Authors' replies to referees' comments for NHESS-2021-206

We thank the Editor Dr. Paolo Tarolli for handling our manuscript and the two anonymous referees for their insightful comments. In what follows, we provide our replies to both referees. Referee comments are in black, our replies in red. At the end of the document, a draft of revised Fig. 1 is included.

Please note that referee #2 had some criticism about the title of the manuscript. We have formulated a possible alternative one: "Brief communication: introducing rainfall thresholds for landslide triggering based on artificial neural networks". We would appreciate also the Editor's advice on whether this title change may represent an improvement respect to the title of the preprint.

Reply to RC 1

Dear Authors, I have read with great attention your brief communication entitled "Rainfall thresholds based on Artificial neural networks can improve landslide early warning". The topic is original and interesting, the research design is robust and innovative, the English is good, the structure is fine. The manuscript surely deserves publication in NHESS. However, before endorsing final publication, I would like to ask you some clarifications and some improvements. I think all the modifications could be considered intermediate between "minor revisions" and "major revisions".

I look forward to receive the revised version of the paper.

Best regards.

Thank you for your positive, timely and detailed review, which will certainly help us to improve our manuscript.

L13: A reference could be useful here.

We will add Froude and Petley (2018).

Froude, M. J. and Petley, D. N.: Global fatal landslide occurrence from 2004 to 2016, Nat. Hazards Earth Syst. Sci., 18, 2161–2181, https://doi.org/10.5194/nhess-18-2161-2018, 2018.

Please note that Brief communications in NHESS allow max 20 references (the preprint presents 18 references). https://www.natural-hazards-and-earth-system-sciences.net/about/manuscript_types.html. We will try to comply with this limitation by better summarizing the literature about applying ANNs to other landslide modeling issues, as suggested by referee #2.

L16. I find odd not citing Caine (1980) who gave start to the research topic about I-D thresholds.

We will add Caine (1980).

Caine, N.: The rainfall intensity-duration control of shallow landslides and debris flows. Geografiska annaler: series A, physical geography, 62(1-2), 23-27, 10.1080/04353676.1980.11879996, 1980.

Fig 1: even zooming the pdf, I cannot distinguish very well red triangles and red points. Could you please change the color of the 2009-2018 rain gauges? Green would be an excellent choice I think.

Thanks for the suggestion. We have drafted a new improved version of the figure (see end of this document).

I recommend to add some details in the methodology description. In particular:

- If possible, I recommend describing the typical landslide typology of your dataset. This is important to understand which is the "target" landslide typology for your model, as different landslide types may be more sensitive to very different rainfall characterisitcs. Since you used FraneItalia, which is basically derived from newspapers, I guess you cannot exactly assess the typology of each landslide of your dataset, therefore your model is aimed to model and predict every landslide tyopology. Is my assumption correct? I don't think that would be wrong, but I think it should be clearly stated in the manuscript.

Thanks for this question. The Franeitalia database presents, among others, the following fields relative to landslides: typology, material and trigger. The events reported for Sicily within the database have been filtered based on the three above-mentioned landslide characteristics. In particular, only events having "rainfall" o "rainfall and other" trigger have been considered, so to exclude landslides due to earthquakes and anthropogenic activities. Also, events of the "fall" typology combined with "rock" material (rockfalls) have been removed from the analysis, as in the case of rockfalls, rainfall may have a triggering role different from the other types of landslides. Indeed, information about the landslide typology is not always available. In our case, among the 144 landslide events we have analysed, detailed information was available only for 18 events, 10 of which were "fall" of "more than one material", 4 "flow" and other 4 "slide". To respond briefly, we have considered all landslide types that have rainfall as the main triggering cause. We will add a statement in the revised manuscript accordingly.

- Both in CTRL+ and ANN: it is not clear how you relate each landsldie to the triggering rainfall. Do you use the nearest rain gauge? Do youconsider all the sorroundign rain guges? In the secnond case: how you decide which rain gauge is selected to characterize the triggering rainfall?).

