Referee #2

We provide here our reply to the comments by Anonymous referee #2, 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.

Thank you for the opportunity to review Marra et al., "Brief communication: Threshold not probability. The conceptual difference between ID thresholds for landslide initiation and IDF curves." (egusphere-2025-3378). In this contribution, the authors explore the conceptual difference between rainfall intensity-duration (ID) thresholds for landslide initiation, which are conventionally fit to ID pairs that consider average intensity over entire landslide-triggering rainfall events with the intention of identifying conditions under which landslides are more likely, and intensity-duration-frequency (IDF) curves, which are fit to ID pairs that consider annual maximum average intensities for windows of defined duration with the intention of estimating annual exceedance probabilities. The authors argue that, because the definition of duration is different, these two curves are not comparable and IDF curves should not be used to estimate the exceedance probability of landslide-triggering rainfall. They use an example dataset of debris flows from the eastern Italian Alps to compare the implications of using the conventional approach based on the entire event duration to define I-D thresholds and an alternative approach that selects the duration with the maximum return period during an event. They show that return periods are much higher for the window with the maximum return period during the event than for the whole event. They also show that the slope of an ID threshold that uses the alternative approach better matches the slope of the regional IDF curves.

(1) Overall, this brief communication is well-written, thought-provoking, and has caused me to reconsider some results of my own research. It points out some important issues with ID thresholds that will be instructive for landslide researchers. In my view, the key contributions are (1) the clear explanation of how ID thresholds and IDF curves differ conceptually, (2) the insight that the return period of the average intensity over an entire event is the lowest possible return period and that much higher return periods may exist for shorter periods within an event, and (3) the recognition that if landslide triggering rainfall events are sampled with duration windows akin to the blocks used to determine IDF curves, the slope of ID threshold matches the regional IDF curves, at least for the case study presented. I believe that points (2) and (3) could be further emphasized in the text and should included in the abstract.

Thank you for the time devoted to our manuscript. We are glad to hear that our contribution was appreciated and considered worthy of publication.

Concerning the inclusion of points 2 and 3 in the abstract, we believe they should not be part of the main message of our communication, because:

- Point 2 is not necessarily true. It is generally likely, but counterexamples can be found, e.g. a synthetic event with two strong hourly peaks (e.g, 100 year return period each) separated by a dry hour. It is likely that the 3-hour return period will be the largest
- Point 3 is correct but including it in the abstract would require substantial background information that cannot be included in an abstract e.g., what do we mean by 'duration windows akin to the blocks used to determine IDF curves' and what the slope of ID and IDF are. Point 3 is, however, addressed in section 5, where we state: *"These latter threshold nicely aligns with the regional scaling of extreme rainfall (dashed-dotted lines in the background), solving the apparent difference in the power-law scaling of ID thresholds and IDF curves discussed by \citet{Bogaard2018}.* ***This result confirms that this apparent difference can be attributed to methodological issues in the choice of rain duration, often made regardless of the physical processes responsible for debris flow or landslide occurrence.*** *In general, IW* pairs are associated with temporal scales W^* that are always smaller than the duration D of ID pairs".*

Because this piece is likely to serve as a primer on this topic for future researchers, there are some arguments that need a more nuanced explanation and it must be made clear which points are the author's opinions and which are supported by the evidence presented. Additional references are needed throughout. In particular:

(2) The authors make the arguments that "IDF curves should not be used to quantify the exceedance probability of ID thresholds" (Lines 5 – 6) and "it is therefore erroneous to quantify the return period Td of a given intensity I in the ID space using probabilities estimated from the IW space of IDF curves." (Lines 70 – 71) From my perspective, it is not wrong, if one has made the conventional choice that W=D, to look up what the exceedance probability of that ID pair is. The key issue lies in making the choice that W=D in the first place, as this choice obscures shorter periods of high intensities that may have much lower exceedance probabilities, as shown in the case study. This distinction needs to be made very clearly. The "should" in the first statement and the "erroneous" in the second are based on the opinion that it would be better to use W* to define the exceedance probability of the event than choosing W=D. While I tend to agree, this short paper does not present evidence that W* is a better predictor of triggering rainfall than W=D, so it needs to be clear that this is the authors' opinion.

We respectfully disagree with the referee on this point. The choice W=D will lead to the inconsistency. We will try to better detail why.

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).

