Andes have a modulating character in the presence of rains. And the average precipitation derived from PISCOp that helped us to establish a similarity of the basins, especially in the transitions in the limits between each region.

Also, we use these maximum rainfall regions because we took as an initial reference the paper from Leonarduzzi et al., 2017 where they use the Maximum Intensity within their regionalization, which was the one that gave the best results in their threshold estimation. Finally, we already had this regionalization of previous studies, which is related to the map of climatic regions of Peru (SENAMHI, https://repositorio.senamhi.gob.pe/handle/20.500.12542/1336).

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.

Comment response: Thanks a lot for the observation. I rewrite and reorganize the entire section for better understand, as you can see below.

"2.6 Calibration and validation of thresholds

Calibration and validation are fundamental processes for objectively defining thresholds. The purpose of calibration is to estimate thresholds based on the maximization of predictive or classifier performance capacity. Validation aims to show the potential of the ability to predict or differentiate those rainfall events that trigger landslides. Among the calibration and validation approaches, the most recommended is to divide the datasets for threshold estimation and another independent set for validation (Segoni et al., 2018). In this work, 377 recorded landslide events were used to define rainfall thresholds in Peru (Figure 1). For the calibration, all events occurring before 2020 were selected, representing approximately 80% of the recorded events. Regarding the validation process, it was consisted of evaluating thresholds calibrated using the landslides events recorded in 2020, which represented approximately 20% of the recorded events. This process was carried out for the year 2020, as we wanted to know how the thresholds would perform when they were assimilated into a regional early warning system. This method of calibration/validation that set one year of the dataset to validation is a method that has been used in another research (e.g., Kirschbaum et al., 2015b; Dikshit et al., 2019).

For the evaluation of the thresholds in calibration and validation was used a confusion matrix (also called as contingency table). The confusion matrix is a tool used to determine the accuracy of binary classification models (triggering and non-triggering rainfall events), and also, used to evaluate the analysis of concordance between the results of the model and the observed data. A confusion matrix was computed for each threshold and was counted the number of true successes or true positives (TP), the number of false positives (FP), the number of true negatives (TN), and the number of false negatives (FN) (Figure 5). From which various performance statistics can be calculated. Some of the most common measures for landslide forecasting are the sensitivity (se = TP/(TP +FN)), specificity (sp = 1−FP/(FP +TN)) and true skill statistic (TSS = se+sp−1) (e.g., Staley et al., 2013; Gariano et al., 2015; Leonarduzzi et al., 2017; Mirus et al., 2018; Leonarduzzi and Molnar, 2020; Hirschberg et al., 2021).

The TSS is an efficiency statistic that helps in the measurement of the goodness-of-threshold models, as it is an integrative measure of the predictive performance of the model. The TSS is more objective than simply a random manual estimate (Frattini et al., 2010). It varies between

1 and –1, with its optimal score equal to 1, which indicates a maximum performance of the model. TSS=se-(1-sp) is the difference between the true positive rate (sensitivity se) and false alarm rate (1-specificity sp), which are the two most important components for providing early warnings (Leonarduzzi et al., 2017). The TSS is also referred as the Peirce skill score (Peirce, 1884), the Youden index (Youden, 1950), or the Hanssen–Kuipers skill score (Hanssen and Kuipers, 1965). The benefit of using the specificity over the false positive rate (FPR=FP/(FP+TN)) is that in a perfect model TSS, sensitivity and specificity all equal 1 (Hirschberg et al., 2021).

For thresholds based on rainfall event properties independently (Imax, E, D or Imean), the overall impression of the predictive power was estimated whit the so-called receiver operating characteristic (ROC) curve (Fawcett, 2006), from which the minimum radial distance to the perfect classificatory test (TSS=1, with se=1 and 1-sp=0) was used to select the individual variable threshold (e.g., Uwihirwe et al.; Gariano et al.; Postance et al.) while for the threshold curve (Imax−D,E−D, Imean−D) the scale parameter a and the shape parameter b are simultaneously tuned to maximize the the true skill statistics (TSS) (e.g., Leonarduzzi et al.; Hirschberg et al.). This maximization was automatically calibrated using the shuffled complex evolutionary algorithm (SCEA-UA) (Duan et al., 1993), considering the TSS as the objective function. The methodology was applied for each region within the analysis area, finding different thresholds for each of them."

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.

Comment response: Thanks a lot for the observation. I avoided the use of double definition for TSS, I simplify the paragraph.