Thank you for this question. CTRL+T reconstructs triggering conditions for each landslide event by a specific algorithm which is explained in detail in Melillo et al. (2018). In particular, for a given landslide, all rain gauges within a circle of radius R_b specified by the user are searched and, when more than one rain gauge is located within the circle, the rainfall events from each rain gauge ("Multiple rainfall conditions", MRC) are weighted based on the following equation:

$$w = f(d, E, D) = d^{-2}E^{2}D^{-1}$$

where w is the weight, d is the distance between the rain gauge and the landslide, E is the cumulated rainfall and D is its duration. For each landslide, w is used to identify the representative rain gauge, considering both geographical and rainfall features, and to determine the "probability" of the multiple rainfall conditions to be adopted for the calculation of rainfall thresholds. The probability, in case of multiple pairs, is computed by normalizing each w to the sum of the individual weights, whereas is set to one in case of a single condition. CTRL-T then determines the triggering rainfall conditions of each landslide as those corresponding to the highest probability.

We do not think it is essential to repeat in our manuscript full details of CTRL-T's algorithm. However, we agree that the manuscript will benefit from adding a short sentence about this point, and also, from specifying the value radius we have adopted, i.e. R_b = 16 km. It is perhaps also relevant to write that seldom this distance was reached, as the average distance between rain gauge and landslide for the 144 events was about 5 km.

- I understand that you train the ANN with 70% of 144 triggering events and 70% of 47398 events. Doesn't it lead to an unbalanced prediction? ANN will be trained to detect non triggering conditions more effectively than triggering conditions.

We understand the concern of the referee, and think that this indeed relates to the possibility to weight differently triggering events respect to non-triggering events, or more specifically a false positive (say, a "false alarm") respect to a false negative (say, a "missed alarm"). By including all events (without "trimming" the non-triggering ones), is equivalent to weighting equally false positives and false negatives. This is a common choice in the literature, as studies applying different weights in real contexts have appeared only very recently (Sala et al., 2021).

We will include a statement about this point in the methodology section explaining this point and cite Sala et al. (2021).

Sala, G., Lanfranconi, C., Frattini, P. et al. Cost-sensitive rainfall thresholds for shallow landslides. Landslides (2021). https://doi.org/10.1007/s10346-021-01707-4

L86 - please make clear the difference between training and validation dataset. I assume one of them is used for internal verification of the model while the other is used as an independent verification. Could you please make it clearer? I am used to say "calibration", "internal testing" and "independent validation", but since the order of your terms is different I guess some confusion may arise.

We understand the confusion of terms related to that fact that they may change based on the specific context (i.e., type of models employed). Let us first say that terms "calibration dataset" and "training dataset" can be considered equivalent. What rises confusion is the distinction of "validation" and "test" datasets. This confusion is also documented in the internet (e.g., https://machinelearningmastery.com/difference-test-validation-datasets/). This is also partially due to the fact that usually only a calibration and a (separate) validation dataset are considered. In our case, however, the validation dataset is somehow used for model calibration. More in details, the early stopping criterion for generalization works with both training and validation datasets: during calibration of the model on the training set, it checks if the error metric in the validation set gets worse. When it starts to get worse, the training is stopped (that explains the name of the criterion), even if in the training dataset the error metric would continue improving. This allows to prevent overfitting. As can be seen, the validation set is not directly used in determining the parameters of the ANN, but it is somehow used implicitly. This is why a third separated "test dataset" is considered: to assess the performance of the selected network when applied to "new" data by measuring its performance on a third fully independent set of data.

Here some suitable definitions that will be included in the revised manuscript:

- Training Dataset: The sample of data used to fit the model.
- Validation Dataset: The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.
- Test Dataset: The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.

Once the network is developed by a procedure considering these three datasets and early stopping, it is "frozen" and the ROC metrics (TSS) can be computed with that network on the entire dataset (this is what is presented in Tab. 1, column "TSS all"). Thus the "TSS all" is a "generalizable"

value. We think that the discussion (comparison with I-D power law) should focus on the performances of the "frozen ANN" with respect to the entire dataset, as the I-D power is determined with respect to the entire dataset, and thus more representative of the performances that can be expected from the approach we propose.

L104 - Here you introduce ROC curve, but then you don't use it (and I agree that is not an useful metric for the objective of this work). I think this part can be deleted.

We agree, the sentence will be modified as follows:

"We then identify the threshold maximizing the TSS."

