Referee #1

We provide here our reply to the comments by Anonymous referee #1, together with the changes we plan to apply to the manuscript in response. The referee's comments are in black fonts and our reply in blue.

(1) This brief communication is very… brief! I mean, in the positive sense of the word. Indeed, it is a clear, concise manuscript that is perfectly written in fluent English - something very rare for a reviewer to find. I thank the authors for that! The paper gets straight to the point: landslide-triggering intensity-duration thresholds and precipitation intensity-duration-frequency curves cannot be confounded, compared, or plotted together. Neither one can be used to quantify the return time of the other.

Thank you for devoting your time to consider our work, this is much appreciated. We are also glad to hear that our conciseness was appreciated.

(2) Frankly, having worked on rainfall analysis and landslide prediction for years, the idea of mixing/comparing ID thresholds and IDF curves is something that never came to my mind. In the few cases I have seen in the extensive literature on these topics, it has always seemed very strange, not to say a downright methodological error. So, I can say that I certainly agree with the authors of this paper, although I do not think the article addresses a relevant scientific and/or technical question. I simply think that mixing ID thresholds and IDF curves is a misconception that does not even require discussion.

We agree the issue is indeed obvious when giving it a more profound thought. However, the misconception is to some extent also 'logical' with the used terminology and has been around implicit for quite some time in conference discussions and literature. Therefore, we think it can be useful to the community to clarify the differences and also indicate the consequences for landslide probabilistic analyses in climate change discussions.

(3) The authors list the differences between ID thresholds and IDF curves, focusing on the different durations (D and W) considered by the two tools, and then analysing the differences in terms of return time referring to these durations. In my opinion, they forgot the main and most important difference. That is: since their definition from pioneering works (Nel Caine and also previous pioneers), ID thresholds have been defined considering ID pairs that are somehow - arbitrarily or not, subjectively or not - linked to the initiation or re-activation of one or more landslides. On the other hand, IDF curves are defined considering IW (using the same terminology as the authors) pairs that are not linked to landslide/debris flow occurrence, referring only to rainfall itself. Indeed, the authors write "IDF are obtained by collecting the highest rainfall intensities observed any year over the time windows of

interest" (lines 45-46). Therefore, the two tools summarise or describe different variables (the ID pairs by which the thresholds are defined are different by definition from the IW pairs with pre-fixed durations of the IDF curves, having different characteristics consequently) and different processes (landslide or debris flow initiation and rainfall severity). This is, in my opinion, the main reason why the two tools must not be compared or mixed. I wouldn't have added anything else to this discussion

Reading this comment and some of the following, we realize that our message was not formulated clearly enough. While we agree that ID thresholds concern landslides/debris flow triggering or reactivation, we tend to disagree on the fact that this is the "main and most important" difference with IDF curves. The entire idea originates once ID pairs or thresholds are defined (in any way), and a perfectly natural question arises: "what is the probability of these conditions to occur?". The answer to this question is independent from the triggering of mass movements and only depends on the precipitation climatology of the area. What we would like to point out with our brief contribution is that the way that is often used to quantify this probability (by using IDF curves) hinges on a misconception. To make our point more clearly, we will include specific text in the introduction, as follows: "*[…] events on the basis of their duration and average intensity (Leonarduzzi et al., 2017).* ***Once these thresholds are defined, it is natural to ask the question "what is the probability of these conditions to occur?'". Answering it is not trivial.*** *IDF curves are [...]*".

However, the authors added more to the discussion, deserving attention. I list below some other comments on this paper.

(4) First, I don't understand the first part of the title "Threshold not probability". Actually, thresholds can be probabilistic. As a matter of fact, the Bayesian thresholds mentioned by the authors are probabilistic. Moreover, the frequentist thresholds also mentioned by the authors allow defining probabilistic diagrams to be used for early warning purposes. Therefore, I would remove this part of the title, which works only for deterministic, binary thresholds.

