Response to Report #1 - Anonymous Referee #1

GENERAL COMMENTS

The quality and readability of the manuscript has significantly improved compared to the first version. It's very clear that the authors made an effort in revising their work based on the reviewer's comments. There is a better structure and most ambiguities have been sorted out. I still have one relevant general comment and some specific comments below which I think should be considered before publication in NHESS. Additionally, I suggest that the manuscript is edited by a native English speaker or maybe just by using Grammarly, because the grammar and vocabulary have room for improvement. I made some corrections but it's probably not comprehensive.

Comment response: Thank you very much for your review, we learned a lot from your comments and the manuscript was significative improved. We edit and use Grammarly in the new version of the mn.

Regarding the ID vs ED thresholds, I am not convinced by your argument on the differences in distribution of variables E and I. Sure, you're right, but this only explains why I and E have different performances on their own, not why ED and ID should be different. When you do the math, you can rewrite I=aD^b to E=aD^(b+1). When you transform from ID to ED plot, and look at rainfall events with the same duration (parallel to y-axis), these points will just scale with duration, but the information or order of the data does not change by scaling with a dependent variable (D). This is maybe not the perfect explanation (maybe another reviewer can do better) but I made a quick test with both ED and ID and the result of the threshold is exactly of the relation above and the performance stayed the same. If you're still convinced you're right, maybe do an example of ID and ED thresholds with a strongly reduced dataset to show that the result is different (or something similar). If I am right, the results should be checked carefully.

Comment response: Thanks for the observation. We agree with you and your explanation was perfect for us. We did a rapid example (Fig. X1) and we understand that I=aD^b are equivalent to E=aD^(b+1), but this occurred when the shape parameter is obtained by linear regression and the scale parameter is selected by other methods (e.g., frequentist method). In our study, as we mentioned, we optimized both parameters a and b ("*this optimization was automatically calibrated using the shuffled complex evolutionary algorithm (SCEA-UA) (Duan et al., 1993), considering the TSS as the objective function.*"), so it´s not necessary to coincide with the regression, we assume that there is a threshold that not adjust necessary with the linear regression.

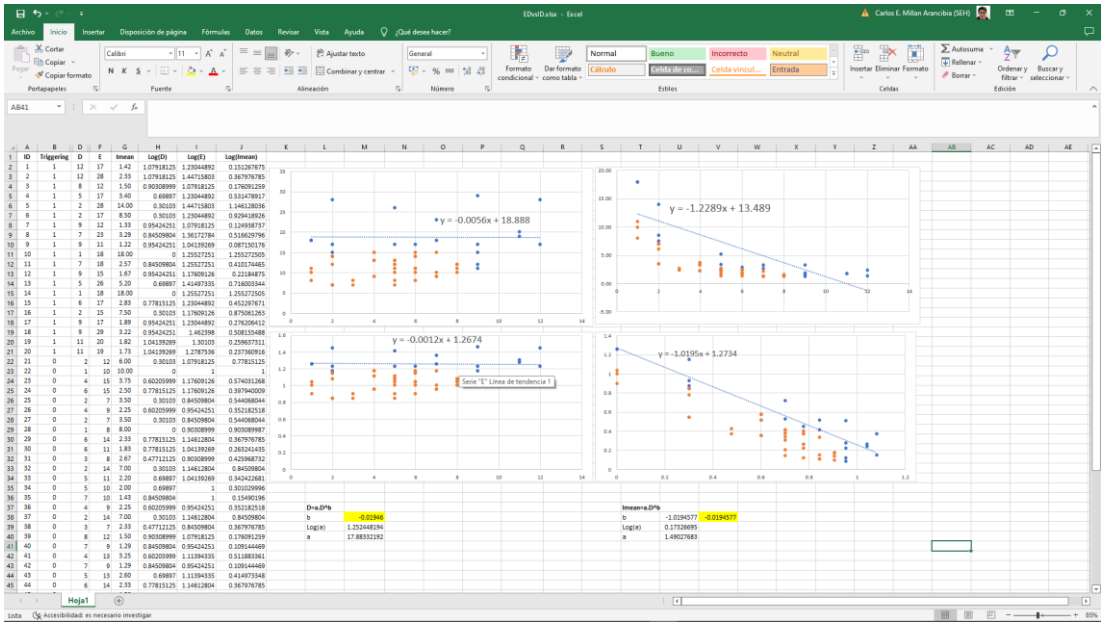


Fig. X1. Example of I=aD^b and E=aD^(b+1) where the parameter b its yellow highlight.