Equations 7,10. These thresholds seem very low. I think in an operational use they would be regularly exceeded, expecially for short durations (think about how many times it rains 5.6mm in one hour). I understand your reasoning about the exponent, which makes the threshold higer for longer durations, but maybe you should state which is the duration range for which the thresholds are valid (e.g. the equation is enpirically defined for duratinos between 10 hours and 100 hours: the rest is an extrapolation where empirical data do not exist). Moreover, you can link this issue with the following discussion (around line 150) about the effectiveness of the ANN: the shortcoming of a power law is that the same equation is assumed valid for all the durations, while ANN could be more flexible).

Many researchers state that the lower boundary of durations for which thresholds are valid coincides with the minimum event duration available in the dataset. We use hourly data, hence, according to this criterion, the minimum duration of the threshold given in the preprint is >1 hour and < 10 days (240 h). We agree that the fact that for short durations the thresholds may be frequently exceeded is one of the drawbacks of traditional thresholds that can be overcome through ANNs, thanks to their flexibility. A sentence will be added on the discussion on this point.

L127-130 and table 2. Since I = H/D, performances of D-H and D-I should be identical. I think the reason of the differences in Tab 2 is the number of hidden neurons. Results of table 2 are influenced by the reinfall parameters (first column) and by the model metaparameters (e.g. hidden neurons - second colums). This complicates the discussion and interpretation of the results.

We agree with the referee that, given the same information content, one should ideally obtain identical performances and meta-parameters from the *D-H* (*D-E*) and *D-I* networks. However, ANNs can be slightly sensitive to how a set of variables having together the same information content of another are presented to the network. This is the main reason of the difference in the optimal number of hidden neurons and in the performances. Indeed, as specified in the preprint (LL 130-131), the networks are trained under the same conditions, i.e. by searching the optimal number of neurons: to fix a predetermined number of hidden neurons may penalize a given set of input variables respect to another.

TAb.2 I would add to the table two colums showing TPR and FPR: TSS alone is not very inforative about the effectiveness of the thresholds (e.g. 0.3 could be derived by the couple of values 0.9 and 0.6 or by the couple 0.4 and 0.1). I suggest to keep the table simple and to add TPR and FPR only for the independent verification dataset (the test dataset? See previous comment).

We will add the TPR and FPR for the entire dataset, i.e., those corresponding to "TSS all", which we deem more representative of the performance of our proposed models, as explained above. We

anticipate that, across the various ANN models, TPR is in the range 0.75 ÷ 0.85 while FPR in 0.20 ÷ 0.45

L150 I think this is a good point to add a couple of lines about what I mentioned in one of my previous comments.

We will add the following comment (draft): one shortcoming of traditional power-law thresholds is that the same equation is assumed valid for all the durations, while ANN can be more flexible.

L152-155. This point is very important. I-D and E-D thresholds work very well in case of shallow landslides in permeable soil (that's how Caine introduced them back in 1980). Later, researchers tried to estend the applicability of the techniques also to other settings, but the methodology shows evident theoretical and practical limitations (expecially when case studies are tested against a rigourous validation procedure). At present, research focuses on inoovations to increase the effectiveness of the technique proposing enhancements to better adapt to complex case studies. The idea of adding a third variable to the model is one of this innovations and others (e.g. Rosi et al. 2021 - even if they used antecend rainfall as third variable) obtained an increased effectiveness. I suggest adding this reference to your reasoning to better stress the results you obtained.

We will add the following comment (draft): the addition of a third variable, in general, has been seen as a possible way to derive thresholds that better adapt to complex case studies (e.g. Rosi et al., 2021).

Rosi, A., Segoni, S., Canavesi, V. *et al.* Definition of 3D rainfall thresholds to increase operative landslide early warning system performances. *Landslides* **18,** 1045–1057 (2021). https://doi.org/10.1007/s10346-020-01523-2

L156 - I find very interesting your work and I think this use of ANN is very promising. However, I suggest to mention some limitation. For instance, I think ANN would be difficult to operate and this aspect is still open to future research (I think it is why you presented a brief communication instead of a research paper). You add something similar in the conclusion but in my opinion the conclusion section should not contain new concepts and this comment would be better placed at the end of the discussion.

We have presented our work in the form of a “Brief communication” for various reasons, among which: many possible extensions of the research and immediateness for readers. This manuscript type, in our opinion also allows to slightly deviate from a “traditional” presentation of the material. Maybe it is just a matter of taste, but we believe that the part you have mentioned stands out better in the ending section of the manuscript.

