Dear Editor and Reviewers,

Thank you very much for your useful comments and suggestions.

In this document, you will find a detailed explanation of the changes made to the original manuscript to meet your suggestions.

For the sake of clarity, we used the following text styles:

black, italics:	reviewer comment
blue, plain text:	our reply
blue, italics:	revised text

Best regards

Elena Ioriatti

Mauro Reguzzoni

Edoardo Reguzzoni

Andreas Schimmel

Luca Beretta

Massimo Ceriani

Matteo Berti

Reviewer 1

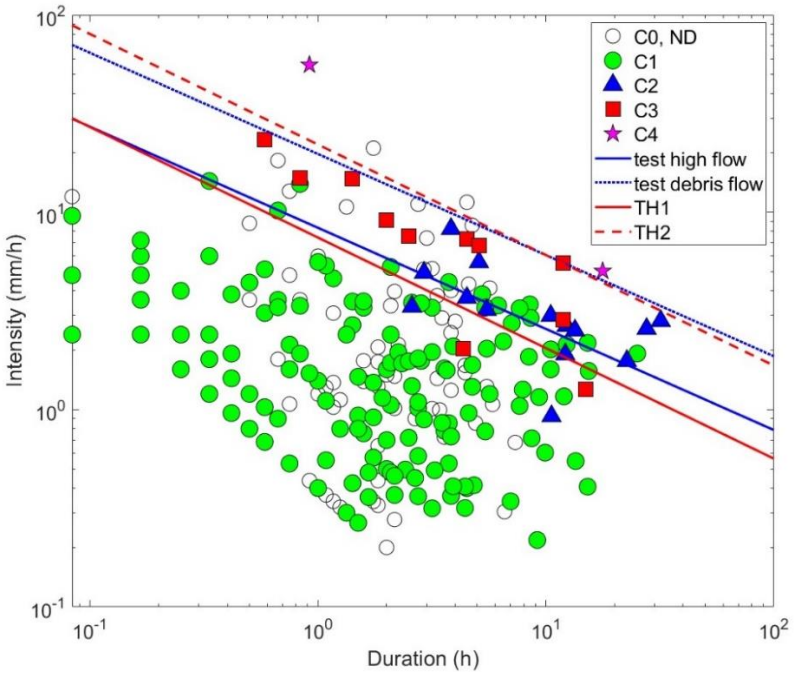
1. Line 67: the term "propose" here can be misleading since the use of non-triggering events is already established in literature. I suggest "use" or similar terms.

Thank you for your comment, we have changed *'propose'* to *'use'*. However, we wish to note that this approach is common for historical data at larger scales but remains uncommon in catchment-scale monitoring.

To overcome the limitation posed by the small number of debris-flow events, we use two complementary strategies. First, we consider not only triggering but also non-triggering rainfall events, applying statistical analysis to distinguish between the two classes. While this practice is established at regional scales, it has rarely been implemented within catchment-scale monitoring. Second, we draw on the larger set of high-flow and sediment-transport events to establish a robust lower threshold, which then serves as a reference for isolating debris-flow conditions and defining the debris-flow threshold.

2. Lines 202-205: I still think the reproducibility of the study can be affected by this approach. This should be declared in the manuscript, and the potential impacts on the results should be discussed or evaluated.

Thank you for your the comment. To assess reproducibility we performed a sensitivity analysis on 11 events with uncertain classification, nine involving the transition from low flow (C1) to high flow (C2), and two from high flow (C2) to high flow with sediment transport (C3). We reassigned each to the alternative plausible class and recomputed the rainfall thresholds; the results under this worst-case reassignment are reported in the revised manuscript. We provide below the figure showing this alternative classification and comparisons between these thresholds and TH1/TH2.



4.2 Rainfall thresholds

[…]

In addition, operator-based classification of events may have introduced uncertainty in the derived thresholds. To assess reproducibility, we ran a sensitivity analysis on 11 events whose classification was uncertain, 9 involving the transition from low flow (C1) to high flow (C2) and 2 from high flow (C2) to high flow with sediment transport (C3). Each event was reassigned to its alternative plausible class, and the rainfall thresholds were recomputed. For the high-flow threshold, the refitted parameters are slope $\beta = -0.51$ and $\alpha = 8.34$ (−7.9% and +12.0% relative to TH1). For the debris-flow threshold, the coefficient α is 19.74, corresponding to −10.5% relative to TH2 (β fixed equal to the high-flow threshold). For a 30-min storm, the associated critical rainfalls are 5.9 mm for high-flow conditions and 14.1 mm for debris-flow initiation (compared to 5.5 mm for TH1 and 16.2 mm for TH2). These uncertainties are minor relative to the overall uncertainty sources, and our conclusions remain robust even under a worst-case reclassification in which all 11 ambiguous events were simultaneously reassigned to their alternative plausible class.

