

# Response to comment on nhess-2022-199 - Anonymous Referee #1

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 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 manuscript we have tried to make it not a bit difficult to read and also not seem 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 manuscript.

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 manuscript. We have made an exhaustive revision of the manuscript 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 manuscript. 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 concerning

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 includes more shallow landslides events data.”

Just to comment, 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 the level of the small basins.

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

“The calibration/validation methodology, based on taking one year of observations for the 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 I<sub>mean</sub>-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 the observation. We have taken into account your comment. For better understanding, according to the way we have defined the variables for a dataset, I<sub>mean</sub>, which 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 I<sub>mean</sub> of both resulting events are different (I<sub>mean</sub> 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 to saturation

**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 understanding according to 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. We 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 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 fewer uncertainties that we will use to define rainfall thresholds at a 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 is a daily scale. It was edited.

L. 60: gridded

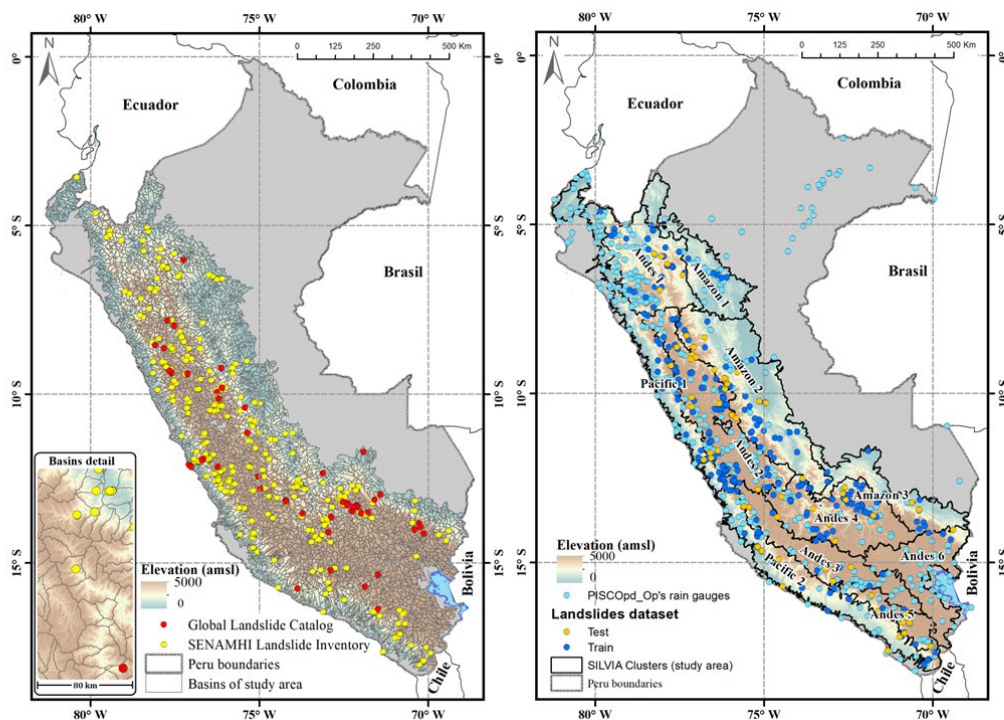
**Comment response:** Thanks for the observation. It was edited.

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-**O**perative 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 they 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 the 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 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

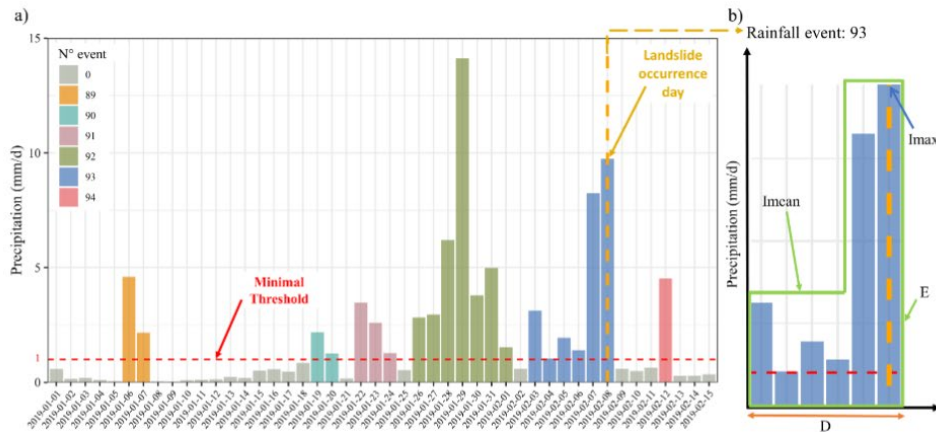
**Comment response:** Thanks for the observation. It was edited.

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 to the previous comment, the period was 2007-2020. The number of events was edited. The figure is just an extracted period to 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 is 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 is just an extracted period to 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 in the edited manuscript, moreover we put the Regionalization subsection before the Rainfall threshold model subsection because we think it helps 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 reason 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 other rainfall products that have the purpose of improving 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 into 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  $I_{mean}-D$  should result in the same thresholds, only that  $b(E-D) = b(I_{mean}-D)+1$ . That's what I get when substituting  $I_{mean}$  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 the observation. We have taken into account your comment. For better understanding, according to the way we have defined the variables for a dataset,  $I_{mean}$ , which 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  $I_{mean}$  of both resulting events are different ( $I_{mean}$  equal to 5 and 2.5 respectively), so the threshold could not be defined as the division of both. Fig. X1 shows what is mentioned for an example dataset, where it is observed that  $E$  and  $I_{mean}$  have different density distributions and therefore their predictive potentials also change (i.e., the thresholds do not have the same  $I_{mean}$  relationship  $=E/D$ ).