Reply to RC 2

General Comments

The theme addressed in the manuscript is of interest and relevant within the scope of NHESS , particularly regarding the definition of landslide-triggering rainfall thresholds to be possible to be included in an early warning system for landslides. The manuscript, in my opinion presents some aspects that must be better addressed, modified, or discussed in more detail. Strengths: the methodology /the scientific method. Weaknesses: the analysis of the results /discussion.

Thank you for your positive, timely and detailed review, which will certainly help us to improve our manuscript.

Specific comments

Scientific Significance/Originality:

The manuscript presents and interesting approach to evaluate the possible contribute of ANN's to determine rainfall thresholds for landslide occurrence by comparison with rainfall thresholds determined by more frequentist methods. One of the main contributes of the work is the possibility to explore different variables related with the rainfall events that triggered and not triggered landslides to better characterize the rainfall critical conditions responsible for the landslide initiation and improve the predictive performance of the rainfall thresholds and its possible application to a landslide early warning system.

Thank you for appreciating the general approach proposed within our manuscript.

Scientific Quality:

The manuscript reports scientific and technical subjects relevant within the scope of NHESS, nevertheless, and if I made a correct interpretation of the approach used and results, is my understanding that some aspects must be better addressed, modified, or discussed in more detail.

Thank you for again for your comments. In the following we reply at each one.

- The title of the manuscript is suitable to be improved. The early warning component was not effectively explored or sufficiently described in the manuscript and therefore my suggestion is to reformulate or to remove the reference to "can improve landslide early warning" from the title and address it in the conclusions as future work.

Thank you for your suggestion. We agree with the fact that an early warning system is something more complex that just a threshold or an ANN providing the relationship between precursor variables and landslide occurrence. Nevertheless, we think that such a relationship is one of the most important components of an early warning system, and if it is improved, the whole early warning system is consequently improved. We have however thought of a possible alternative title, that can be as follows:

"Brief communication: introducing rainfall thresholds for landslide triggering based on artificial neural networks"

We would appreciate also the Editor's advice on whether this title change represents an improvement respect to the title of the preprint.

- Introduction section: The discussion regarding the different types of rainfall thresholds, explanatory variables, constraints, and strategies to improve their predictive performance available in the literature is very limited and needs to be ameliorated. This, to be perfectly understandable the methodological aspects that the ANN approach intends to fulfil. In addition, from lines 23-40 most of the examples cited by authors regarding the application of ANN are more related with the assessment of landslide susceptibility, which is not the scope of the work, than with rainfall thresholds definition or applications of rainfall thresholds to early warning systems. Consequently, is my understanding that this part should be completely reformulated.

Thank you for this comment. We agree that the content of the introduction can be improved if a greater focus is given on "why" we propose ANNs for estimating the rainfall conditions leading to possible landslide occurrence. At the same time, we think that it is important to stress the novelty of our manuscript in proposing a new way of using the tool of ANNs. However, the brief communication manuscript type allows only 20 references. We will make all possible efforts to add in the revised manuscript some sentences with some literature about the different types of rainfall thresholds, explanatory variables, constraints, and strategies to improve their predictive performance. Given the limit of 20 references, we will, on the other hand, reduce the part regarding ANN applications in the broad topic of landslides.

- Lines 38-40. The work objectives could be better defined. Although I recognize that the definition of rainfall thresholds can contribute for the development of landslide early warning thresholds, I think that was not necessarily explored/describe in the present work. To include this topic, it should be clearly presented in the methods and in the results section the link between them.

We agree that an actual early warning system is more complex than the relationship between precursors and possible landslide occurrence, though this is perhaps its most important component (see reply to previous comment). The manuscript will be amended with a sentence on this point.