3. Lines 252-256: this part is indeed an optimization (specifically, a maximization). I think the text here is still overly complex. One could write: "TH2 was set to have the same scaling exponent \beta (slope) of TH1, and the coefficient \alpha that maximises the Area Under the Receiver Operating Characteristic Curve (AUC). Although…"

Thank you for the comment, the lines have been revised in the manuscript.

[…]

TH2 was defined by keeping the same scaling exponent β (slope) as TH1 and selecting the coefficient α that maximises the Area Under the Receiver Operating Characteristic Curve (AUC). Although…

4. Figure 12: I agree with the authors when they say that the UNIBO gauge was treated as ground truth. But what I question is whether this is appropriate here, given that rainfall spatial variability plays a major role (and not only the measurement accuracy of the gauge). I think a regression method that allows for uncertainty in both the variables should be considered. There are several available. Given the situation (spatial variability is important, likely the main factor), at a first approximation one could assume the same error variance in the two variables.

Thank you for this helpful suggestion. In the revision we re-fit all comparisons using Deming regression, which allows error in both axes. Consistent with your point that the gauges have comparable accuracy, we set the error-variance ratio to $\lambda = 1$. We quantified uncertainty via a nonparametric pairs bootstrap (2,000 resamples): BCa (Bias-Corrected and Accelerated) 95% confidence intervals were computed for the Deming slope and

intercept, while percentile 95% intervals were used to form the confidence bands of the regression line. Figure 12 and the corresponding text have been updated accordingly.

5.1 Impact of spatial variability of rainfall on threshold estimation

[…]

A direct comparison between the UNIBO station and the Hortus stations located upslope (H1, H2, H3) and downslope (H4) of UNIBO provides insight into this critical aspect. Using Deming errors-in-variables regression with equal error variances on both axes (Francq and Govaerts, 2014), we compared the Hortus measurements with those from the UNIBO reference. Figure 12 presents differences in precipitation amount (a), duration (b), and mean intensity (c) reporting for each gauge the 95% confidence intervals estimated via a nonparametric bootstrap.

Precipitation totals show that the 95% confidence bands for H1 and H2 lie above the 1:1 line, whereas H3 largely overlaps the line and H4 lies below it. This pattern is consistent with an elevation effect: at higher elevations (H1, 1,330 m; H2, 1,248 m), greater precipitation depths are measured, whereas at lower elevations (H3, 770 m; H4, 695 m) totals are close to or smaller than those recorded at UNIBO (Fig. 12a).

For rainfall duration (Fig. 12b), the 95% confidence bands for H2 and H3 overlap with the 1:1 line, indicating good agreement with the reference gauge. H1 tends to overestimate event duration compared to UNIBO, while H4 tends to underestimate it. The band for H1 is also wider than the others, reflecting greater uncertainty in the relationship between H1 and the reference gauge. In contrast, H4 shows a narrower confidence band, indicating a tighter relationship with the reference gauge.

For rainfall intensity (Fig. 12c), the 95% confidence bands for H2 consistently lie below the 1:1 line, while those for H1 also fall below the line for UNIBO intensities exceeding 8 mm h^{-1}. This indicates that both H1 and H2 tend to record lower rainfall intensities compared to the reference gauge, despite generally measuring greater precipitation depths. This discrepancy is likely explained by longer event durations recorded at these stations, which reduce the computed mean intensity. H3 and H4 show intensity measurements consistent with those recorded at the UNIBO station, although the confidence bands are comparatively wide, indicating appreciable inter-event variability rather than a systematic bias.

