Reviewer 2 - Anonymous

The manuscript addresses a common issue when comparing different publications regarding rainfall events and their characteristics, namely that different authors and research groups use a wide range of metrics to characterize rainfall events. Therefore, this analysis, comparing and evaluating a range of metrics used in prior publications is a useful tool for researchers to interpret and compare results from various papers, as well as for authors and researchers to select meaningful metrics for their analysis in future work. As such this represents a valuable contribution to the field of rainfall event analysis.

However, the current manuscript contains certain limitations that require further analysis prior to publication. The decision to use 10 time points when calculating/representing dimensionless mass curves (DMCs) to compare the effect normalisation has on metrics seems flawed and represents a seemingly unnecessary own goal. Using 10 time points aggregates and summarizes the event time series, hence it is unsurprising that DMCs results for metrics that are sensitive to time discretization and aggregation show significant differences compared to the original time series data. Therefore, in its current state, the tests which are meant to examine whether normalization affects metrics seem to be flawed, as the method used both normalizes and aggregates the data. I strongly recommend that the effect of normalization be tested independently of any aggregation, by using the normalized, dimensionless mass curves at their original 5-minute resolution.

General reply - Reviewer 2

> Thank you for taking the time to review the manuscript and for your constructive comments. We are pleased that you consider it a useful resource for both evaluating existing research and informing future studies. We have carefully considered your comments regarding the limitations associated with using DMCs interpolated to ten data points. We acknowledge that differences observed when computing metrics on DMCs rather than raw rainfall time series reflect the combined effects of double normalisation and interpolation to ten points. However, we do not view this as a methodological "own goal", but rather as a deliberate methodological choice that reflects the widespread use of DMCs processed in this manner within the literature. DMCs are typically employed to enable comparison or aggregation across events of differing durations, magnitudes, and numbers of observations, for which interpolation to a common number of time steps is a fundamental step. Our literature review suggests that 10 timesteps is a frequently selected choice. We recognise that additional context on the development and typical use of DMCs would help clarify this processing decision, and we have therefore expanded the discussion in Section 2.2.
>
> While we have defended the decision to focus primarily on metric performance calculated on DMCs interpolated to ten data points due to it being common practice, we share the reviewer's curiosity regarding the influence of the normalisation step itself. Accordingly, we have run the same experiments as in the original version of the manuscript but without the aggregation to 10 time steps. These additional analyses are based

on events that have been double normalised but retain their original number of data points, thus isolating the effect of normalisation as requested by the reviewer. The revised version of the manuscript will include an extra appendix presenting additional versions of Figures 7 and 8, accompanied by text describing how these results differ from the original figures.

We also recognise that aspects of our terminology contributed to confusion on this point. In particular, we variously referred to the construction of DMCs as 'normalisation' or 'double normalisation', and to the resulting profiles as 'double normalised' or 'normalised' representations. This was intended as shorthand to avoid repeated use of the DMC acronym, which may be unfamiliar to some readers. However, to improve clarity, we have revised the language throughout the manuscript to more explicitly describe the DMC construction process.

Question 2.1. *Line 175: "based on the timing of the peak" intensity? While the context in the text makes it clear that you refer to peak intensity, it is worth being explicit and adopting this phrase here and throughout to avoid confusion with peak volume.*

Response 2.1.

Thank you for highlighting this. We agree the statement could be made more clear, and have adjusted the text accordingly.

Question 2.2. *Figure 1: The labels in the figure do not match those in the caption (b and c seem to be flipped)*

Response 2.2.

Thank you for highlighting this error. The figure has been updated and now shows the correct labels.

Question 2.3. *Line 186: I believe there is an opportunity to make this statement more forceful as categorical classifications do mask finer differences between events. From a design rainfall point of view, categorisation may mask other more relevant characteristics such as peak intensities across a range of relevant durations.*

Response 2.3.

Thank you for this comment. As suggested we have amended this paragraph to add an extra sentence more strongly highlighting the limitations of classification metrics.

The paragraph now reads: **Classification metrics assign discrete labels to rainfall events. Events are often categorised based on the fraction of the storm containing the greatest rainfall, typically a third**

(..), quarter (..), or fifth (..). While classification metrics offer an intuitive summary of storm structure, their categorical nature masks within-class variability. Events assigned to the same class may differ substantially in peak intensity, duration, or short-timescale concentration, despite sharing a similar overall classification. Furthermore, the classification assigned depends heavily on implementation choices. Firstly, how events are divided into fractions. Secondly, which aspect of rainfall is focused on. The initial third-based classification system was based on the timing of the peak, whereas more contemporary research has based classification on the third containing the highest total rainfall (Liang et al., 2023) or the third containing a summary statistic, e.g. the centre of gravity or D_{50} (see Sect. 2.3.2)

Question 2.4. *Line 276: This exclusion is a shame as event definition is probably the single most important factor in determining the ultimate event profile which is being analysed, and while peak intensity is unlikely to be affected by different start/end times, the timing, position, and relative magnitude of the peak intensity is likely to be affected, as will event metrics. Given the importance of this, a lengthier discussion is suggested*

Response 2.4.

We agree with the reviewer that event definition is a critically important factor in shaping rainfall event profiles and can substantially influence the timing, position, and relative magnitude of peak intensity, as well as derived temporal loading metrics. We also agree that different choices of event start and end times have the potential to affect many of the metrics analysed in this study.