Thank you for this consideration. To our view, the fact that ID thresholds can be probabilistic does not make the title wrong. For that matter, IDF curves technically are thresholds: just like the definition of cumulative distribution function, that is non-exceedance probability of a given threshold. To address the reviewer concern, we modified the title as follows: *"Threshold* ***and*** *probability. The conceptual difference between ID thresholds for landslide initiation and IDF curves."*. Further, we will amend the text as described in our reply to the previous comment 3. This will be done early in the introduction, to make this aspect clearer right from the beginning: "*[…] events on the basis of their duration and average intensity (Leonarduzzi et al., 2017).* ***Once these thresholds are defined, it is natural to ask the***

question "what is the probability of these conditions to occur?'". Answering it is not trivial. *IDF curves are [...]*".

(5) In several parts of the text, the authors write that quantifying the return period of a given intensity used to define ID thresholds using probabilities estimated from the IW space is erroneous and causes an underestimation of the severity of the triggering rainfall. I agree with the authors, totally. However, I'd suggest mentioning some works in which this erroneous approach was adopted, also because these are cited again in the last sentence of the paper ("Some results in the literature may thus be quantitatively inexact").

Thanks for this suggestion. References to some of these works are indeed reported in the manuscript, including the ones authored by ourselves (line 15 of the discussion paper). We intentionally did not provide a list of other instances to avoid the (unintended) implication that we want to challenge the main results of those studies, and we therefore removed the sentence mentioned by the referee.

(6) Moreover, I would add that the return period of a given ID thresholds should not be calculated at all. Indeed, rather than adopting dichotomous approaches (above/below threshold), using statistical and probabilistic approaches, as the two mentioned above, allows the probabilistic characterisation of the thresholds without introducing (erroneously) the concept of return time, which is also highly questionable for a variable not easily measurable as landslide or debris flow occurrence/triggering. In addition, as the authors certainly know, the concept of return time and how it changes in relation to non-stationarity is a topic of discussion in the scientific community.

We fully agree with the reviewer on these considerations. However, they seem to tackle a different, and much wider, problem: the one of how landslide triggering thresholds should be defined. Although most important, in this brief communication, we highlight a conceptual difference that hinges from theoretical arguments, with the aim of stimulating discussion within the community.

Moving to sections 2 and 3, the differences between ID thresholds and IDF curves are listed, focusing in particular on the different ways to define the duration of the ID/IW pairs.

(7) According to the authors' view, the durations D are user- (or arbitrary-) defined while the durations W are not. But, actually, W are also user- (or arbitrary-) defined using running windows of x minutes or hours: 5, 10, … 45 minutes or 1, 2, … 48 hours were also defined

by a user. Moreover, the authors didn't mention that IDF curves can be defined using the partial duration series approach as well, so introducing another point of discussion.

Thank you for this suggestion. We will rephrase this portion of the text to improve the clarity of our message. In IDF curves, the window W is set a priori based on some considerations (which we do not need to discuss here). At this point the probability of exceeding a given intensity over that window is estimated. What is relevant is that once W is chosen, the probability distribution of the extreme intensities observed over that time interval is well defined from a theoretical perspective (and it does not depend on the way we quantify it, annual maxima, partial duration series or other approaches). Conversely, the duration D in ID thresholds is defined *conditionally* to (a) the occurrence of a landslide, as correctly pointed out by the reviewer in comment 1, and also by (b) the identification of a triggering event. While statistically, it is possible to objectively define the population of ID pairs conditioned on (a), the population of ID pairs conditioned on (b) is not well defined because it depends on user-defined choices concerning the triggering event definition (different event definitions will lead to different populations and, hence, different statistics). This is because different definitions will cause the intensity to be averaged over a different time interval (or the depth to be accumulated).

We amended the text in section 4 to emphasize these aspects: "*In fact, duration refers to the total length of an* ***user-defined*** *rain event D, on the one hand, and to a fixed-length temporal running window W, on the other.* ***Once W is chosen in IDF curves, the population of the corresponding intensities is theoretically well defined, and hence their probability distribution. The duration D in ID thresholds, instead, is defined conditionally to (a) the occurrence of a landslide and to (b) the way triggering events are identified. While it is possible to objectively define the population of ID pairs conditioned on (a), the population of ID pairs conditioned on (b) depends on user-defined choices concerning the identification of the triggering events****. The use of the same term [...]*"

