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 show how the manuscript has been amended to take into account referees' comments. Referee comments are in black, our replies in red. Line numbers refer to the revised manuscript with changes not tracked.

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 has helped us to improve our manuscript.

L13: A reference could be useful here.

L14 We have added citation to 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(8), 2161–2181, doi:10.5194/nhess-18-2161-2018, 2018.

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

L17: We have added citation to Caine (1980).

Caine, N.: The Rainfall Intensity-Duration Control of Shallow Landslides and Debris Flows, Geogr. Ann. Ser. A, Phys. Geogr., 62(1), 23–27, 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 improved figure 1, according to this and other comments.

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.

We have added more information on FraneItalia database and landslide typology considered in our study in the Data and Methods section at LL 55 – 61 and LL 94 – 97:

"Thus, our analysis is based on the period from January 2010 to October 2018, where both rainfall and landslide information is available. Some landslide events have been removed from the analysis. In particular, this was done based on the fields included in FraneItalia that characterize the observed landslide events – typology, material and trigger. 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 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."

"Furthermore, for the 144 landslide events detailed information on the typology was available only in 18 cases, 10 of which were "fall" of "more than one material", 4 "flow" and other 4 "slide". The average distance between rain gauge and landslide for the 144 events is about 5 km, thus seldom the maximum value of R_b= 16 km was reached."

- 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?).

We have added more information on how triggering rainfall is reconstructed by CTRL-T at LL 75-81

"In particular, for a given landslide, all rain gauges within a circle of radius Rb 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 are weighted based on the rain gauge-landslide distance and the rainfall event characteristics (cumulated rainfall and duration). The weight is used to estimate the "probability" associated to each rainfall condition potentially attributable to each landslide event. In case of multiple rainfall conditions, the probability of the individual event is computed by dividing its weight to the sum of concurrent events' weights. CTRL-T then determines the triggering rainfall conditions of each landslide as those corresponding to the highest probability."

- 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 have added a sentence on this point at LL 118 – 121:

"We then identify the threshold maximizing the TSS. Maximization of TSS implicitly assumes that all entries of the confusion matrix have the same importance. Quantifying how more important is a

false negative respect to a false positive, is a complex task that goes beyond the aim of the present analysis (cf. 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 have added the following part to clarify the difference between training, validation and test dataset at LL 133 - 138:

"The entire dataset of rainfall events was divided into a training, a validation, and a test data set, selected randomly from the entire dataset, in the proportions of 70%, 15% and 15%. The training dataset is data used to fit the model, whereas the validation provides an unbiased evaluation of a model fit on the training dataset while tuning model hyper-parameters, such as the number of training iterations. Finally, the test dataset provides an unbiased evaluation of a final model fit. This subdivision allowed to apply the early-stopping criterion to prevent overfitting. According to this criterion, the training of the neural network is stopped when the values of the performance function calculated on the validation dataset start to get worse."

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 has been modified as follows (LL 118-119):

"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. The following sentence has been added on the discussion on this point (LL 186-187):

"In other words, one of the shortcomings of a power law is that the same equation is usually assumed valid for all the durations, while ANNs are more flexible"

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 already stated in the preprint, 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.

The discussion on this point has been expanded as follows at LL 126-130:

"This has been done because the two pairs D-I and D-E have the same informative content by construction, as confirmed by the fact that the performances of the D-I and D-E neural networks do not differ significantly (see later, Tab. 1) – slight differences may occur as ANNs can be sensitive to how a set of variables is presented to the network even though the information content is equivalent under a mathematical point of view."

TAb.2 I would add to the table two colums showing TPR and FPR (all): 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 have added two columns in the Table with TPR (all) and FPR (all) and showed the values for the power-law. Some discussion has been added about the comparison of this values for the power law and the ANN, e.g. at LL 181-183:

"Notably, in the case of the pairs D-I and D-E – i.e., the same variables used for the power law – the TSS = 0.59 (0.60), which is significantly higher than TSS_0. This is obtained by both an increase of the TPR (true positives) and a decrease of the FPR (false positives)."

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

The following sentence has been added on the discussion on this point (LL 186-187):

"In other words, one of the shortcomings of a power law is that the same equation is usually assumed valid for all the durations, while ANN are 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 added the following part at LL 189-192:

"In particular, in this case, it has been shown that peak intensity may have an important informative content, an aspect that has not been perhaps sufficiently investigated in the literature, even though some researchers have found that the addition of a third variable is 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, which now has been entitled "Concluding remarks" to take into account your comment.