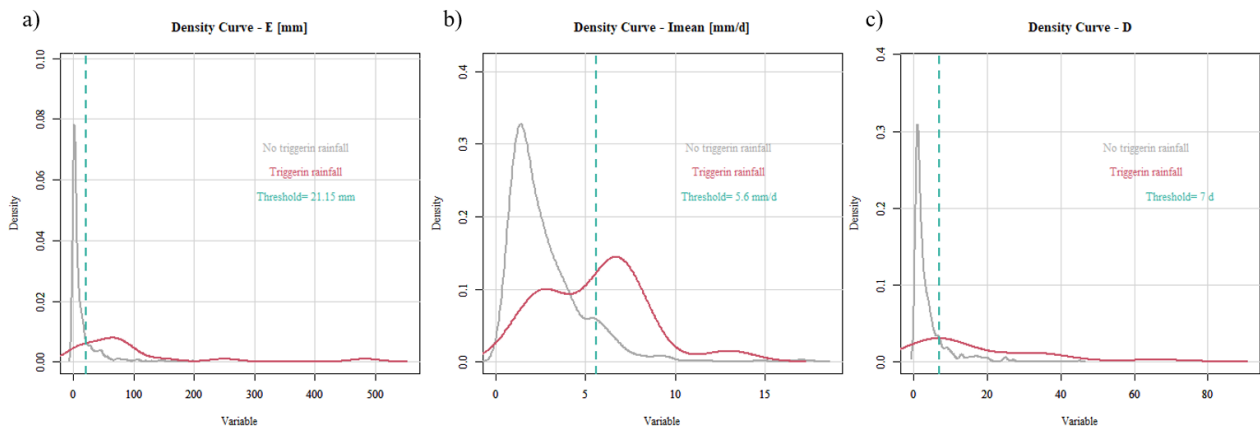


Fig. X1: Density plot of the variables  $E$  (a),  $I_{mean}$  (b), and  $D$  (c) for the same data set, where it is observed that the distributions of the variables  $E$  and  $I_{mean}$  are different.

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 separate variables, before determining a threshold, and how it changes 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 established 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/non-trigger shallow landslides. Also, the plot shows the regional 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 a daily scale (it was edited). We use this Max Daily 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 maintains an orographic similarity, since, in Peru, and generally, in South America, the Andes have a modulating character in the presence of rain. And the average precipitation derived from PISCOp 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 rewrote and reorganize the entire section for better understanding, 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 consisted of evaluating thresholds calibrated using the landslide 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 other 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 a 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 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 the 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 ( $I_{max}$ ,  $E$ ,  $D$  or  $I_{mean}$ ), the overall impression of the predictive power was estimated with 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 ( $I_{max} - D$ ,  $E - D$ ,  $I_{mean} - D$ ) the scale parameter  $a$  and the shape parameter  $b$  are simultaneously tuned to maximize 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 ( $I_{max}$ ,  $E$ ,  $D$  or  $I_{mean}$ ), the overall impression of the predictive power was estimated with 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 ( $I_{max} - D$ ,  $E - D$ ,  $I_{mean} - D$ ) the scale parameter  $a$  and the shape parameter  $b$  are simultaneously tuned to maximize 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 that occurred in 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 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 I<sub>mean</sub>-D and E-D don’t have the same performances. See comment L. 134.

**Comment response:** Thanks for the observation. It was responded to in the observation of 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 show the thresholds estimated with the calibration set. The number of events is specified in 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 a high-impact stream as a basin with a constant occurrence of landslides. But it's a local phrase, so it was removed for better understanding.

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 it. The paragraph was edited, as you can see:

"The calibration/validation methodology, based on taking one year of observations for the 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 that occurred before 2020 and validation with landslides that 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 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 rainfall thresholds.

**Comment response:** Thanks for the observation. This recommendation 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 to a body of flowing water. Regarding the processes, we included the debris flow category which is shallow in nature (Naidu et al., 2018) in shallow landslide term. A clarification of 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 other available sources. Most of them belong to the debris flow category which is shallow in nature

(Naidu et al., 2018). In this sense, this study 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 high 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 non-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 scarce data on shallow landslide events in 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 for the thresholds. The few landslides recorded made the validation performance 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 concerning 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 includes more shallow landslides events data.”



# Response to comment on nhes-2022-199 - Anonymous Referee #2

The manuscript presents an interesting application of methods for the definition of empirical rainfall thresholds for landslide occurrence at a national scale. The aim of the paper is clear and the results are also well-presented. Despite some points not very clear, I found the manuscript clear and sufficiently well-organized. From a methodological point of view, I found some problems in the work, which should be addressed before the paper can be reconsidered for publication.

I list in the following some general comments and a few specific technical corrections and other suggestions.

**Comment response:** Thank you very much for your review, we have tried to make it more clear and correct the problems in the work considering your comments, corrections, and suggestions in the new manuscript version. We are very grateful and sure that each of your comments contributed to the improvement of our work. 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). In this sense, this study is the first to be carried out on a national scale in Peru and its objective is to support the operational monitoring system of shallow landslides in Peru (<https://www.senamhi.gob.pe/?p=monitoreo-silvia>), and since our institution (SENAMHI) is responsible for maintaining this system, this work will contribute to giving it scientific validity, understanding its limitations but which will continue to be improved over time.

---

## Response to General comments

The main problem of the work lies in the validation procedure. In particular, the use of only one year of data as validation set is inconvenient. This choice was proved to be not effective cause is too much linked to the variability of the selected year. Indeed, you found that the performances decreased in the validation, "which may be due to the fact that, in the year 2020, there were no extreme rainfall events as in other years, and the number of landslides was lower than in other years". A more reliable procedure would consider a random selection of triggering and non-triggering rainfall conditions in a calibration (e.g. 80% of the total) and a validation set (remaining 20%). You can found examples in: <https://doi.org/10.1007/s11069-019-03830-x> or <https://doi.org/10.5194/hess-25-3267-2021>

**Comment response:** Thank you so much for the comments. In this regard, we have made the comparison between our calibration/validation methodology and the methodology based on random selection, which has required us more time, but we managed to do what was suggested. We attach the comparative table of both validation approaches (Table X1) and a summary of the best threshold TSS comparison (Fig. X1), in this regard, we can indicate that using the random method we did not find substantial changes, concerning our method used, it is observed that, although it is true, there is an improvement in some regions, in other regions it worsens and seeing a global average of all the regions (Table X2), we observe that the statistics do not suffer significant changes, and rather in some cases the performance of our approach has better results. These methods probably have similar results since, as we have mentioned, our sample size is not large enough to note the variations between the methods.