SPECIFIC COMMENTS

L1: I would avoid using abbreviations like PISCOpd_Op in the abstract
Comment response: Thanks for the observation. We avoid using abbreviations in the new version of the mn.

L1: obtained
Comment response: Thanks for the observation. It was edited in the new version of the mn.

L9: followed
Comment response: Thanks for the observation. It was edited in the new version of the mn.

L25: "flow in channels, streams and rivers" I would just say "stream flow"
Comment response: Thanks for the observation. It was edited in the new version of the mn.

L36: change to "…, empirical-statistical approaches for the estimation of global (citations) and national (citations) thresholds been developed."
Comment response: Thanks for the observation. It was edited in the new version of the mn.

L38: to forecast or for forecasting
Comment response: Thanks for the observation. It was edited in the new version of the mn.

L42: "Empirical approaches are widely applied …"
Comment response: Thanks for the observation. It was edited in the new version of the mn.

L48: This is kind of repetitive because you say that in L35 already
Comment response: Thanks for the observation. It was deleted in the new version of the mn.

L58-59: is this decision based on the findings of other studies?
Comment response: Thanks for the observation. Yes, based on part of the methodology of Leonarduzzi et al., 2017, and a previous work developed on SENAMHI (Yupanqui et al., 2017).

L60: is it really for the purpose of monitoring? Not to test the feasibility of a potential early-warning system?
Comment response: Thanks for the recommendation. It was edited in the new version of the mn, as you can see below.

"The main objective of this work is to estimate rainfall thresholds to test the feasibility of a potential early-warning system of shallow landslides generated by rainfall from a gridded rainfall database and shallow landslide inventory."

L61-62: I think it's rather "…implementing an objective methodology for empirical rainfall-based landslide early warning at a national scale" or something like that
Comment response: Thanks for the recommendation. It was edited in the new version of the mn, as you can see below.

"Additionally, this work focuses on implementing an objective methodology for empirical rainfall-based landslide early warning at a regional scale combining a gridded rainfall database and shallow landslide inventory."

L75-80: I don't understand what the susceptibility map and the basin map were used for. Before you said you said that regions are discretized by max rainfall (L59).
Comment response: Thanks for the observation. These lines are just the explanation of the delimitation of the study area (used by SENAMHI for the landslide monitoring service). This area was discretized by max rainfall.

L85: is reference ET the same as potential ET?

Comment response: Thanks for the observation. The **reference evapotranspiration** (ETo) is the evapotranspiration rate of a reference surface (a hypothetical grass reference crop with specific characteristics) which occurs without water restrictions (Allen, R. G. et al. 1998). For more information about the differences between reference evapotranspiration (ETo) and potential evapotranspiration (ETp) you can see Xiang et al. 2020 (Similarity and difference of potential evapotranspiration and reference crop evapotranspiration – a review - https://doi.org/10.1016/j.agwat.2020.106043).

L86: how much is 0.1° approx. in m or km?
Comment response: Thanks for the observation. 0.1° is approx. ~10 km.

L100: there's a repetition here with regard to the certainty in recent years
Comment response: Thanks a lot for the observation. It was deleted in the new version of the mn.

L103: what do you mean by "geospatial analysis" and based on what was the one event excluded?
Comment response: Thanks a lot for the observation. It refers to the use of geospatial tools: spatial sub-setting. It was changed for better understanding.