One can choose W=D, but the statistics of W are the same no matter how events are defined, but the statistics of D depend on when the event starts. Let's see an example: we have one wet hour with 10 mm, then 12 dry hours and then 1 wet hour with 18 mm that leads to a landslide. If we separate our events with 24 dry hours, the two wet periods would constitute one event with D=14h and I=28mm/14h=2mm/h. If we separate the events with 6 dry hours, the two wet periods would be two distinct events, and the triggering ID pair would be D=1h and I=18mm/h. We can continue the reasoning with different separations without finding a solution.

We amended the text in section 4 to emphasize this critical aspect: "*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 [...]*"

Further, we added new text in section 5 as follows: "***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.***"

It follows that our "should" and "erroneous" mentioned by the referee are not based on an opinion, rather on the difference between probability concepts highlighted above.

(3) The analysis shows that an ID threshold fit to IW* pairs better matches the slope of the regional IDF curves than a conventional threshold and the authors argue that this solves

“the apparent difference in the power-law scaling of ID thresholds and IDF curves discussed by Bogaard and Greco (2018).” This is an interesting result, my interpretation of which is that when time series of debris flow triggering rain are sampled with W*, the method is similar enough to using block maxima that the distribution of extreme rainfall is similarly represented. That would suggest that the difference between ID and IDF slopes can be attributed to methodological differences in how rainfall time series are sampled rather than any physical processes. If the authors agree that this is the case, I recommend making this point explicitly to avoid any further confusion. But then I have to wonder – what about filling-storing-draining?

Thank you for this thought-provoking point. Replacing D with W* leads to events that belong to the same type of population, so that they align with the slope of the IDF curve. This is an argument from a purely statistical viewpoint and as such, correct. This particular result comes from the fact that the maxima over a duration W of independent events share the tail statistics of the annual maxima (see Marra & al 2020, cited in the discussion paper).

Clearly, this brief communication focuses on the incorrect link between the statistics of rainfall time series versus the statistics of the amount of water needed to (re-)activate a landslide. However, the hydrological cause is – in a physical sense – responsible for the variation in meteorological triggers for initiating landslides. Filling-storing-draining is a conceptual framework for the physics of delayed response, variation in threshold volumes of water and the timing of landslides. Phrased differently: The probabilistic mismatch between IDF and ID threshold has nothing to do with the physical interpretation of the landslide triggering thresholds, but at the same time it depends on the physical processes behind the ID (or depth-duration ED) pairs. The point here is that the exceedance probability of an ID pair, even assuming it can be unambiguously calculated, is not the probability of a landslide or debris flow triggered by a rainfall of duration D. This latter, in fact, depends also on the physical response of the slope (as we already pointed out in our reply to a previous comment, it is indeed a conditional probability). As the focus of this brief communication is on the statistical consequences, we then made a few changes/additions to the text to avoid misunderstanding about the role of the physical processes.

We made this aspect clearer right from the beginning of the introduction, as follows: *“These thresholds are derived from landslide archives and rainfall observations with the aim of separating triggering and non-triggering events on the basis of their duration and average intensity \citep{Leonarduzzi2017}.* ***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 the standard tool to calculate the annual exceedance probability of extreme rainfall intensities over a duration of interest \citep{Kottegoda2008}.”*

We amended the text to explicitly mention the probabilistic mismatch 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."

Further, following the referee's suggestion, we included a comment on the slope, as follows: *"These latter threshold nicely aligns with the regional scaling of extreme rainfall (dashed-dotted lines in the background), solving the apparent difference in the power-law scaling of ID thresholds and IDF curves discussed by \citet{Bogaard2018}.* ***This result confirms that this apparent difference can be attributed to methodological issues in the choice of rain duration, often made regardless of the physical processes responsible for debris flow or landslide occurrence.****"*

(4) The authors note that corresponding intensities for W* are systematically higher than W=D, which they argue "Implies that what is really important for triggering are 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" (Lines 99 – 102). I do not understand how the first point implies the second. There is a logical gap here that needs to be addressed.

Following the recommendations from referee #1, we softened this sentence to fully clarify that we are dealing with a hypothetical case: "*[...] 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 …"*

I have some additional suggestions that I believe could make the manuscript more instructive, particularly for readers who are less familiar with ID, IDF, or both:

(5) In Figure 1, I suggest labeling the IDF scaling lines with return periods to make it more clear what these refer to.

Unfortunately this cannot be done because the 133 debris flows examined occurred in different places, where the statistics of extremes differ.

(6) I also suggest adding a panel to this figure that shows a time series of a debris flow triggering event with windows that show W* and W=D and the average intensities and their return periods over each of these windows. This will help readers to better understand the difference between the ID pairs and IW* pairs.

Thank you for the suggestion, we included a new panel in figure 1 as recommended.