- On the Data and methods section a brief description about the overall quality/accuracy of the landslide data (particularly about the landslides date of occurrence), and about the completeness of the rainfall database is critical since it can affect the prediction capability of the rainfall thresholds. In addition, authors mentioned that the rainfall data is available until 2018 but landslide inventory include landslides updated for the years 2018-2019. How many landslides fall outside the rainfall data? Moreover, is also mentioned that part of this landslides associated to the FraneItalia database are triggered by anthropogenic causes or earthquakes. A better description of the landslide inventory is needed. In fact, this information is not so much relevant and is unnecessary since the work deals only with rainfall thresholds for landslide initiation. My suggestion is to consider only the landslides that were effectively used for the study. What are the landslide types? Are they shallow or deep-seated? That information should be better addressed and it's important to understand the critical conditions defined for their occurrence, as largely explored in literature.

Thanks for asking this clarification. We agree that the paper may benefit from a few sentences with more details on the landslide data accuracy and types in the manuscript. Regarding the accuracy of the landslide data, for 103 events only the day of triggering was known, while for the remainder a more precise indication of the triggering instant was available. In the first case the triggering instant

was attributed to the end of the day, in the second case to the peak rainfall within the time interval when the triggering has occurred.

Rainfall data was available up to October 2018, so the landslide information after this date could not be exploited. Writing about the update of Franeitalia was just an additional information given for completeness, but we understand that this can lead to some confusion for the reader. We will write better this part, in the case that we will have the opportunity to send a revised manuscript.

As replied to a similar comment by referee # 1, we have removed rockfalls from the database, as well as landslides triggered by earthquakes and anthropogenic activities. Information on landslide type was available only for 18 out 144 landslide events in the database.

- Lines 56-60. In addition, regarding the rainfall data, how was determined the limit of 250 mm for considering an error in the rainfall record? What was the maximum hourly rainfall registered historically? It's possible to address better what is an evident error /rain gauge malfunction? It's possible to characterize the fraction of gaps and errors detected in the rainfall records based on visual inspection of the rainfall time series, and the reliability of the procedure? If an error in the rainfall records exist, and if the whole rainfall event surrounding the peak has been removed, what was the criteria to stablish the critical rainfall conditions for triggering the landslide associated to that rainfall event? If exists!

We understand that the referee has some concerns about the value of 250 mm as an indicator of a possible corrupted value in the rainfall record. This value derives from the analysis of annual maxima recorded historically in Sicily, for durations of 1, 3, 6, 12, and 24 h, in the period 1921-2015. This information is available from the Hydrological yearly bulletins (*Annali idrologici*, in italian).

Historically, a value of about 125 mm in one hour is the maximum value recorded in Sicily (in 1965 at Lentina, near Trapani), while, on 3 hours, an amount of 254 mm has been recorded (in 1979 at Fleri, near Catania). We then thought as reasonable to assume that the double of the maximum hourly observed precipitation (2 x 125 = 250 mm) as an upper limit for reliability of the recordings, also taking into account the maximum value for the closest greatest duration (3 hours) for which information on this issue was known.

The referee is also concerned about attribution of critical rainfall conditions to a landslide event when corrupted data is present (and thus the rain gauge is not considered). This is indeed not a big issue, as the next closest rain gauge with correct data will be used (average landslide-rain gauge distance is about 5 km – see a previous comment).

- The application used by Melillo et al (2018) it's used to derive a set of variables to be used for computing rainfall thresholds for landslide occurrence. These variables are indicated in lines 67-68. Can authors address better the mean intensity and total depth variables and in what differs the total depth from the ID threshold? In addition, regarding those variables and thinking on triggering and non-triggering rainfall events how related with the rainfall events are the critical rainfall conditions (rainfall events that triggered the landslides), they exactly match? I'm asking this, because, if I understand well, the algorithm identified a variable number of rainfall conditions responsible for the failures.

Regarding "mean intensity and total depth variables and in what differs the total depth from the ID threshold?", we think that the notation has generated some confusion. First of all, we meant *E* and

not H. Second $E = I \times D$. In other words, we will explicitly insert the multiplication symbol between I and D. The ID threshold/ANN considers both I and D, while the E threshold (H) only E.

Regarding the second part, we will add to the manuscript more details on the CTRL-T software proposed by Melillo et al (2018), explaining how triggering and non-triggering events are obtained from the original hourly rainfall and landslide data. In particular, in the case of multiple suitable rain gauges near to the given landslide, a "most probable (triggering) rainfall condition" is computed by a specific algorithm that attributes a weight to rainfall events from different rain gauges based on the distance from the landslide and the rainfall event characteristics (see a previous reply to referee #1).