**Table X1: TSS comparison between validation approaches**

Procedure	Region	Region		TSS comparison between two validation approaches: 1-year selection vs. random selection								
				Imean-D			Imax-D			E-D		
				1 year	Random	$\Delta$ TSS (%)	1 year	Random	$\Delta$ TSS (%)	1 year	Random	$\Delta$ TSS (%)
Calibration	1	Pacific 1	PN	0.68	0.53	-0.14	0.71	0.54	-0.17	0.66	0.52	-0.14
	2	Pacific 2	PS	0.61	0.54	-0.07	0.51	0.50	-0.01	0.44	0.50	0.06
	3	Andes 1	AN	0.44	0.42	-0.02	0.36	0.42	0.05	0.39	0.39	0.00
	4	Andes 2	AC	0.62	0.66	0.03	0.64	0.65	0.01	0.57	0.62	0.05
	5	Andes 3	AS1	0.89	0.75	-0.14	0.91	0.72	-0.19	0.77	0.75	-0.02
	6	Andes 4	AS2	0.52	0.56	0.04	0.49	0.51	0.02	0.48	0.52	0.04
	7	Andes 5	AS3	0.66	0.62	-0.05	0.64	0.59	-0.06	0.66	0.53	-0.14
	8	Andes 6	AS4	0.62	0.60	-0.02	0.56	0.58	0.02	0.65	0.65	0.01
	9	Amazon 1	SN	0.77	0.79	0.02	0.66	0.67	0.01	0.64	0.67	0.04
	10	Amazon 2	SC	0.57	0.54	-0.03	0.57	0.58	0.00	0.58	0.57	-0.01
	11	Amazon 3	SS	0.73	0.71	-0.02	0.73	0.71	-0.02	0.68	0.66	-0.02
Validation	1	Pacific 1	PN	0.26	0.43	0.17	0.28	0.46	0.18	0.21	0.38	0.17
	2	Pacific 2	PS	0.20	0.32	0.11	0.27	0.26	0.00	0.38	0.24	-0.14
	3	Andes 1	AN	0.19	0.33	0.14	0.09	0.20	0.11	0.11	0.23	0.12
	4	Andes 2	AC	0.28	0.29	0.00	0.34	0.42	0.08	0.34	0.31	-0.03
	5	Andes 3	AS1	0.33	0.41	0.08	0.34	0.37	0.03	0.57	0.42	-0.15
	6	Andes 4	AS2	0.68	0.63	-0.05	0.70	0.55	-0.15	0.61	0.58	-0.03
	7	Andes 5	AS3	0.39	0.33	-0.05	0.38	0.29	-0.09	0.26	0.75	0.50
	8	Andes 6	AS4	0.63	0.63	0.01	0.63	0.61	-0.02	0.67	0.70	0.03
	9	Amazon 1	SN	-	0.78	-	-	0.60	-	-	-0.36	-
	10	Amazon 2	SC	0.53	0.56	0.03	0.48	0.39	-0.09	0.44	0.45	0.00
	11	Amazon 3	SS	0.68	0.75	0.06	0.70	0.77	0.08	0.66	0.72	0.06

**Table X2: TSS comparison summary between validation approaches**

Procedure	TSS comparison summary between two validation approaches: 1-year selection vs. random selection								
	Imean-D			Imax-D			E-D		
	1 year	Random	$\Delta$ TSS	1 year	Random	$\Delta$ TSS	1 year	Random	$\Delta$ TSS
Calibration	0.65	0.61	-0.04	0.62	0.59	-0.03	0.59	0.58	-0.01
Validation	0.42	0.50	0.08	0.42	0.45	0.03	0.43	0.40	-0.02



**Fig X1: Comparison between validation approaches for the best threshold (Imean-D) for calibration (above) and validation (bottom) procedure.**

In this sense, this is one of the discussions that we added taking into account your comments and observations. We take this methodology that has already been used in

other investigations (e.g., Dikshit et al., 2019; Kirschbaum et al., 2015), however, as we conclude, it did not obtain good results for few data, so we add this discussion so that be taken into account in future research in Peru. We have added your observation to the new version of the manuscript, as you can see below:

“The calibration/validation methodology, based on taking one year of observations for the 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 and the validation based on random selection.

--

The use of daily rainfall data is also not the best choice for defining rainfall thresholds, particularly for shallow landslides, given the high uncertainties related to this temporal resolution as highlighted by <https://doi.org/10.1007/s11069-018-3508-4> <https://doi.org/10.1007/s11069-019-03830-x>. This should be pointed out and discussed better. I would add that there are currently other satellite-based rainfall products with better temporal resolutions (e.g GPM), which could be employed in such analyses.

Comment response: Thanks a lot for the observation. We agree that more exact thresholds could be defined with sub-daily rainfall data, however, we chose to use these daily rainfall data for different reasons, the first is that this work is the first approximation of regional rainfall thresholds in Peru from which new and better thresholds will be generated, in addition to the fact that we take into account different investigations that developed thresholds from daily rainfall data (e.g. (Berti et al. 2012; Kirschbaum and Stanley 2018; Leonarduzzi and Molnar 2020; Leonarduzzi, Molnar, and McArdell 2017; Monsieurs et al. 2019), in addition to the fact that these thresholds have the objective of improving landslide monitoring services triggered by rainfall that already exists in Peru (<https://www.senamhi.gob.pe/?p=monitoreo-silvia>) and that our institution, SENAMHI (the hydrometeorological service of Peru), is responsible for monitoring and improving it. Finally, until 2017 we used TRMM data for our hydroclimatic services, however, for a period, the TRMM data was not maintained, and all our hydrological services that depended on this data had to stop, for this reason, at SENAMHI we choose to generate operational data (PISCO) that takes into account the assisted climatology data (e.g. PISCO monthly mean) but does not depend of an external database.