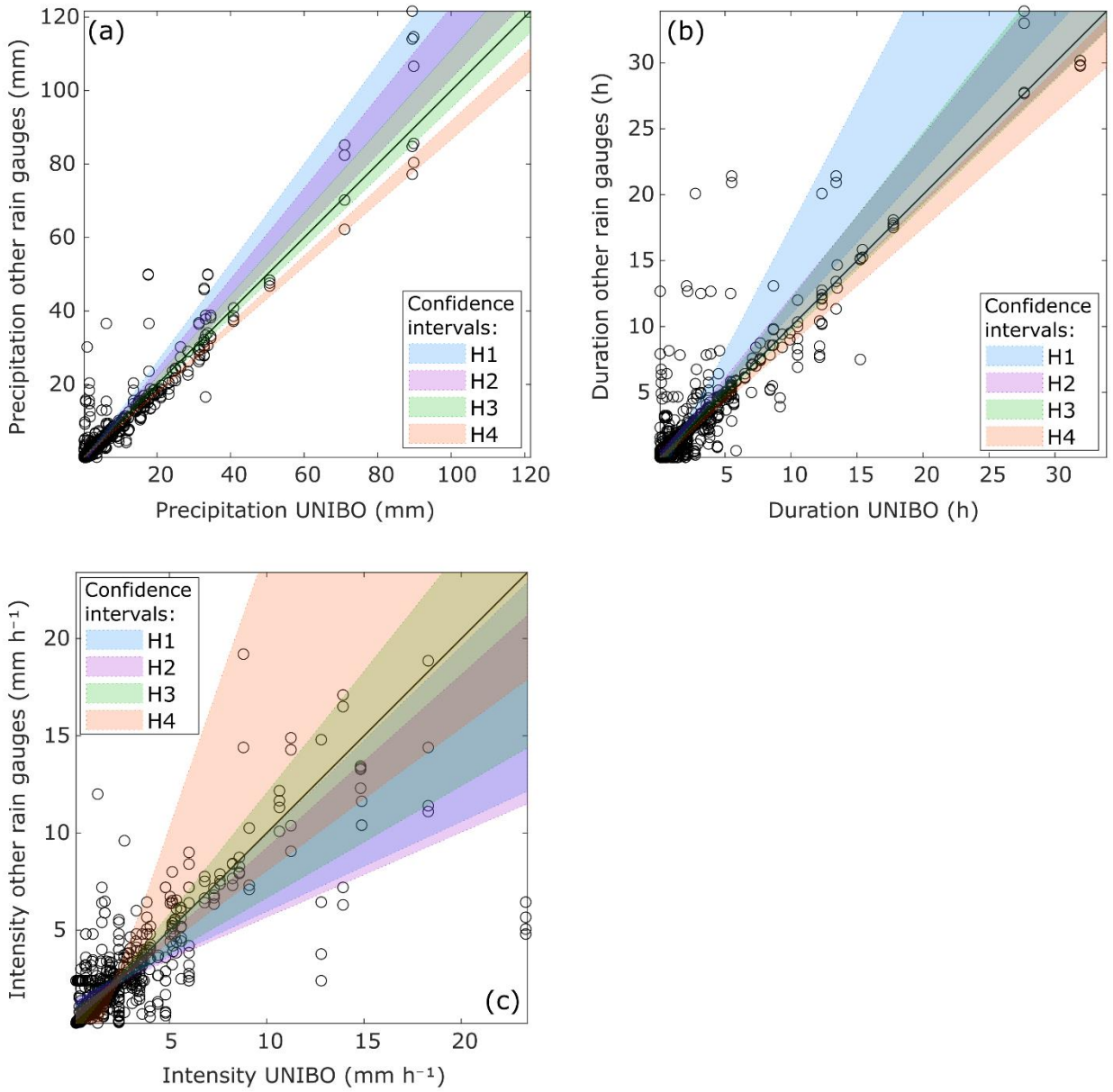


Figure 1: Comparison of rainfall event characteristics recorded at the UNIBO rain gauge and each of the four Hortus stations, where each point represents a single event measured at both locations. (a) Precipitation; (b) duration; (c) mean intensity. Shaded bands show 95% confidence intervals for the fitted Deming relationships (λ=1), obtained via nonparametric bootstrap (2,000 resamples). The black 1:1 line indicates perfect equivalence.

5. Line 441-449: please specify that the min-max value are min-max values among the ones observed by the available rain gauges (as done in the caption). Also, please specify how the UBR is obtained from these samples: is it the envelope of all the obtained values or is it some percentile interval? (Given the use of uniform distributions, I think percentile intervals should not be used here.) I suggest adding some caveats on this due to the independence assumption taken on I and D. One would expect to under-estimate variability when the two variables are assumed independent.

Thank you for the comment. To avoid this assumption of independence between D and I, we replaced the old method with a bootstrap-by-gauge scheme: for each event observed by UNIBO and at least one Hortus station,

we randomly select with replacement one of the gauges that recorded the event (UNIBO, H1–H4) and take that gauge's paired (D,I) values. This preserves the event-level dependence between duration and intensity and propagates inter-gauge variability. We recomputed TH1 for 10,000 bootstrap replicates and define the UBR as the outer envelope of all simulated thresholds (not a percentile band). Figure 13b and the manuscript text have been updated accordingly.