In addition, we amended the text to explicitly mention partial duration series: "*IDF are obtained by collecting the highest rainfall intensities observed any year over the time windows of interest. To do so, usually a running window of the desired length is moved across the timeseries and the largest values are extracted****, for example the annual maxima or the exceedances of very high thresholds****. Extreme value distributions are then used to describe these* ***values****, and intensities corresponding to an assigned cumulative probability are extracted for the required duration.*"

(8) In section 2 (lines 29-32) the authors write "rainfall records are often not available at hourly resolutions nor in close range of the landslide (Marra et al., 2016; Marra, 2019), which makes the events separation dependent also on these aspects.". Actually, this issue

affects the definitions of W too. Indeed, if only daily measurements are available in a given area, sub-daily values of W (e.g. the classical 1, 3, 6, 12, 24 hours) can't be defined, and the IDF curves cannot be drawn for sub-daily durations.

Thank you for commenting on this. Once again, here the referee focuses on the practical aspects while we are addressing a theoretical standpoint. In the case mentioned by the referee, it is not possible to *practically* calculate sub-daily values of *W* and draw the IDF curves, but the population and its statistics are well defined. Indeed, there are several approaches to estimate sub-daily IDF curves from daily observations, as well as sub-hourly IDF curves from hourly observations, via assumptions on the statistics of extremes across scales (e.g., Aronica & Freni, 2004; https://doi.org/10.1016/j.atmosres.2004.10.025).

(9) In section 4 (lines 62-63), the authors write "In a univariate framework, the return period T* of a rainfall event can reasonably be defined as the maximum among the return periods Tw associated with all possible temporal scales". I think that some examples should be provided to support this statement.

Thank you for this comment, which allows us to better specify some aspects of our reasoning that were not fully clarified. In theory, there is not one unique probability for a rainfall event, because this will depend on the examined scale (*W*, but one can think of areal scales as well, getting to the IDAF curves, with A standing for area). What we do here is provide a practical univariate solution, which, differently from the above points, cannot be backed theoretically. To make our study clearer, we reorganized the manuscript as follows.

(A) In section 4, we define W* as the true, unknown, triggering interval: ***"****Indeed, precipitation events are characterized by different return periods at different temporal windows****, and any temporal interval during the event could be the true, unknown triggering interval W^*. It follows that the length of the user-defined triggering rainfall D does not necessarily coincide with the unknown temporal window W^* that triggered the landslide (the equality $W^*= D$ only holds in very peculiar cases).*** *Even more crucially, [...]****"*** and "***Using the entire event length as D, for example when the exact time of occurrence of the landslide is unknown, almost always gives*** *$T_D \le T^*$, this erroneous approach* ***may often*** *cause a systematic underestimate of the severity of the triggering rainfall, leading to false alarms when the information is used in real-time early warning systems.*"

(B) We move to the study case (section 5) our subjective choice in which W* is defined, only for the example case, as the one that maximizes the severity, rephrasing it as follows: ***"In a univariate framework, a possible choice of a return period T^* representative of a rainfall event could be the maximum among the return periods T_W associated with all possible temporal scales $W\le D$: $T^*=\mathrm{max}(T_W)$. Here we will provide an example in which the triggering time interval W^* is assumed to be the***

time interval during which the most severe intensity was observed. It is important to note that this W^* also depends on these user-defined parameters, because temporally-close heavy events may or may not be aggregated into one depending on these choices. For example two convective events occurring at a distance of 12 hours one from the other may be considered as one event when events are separated by dry periods of 24 hours and as two distinct events if separated by dry periods of 6 hours. Notably, \citet{marra2020} showed that once independence is granted, this definition allows one to directly link the statistics of the event maxima to the statistics of the annual maxima, thereby removing the ambiguity of rainfall event definition from IDF calculations."

Moving to section 5, I have some comments regarding the dataset used.

(10) First, it should be noted (and somewhere acknowledged by the authors) that the dataset is quite dated, having been collected over ten years ago.

Thank you for this observation. We added a sentence to section 5 to clarify this aspect: "***The dataset dates back over 10 years, but it remains among the few available datasets with quality-controlled continuous high-resolution rainfall estimates for many known landslides/debris-flows.*** "

(11) Second, spatial and temporal information of the debris flow records is missing. In particular, authors should specify whether the time of occurrence is known for the debris flows included in the dataset used. This is extremely important information for a dataset to be used for the definition of rainfall thresholds. Moreover, it is relevant for another issue that I write further on in my comment.