Currently, as SENAMHI we are also focused on the generation of hourly rainfall product (e.g. Huerta et al., 2017 <https://doi.org/10.1016/j.dib.2022.108570>), but that it be updated in real-time for our monitoring services, with which our next investigations regarding thresholds will take into account these hourly data.

--

The whole section 2.4 misses several information and needs a check and a huge review.

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

## 2.4 Rainfall threshold model

An empirical–statistical approach was used to define rainfall thresholds for landslide-susceptible regions, consisting of the following steps: (1) determination of rainfall events from a historical rainfall series, (2) definition of the variables of rainfall events, (3) define landslides regions from maximum daily rainfall region and GEOGloWS basins for the area studio, (4) threshold estimation for individual rainfall event variables for calibration period based on an objective maximization of predictive performance, (5) threshold estimation for combination of rainfall event variables for calibration period based on an objective maximization of predictive performance, and (6) run thresholds models and get metrics for analysis and discussions 110 (methodology is presented in Figure 2). Below are the details of the method.

The first step was the construction of a historical rainfall series from gridded rainfall data (PISCOpd\_Op) for each basin that had a minimum of one landslide event. After obtaining the rainfall series, rainfall events were defined along with a historical series for each selected basin. 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). Many authors use minimum thresholds of 1 mm to define rainy days (Dai, 2006; Dai et al., 2007; Han et al., 2016; Leonarduzzi et al., 2017; Shen et al., 2021; Tian et al., 2007; Yong et al., 2010). However, given the great climatological spatial variability in the study area, it was determined that there was not a single minimum threshold for the entire territory, but a minimum threshold was discretized from the bias of PISCOpd\_Op for non-rainy days. The PISCOpd\_Op bias was determined when rain gauges did not report rain (0 mm), and the discretized minimum threshold ( $U_{min}$ ) of rain was defined according to the following Equation 1:

$$U_{min} = \begin{cases} U_0 & \text{if } s \leq U_0 \\ s & \text{if } s > U_0 \end{cases} \quad (1)$$

where  $s$  is the average of simple bias when rainfall stations reported a value of 0 rainfall compared with the estimation in PISCOpd\_Op. And  $U_0$  is the initial minimum rainfall threshold, and it is established as 1 mm for all regions with exception of coastal Pacific regions which is considered 0.5 mm. Once rainfall events were defined, whether they were triggering or non-triggering events were established. A rainfall event is considered a rainfall trigger event if it is associated with a landslide event, i.e. if during the duration of the rainfall event a shallow landslide has occurred.

The second step was to determine analysis variables for each rainfall event, for which the maximum daily intensity  $I_{max}$  (mm/day), the accumulated rainfall  $E$  (mm), the duration  $D$  (day), and the mean daily intensity  $I_{mean} = E/D$  (mm/day) were calculated. Concerning the triggering rain events, two scenarios were considered. For the first scenario (entire event - EE), the properties of the rainfall event (Figure 3) were defined considering the rainfall rate of the landslide occurrence day. The second scenario (antecedent event - AE) defined the properties up to one day before the occurrence, i.e., it did not consider the rainfall rate of the landslide occurrence day. The reason for analyzing the second scenario was to evaluate the level of incidence that is attributed only to antecedent conditions for landslide occurrence, as this allows us to evaluate if it is possible to forecast or warn landslides based only on the antecedent conditions. The temporal evolution of hydrometeorological variables provides an idea of how the critical conditions of the activation of landslides develop (Prenner et al., 2018; Segoni et al., 2018).

The third step consisted in divide the study area into regions based on clustering techniques (this step is explained in more detail in section 2.5). Next, GEOGloWS basins were merged with regions to determine their spatial correspondence. The fourth and fifth step was to objectively select a rainfall threshold that separates triggering rainfall events

from non-triggering rainfall events with the best level of predictive performance. Rainfall thresholds were established by maximizing predictive performance in two ways: the first one only included variables independent of rainfall properties ( $I_{max}$ ,  $E$ ,  $D$ ,  $I_{mean}$ ), and the second one determined was through curve-like thresholds that related two properties ( $I_{max} - D$ ,  $E - D$ ,  $I_{mean} - D$ ) in the form of  $V = a \cdot D^{-b}$ , where  $V$  represents the variables  $I_{max}$ ,  $E$ , and  $I_{mean}$ ;  $a$  and  $b$  are the scale and shape parameters of the curve (while for logarithmic space,  $a$  is the intersection parameter and  $b$  denotes the slope of the linear curve). Finally, the sixth step consisted in apply the model to the rainfall events and compare with the observed landslides events and get the predictive performance metrics for each region at calibration and validation periods.”

It is not clear how the association between a rainfall event and a landslide is done (Line 120), in order to classify an event as a triggering rainfall event.

Comment response: Thanks for the observation. A rainfall event is considered a triggering event if during the duration of the rainfall event a shallow landslide has occurred. We edit the sentence for better understanding, as you can see below.

“A rainfall event is considered a rainfall trigger event if it is associated with a landslide event, i.e., if throughout the rainfall event duration a shallow landslide has occurred.”

Moreover, at line 118 it is reported that “For coastal Pacific regions, 0.5 mm was considered the minimum rainfall threshold”. What about the other regions?

Comment response: Thanks for the observation. The minimum rainfall threshold considered for other regions is 1 mm, and only for the coastal Pacific region is 0.5 mm. We edit the equation and sentence for better understanding, as you can see below.

“... where  $s$  is the average of simple bias when rainfall stations reported a value of 0 rainfall compared with the estimation in PISCOpd\_Op. And  $U_0$  is the initial minimum rainfall threshold, and it is established as 1 mm for all regions with exception of coastal Pacific regions which is considered 0.5 mm.”