5.1 Impact of spatial variability of rainfall on threshold estimation

[…]

To further explore this aspect, 10,000 random simulations were performed via a bootstrap procedure with replacement: for each rainfall event recorded by UNIBO and at least one additional Hortus station, one of the gauges that recorded the event (UNIBO, H1, H2, H3, or H4) was randomly selected, and its paired (D, I) values were taken. For each simulation, the corresponding TH1 threshold was calculated using LDA, resulting in an uncertainty band (Uncertainty Band for Rainfall variability, UBR; Fig. 13b) given by the envelope of all thresholds.

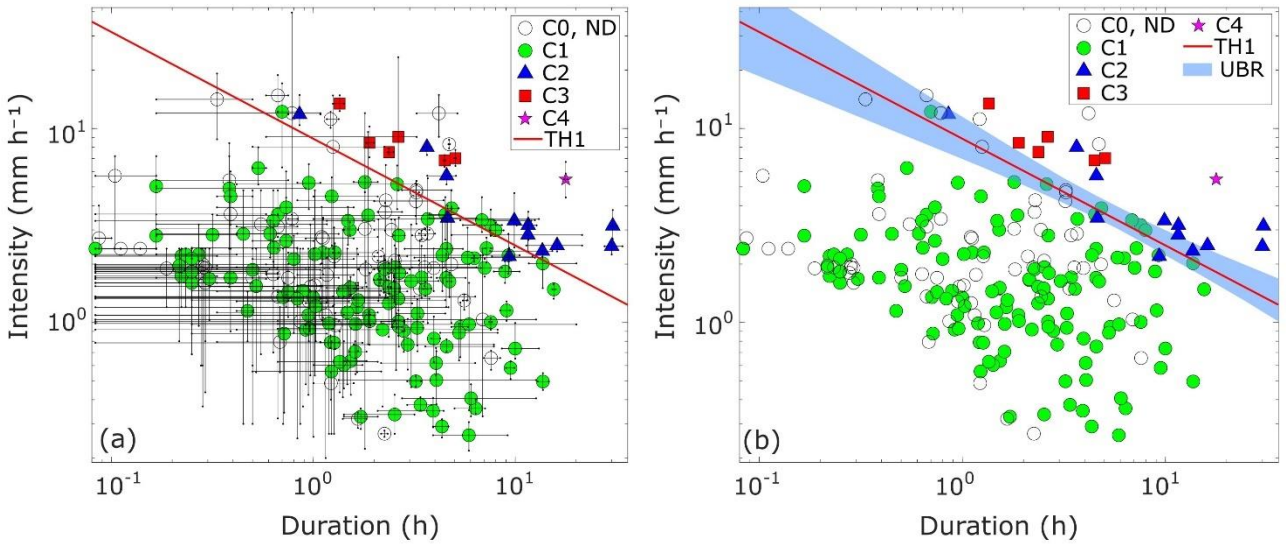


Figure 2: (a) Ranges (min-max bars) and mean values (dots) of rainfall duration and intensity for each event recorded by UNIBO and at least one additional Hortus station. Values are calculated across five stations: UNIBO, H1, H2, H3, and H4. Note that, due to the logarithmic scale of both axes, bar lengths are also represented on a logarithmic scale. (b) The blue band (UBR) represents the envelope of TH1 curves derived from 10,000 bootstrap simulations, selecting a random gauge per event and taking its paired (D, I). The red line TH1, shown in both panels, is the threshold calculated using UNIBO data for rainfall events that were also recorded by at least one Hortus station (monitoring periods 2022 and 2023 only).

6. Figure 15: the blue triangles are hard to see over the blue background, I suggest changing the color of the symbols.

Thank you for the suggestion, the colours of the figure have been changed.