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 has helped 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 show how we modified the manuscript to take into account of them.

- 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.

We have changed the title as follows:

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

We hope that 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 we have modified the introduction by we have revised the introduction adding a discussion about the different types of rainfall thresholds linking ANN as a strategy to improve their performance at LL19-34. At the same time, we have reduced the part about ANNs for modeling landslide susceptibility at LL 35-39. We have cited the most essential papers in our view, as Brief communication do not allow more than 20 reference entries. In the following we paste the modified parts:

"Such a constraint can potentially limit the predictive performance of the thresholds, because the informative content of the considered explanatory variables may not be exploited at fullest. This holds true all the more so when searching for alternative or additional variables with the aim at improving the performances of the thresholds, such as antecedent rainfall conditions (Glade et al., 2000), water storage and soil moisture data (Bogaard and Greco, 2018; Marino et al., 2020). For the case of E-D or I-D thresholds the use of a power law is customary and its rationale has been also verified based on a combined stochastic and physics-based approach (Peres and Cancelliere, 2014). On the contrary, either in the case of a different pair of variables or the analysis of more than two variables, there is no analogous prominent parametric form of the threshold equation. For instance, as reported by Conrad et al., (2021), alternative formulas have been considered for hydrometeorological thresholds – i.e., based on rainfall and soil moisture or catchment storage – , including linear and bilinear functions, interpolated line segments without a mathematical function, cut-off values for integration of antecedent conditions with traditional rainfall thresholds, and more complex logical operators. The choice of a predetermined threshold equation form can potentially limit the performance of the threshold derivable from the given set of variables, and thus also limit the scientific soundness of the comparison between different approaches for deriving landslide triggering thresholds. Artificial Neural Networks (ANNs), belonging to Artificial intelligence or Machine learning techniques, allow to potentially remove the mentioned limitation of having to choose a predetermined parametric threshold form, as they are universal approximators, i.e. capable of reproducing any continuous function (Haykin, 1999)."

- 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).

Manuscript has been amended as follows at LL14-17:

"Commonly, rainfall thresholds indicating the conditions under which landslides can potentially occur, are a key component of warning systems aimed at protecting the population from a possible landslide event. In most of the cases, thresholds are determined using empirical methods that link characteristics of precipitation, such as duration D and mean intensity I or cumulated rainfall E = I×D, to landslide occurrence (Caine, 1980)"

- 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.

Manuscript has been amended as follows at LL55-61:

"Thus, our analysis is based on the period from January 2010 to October 2018, where both rainfall and landslide information is available. Some landslide events have been removed from the analysis. In particular, this was done based on the fields included in FraneItalia that characterize the observed landslide events – typology, material and trigger. 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 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."

and LL90-97:

"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. 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 instant of peak rainfall within the time interval when the triggering has occurred. Furthermore, for the 144 landslide events detailed information on the typology was available only in 18 cases, 10 of which were "fall" of "more than one material", 4 "flow" and other 4 "slide". The average distance between rain gauge and landslide for the 144 events is about 5 km, thus seldom the maximum value of Rb= 16 km was reached."

- 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).

Manuscript has been amended as follows at LL61-64:

"Rainfall data have been checked so to remove suspicious rainfall data. In particular, where hourly rainfall exceeded 250 mm – corresponding to about one third of mean yearly rainfall for Sicily and to about two times the maximum rainfall ever recorded in 1 hour – the series has been visually inspected, and in the case of an evident error (rain gauge malfunction) the whole rainfall event surrounding the peak has been removed."

and LL75-81

"In particular, for a given landslide, all rain gauges within a circle of radius Rb 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 are weighted based on the rain gauge-landslide distance and the rainfall event characteristics (cumulated rainfall and duration). The weight is used to estimate the "probability" associated to each rainfall condition potentially attributable to each landslide event. In case of multiple rainfall conditions, the probability of the individual event is computed by dividing its weight to the sum of concurrent events' weights. CTRL-T then determines the triggering rainfall conditions of each landslide as those corresponding to the highest probability."

LL 88-89: "The rain gauge search radius has been fixed to R_b =16 km."

LL 96-97: "The average distance between rain gauge and landslide for the 144 events is about 5 km, thus seldom the maximum value of R_b= 16 km was reached."