At lines 131-136, it is not clear the actual method used to define the thresholds, based both on 1 or 2 variables. How the parameters and the equations were obtained? Before “maximizing predictive performance” a threshold should be calculated using a method. Which method was used? This issue needs to be better explained.

Comment response: Thanks for the observation. The paragraph was edited and corrected for a better explanation, as you can see below:

“... Rainfall thresholds were established by maximizing predictive performance in two ways: the first way includes every rainfall event property independently ( $I_{max}$ ,  $E$ ,  $D$  or  $I_{mean}$ ), and the second one determined was through curve-like thresholds that related two properties ( $I_{max} - D$ ,  $E - D$ ,  $I_{mean} - D$ ) in the form of  $V = a \cdot D^{-b}$ , where  $V$  represents the rainfall properties ( $I_{max}$ ,  $E$ , and  $I_{mean}$ );  $a$  and  $b$  are the scale and shape parameters of the curve (while for logarithmic space,  $a$  is the intersection parameter and  $b$  denotes the slope of the linear curve). The approximation of the first form, thresholds based on only one of the rainfall event properties ( $I_{max}$ ,  $E$ ,  $D$  or  $I_{mean}$ ), was estimated whit the minimum radial distance to the perfect classificatory test ( $TSS=1$ , with  $se=1$  and  $1-sp=0$ ) from the ROC space (e.g., Uwihirwe et al., 2020; Gariano et al., 2015) and the approximation of the second form, curve-like thresholds, was established with the optimization of  $a$  and  $b$  parameters of the curve model ( $V = a \cdot D^{-b}$ ) with an initial approximation of the curve based on  $a$ =average of the variable  $V$

of the triggering rainfall events and  $b=0$ . ....”

Moreover, at line 133 is written “variables independent of rainfall properties ( $I_{max}, E, D, I_{mean}$ )”; actually,  $I_{mean}$  and  $D$  are not independent on each other, being  $I_{mean}=E/D$ . Please explain also this point.

Comment response: Thanks for the observation. We refer as variable independent to only one of the rainfall event properties ( $I_{max}, E, D, I_{mean}$ ). The sentence was corrected for better understanding as you can see below:

“... the first way includes every rainfall event property independently ( $I_{max}, E, D$  or  $I_{mean}$ ), ...”

Finally, I believe that proposing thresholds based only on one parameter (e.g.  $E, D, I_{mean}$ , or  $I_{max}$ ) is now anachronistic, given the huge literature on rainfall thresholds based on two variables.

Comment response: Thanks for the observation. We agree that there is a large amount of literature on thresholds based on two variables, although there is also literature that evaluates one-parameter variables and/or how they impact when combined with other variables (Hirschberg et al. 2021; Leonarduzzi et al. 2017; Uwihirwe, Hrachowitz, and Bogaard 2020), in this sense, our approach, being a novel work in Peru, is to provide variables that could be beneficial for certain regions and in future research combine or improve them in greater detail. Additionally, this paper 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.

Regarding the thresholds based on two variables, actually, there is no need to calculate both  $E-D$  and  $I_{mean}-D$  thresholds, given that they are analytically equivalent, being  $I_{mean}=E/D$ . I can't figure out how different results are obtained for the two types of thresholds ( $I-D$  and  $E-D$ ); they should have the same performances).

Comment response: Thanks for the observation. We have taken into account your comment. According to the way we have defined the variables for a dataset,  $I_{mean}$ , which 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  $I_{mean}$  of both resulting events are different ( $I_{mean}$  equal to 5 and 2.5 respectively), so the threshold could not be defined as the division of both. The next Fig. X2 shows what is mentioned for an example dataset, where it is observed that  $E$  and  $I_{mean}$  have different density distributions and therefore their predictive potentials also change (i.e., the thresholds do not have the same  $I_{mean}$  relationship  $=E/D$ ).



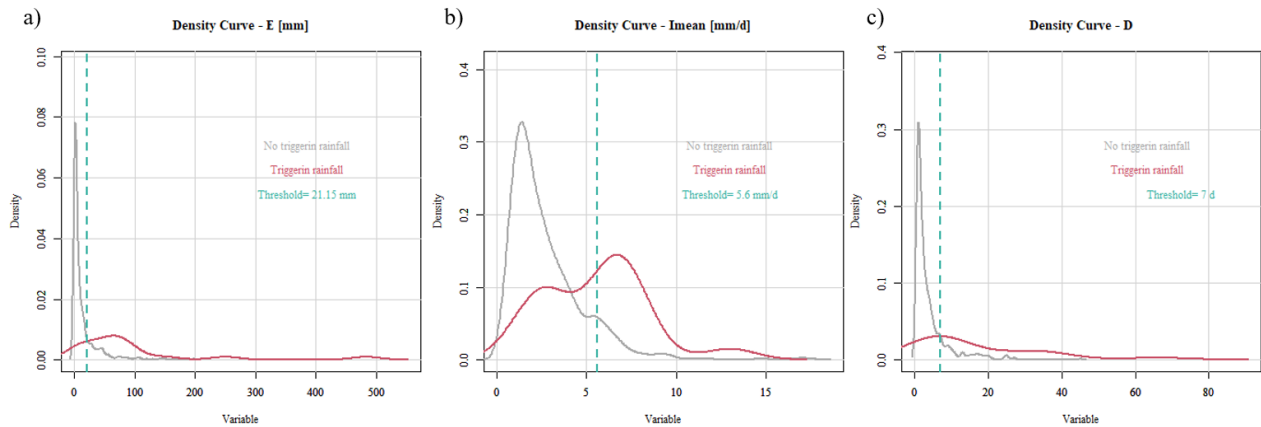


Fig. X2: Density plot of the variables E (a), Imean (b), and D (c) for the same data set, where it is observed that the distributions of density of the variables E and Imean are different.

