Manuscript reference number: hess-2017-308 - Response to Anonymous referee #1

We would like to thank the referee for his review and for the helpful comments. We provide here a response to his comments together with our proposed edits to the manuscript. The referee's comments are reported in black and denoted as RXCY where X is the reviewer number and Y is the corresponding comment number whereas our response is in blue.

Consistently with its title, the paper analyses how the spatial scale of aggregation of rainfall can influence the determination of intensity-duration debris flow occurrence thresholds. It distinguishes between the two cases of (i) regular grids, and (hypothetical) (ii) rain gauge networks. It capitalizes on a data set of 1 km/5 min radar rainfall and 99 debris flow events. The paper fits within the scope of HESS, it is well written, scientific questions are clear and relevant, and conclusions are supported by the results.
Nevertheless, the paper may benefit from a more in-depth analysis regarding the available methods for threshold determination, as I describe in the first point of the "specific comments". For this reason, I suggest moderate revisions for the manuscript to be finally published in HESS.
We would like to thank the referee for his review.

R1C1
Section 3: The so-called frequentist method for threshold determination originally proposed by Brunetti et al. (2010), involves only triggering events. Relatively recent research has highlighted the importance to take into account also non-triggering events. To consider only triggering events generally brings to thresholds that are lower than many non-triggering events, and thus a high number of false alarms, which may generate a disbelief in the early warning system (e.g. Berti et al., 2013 doi:10.1029/2012JF002367). The authors should at least should discuss the drawbacks of the method proposed by Brunetti et al. (2010). Possibly, the authors should add to the paper the same analysis they have conducted, but for the case that threshold determination is conducted by taking into account both triggering and non-triggering events.
Thank you for bringing up this point. Oftentimes, the frequentist method is thought to have issues with false alarms. However, in our view, we should more appropriately talk about positives/negatives rather than 'alarms'. In fact, going from 'thresholds' to 'alarm' requires the definition of rules, that represent a complicated and often overlooked issue (see e.g., Piciullo et al. 2017, doi:10.1007/s10346-016-0750-2, and references therein). In addition, unfortunately, an unknown number of so called 'non-triggering' events may actually have caused landslides that simply went unnoticed or unreported (see e.g., Gariano et al. 2015, doi:10.1016/j.geomorph.2014.10.019). This is a well-known problem for non-instrumental measures such as 'landslide' vs. 'no landslide'. For this reason, using 'non-triggering' events may result in too high thresholds and consequently in an increased number of false negatives which, in an alarm system, are generally much worse than false positives.
In addition, in operational environment, the spatial-temporal characteristics of rainfall are rarely considered. Conversely, our research (here and in Marra et al. 2016, doi:10.1016/j.jhydrol.2015.10.010) confirms the presence of (spatial) non-stationarity of the rainfall fields around the debris flow triggering locations, meaning that the spatial organization characteristics of triggering and non-triggering rainfall are completely different (see Fig. R.1). It would be thus misleading to include such different patterns in the analysis.
To conclude, given the generality of the findings ("qualitatively transferrable to those situations in which lower envelop curves are used to predict the occurrence of point-like events in presence of non-stationary fields" [P6L26]), we think that using an (i) objective and (ii) straightforward method, such as the frequentist, is crucial to allow a clear interpretation of the study.
As suggested by the referee we think it is worth discussing this choice in the updated manuscript: "*In operational environment, the spatial-temporal characteristics of rainfall have rarely been considered,*

despite the observed non-stationarity of the rainfall fields around the debris flow triggering locations reported by Marra et al. (2016). Since our objective is to analyse the impact of this non-stationarity on the use of spatially aggregated rainfall information, it is crucial for us to focus on the triggering events, i.e. on the events in which the systematic spatial feature is observed, and to use an objective and straightforward method, such as the frequentist, that allows a clear interpretation of the results".