- 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 $E = I \times D$. We have explicitly inserted 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 added 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 at LL 75-83:

"In particular, for a given landslide, all rain gauges within a circle of radius Rb 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 are weighted based on the rain gauge-landslide distance and the rainfall event characteristics (cumulated rainfall and duration). The weight is used to estimate the "probability" associated to each rainfall condition potentially attributable to each landslide event. In case of multiple rainfall conditions, the probability of the individual event is computed by dividing its weight to the sum of concurrent events' weights. CTRL-T then determines the triggering rainfall conditions of each landslide as those corresponding to the highest probability. When the triggering instant is after the end of the rainfall event, the most probable triggering rainfall conditions are computed considering the whole event, otherwise the event is truncated at the triggering instant."

- 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.

Sentence has been changed 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. 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 instant of peak rainfall within the time interval when the triggering has occurred."* (LL 90-94)

- 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 statistical terms, the same proportions of triggering vs. non-triggering events are present in the each of the three subsets.

Regarding the distribution of rainfall events characteristics, we have re-run our ANN model development by adding a consistency constraint, which requires that the performances in the validation and test sets are not greater than in the training set. This ensures that the level of difficulty of classifying triggering from non-triggering events across the three sub-datasets is similar (LL 138-141).

"In order to ensure representativeness of the training, validation and test datasets, when the TSS in the test or the validation data set is greater than the TSS in the training data set, a new training is carried out with a different random data split."

- 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. A sentence has been added on this point:

LL161-162: “We have hence derived the power-law threshold corresponding to the maximum TSS – externally to the CTRL-T software, via MATLAB® global optimization toolbox –, obtaining the following result: …”

- 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?

We have added details on the differences between the methods and datasets with other thresholds in literature (LL 156-159):

“It should be mentioned that these thresholds were both derived from rainfall datasets covering the period July 2002-December 2012, which is different from the one we have considered in our analysis. The first threshold has been derived with an earlier version of the CTRL-T code, which required manual selection of the most representative rain gauge (Melillo et al., 2015), while the second study derives from the updated algorithm, where this selection is made automatically.”

- 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. 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).

- 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).

We have moved some parts on ANN model development from the “Results and discussion” to the “Data and Methods” section (LL123-132):

“For our analysis, different combination of input data (duration D, intensity I, total depth E and peak intensity Ip) and different architectures, changing number of hidden neurons were tested. In particular, the following input variable configurations have been investigated: 1) D; 2) E; 3) I; 4) Ip; 5) D and E; 6) D and I; 7) D and Ip; 8) E and Ip; 9) I and Ip; 10) D, E and Ip. The listed input

configurations are indeed all possible ones, except those combining both E and I with duration D. This has been done because the two pairs D-I and D-E have the same informative content by construction, as confirmed by the fact that the performances of the D-I and D-E neural networks do not differ significantly (see later, Tab. 1) – slight differences may occur as ANNs can be sensitive to how a set of variables having together the same information content of another are presented to the network. All the data have been inputted taking their natural logarithms. Different networks have been considered varying the number of hidden neurons from 5 to 20, in order to search for the best value, i.e., the one yielding the highest TSS."

- 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. As replied for a previous comment we have modified our ANN model development scheme adding a consistency check (LL 138-141).

"In order to ensure representativeness of the training, validation and test datasets, when the TSS in the test or the validation data set is greater than the TSS in the training data set, a new training is carried out with a different random data split."

Check also the new Table 1. There are no substantial changes in the TSS all values. Also, the TSS test values are still greater than the TSS of the power law (E-D). For the *D-E-*I_p ANN performances slightly improved respect to previous computation.

- 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%). This allows also a better comparison with the performances of the power-law, as this is derived from the entire dataset. We added the following part to the manuscript (LL 141-145):

"Once the network is developed considering these three datasets and early stopping, it is "frozen" and ROC metrics (e.g., TSS) can be computed with that network on the entire dataset, and the corresponding performances can be considered generalizable. Thus, when comparing our proposed approach with the traditional one, we focus on these last performances (labeled as "all"). This

seems the most appropriate way to proceed, as the I-D power law and its performance is determined with respect to the entire dataset."

- 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 it 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, and have implemented most of them as described when replying to your other 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 added some literature about advances on improving landslide triggering thresholds and reduced 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 have changed "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…."

We improved the Figure according to your suggestions.

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.