--

Line 170: actually, a threshold is represented by a point in the ROC space (the point is the TRP, FPR couple), so I believe that the area under the curve is only a quadrangle. Please explain better this point. Being the thresholds represented only by one point in the ROC space, I would suggest using the distance of this point from the perfect classificatory point (upper left corner of the space, TPR=1, FPR=0) instead of the area under curve. You can find more details in <https://doi.org/10.1016/j.geomorph.2014.10.019>

Comment response: Thanks a lot for the suggestion and clarification of the topics. We use this method to estimate the thresholds of one variable, and we edit and correct the paragraph and explanation of the calibration methods for better understanding, 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., 2020; Postance et al., 2018; Gariano et al., 2015) ...”

Lines 179-182: actually, more simple and useful definitions are:  $TPR = TP/(TP + FN)$ ;  $FPR = FP/(FP + TN)$ . I would suggest using these definitions instead of mentioning sensitivity and specificity.

Comment response: Thanks for the suggestion. We use these definitions as we review that they were also used in many other current publications, but we have edited and added these citations for better understanding, as you can see:

“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 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).”

Passing to Section 3, regarding the regionalization, it is not clear how many empirical points are employed for calculating the thresholds in each of the 11 regions. Please add this information and discuss possible limitations in case of thresholds based on too few points.

Comment response: Thanks for the observation. We add this information, and add a discussion on the new version of the manuscript, as you can see below:

“Hirschberg et al. (2021) found that 25 events are enough to limit the uncertainties in the ID threshold parameters to  $\pm 30\%$  in his study, based on this, it is observed that there are several regions (Andes 3, 5, 6 and Amazon 1, Amazon 3 and Pacific 2) that do not exceed that quantity, so these regions have a greater source of uncertainty due to the quantity of the data. A summary of the number of shallow landslide events used for the research and the thresholds with best performances per region is presented in Table 3.”

**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

Figure 6. Please note that the thresholds should have duration ranges based on the minimum and maximum durations of the triggering events. Theoretically, you can't draw a threshold in a duration value when you don't have a triggering event. This allow also avoiding having very low values of thresholds at long durations (see thresholds for Andes 4, 5, 6). Moreover, I would suggest correcting all the equations replacing Y and X with I<sub>mean</sub> and D, and replacing the “^” with a proper superscript.

Comment response: Thanks for the observation. The figure was edited taking into account your suggestions on the new version of the manuscript as you can see below:

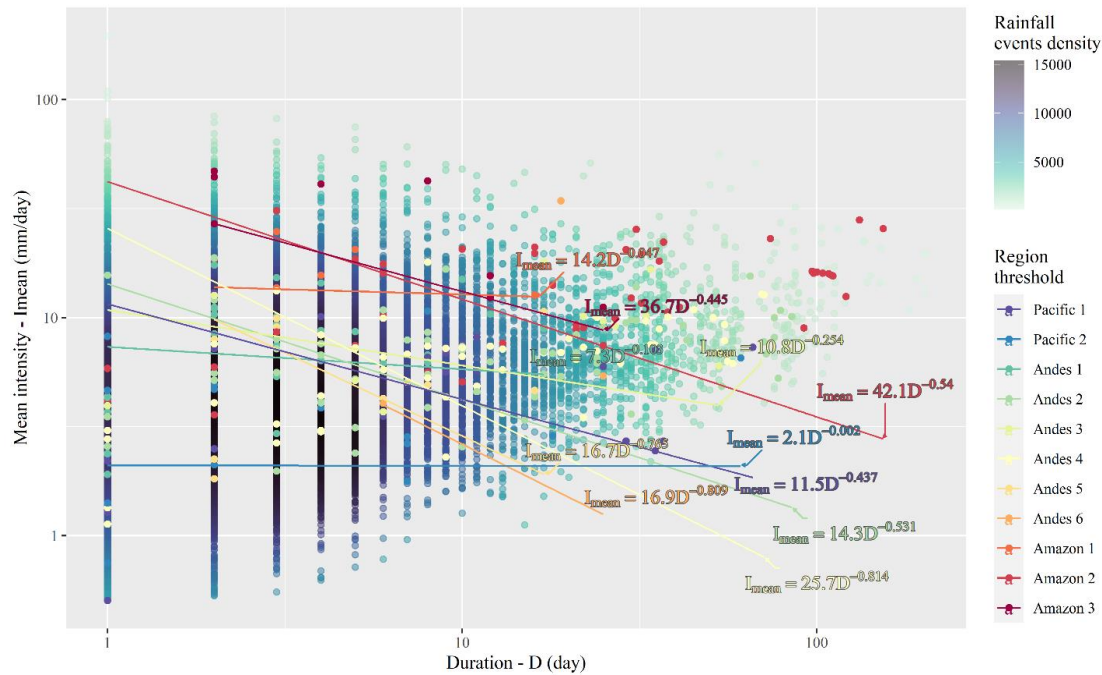


Figure 7. Is there some physical explanation for the variation of the values of the 1-variable thresholds? In some cases, I see differences that seem not related to morphology or other environmental factors.

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 manuscript, 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 concerning 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 includes more shallow landslides events data.”

---

### Response to technical corrections and suggestions

Abstract: I would use the present tense in the abstract

Comment response: Thanks for the suggestion. It was edited on the new version of the manuscript, as you can see:

“Abstract. The objective of this work is to generate and evaluate regional rainfall thresholds obtain from a combination of high-resolution gridded precipitation data (PISCOpd\_Op) developed by the National Service of Meteorology and Hydrology of Peru (SENAMHI), and information from observed shallow landslide events. The landslide data were associated with rainfall data, determining triggering and non-triggering rainfall events with rainfall properties from which rainfall thresholds are determined. The validation of the performance of the thresholds is carried out with events that occurred

during 2020 and focuses on evaluating the operability of these thresholds in landslide warning systems in Peru. The thresholds are determined for 11 rainfall regions. The method of determining the thresholds is based on an empirical–statistical approach, and the predictive performance of the thresholds is evaluated with the “true skill statistics” (TSS). The best predictive performance is the mean daily intensity-duration (I<sub>mean</sub> – D) threshold curve, followed by accumulated rainfall E. This work is the first attempt to estimate regional thresholds on a country scale to better understand landslides in Peru, and the results obtained reveal the potential of using thresholds in the monitoring and forecasting of shallow landslides caused by intense rainfall and in supporting the actions of disaster risk management.”