Figure2: in sttep1 "don't trigger" instead of "no trigger"

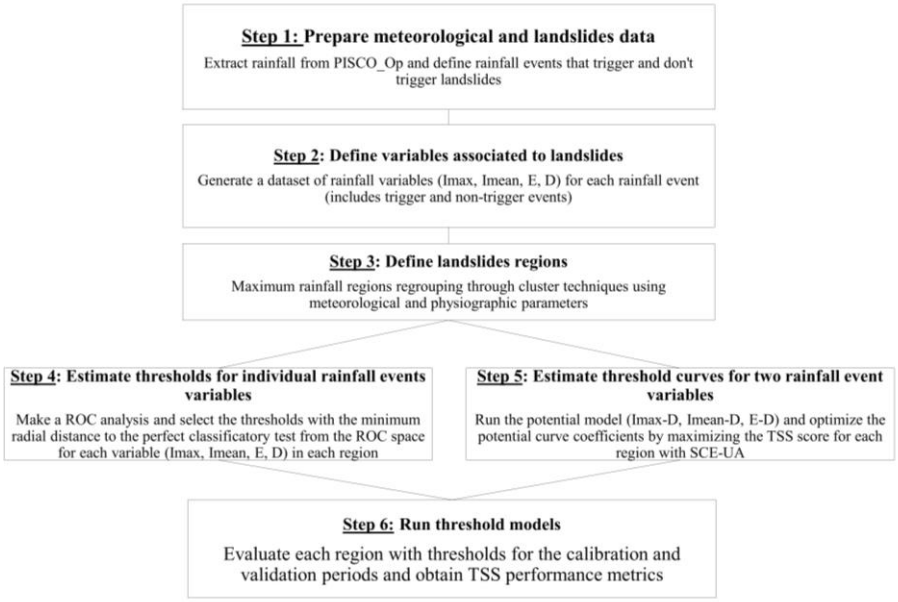
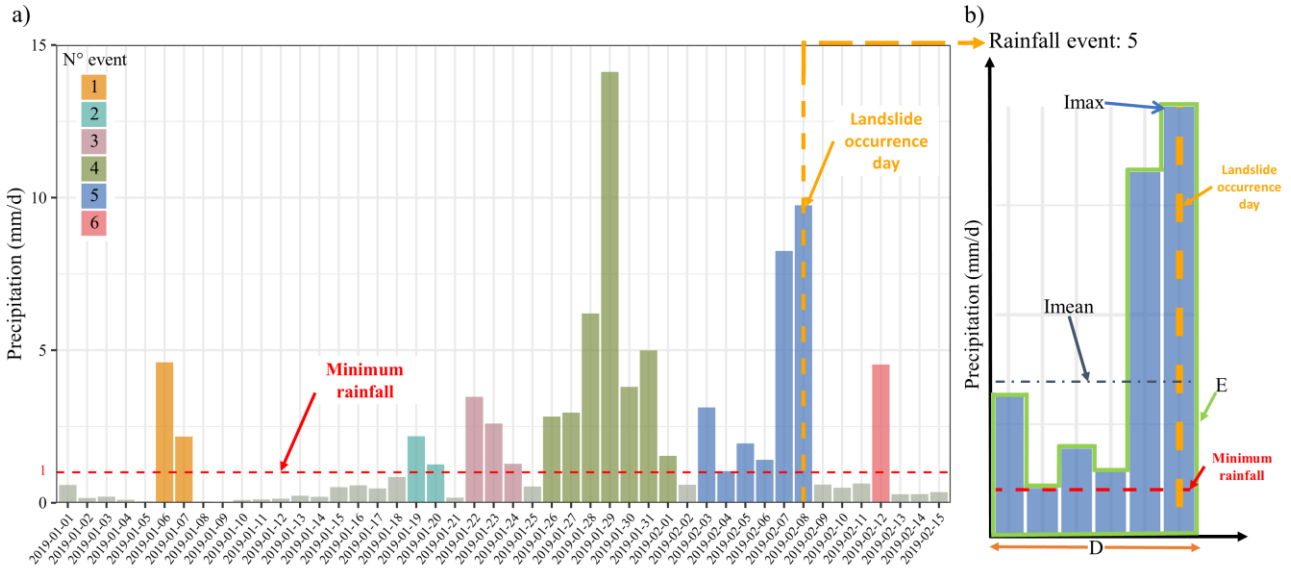


Figure3b: Please draw a clear line for Imean and E. Is Imean the green or the red dashed line? It's not clear to me what the green line is around the bars. E should be much higher as it's the sum of the six bars, shouldn't it?
Comment response: Thanks a lot for the observation. Yes, Imean is the green line that involves all the event representing the sum of the six bars, and for better understanding, we add a new blue line for Imean. The new version of the figure is shown below.