Thank you for this comment. Indeed, information on the exact time of triggering was not available for all events, unfortunately. We amended section 5 to include it: "*High-quality weather radar observations with resolutions of 1 km and 5 minutes are available for these events from Marra et al. (2014)****, but only information on the day of occurrence of the debris flows was available, while the exact time of occurrence was unknown.*** *[...].* ***Since information on the exact triggering time was not available, we used as D*** *the total length D of the precipitation event and the average precipitation intensity I observed over this time interval (duration concept of ID thresholds).*" For what concerns the spatial information, we prefer to leave the reference to the studies in which the dataset was presented, as it is not important for our example and conclusions.

(12) Third, it is not described how the triggering precipitation events used to draw the thresholds were defined. This is also very relevant, given the comparison with IDF curves done in the paper.

Thank you for pointing out this critical aspect. We included the information as follows: "***We identified the triggering events isolating them with at least 24 dry hours.***"

(13) Further on in section 5, the authors describe the procedure used to calculate W* (lines 88-92). It should be acknowledged that the outcomes of this procedure are not related to debris flow triggering. Indeed, the fact that they have the highest return time among all IW pairs does not mean that they triggered debris flow. It would be useful to know when these IW* pairs occurred within the whole event duration, in order to establish whether they are relevant to the triggering of debris flows or not. If the IW* pairs occurred many hours (or days) before the occurrence of the debris flows, it cannot be said that they were certainly relevant to the initiation; at least, not more important than the entire event. This is the reason why knowing the exact time of occurrence of the debris flows is essential to prove that "what is really important for triggering are the rain intensities over time scales that can be much shorter than the total length of the identified rainfall event in combination with the hydrological antecedent conditions". In my opinion, selecting IW* pairs using the maximum return time as the only constraint is not sufficient to prove this hypothesis, and adds subjectivity in the process.

Thank you for this comment, which makes some very good points. The referee is correct: (i) the procedure is unrelated to the triggering, and (ii) the interval W* over which T is maximized could be unrelated to the triggering as well (as a matter of fact, in our study case it could even happen after the triggering, given that we are considering the entire event and we only have information about the day of the triggering. The objective of our data analyses is to show there is an inconsistency and demonstrate that an objective definition of W* leads to consistent representations of probability in the ID (or IW) space and we did not intend to provide a framework to assess these probabilities or to define thresholds.

These are important aspects that should be better explained in the manuscript, and we did extensive edits to the storyline. As detailed in our reply to comment 9 above (formal definition of W* in section 4 and practical definition for the example case).

Further, we rephrased section 2 to clarify some aspects and put more weight on the difference between unknown and known triggering times (note that knowing the triggering times helps in practice, but does not change our theoretical arguments) as follows: "*Following pioneering work by \citet{Caine1980}, the intensity I and the length* ***D*** *of the rain* ***period that led to the triggering*** *began to be used to determine the triggering conditions.* ***Often, the total precipitation depth (E) is used instead of the intensity, with no difference in the generality of our arguments since depth and intensity of any***

event are directly linked ($I=E/D$). *Therefore, the \emph{duration} D in this ID space is defined as the length of the wet period (that is, a user-defined event) that leads to the triggering, and the intensity I refers to the average rain intensity observed during this period. Although landslides* ***can be*** *triggered by periods of high intensities that occur within rainfall events \citep{DOdorico2005, Moreno2025}, the entire length of the events****, or the length until the triggering time, if known,*** *is used to build ID thresholds*".

Last, to fully clarify that we are dealing with a hypothetical case, we rephrased the part of section 5 highlighted by the referee as follows: "*[...] and* ***suggests*** *that what is really important for triggering* ***can be*** *rain intensities over time scales that can be much shorter than the total length of the identified rainfall event****, as they are related with the response time of the system*** *in combination with the hydrological antecedent conditions. Indeed, for the 133 debris flows examined, the most severe intensities were observed for temporal windows W between 30 minutes and 6 hours (Fig \ref{fig:f2}a). The severity on these time scales is about an order of magnitude higher than at other windows.* ***Interestingly, these durations align with the critical durations for runoff generation in the catchments of the study area \citep{Penna2017}, where intense runoff is indeed the triggering mechanism of debris flows****. These are* ***also*** *the time scales of convection …"*

(14) Then (lines 98-99), the authors write that "IW* pairs are associated with temporal scales W* that are always smaller than the duration D of ID pairs. In addition, by design, the corresponding intensities are systematically higher". This is tautological and led to what is written in lines 109-111 (i.e., the underestimation of the return times of the whole events compared to the IW* pairs). Again, having a lower return time does not imply that an ID pair is less severe in terms of landslide/debris flow triggering. This is another point to be added in the conceptual difference between ID thresholds and IDF curves.