I would use rainfall instead of precipitation everywhere in the text.

Comment response: Thanks for the suggestion. We use rainfall instead of precipitation in the new version of the manuscript.

Line 24: “Terrain saturation is the original cause of landslide occurrence”. Actually, this depends on the type of landslides.

Comment response: Thanks for the observation. This sentence was removed for better understanding according to the comment.

Line 33: perhaps the correct reference is Segoni et al 2018 (already mentioned), not Segoni et al 2014

Comment response: Thanks for the observation. The correct reference is Segoni et al., 2018, and it was corrected in the new version of the manuscript.

Line 36-39: The sentence “For example, global thresholds have been developed based on antecedent precipitation indices (Caine, 1980; Guzzetti et al., 2008; Kirschbaum and Stanley, 2018), and national thresholds have been established under an empirical–statistical approach (Leonarduzzi et al., 2017; Peruccacci et al., 2017a; Uwihirwe et al., 2020).” is not correct. Actually, both the mentioned thresholds based on antecedent precipitation and the cited national thresholds are established using an empirical–statistical approach. Please review and correct.

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

“For example, there has been developed an 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).”

Line 38: Note that there are two references to the work “Peruccacci et al. (2017)” a and b, which are actually the same.

Comment response: Thanks for the observation. The reference has been corrected.

Line 47: I would suggest using “slope” instead of “hillside”

Comment response: Thanks for the observation. The text has been changed, as you can see below.

"Thresholds can be set for different spatial scales depending on the extent of the analysis, and these can be categorized into six classes: global, national, regional, basin, local, and slope scales. ..."

Line 51: in relation to environmental subdivisions within a national territory, please consider also the work of Peruccacci et al. (2017) – already mentioned – which present several morphological, geological, meteorological, climatic subdivision of the Italian territory with the aim of defining rainfall thresholds.

Comment response: Thanks for the suggestion. The reference was added, as you can see below.

"..., as well as an environmental subdivision within a national territory based on erodibility and climatology represented by the maximum daily intensity of a rainfall event (Leonarduzzi et al., 2017) or on topography, lithology, land-use, land cover, climate, and meteorology (Peruccacci et al., 2017)."

Caption of figure 2. Delete "Methodology six steps"

Comment response: Thanks for the observation. The caption has been corrected.

Line 101: I suppose you wanted to write "is shown in Figure 2".

Comment response: Thanks for the observation. It was edited in the new manuscript.

Line 185-186: please check syntax and grammar.

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

"For thresholds based on rainfall event properties independently ( $I_{max}$ ,  $E$ ,  $D$ ,  $I_{mean}$ ), the overall impression of the predictive power was estimated with 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.) while for the threshold curve ( $I_{max}-D$ ,  $E-D$ ,  $I_{mean}-D$ ) the scale parameter  $a$  and the shape parameter  $b$  are simultaneously tuned to maximize the true skill statistics (TSS) (e.g., Leonarduzzi et al.; Hirschberg et al.). This maximization was carried out automatically 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."

Line 196: actually, TSS varies between -1 and 1, as you correctly mentioned some lines above.

Comment response: Thanks a lot for the observation. It was edited in the manuscript.

Table 3: I would suggest using always the same number of decimal places.

Comment response: Thanks a lot for the suggestion. It was edited in the manuscript using 2 decimal places.

# Response to comment on nheSS-2022-199 - Anonymous Referee #3

## Response to general comments

The manuscript deals with the development and evaluation of regional landslide precipitation thresholds in Peru. The Authors used the available high-resolution gridded precipitation and landslide events data to define empirical thresholds which is an important step towards the development of landslide early warning system in Peru (a country with limited landslide studies).

The study seems very important especially in a country with limited landslide studies yet with frequent landslide hazards problems. However, some sections of the manuscripts need to be polished for a better flow of the manuscript. Some points also need to be corrected:

Comment response: Thank you very much for your review and general comments, we have tried to make it not a bit difficult to read and also not seem unorganized, considering all your comments in the new version of the manuscript. 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 also 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).

## Specific comments

**Section 2** This section presents the methodology used. Figure 1 summarises the methodology in 6 steps which is really good. However, from sub\_sect. 2.1 to 2.6 one would expect the details from step 1 to step 6. These steps are not outlined clearly in these sections and may break the flow of the manuscript not only in Methodology section but also the Results section.

Comment response: Thanks for the observation. This observation was taken into account in the new manuscript, we reorder and organized the methodology on subsection 2.4 as you can see below:

### "2.4 Rainfall threshold model

An empirical–statistical approach was used to define rainfall thresholds for landslide-susceptible regions, consisting of the following steps: (1) determination of rainfall events from a historical rainfall series, (2) definition of the variables of rainfall events, (3) define landslides regions from maximum daily rainfall region and GEOGloWS basins for the area studio, (4) threshold estimation for individual rainfall event variables for calibration period based on an objective maximization of predictive performance, (5) threshold estimation for combination of rainfall event variables for calibration period based on an objective maximization of predictive performance, and (6) run thresholds models and get metrics for analysis and discussions 110 (methodology is presented in Figure 2). Below are the details of the method.

The first step was the construction of a historical rainfall series from gridded rainfall data (PISCOpd\_Op) for each basin that had a minimum of one landslide event. After obtaining the rainfall series, rainfall events were defined along with a historical series for each selected basin. 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).