L129: "…, they were classified into triggering and non-trigger events, i.e. if a landslide occurred during the rainfall event."
Comment response: Thanks for the observation. It was edited in the new version of the mn.

L134-137: Maybe I'm misunderstanding something. Maybe you could draw these two scenarios in Figure 3 so it becomes clear. One scenario is just the daily rainfall (independent from the events you defined) and the other is the defined event minus the day of landslide occurrence, right? Why is not one scenario the defined event including the day of landslide triggering?
Comment response: Thanks for the observation. It was edited in the new version of the mn for better understanding.

"The first scenario (entire event - EE) considers all the rainy days of the rainfall event including the rainfall of the landslide occurrence day to determine the properties of the rainfall event (Figure 3). The second scenario (antecedent event - AE) considers only the antecedent rainy days of landslide occurrence to determine the properties of the rainfall event, i.e., AE does not consider the rainfall of the landslide occurrence day."

L141: dividing
Comment response: Thanks for the observation. It was edited in the new version of the mn.

L150: Why do you use different scores for uni- and multivariate predictors?
Comment response: Thanks for the observation. It's because we use different methodologies to define the thresholds, in the case of univariate predictors, the thresholds it was selected directly from the ROC space, and for multivariate predictors, we use an automatic optimization, which needs an optimization function based on TSS.

L154: applying
Comment response: Thanks for the observation. It was edited in the new version of the mn.

L188: change to "…a confusion matrix was used…"
Comment response: Thanks for the observation. It was edited in the new version of the mn.

L210: the year is missing in the citation
Comment response: Thanks for the observation. It was edited in the new version of the mn.

L287-291: This last paragraph I think should go into the discussion section. The results overall read very well!
Comment response: Thanks a lot for the recommendation. It was edited in the new version of

the mn.

L294: Technically, because your thresholds are based on TSS and radial distance, the thresholds are not only based on rainfall events associated with landslides, but also with non-triggering rainfall events
Comment response: Thanks a lot for the recommendation. It was edited in the new version of the mn.

L299: 2x shown
Comment response: Thanks a lot for the recommendation. It was edited in the new version of the mn.
"The estimated thresholds are shown in Table 1 for independent variables and Table 2 for curve thresholds."

L303: is the background rain the same as antecedent rain? I think you used this term earlier. Please be consistent.
Comment response: Thanks a lot for the recommendation. It was edited in the new version of the mn.
"However, it allows us to associate landslide events with the antecedent rain conditions of the last 8 days, an association that can be used for future research."

L322: …in their study. Based on…
Comment response: Thanks a lot for the recommendation. It was edited in the new version of the mn.

L334-339: I think it's good to have this listing but some points need to be more specific to avoid misunderstandings. Not necessarily more explanation, but more precise. ii) isn't this the same as i) because a 10 km cell may cover several streams, or what exactly do you mean? Iii) you mean because your landslide record may not be complete? Iv) because thresholds based on short records may still be uncertain? V) is unclear to me what is meant.
Comment response: Thanks a lot for the recommendation. It was edited in the new version of the mn.

"There are still many limitations to rainfall threshold study at the regional scale in Peru. Mainly the landslide short records are not enough to limit uncertainty in the threshold definition (Peruccacci et al., 2012; Hirschberg et al., 2021). Another important source of uncertainty was the use of coarse temporal rainfall data resolution that caused a systematic underestimation of the thresholds (Gariano et al., 2020; Marra, 2019). Another is the spatial rainfall data resolution because a 10 km cell may cover several streams. And finally, the regionalization can be not enough representative of the high variability of descriptor landslide variables. These limitations must be taken into account in future research."

L344: "relationship" instead of "interrelation"
Comment response: Thanks a lot for the recommendation. It was edited in the new version of the mn.

L346-349: very long sentence, consider splitting it in two.
Comment response: Thanks a lot for the recommendation. It was edited in all parts of the new mn.
"Daily gridded rainfall data and landslide data were used to estimate landslide-triggering and non-triggering rainfall events. With this data was possible to estimate and validate rainfall thresholds for the activation of shallow landslides triggered by rainfall."

L353: I would change "accumulated daily intensity" to "daily rainfall"
Comment response: Thanks a lot for the recommendation. It was edited in the new version of the mn.