(7) As an outlook in the conclusions, the authors may want to consider mentioning the variety of alternative approaches to determining thresholds or estimating continuous probabilities that are better able to capture intense periods in landslide triggering time series than averaging over the entire event. For example, both (Staley et al., 2017) and (Patton et al., 2023) compared models trained with accumulations over different windows to select a model that best separated triggering from non-triggering events for post-fire debris flows in the western United States and shallow landslides in Alaska. The (Moreno et al., 2025) study that was already cited is a nice example of how we can move away from the need to bin time series entirely.

Thank you for this interesting suggestion. We added a sentence to the conclusion in this regard: ***"Alternative approaches to the definition of landslides triggering conditions which may be able to better capture intense rainfall periods during triggering events \citep[e.g.,][]{Staley2017,Patton2023,Moreno2025} may be useful in this sense."***

Line by line comments:

Line 9 – suggest citing (Guzzetti et al., 2020)

Thank you, we added the reference.

Line 58 – this statement needs a reference

The references to this were mentioned in the introduction, including references to works we authored (line 15 of the discussion paper). We prefer not to provide a list of other instances to avoid the feeling we want to challenge the results of other researchers. It is likely that in many cases only minor parts of the publications may be affected by this issue, and we trust that the reader will have enough information to understand which parts, should there be any, of a paper of interest could be less reliable.

Line 63 – this statement needs a reference and possibly more context. Is this choice a convention in meteorology or is this an argument that the authors are making here?

Thank you for raising this good point. This is a subjective argument of ours. 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."

Line 72 – Please add ~one sentence clarifying how this leads to false alarms.

We rephrased the sentence as follows: "**As, by definition, events with lower severity are observed more frequently than events with higher severity, this underestimation may lead** to false alarms when the information is used in real-time early warning systems"

Line 74 – this statement needs a reference

The sentence will be removed.

Line 81 – How did you select these 12 storms as opposed to considering all debris flow triggering storms? Please clarify.

Preparing quality controlled and corrected weather radar data using physically-based corrections is a time consuming activity. These events constitute the database prepared over the course of a PhD on radar hydrology (from the first author of this manuscript). We preferred events that triggered several debris flows, for which weather radar visibility in the area of interest was good, radar data was complete and with high quality. As mentioned in the discussion paper, these storms triggered 40% of the debris flows in the region during that period, and can therefore be considered a very good sample.

Line 87 – Please add one sentence detailing how you defined the events (e.g. length of dry period). As you noted earlier, the derived ID pairs are sensitive to these choices.

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.***"

Line 94, Figure 1 – Please add an estimate of statistical uncertainty to these thresholds.

While uncertainty is fundamental in practical applications, here we are trying to communicate a theoretical point. Adding the uncertainty would not add useful information for our message.

Line 127 – surely, Moreno et al., 2025 aren't the first to point this out. Earlier reference?

We removed this reference and included the one recommended in the next comment.

Line 129 - (Iida, 1999) also noted this

Thanks, the reference was added.

References

Guzzetti, F., Gariano, S. L., Peruccacci, S., Brunetti, M. T., Marchesini, I., Rossi, M., and Melillo, M.: Geographical landslide early warning systems, Earth-Science Reviews, 200, 102973, https://doi.org/10.1016/j.earscirev.2019.102973, 2020.

Iida, T.: A stochastic hydro-geomorphological model for shallow landsliding due to rainstorm, CATENA, 34, 293–313, https://doi.org/10.1016/S0341-8162(98)00093-9, 1999.

Moreno, M., Lombardo, L., Steger, S., De Vugt, L., Zieher, T., Crespi, A., Marra, F., Van Westen, C., and Opitz, T.: Functional Regression for Space-Time Prediction of Precipitation-Induced Shallow Landslides in South Tyrol, Italy, JGR Earth Surface, 130, e2024JF008219, https://doi.org/10.1029/2024JF008219, 2025.

Patton, A. I., Luna, L. V., Roering, J. J., Jacobs, A., Korup, O., and Mirus, B. B.: Landslide initiation thresholds in data-sparse regions: application to landslide early warning criteria in Sitka, Alaska, USA, Natural Hazards and Earth System Sciences, 23, 3261–3284, https://doi.org/10.5194/nhess-23-3261-2023, 2023.

Staley, D. M., Negri, J. A., Kean, J. W., Laber, J. L., Tillery, A. C., and Youberg, A. M.: Prediction of spatially explicit rainfall intensity–duration thresholds for post-fire debris-flow generation in the western United States, Geomorphology, 278, 149–162, https://doi.org/10.1016/j.geomorph.2016.10.019, 2017.