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

Comment response: Thanks a lot for the observation. I checked the information and simplified the paragraph, as you can see below:

"For thresholds based on rainfall event properties independently (Imax, E, D or Imean), the overall impression of the predictive power was estimated whit the so-called receiver operating characteristic (ROC) curve (Fawcett, 2006), from which the minimum radial distance to the perfect classificatory test (TSS=1, with se=1 and 1-sp=0) was used to select the individual variable threshold (e.g., Uwihirwe et al.; Gariano et al.; Postance et al.) while for the threshold curve (Imax−D,E−D, Imean−D) the scale parameter a and the shape parameter b are simultaneously tuned to maximize the the true skill statistics (TSS) (e.g., Leonarduzzi et al.; Hirschberg et al.). This maximization was automatically calibrated using the shuffled complex evolutionary algorithm (SCEA-UA) (Duan et al., 1993), considering the TSS as the objective function. The methodology was applied for each region within the analysis area, finding different thresholds for each of them."

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?

Comment response: Thanks a lot for the observation. This validation process was computed for landslides occurred on 2020 year using the thresholds calibrated to get the metric for this period and compare the capacity of thresholds to separate rainfall events that trigger shallow

landslides.

"Regarding the validation process, it was consisted of evaluating thresholds calibrated (both individual and curve thresholds) using the landslides events recorded in 2020, which represented approximately 20% of the recorded events. This process was carried out for the year 2020, as we wanted to know how the thresholds would perform when they were assimilated into a regional early warning system."

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?

Comment response: Thanks for the observation. Considering your comments, we agree with the observation. The values are not standardized, in addition to the fact that they were not taken into account in the discussion carried out, so we decided to remove the sentence.

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

Comment response: Thanks for the observation. It was responded in the observation of the L. 134 from the present text.

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

Comment response: Thanks for the observation. It was corrected. The tables shown the thresholds estimated with the calibration set. The number of events is specified on the Table 3 of the new version of the manuscript (previously table 4). As you can see below:

Table 3. Number of SL events and best thresholds for one and two variables for each region (Th: threshold, SL: number of landslides per region, Cal: Calibration, Val: Validation)

Region	SL total	SL Cal	SL Val	Best Th - 1 variable	TSS	Best Th - 2 variables	TSS
Pacific 1	46	43	3	I_{max}	0.68	$I_{max} - D$	0.71
Pacific 2	27	20	7	I_{mean}	0.61	$I_{mean} - D$	0.61
Andes 1	34	28	6	I_{mean}	0.43	$I_{mean} - D$	0.44
Andes 2	98	83	15	E and I_{mean}	0.58	$I_{max} - D$	0.64
Andes 3	17	10	7	I_{max}	0.92	$I_{max} - D$	0.91
Andes 4	65	54	11	E	0.51	$I_{mean} - D$	0.52
Andes 5	14	7	7	E	0.67	$I_{mean} - D$ and $E - D$	0.66
Andes 6	4	3	1	D	0.68	$E - D$	0.65
Amazon 1	6	6	-	I_{mean}	0.74	$I_{mean} - D$	0.77
Amazon 2	54	41	13	E	0.57	$E - D$	0.58
Amazon 3	12	10	2	E	0.68	$I_{mean} - D$ and $I_{max} - D$	0.73

L. 270: do you mean first in Peru? Please specify.
Comment response: Thanks for the observation. Yes, the first approximation in Peru. It was edited.

L. 277: Table 3
Comment response: Thanks for the observation. It was edited.

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.
Comment response: Thanks for the comment. The paragraph was edited, as you can see:

"However, it was observed that to seek this optimization, the detection of landslides is sacrificed (giving false negatives), though false alarms are reduced, and this is a dilemma in terms of alert systems, but TSS is a good balance between landslides detection and false alarms."

L. 283: What is a high-impact stream?
Comment response: Thanks for the question. We refer to high-impact stream a basin with a constant occurrence of landslides. But it's a local phrase, so it was removed for better understand.

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

Comment response: Thanks for the question. The paragraph was simplified, as you can see below:

"The Pacific 1 region is constantly impacted by shallow landslides and also contains most of the cities with the highest population density in Peru, so their evaluation is highly relevant."

L. 284: Imax-D-D?

Comment response: Thanks for the observation. Its Imax-D. It was edited.

L. 285: do you mean entire event?

Comment response: Thanks for the observation. Yes it´s the entire event. It was edited.

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

Comment response: Thanks for the observation. Yes it´s the entire e antecedent event scenario. It was edited.

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.

Comment response: Thanks a lot for the observation. The paragraph was edited, as you can see:

"The Imax variable had the best performance, which suggests that high-intensity rains have a high conditioning impact on landslide development. Regarding the validation performances in the antecedent conditions scenario were higher in the calibration performances, it may be because the validation set is too small."

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.