L354: maybe "variability" instead of "differentiation"? and then "This differences in performance are…"
Comment response: Thanks a lot for the recommendation. It was edited in the new version of the mn.

L354-361: These lines are hard to read and understand. Shouldn't you regionalization improve the situation? Or is the regionalization not that good so within each region there is still high variability? What could you do about it in the future? What do you mean by greater incidence of lithology and geology, do these affect the thresholds or is it a question of sediment supply? What is SL (I would avoid abbreviation in conclusions)?
Comment response: Thanks a lot for the observation. Regionalization improves the climate context of Peru, but it is not conclusive, we cannot divide into more regions for the landslides little data. In some regions with lower performances, we think that other geological components influence the occurrence of shallow landslides, and future studies can explore new regionalization based on lithology (e.g., Peruccacci et al. 2011).

The conclusion was edited for better understanding in the new version of the mn, as you can see below.

"The performances of the calibrated thresholds had a high variability between regions. These differences in performance are associated with the high variability of rainfall events in each region, where best performances occur in areas where it is easier to separate rainfall events that trigger and non-trigger shallow landslides (e.g., 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 a lot of rainfall events with high magnitudes that do not trigger landslides, reflecting in lower performances (e.g., Andes 1, Andes 4 and Amazon 2). Thus, the regionalization shows that exists regions where the climate component had more predominance in the shallow landslide occurrence in comparison with other regions where lithology could have more influence in the occurrence of shallow landslides than just the rains. Future studies can explore regionalization based on lithology."

L362: you could add that despite these uncertainties, the framework you've set up allows for systematically updating the thresholds as the records grow.
Comment response: Thanks a lot for the recommendation. It was edited in the new version of the mn.
"Despite these uncertainties, the framework set up of this work allows to systematically update the thresholds as the records grow."

L372: Thumb up for making the code available!
Comment response: Thanks a lot for the comment.

Response to Report #2 - Anonymous Referee #2

Dear Authors,

my sincere apologies for my very late reply!

I've read your replies to my comments and the revised version of the manuscript. I found a relevant improvement from the original version, despite some limitations remains, as acknowledged by you.

Comment response: Thank you very much for your review, we learned a lot from your comments and the manuscript was significative improved. We hope that the answers in this document satisfy your observations.

The most crucial point is the number of empirical points used to calculate the thresholds in each pre-defined region. This number is depicted only now after my comment. Overall, looking at table 3, the number of points is low (except for Andes 2 and Andes 4, and to a minor extent Pacific 1 and Amazon 2). In some cases, the number of empirical points is not acceptable. The mentioned work (Hirschberg et al. 2021) refers to a little basin affected by debris flows, covering a few square kilometers. Therefore, I think it can't be taken as a reference, being the regions defined in this work very wide. Other works, on a regional scale (e.g. https://doi.org/10.1016/j.geomorph.2011.10.005) suggested higher values for obtaining reliable thresholds with acceptable uncertainties. Moreover, an uncertainty of 30% is too high in my opinion for an operational tool like the one proposed in the work. I suggest merging the regions into only three macro-regions (i.e. Pacific, Andes, and Amazon) in order to re-calculate the thresholds and obtain more reliable results. Alternatively, another possible merging might be: Pacific (former Pacific 1 and Pacific 2; 73 landslides); Andes 01 (former Andes 1 and Andes 2; 132 landslides); Andes 03 (former Andes 3, Andes 4, Andes 5, and Andes 6; 100 landslides); Amazon (former Amazon 1, Amazon 2, and Amazon 3; 72 landslides).

Other issues were solved, even if I think they deserve additional comment/discussion.

Comment response: Thanks for the observation. We applied the suggestion, merging Pacific (former Pacific 1 and Pacific 2; 73 landslides); Andes 01 (former Andes 1 and Andes 2; 132 landslides); Andes 02 (former Andes 3, Andes 4, Andes 5, and Andes 6; 100 landslides); Amazon (former Amazon 1, Amazon 2, and Amazon 3; 72 landslides).

Regarding, we noted that metrics worsen, so we opted, based on your comment, to taking account the 4 regions with a high number of events (Andes 2 and Andes 4, Pacific 1 and Amazon 2), and we are a pleasure to see that the metrics slightly better for validation procedure (Table X1).