Many authors use minimum thresholds of 1 mm to define rainy days (Dai, 2006; Dai et al., 2007; Han et al., 2016; Leonarduzzi et al., 2017; Shen et al., 2021; Tian et al., 2007; Yong et al., 2010). However, given the great climatological spatial variability in the study area, it was determined that there was not a single minimum threshold for the entire territory, but a minimum threshold was discretized from the bias of PISCOpd\_Op for non-rainy days. The PISCOpd\_Op bias was determined when rain gauges did not report rain (0 mm), and the discretized minimum threshold ( $U_{min}$ ) of rain was defined according to the following Equation 1:

$$U_{min} = \begin{cases} U_0 & \text{if } s \leq U_0 \\ s & \text{if } s > U_0 \end{cases} \quad (1)$$

where  $s$  is the average of simple bias when rainfall stations reported a value of 0 rainfall compared with the estimation in PISCOpd\_Op. And  $U_0$  is the initial minimum rainfall threshold, and it is established as 1 mm for all regions with exception of coastal Pacific regions which is considered 0.5 mm. Once rainfall events were defined, whether they were triggering or non-triggering events were established. A rainfall event is considered a rainfall trigger event if it is associated with a landslide event, i.e., if during the duration of the rainfall event a shallow landslide has occurred.

The second step was to determine analysis variables for each rainfall event, for which the maximum daily intensity  $I_{max}$  (mm/day), the accumulated rainfall  $E$  (mm), the duration  $D$  (day), and the mean daily intensity  $I_{mean} = E/D$  (mm/day) were calculated. Concerning the triggering rain events, two scenarios were considered. For the first scenario (entire event - EE), the properties of the rainfall event (Figure 3) were defined considering the rainfall rate of the landslide occurrence day. The second scenario (antecedent event - AE) defined the properties up to one day before the occurrence, i.e., it did not consider the rainfall rate of the landslide occurrence day. The reason for analyzing the second scenario was to evaluate the level of incidence that is attributed only to antecedent conditions for landslide occurrence, as this allows us to evaluate if it is possible to forecast or warn landslides based only on the antecedent conditions. The temporal evolution of hydrometeorological variables provides an idea of how the critical conditions of the activation of landslides develop (Prenner et al., 2018; Segoni et al., 2018).

The third step consisted in divide the study area into regions based on clustering techniques (this step is explained in more detail in section 2.5). Next, GEOGloWS basins were merged with regions to determine their spatial correspondence. The fourth and fifth step was to objectively select a rainfall threshold that separates triggering rainfall events from Non-triggering rainfall events with the best level of predictive performance. Rainfall thresholds were established by maximizing predictive performance in two ways: the first one only included variables independent of rainfall properties ( $I_{max}, E, D, I_{mean}$ ), and the second one determined was through curve-like thresholds that related two properties ( $I_{max} - D, E - D, I_{mean} - D$ ) in the form of  $V = a \cdot D - b$ , where  $V$  represents the variables  $I_{max}, E,$  and  $I_{mean}$ ;  $a$  and  $b$  are the scale and shape parameters of the curve (while for logarithmic space,  $a$  is the intersection parameter and  $b$  denotes the slope of the linear curve). Finally, the sixth step consisted in apply the model to the rainfall events and compare with the observed landslides events and get the predictive performance metrics for each region at calibration and validation periods."

### Minor comments/technical corrections

Figure 2 caption. "Methodology six steps" is not relevant for the Figure. I would suggest to correct the Caption as: "Study area. Left: Spatial distribution of the Global Landslide Catalog (red) and SENAMHI landslide inventory (yellow). Right: Eleven landslide-susceptibility regions for Peru and distribution of calibration (blue) and validation (yellow) landslides" .

Comment response: Thanks for the suggestion. It was added to the new version of the manuscript, as you can see:

“Study area. Left: Spatial distribution of the Global Landslide Catalog (red) and SENAMHI landslide inventory (yellow). Right: Eleven landslide-susceptibility regions for Peru and distribution of calibration (blue) and validation (yellow) landslides.”

LL101. .... Is shown in 3. There is something missing. Is it Figure 2? Or sect. 3?

Comment response: Thanks for the observation. It is Map Figure (Fig. 2), and it was edited in the new manuscript.

LL126-127. “If it is possible to forecast or warn of possible landslides”. To be corrected as “If it is possible to forecast or warn landslides”

Comment response: Thanks for the observation. It was corrected on the new version of the manuscript, as you can see:

“The reason for analyzing the second scenario was to evaluate the level of incidence that is attributed only to antecedent conditions for landslide occurrence, as this allows us to evaluate if it is possible to forecast or warn landslides based only on the antecedent conditions.”

LL 131. “triggering rain evens” to be corrected as “triggering rain events”

Comment response: Thanks for the observation. It was corrected on the new version of the manuscript, as you can see:

“... objectively select a rainfall threshold that separates triggering rainfall events from non-triggering rainfall events with the best level of predictive performance.”

Figure 7 caption is a little bit messy. May be this: The first column shows the spatial distribution of Rainfall thresholds for independent variables magnitude for Peru: (a) D (days), (b) total cumulative rainfall E (mm), (c) mean daily intensity  $I_{\text{mean}}$  (mm/day) and (d) maximum daily intensity  $I_{\text{max}}$  (mm/day). The second and third columns show the bivariate maps indicating the spatial distribution of the sensibility (probability of correctly predicting landslide triggering rainfall events) and specificity (probability of correctly predicting non-triggering rain events from landslide) of the thresholds for calibration and Validation.

Comment response: Thanks a lot for the observation and recommendation. It was corrected on the new version of the manuscript, as you can see:

“Figure 7. The first column shows the spatial distribution of Rainfall thresholds for independent variables magnitude for Peru: (a) day D (days), (b) total cumulative rainfall E (mm), (c) mean daily intensity  $I_{\text{mean}}$  (mm/day) and (d) maximum daily intensity  $I_{\text{max}}$  (mm/day). The second and third columns show the bivariate maps indicating the spatial distribution of the sensitivity (probability of correctly predicting landslide triggering rainfall events) and specificity (probability of correctly predicting non-triggering rainfall events from landslide) of the thresholds for calibration and validation.”