Comment response: Thanks for the observation. Regarding calibration/validation we have added your observation and we have discussed about it. The paragraph was edited, as you can see:

"The calibration/validation methodology, based on take one year of observations for validation

set, which was used in other research works (e.g., Dikshit et al., 2019; Kirschbaum et al., 2015), is quite short and there is the risk of overinterpretation. It is therefore highly recommended for future research to expand the dataset and explore other calibration/validation methods, for example, a random selection of the differentiated data set for the calibration and validation (e.g., 70% for calibration and 30% for validation) (Brunetti et al., 2021; Gariano et al., 2020)".

In addition, we add the recommendation that taking only one year for validation may be inconclusive due to the little data, so it should be taken into account in future studies and explore 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?

Comment response: Thanks for the observation. The calibration was made with landslides occurred before 2020 and validation with landslide occurred in 2020.

The paragraph was edited in the new version of mn as you can see:

"For the calibration, all events occurring before 2020 were selected, representing approximately 70% of the recorded events. Regarding the validation process, it was consisted of evaluating thresholds calibrated using the landslides events recorded in 2020, which represented approximately 30% of the recorded events."

L. 313: Again, you mean first in Peru, right? Please specify.
Comment response: Thanks for the observation. Yes, the first approximation in Peru. It was edited:

"This study is the first approximation of the regional rainfall thresholds that trigger landslides in Peru."

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 rainfallthresholds.
Comment response: Thanks for the observation. This recommend was incorporated, as you can see:
"The advantage of this study is the use of landslides 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.

Comment response: Thanks for the observation. We mention streams only to refer a body of flowing water. Regard the processes, we included debris flow category which are shallow in nature (Naidu et al., 2018) into shallow landslide term. A clarification to this was added to the new version of the mn.

"The second main source of information used for this research was two inventories of observed and collected landslide events: SENAMHI's of Rainfall-Triggered Shallow Landslides Inventory of Peru (SLIP) and NASA's Global Landslide Catalog (GLC) (Kirschbaum et al., 2015a). Both catalogs consider all types of shallow landslides triggered by rainfall that have been reported in the media, in databases of agencies associated with disasters, in scientific reports, and in other available sources. Most of them belong to the debris flow category which are shallow in nature

(Naidu et al., 2018). In this sense, in this study was used shallow landslide (SL) for all types of shallow landslide processes."

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

Comment response: Thanks a lot for the observation. The differentiation of threshold yields for each region responds to the high variability of rainfall events and their properties (see Figure 4 Boxplot and Figure 7 threshold plots) in each region, we explain this topic and add the next conclusion, as you can see below:

"The performances of the calibrated thresholds had a high differentiation between regions. This performances difference is associated with the highly variability of rainfall events and their properties in each region, where it is observed that the best performances occur in areas where it is easier to separate rainfall events that trigger and no trigger shallow landslides, which is reflected in high performances (Andes 3, Amazon 1, Amazon 3 and Pacific 1 regions). However, in other regions, this separation between rainfall events is more complex to carry out, since there are more rainfall events with high magnitudes that do not trigger landslides but that exceed the thresholds, reflecting in lower performances (Andes 1, Andes 4 and Amazon 2). Thus, we could assume that in these regions there is a greater incidence of lithology and geology in the occurrence of SL than just the rains."

L. 329: high sensitivity to what?

Comment response: Thanks for the observation. High sensitivity to the little data, in the context of scare data of shallow landslide events of Peru. The text was edited for better understanding, as you can see:

"However, the uncertainties associated with these databases are the main source of uncertainty fo the thresholds. The few landslides recorded made that the validation performance had highly sensitive to the few data (i.e., a single event could lead to a high or low value of the performance statistics)."

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

Comment response: Thanks for the observation. Our reason for showing the sensitivity/specificity was to show which parameter had a greater incidence in the TSS, whether it was the good detection of triggering events (sensitivity) or the good detection of non-triggering events (specificity).

We have taken into account your observations and recommendations and have included them in the discussions of the new version of the mn, as you can see below:

"Regarding the variability of the thresholds, we can explain it mainly to the rainfall climatology in Peru. It can be seen that the magnitudes have a relationship with respect to the spatial distribution of rainfall in Peru, that is, low thresholds related to rainfall of lesser magnitude in the arid zones in the western part of Peru (Pacific region), thresholds intermediates related to

the increase in the magnitude of rainfall in the middle part or mountainous region (Andes region) and the highest thresholds related to wet regions (Amazon region). However, the Andes 1, Andes 3 and Andes 6 regions do not have this relationship, so this discussion is not conclusive and is considered to be related to limited data, so it is suggested that this variability be discussed in future research that include more shallow landslides events data.”