Thank you for raising this consideration. The aim of our case study is to demonstrate with some numbers the impact of the misconception, and to introduce a possibly questionable but objective way to extract intensity-duration pairs (IW*) that belong to well defined populations. Indeed, using such pairs the mismatch between ID thresholds and IDF curve scaling shown in Bogaard & Greco, 2018 (reference in the manuscript) is solved. We believe that the restructuring detailed in our replies to comments 9 and 13 fully addresses this concern.

(15) Moreover, the authors assumed that ID thresholds are always defined considering D as the whole duration of the rainfall events. This is not always true. There are several examples in the literature in which sub-events are distinguished (automatically or not) within the entire rainfall events and used to define rainfall thresholds. This can be considered a

solution to the issues about durations being too long. I'd suggest mentioning it in the discussion.

Thank you for pointing this out, we amended the text in several places to address this issue. We'd like to underline that, while the practical issue is important, it has no implications toward our discussion.

(16) Before moving to the conclusions, two comments on Figs. 2 and 3. In Fig. 2, the (a) and (b) labels are missing. Fig. 3 and its description are not very clear; a better description a more discussion are needed.

Thank you for the suggestion, we added the labels to Figure 2. The caption of Figure 3 has been updated to: "Temporal autocorrelation of the precipitation time series for the 133 debris-flow triggering events. **Red and blue shaded areas show the 50\% and 90\% ranges across the 133 debris flow triggering events. The horizontal boxplot shows the distribution of the decorrelation times, calculated as the lag time at which the autocorrelation drops below e^{-1}.** "

(17) Going to the conclusions of the work, I totally agree that the calculation of return times of triggering conditions should be avoided, for several reasons including the ones described by the authors. However, the main motivation should be that it's better to use statistical/probabilistic approaches to define rainfall thresholds rather than calculating return times of the triggering conditions.

Our objective of this brief communication is to highlight a misconception in ID-IDF interpretation and consequently, we conclude that ID thresholds cannot be used to assess return times of triggering conditions. However, we do not aim to discuss how we can 'best' calculate landslide/debris flow activation probability. In short, the motivation proposed by the referee to also discuss statistical/probabilistic approaches to define rainfall thresholds is beyond the scope of the brief communication, although we look forward to a scientific debate on that.

(18) Moreover, the underestimation of the return periods should be better evaluated considering the time of occurrence of the IW pairs and landslides/debris flows.

Thank you for this suggestion. We rephrased this part of the conclusions as follows: *"Estimating the probability of occurrence of these triggering conditions using IDF curves on ID pairs may cause an underestimation of the rainfall return period,* ***especially when the***

exact time of triggering is not known and the entire event duration is used. This may *lead to false alarms in early warning systems that operate in real time."*

(19) Overall, I think that the main message of the work is clear and shareable. However, I believe that the conclusions would need results based on an accurate dataset and improved methodology. In my opinion, more temporal details on the dataset are needed, in order to allow the most important methodological improvement needed in the work: that is, find the time of occurrence of the IW* pairs and their temporal distance from the debris flow occurrences. Only in this way will the conclusions be adequately justified by the data and results. So, my suggestion is that the work needs major revisions before being reconsidered for publication. The revised version of the paper should include an analysis of the temporal instants of the IW* pairs, so as to say with certainty that they can be considered the cause for debris-flow-triggering. This may be done using information from the proposed dataset (if any) or using other datasets. Moreover, I'd kindly suggest taking into consideration all my comments regarding theoretical and methodological aspects of the work.

Thanks again for the time devoted to our work and for the useful suggestions. We believe the amendments we proposed to the manuscript allowed us to clarify several important points and address the referee's concerns.