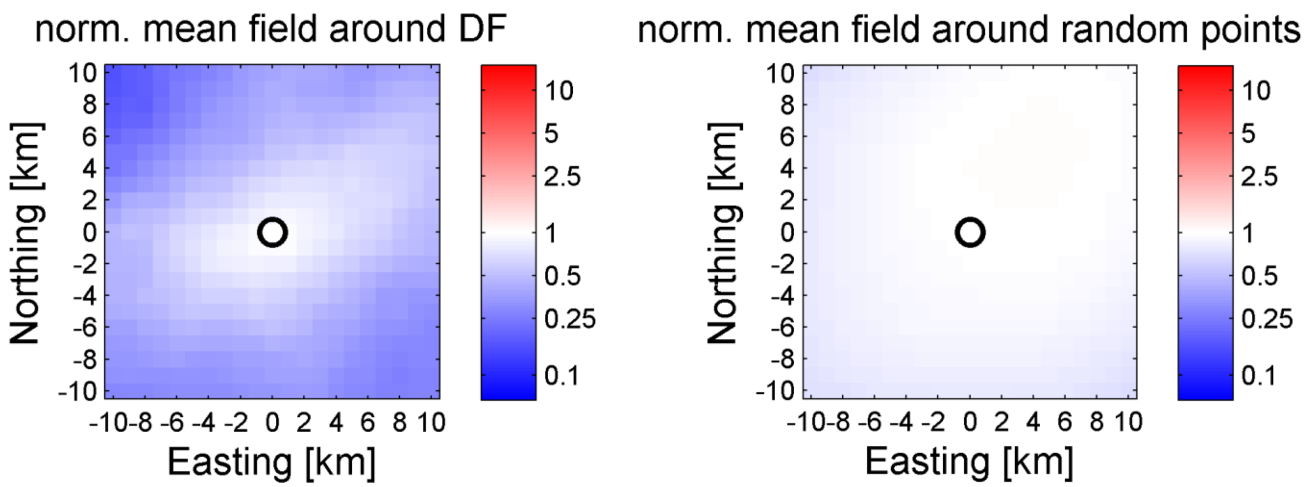

Figure R.1: mean normalized rainfall fields (i.e. the central point is 1) calculated for the events included in Marra et al. (2016) for (a) debris flow locations ('triggering') and (b) random points within the radar field ('non-triggering'). The non-stationarity of the triggering rainfall fields is clear. Note that (a) is analogous to Fig. 6 in Marra et al. (2016) with different color scale (here is log-scaled) and color bar

R1C2

P5 L4: Few details are given about the method for generating the synthetic rainfall fields. Since the method may affect the results, please provide these details.

Thank you for the question. The method is based on generating rain gauge networks (i.e. coordinates of hypothetical rain gauges) rather than rainfall fields. The synthetic rain gauge estimates are defined as the radar measurements on the corresponding pixels (i.e. the pixels containing the location of the rain gauge). The approach of using the radar rainfall fields as the 'true' rainfall fields follows exactly what was done for the analysis of spatial aggregation. The triggering rainfall is then defined by the measurement of the rain gauge closest to each debris flow (nearest neighbor 'interpolation' method). This approach strictly follows what previously used by Nikolopoulos et al. (2015, doi:10.5194/nhess-15-647-2015) and Destro et al. (2017, doi:10.1016/j.geomorph.2016.11.019). We propose to update this portion of the manuscript to improve its clarity: *"Synthetic rain gauge networks were produced using the procedure proposed by Nikolopoulos et al. (2015a) and Destro et al. (2017). The location of the rain gauges was randomly generated to obtain densities of* $1/A$*, with* A *set to 10, 20, 50, and 100* km^2*. To avoid clustering of the rain gauges, a minimum distance between two synthetic stations was set to* $0.5\sqrt{A}$*. Rainfall estimates of the synthetic rain gauges were defined as the value of the radar rainfall fields for the pixels corresponding to the simulated gauge locations. The rain gauge estimation of triggering rainfall was then defined as the value reported by the rain gauge closest to the triggering location."*

R1C3

P3 L23-24: "The severity of debris flow events was classified as mild (T<=2y, 21 debris flows), moderate (2<T< 50y, 41 debris flows)…". The authors should mention that return period of a debris flow depends in general from both initial conditions and triggering rainfall, and not only from the latter (Peres and Cancelliere 2016, http://dx.doi.org/10.1016/j.jhydrol.2016.03.036). Furthermore, intra-event rainfall intensity variability, which also affects return period, may not be properly taken into account with depth-duration-frequency curves (see D'Odorico et al. 2005, doi:10.1029/2004JF000127).

Thank you for pointing this out. The use of 'severity of debris flow events' is actually misleading, since we only refer to the return period of the triggering rainfall and do not include other sources of information. We propose to update this part of the text to: "*The severity of the debris-flow triggering rainfall was classified as…*".

Technical corrections

R1C4

P2 L8. Perhaps "variability" is more appropriate than "non-stationary"

We respectfully disagree with the reviewer on this suggestion. What is meaningful reporting from Marra et al. (2016) is the presence of systematic patterns in the triggering rainfall fields around the debris flows, meaning a non-stationarity, observed at least for the first moment. The use of 'variability' is not sufficient to explain these findings and, in fact, this observed non-stationarity is among the motivations behind this study (see the abstract: "The systematic underestimation observed in debris flows early warning thresholds has been associated to the use of sparse rain gauge networks to represent highly non-stationary rainfall fields" [P1L11]). Please, see also Fig. R.1 and the related text in the response to R1C1 for more details.