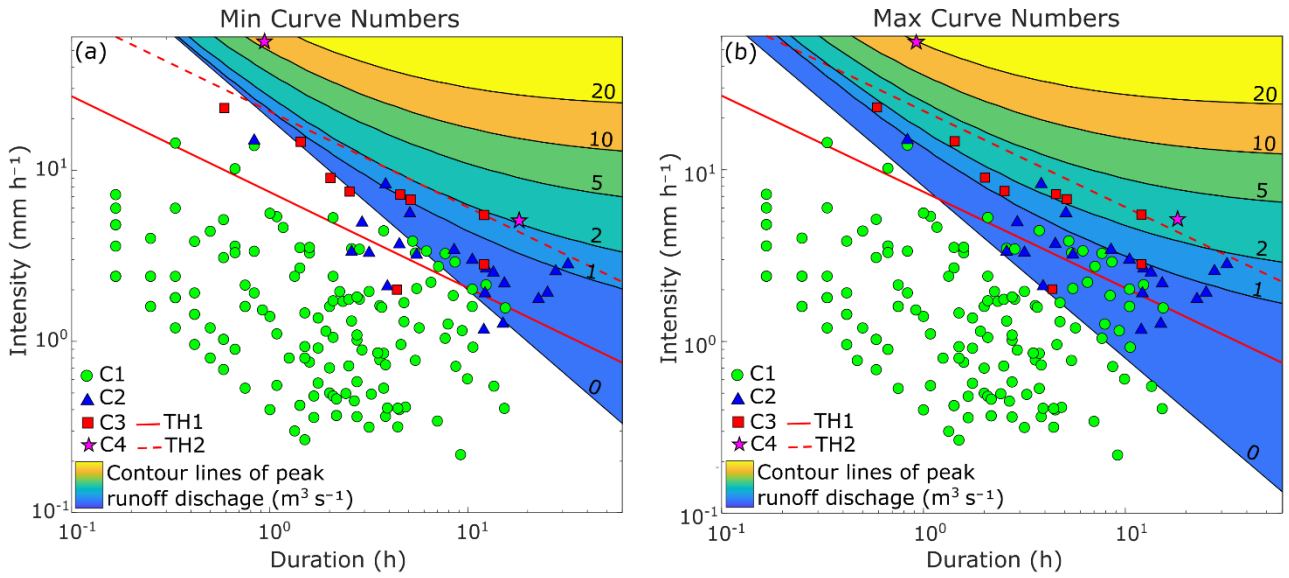


Figure 15. Contour maps of peak runoff discharge obtained with the SCS–UH method for (a) minimum CN values and (b) maximum CN values. Empirical observations of rainfall events are superimposed, with symbols indicating event classification. The comparison illustrates the sensitivity of theoretical runoff estimates to Curve Number selection.

New References:

Francq, B. G. and Govaerts, B. B.: Measurement methods comparison with errors-in-variables regressions. From horizontal to vertical OLS regression, review and new perspectives, Chemometrics and Intelligent Laboratory Systems, 134, 123–139, https://doi.org/10.1016/j.chemolab.2014.03.006, 2014.

Reviewer 2

The authors state: "To overcome the limitation posed by the small number of debris-flow events, we propose two complementary strategies. First, we consider not only triggering but also non-triggering rainfall events, applying statistical analysis to distinguish between the two classes."

However, it is standard practice—and indeed necessary—to derive rainfall thresholds using datasets that include both triggering and non-triggering rainfall events. Prior research has clearly shown that thresholds derived solely from triggering events tend to be unrealistically high. Therefore, I recommend removing the aforementioned statement to avoid overemphasizing what is already a widely accepted methodological requirement.

Thank you for your comment; we have changed the verb 'propose' to 'use' so as not to overemphasise the novelty of this approach. However, we wish to highlight that in monitored basins it is uncommon to include non-triggering rainfall events, whereas this is common when non-triggering events are drawn from larger-scale historical series.

To overcome the limitation posed by the small number of debris-flow events, we use two complementary strategies. First, we consider not only triggering but also non-triggering rainfall events, applying statistical analysis to distinguish between the two classes. While this practice is established at regional scales, it has rarely been implemented within catchment-scale monitoring. Second, we draw on the larger set of high-flow and sediment-transport events to establish a robust lower threshold, which then serves as a reference for isolating debris-flow conditions and defining the debris-flow threshold.

In the hydrological simulations (Table 5), the initial abstraction is set as Ia = 0.2S. Yet, several related studies conducted in alpine catchments—including Berti et al. (2020) and Bernard et al. (2025), both cited by the authors—adopt Ia = 0.1S. The discrepancy between these values and the authors' choice should be explicitly discussed, along with a justification for the selected parameter.

Thank you very much for this insightful comment. In the studies conducted by Berti et al. (2020) and Bernard et al. (2025), the value Ia = 0.1S was calibrated using monitored events in which the hydrological response was known, thanks to the configuration of their monitoring sites, which included a sharp-crested weir measuring discharge at the outlet of the headwater catchment. In this study, since such information was not available, the commonly used value of Ia = 0.2S from the SCS Curve Number (CN) rainfall-excess model was adopted.

5.3 Hydrological interpretation of rainfall thresholds

[…] Berti et al. (2020) and Bernard et al. (2025) used an initial abstraction Ia = 0.1S (where S is the potential maximum retention), calibrated on the known hydrological response of the catchment. In this study, the standard SCS-CN value of Ia = 0.2S was adopted due to the lack of this basin-specific hydrological information.