- Line 77. The 144 triggering rainfall events are defined for, I suppose, the 144 landslide events /landslide cases included in this study from the FraneItalia database. Please turn clear here and in the data section.

"Application of the CTRL-T software yielded 144 triggering rainfall events and 47398 non-triggering events." will be change in "Application of the CTRL-T software allowed to reconstruct the rainfall events associated to the 144 landslide events in the inventory (triggering events) and 47398 non-triggering events".

- Line 85: Could authors be more specific regarding the specific objectives of partition of the rainfall events into validation and test dataset? Are the groups similar with respect the distribution of rainfall events characteristics? How the partition considers triggering and non-triggering rainfall events? Please address better this issue.

The training, validation and test datasets are randomly taken from the entire dataset. Each triggering and non-triggering event is sampled with the same probability from the entire dataset. Hence, in probabilistic terms, it is expected that the same proportions of triggering vs. non-triggering events are present in the each of the three subsets. Given your comment, we have done a check and found that the triggering/non-triggering ratios are respected across the different subsamples (in a probabilistic sense). Regarding representativeness of the three (training, validation and test) subsets, please see our reply to one of the next comments.

- Line: 106-107 "Results from ANNs are compared with rainfall duration-depth power-law thresholds derived through the maximization of TSS analysing both triggering and non-triggering events". How was that done – derived by the CTRL-T software? Please turn clear in the methods section.

CTRL-T software does not implement TSS maximization. That was done externally by loading the rainfall data reconstructed by CTRL-T in a MATLAB code using the global optimization toolbox to find the *E-D* threshold corresponding to maximum TSS. In particular global optimization was obtained via the particle swarm algorithm. We will add a few details on this point in the revised manuscript.

- Lines 109-115: I acknowledge the examples with other thresholds available in literature, but I think that differences/similarities could be additionally explored. Are thresholds based on the same datasets? For the same periods? For the same landslide types?

Thresholds of equation (8) and (9) were derived from rainfall datasets different from the one we have considered in our paper. In particular: Both studies – i.e. Gariano et al. (2015) Melillo et al.

(2016) used rainfall hourly data collected between a July 2002 and December 2012 that comes from 2 different monitoring networks and 229 landslide events triggered in the same period. The main difference between the two studies seems in the algorithm for reconstructing rainfall events: in the first study an earlier version of the CTRL-T algorithm was used which required manual selection of the most representative rain gauge in case of "multiple rainfall conditions", while in the second study automatic selection was implemented. A sentence on this point will be added to the manuscript.

- Lines 119-123: This step is not suitable to be included in the methods section? If so, please adjust. In addition, what it's possible to conclude regarding this comparison by the fact that after 5 hours the threshold defined by equation 10 be above the ones defined by equation 7? A figure comparing those two thresholds could help understanding the idea.

The explanation of how the I-D threshold corresponding to max TSS are derived has been inserted in the Results section, because we thought it was short and thus more easily understandable there. Also, we made this choice because the methodologies for deriving ID thresholds are not the main focus of our manuscript. However, we will make efforts to have this part in the Methodology section without undermining clarity and immediateness for readers.

Regarding the other point: we think that, though useful, a figure comparing the various ID thresholds is not essential in the context of a Brief communication (which allows max 3 figure/tables). We will instead try to add some more details in the text on how the thresholds compare among themselves, also to take into account a previous comment by Referee # 1.

- Lines 109 – 133: I strongly believe that aspects described in these paragraphs could be better placed in the methods section to become clearer the author's approach: e.g., the reference to the non-exceedance frequency for triggering events equal to 5%; the list of variable configurations and part of the subsequent descriptions (Lines 125-129).

Thank you for these suggestions for improving the structure of our manuscript. Some choices are specific to our application and not general (e.g., to vary the number of hidden neurons from 5 to 20). This is why we have placed them in the results section. However, we will do our best to move these parts in the methodology section.

- In my perspective the fact that the training datasets obtained less TSS than the validations and test datasets need additional discussion. I understand the approach used to preventing overfitting of the thresholds based on the training dataset, but is not supposed to be achieved better results when the predictive capacity is evaluated with the data partition used for training the predictive model? Being, in this sense, the independent validation with the dataset not used for training the model (validation partition) more robust and with general lower predictive results?