Table X1: Comparison of TSS mean between different regionalization (11 reg: all regions, 4 macro-reg: suggested merge, 4 reg: 4 regions with a high number of events)

Number of regions	Imean-D (TSS)		Imax-D (TSS)	
	Cal	Val	Cal	Val
11 reg	65%	42%	62%	42%
4 macro-reg	56%	36%	52%	42%
4 reg	60%	44%	60%	45%

According to this comparison, we decided to add this to the discussion and the conclusions, as you can see below.

Discussion:

"… Peruccacci et al. (2012) found that the number of events must be mayor to 175 to limit the relative uncertainty below 10% but this figure may change for a different data set. Based on this, it is observed that only four regions (Andes 2, Andes 4, Pacific 1 and Amazon 2) have a number of events that are acceptable. The other regions have a greater source of uncertainty due to the quantity of the data. …"

Conclusion:

"Through the rainfall and landslides databases, it is possible to generate daily rainfall thresholds for shallow landslide occurrence. 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). Thus, only four regions (Andes 2, Andes 4, Pacific 1 and Amazon 2) have enough events to limit these uncertainties. Despite these uncertainties, the framework set up of this work allows for systematic updates of the thresholds as the records grow."

Regarding the validation procedure, I think the provided reply is acceptable. However, I would suggest including the results of Table X2 in the discussion, also because they are relevant for the case study.

Comment response: Thanks for the suggestion. We add a brief discussion and Table X2 in the new version of the mn, 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., Kirschbaum et al., 2015b; Dikshit et al., 2019), is quite short and there is the risk of overinterpretation. For this reason, this method was compared with other validation method based on a random selection of the data set (e.g., Brunetti et al., 2021; Gariano et al., 2020). According to this method, the data was divided into 70% for calibration and 30% for validation. The comparison of both validation approaches is shown in the Table 4. In this regard, the comparison between the validation methods did not indicate significant changes between each method. The results are very similar probably because the data size is not large enough to note the variations between the methods. It is highly recommended for future research focus in the expansion of the data-set and then compare the validation methods efficiency."

Table 4: 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	ΔTSS	1 year	Random	ΔTSS	1 year	Random	ΔTSS
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

Regarding the use of rainfall data with the daily temporal resolution, I still think that even

a brief acknowledgment of this limitation is needed.

Comment response: Thanks for the observation. We add this brief acknowledgment of this limitation in the new version of the manuscript, as you can see below.

"There are still many limitations to rainfall threshold study at the regional scale in Peru. Mainly the landslide short records are not enough to limit uncertainty in the threshold estimation (Peruccacci et al., 2012; Hirschberg et al., 2021). Another important source of uncertainty was the use of coarse temporal rainfall data resolution that cause a systematic underestimation of the thresholds (Gariano et al., 2020; Marra, 2019). Additionally, the spatial rainfall data resolution of ~10 km cell may cover several streams. And finally, the regionalization can be not enough representative of the high variability of descriptor landslide variables. It is highly necessary that these limitations must be taken into account in future research."

Regarding the two-variables thresholds, i.e. E-D and Imean-D, I still think that there is no need for calculating both of them. I thank you for the analysis provided; however, it is related to a single-variable case. The literature on two—variables rainfall thresholds is full of examples showing that E-D and Imean-E are analytically equivalent. As an example, if the equation for an E-D threshold is E=a*D^(b), the equation for the Imean-D threshold is Imean=a*D^(1-b).

Comment response: Thanks for the observation. According to the observation of the two referees, we decided to exclude E-D in the new version of the mn.

Overall, a grammar and syntax check is needed, given that there are still some problems (e.g. lines 125-129; line 145 [rainfall trigger event]; line 250 [Rainfall–landslide threshold]; and so on…)

Comment response: Thanks for the observation. We changed [rainfall trigger event] by [triggering rainfall event], and [rainfall–landslide threshold] by [rainfall thresholds for landslides occurrence] in the new version of the manuscript. In addition, we check all the paper to improve the syntax.

Regarding the figures: In figures 2 and 3, precipitation still needs to be corrected into rainfall. In figure 3 please check the units of measurement. Figure 6 in the text is still the old version, not the revised one.

Comment response: Thanks for the observation. We edit the figures it in the new version of the manuscript.