This point is related to the representativeness of the validation and test datasets. In particular, one can introduce a check to ensure that the complexity of classifying triggering conditions from non-triggering conditions is similar across the training, validation and test sub-datasets. One way to do this is to keep (a-posteriori) only the sets yielding TSS values in the validation and test subsets that are not greater than the TSS value in the training subset. We have done some preliminary applications of this principle and we have seen that there are no significant changes in the "TSS all" values of Tab.1 – while the "TSS test" values still remain better than the reference value of 0.5 for the E-D power-law – which supports the reliability of our procedure for developing the ANN models. We will add some discussion on this issue, and update results according to the revised simulations.

- The analysis of results expressed on table 1 (line 140), could be better explored in lines 144-155. Why the analysis is centred on results defined, if I understand well, considering no partition of the rainfall events dataset (TSS all)? Why not considering the more robust evaluation of the predictive performance of the rainfall-thresholds model, this is, with the validation dataset (TSS validation or TSS test), and what explanations could be attributed considering the TSS values obtained for the different variables configuration and for the training and validation datasets.

We understand that the Referee is concerned about the use of "TSS all" instead of "TSS test". This relates to a previous comment by Referee #1 about L86. As explained in our response to that comment, "TSS all" is derived from the ANN trained with early stopping (no recalibration on the entire dataset has been done), and thus the generalization capabilities of the ANN are not undermined. Thus "TSS all" is representative of the performance on the entire dataset which we think is more significant than the performance on TSS test, which refers to only to a portion of the entire dataset (15%). We will add some discussion on these points.

- A part of the conclusions section is suitable to be moved for the results and discussion section, particularly, the text after line 169 related with the drawbacks of the ANN approach.

We agree that in a regular Research article the mentioned lines would have been better placed in the discussion section. Here we think that they are more suitable for the Conclusions. However, an improvement in this sense may be to name this last section as "Concluding remarks".

- Abstract and conclusions should be adjusted accordingly.

Given the 100 words limit for the abstract, we believe that is already includes the most important results of our manuscript. Regarding conclusions, see previous response.

Presentation quality:

Overall, the manuscript is well written. In my opinion, the structure needs some adjustments, by moving some text sections from results to Data and methods and from Conclusion to Results and discussion. The manuscript presents generally a clear language that is understandable and scientifically precise. I made some small changing suggestions for the title and results analysis, even so, the abstract, the subtitles and the figures and tables captions are in general adequate. With respect to figures and tables they are not in large number, but present generally a reasonable quality and adequate to the purpose of the manuscript. Nevertheless, and in respect to those items described above some comments are made in the "Scientific Quality" section.

We have appreciated a lot your suggestions for improving the presentation quality of our manuscript. We will take into account of your suggestions as described in the above responses to your comments.

Technical comments (e.g., typing errors, format)

Some additionally technical comments are listed below.

Thank you for carefully checking our manuscript for typing and format errors.

Lines 22-23: consider adjusting the use of risk analysis in this sentence, since no exploration of consequences are illustrated by authors. Please see my previous comments about this literature examples using ANN for this manuscript in point 2 of my Scientific Quality comments.

We will use "landslide analysis" instead of "landslide risk analysis".

As replied at your previous "Scientific quality comments", we will add some literature about advances on improving landslide triggering thresholds and will slightly reduce the part on ANN applications to susceptibility mapping.

Line 41. Adjust position of citation of figure 1. In the maps its not possible to see the 20 regions of Italy.

We will change "We refer to the case study of Sicily, one of the 20 regions of Italy (Fig. 1)."

into:

"We refer to the case study of Sicily (Fig. 1), one of the 20 regions of Italy."

In figure 1: i) adjust the Europe map (upper right). Although recognizable it's not properly represented. Include some colour differentiation for the Atlantic Ocean/Mediterranean Sea and Europe countries. Include one or two country/ocean toponyms; ii) Eliminate from the figure the word "Legend" and replace "Rain gauges" by "Rain gauge". The differentiation of the rain gauges 2009-2018 and the landslides are sometimes not easy. Adjust caption: consider changing "and landslides from the…" for "and landslides dataset extracted from the FraneItalia…."

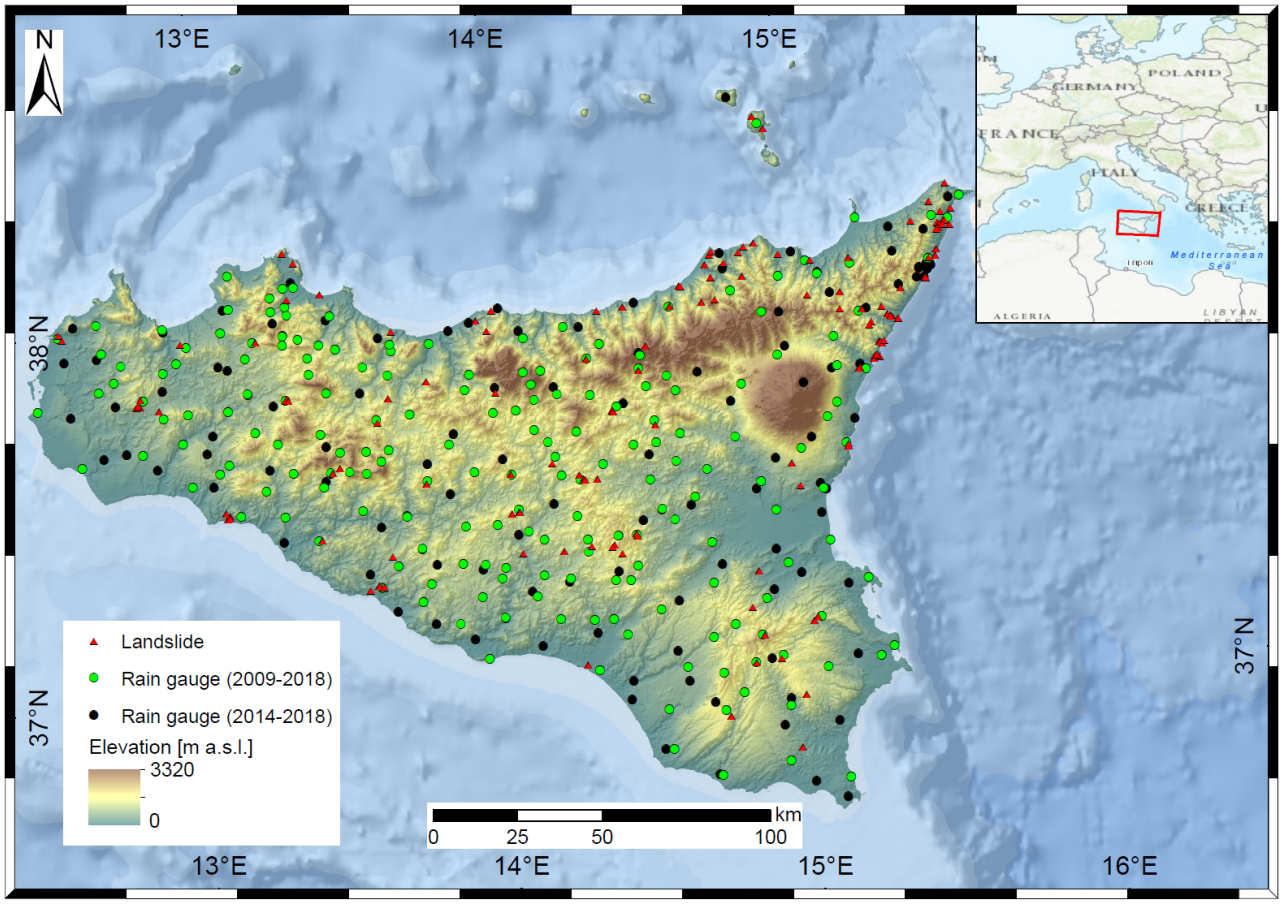
The figure will be improved with your suggestions (see draft at the end of these replies).

In figure 2 what represents the E-D threshold. "E" is used to describe the cumulation rainfall associated to the rainfall or landslide event? Please turns clear in figure and manuscript text. In addition, what differs from H depth in figure 2b from Total Depth (H) in Line 67?

Thank you, this was a typo. We will change all *H*s in *E*s.

In Table 1 verify the number of Table. Should be Table 1.

Thank you, it is Table 1 and not 2, indeed.



Revised Fig.1 (draft).