The decision not to explicitly explore the sensitivity of temporal loading metrics to event definition was therefore not based on a lack of importance, but on considerations of scope. The influence of event boundary definition has been previously examined in the literature, with multiple studies demonstrating its effects on rainfall intensity, duration, depth, and inferred event characteristics. These studies collectively show that event definition alone can materially alter conclusions drawn from rainfall analyses.

In contrast, the focus of this paper is on a set of methodological choices that have received comparatively little systematic attention, namely the definition of the analytical objective (i.e. which aspect of temporal loading is of interest), the selection of metrics used to represent that objective, and the sensitivity of those metrics to rainfall processing choices such as temporal aggregation and normalisation. Expanding the analysis to include multiple alternative event definitions would have substantially increased the dimensionality of the problem and risked obscuring the specific effects this study seeks to isolate.

To address the reviewer's concern, we have expanded the discussion in the manuscript so the paragraph now reads as follows:

To ensure event independence, we extract events using a minimum inter-event time (MIT) threshold (Restrepo-Posada and Eagleson, 1982; Molina-Sanchis et al., 2016). An 'event' thus constitutes any rainfall separated by at least 11 hours of rain-free conditions, following practice in several Danish hydrological studies (Gregersen et al., 2013; Thomassen et al., 2023). This approach ensures that each event begins and ends with non-zero rainfall. The choice of MIT has been shown to play an important role in determining both the number and properties of rainfall events identified (Dunkerley, 2008). In this study, event definition is treated as a fixed preprocessing choice rather than a variable of investigation, reflecting a deliberate scoping decision. While the delineation of event boundaries can influence the timing and relative prominence of peak intensity, and hence derived temporal loading metrics, its effects have been examined in several previous studies, e.g. Dunkerley (2008, 2010, 2015); Wang et al. (2019); Freitas et al. (2020); Molina-Sanchis et al. (2016); Haile et al. (2011); Medina-Cobo et al. (2016); Meier et al. (2016). In contrast, this study focuses on methodological choices that have received less systematic attention, namely the selection and interpretation of temporal loading metrics and their sensitivity to rainfall representation and aggregation.

Question 2.5. *Line 287: Was the impact of using 10 time points examined, say, against using the DMC calculated using the original 5-min resolution? In my experience the number of points used can significantly alter summary metrics if they are calculated using this data rather than on the original measurement intervals*

Response 2.5.

This point relates directly to the broader concern addressed in our response to the general comments regarding the use of DMCs interpolated to ten time points. As discussed there, differences observed when calculating metrics on DMCs reflect the combined effects of double normalisation and interpolation, both of which are intrinsic to how DMCs are commonly constructed and used in the literature.

In the original manuscript, we did not explicitly isolate the effect of interpolation by comparing ten-point DMCs against double-normalised events retained at their native 5-minute resolution. In response to this comment, we have now undertaken this additional analysis and present the results in a new appendix, including supplementary versions of Figures 7 and 8. These allow the specific influence of the interpolation step on metric behaviour to be assessed and are discussed briefly in the accompanying text.

Question 2.6. *Line 345: Continuing the point made above, this seems like an issue that could have been solved using a number of time points that is divisible by 4.*

Response 2.6.

We thank the reviewer for this suggestion. As discussed in our broader response on the choice to interpolate DMCs to ten points, this decision was made to provide a standardised representation across events of different lengths, magnitudes, and numbers of observations. We specifically chose ten points to follow common implementations of DMCs; in our review of the literature, we have not seen DMCs constructed with only four points. While using a number of points divisible by four would address the calculation of the 'Fraction. in Q1–4' metrics, it would not resolve similar issues for other metrics, such as those based on fifths or other fractional divisions. Creating multiple custom versions of DMCs to suit different metrics would defeat the purpose of having a consistent, generalisable implementation. Hence, interpolation to a shared number of points remains the most consistent and practical approach. This rationale is now clarified in the revised manuscript in Section 2.2.

Question 2.7. *Section 4.2: Again, the decision to limit DMCs to 10 points is shown to strongly influence results, I suspect that several of the discrepancies arise more as a result of the use of 10 points than of the normalisation, after all, events can be normalised and presented at their original resolution, which would smooth out the histograms shown in Figure 8.*

Response 2.7.

We thank the reviewer for this observation. As noted in our broader response on the use of DMCs interpolated to ten points, it is correct that the differences observed reflect the combined effects of double normalisation and interpolation. To clarify the specific influence of using ten points, we have included additional analyses in a new appendix, where DMCs are double normalised but retain their original 5-minute resolution. These supplementary figures allow the smoothing effect of interpolation on the histograms in Figure 8 to be directly assessed.

Question 2.8. *Figure 7: The text in the image is small and difficult to read.*

Response 2.8.

Thank you for highlighting this. We have reformatted the figure so that the text is larger and more accessible to readers.

Question 2.9. *Line 338: The effects discussed in this and the following paragraph are compelling.*

Response 2.9.

We assume that this comment is intended to refer to L438, rather than L338. If so, we thank the reviewer for the positive comment and are pleased that the discussion of DMC transformations and their implications is found to be compelling.

Question 2.10. *Figure 9: As with Figure 7, the text is small and difficult to read.*

Response 2.10.

Thank you for highlighting this. We have reformatted the figure so that the text is larger and more accessible to readers.

Question 2.11. *Line 475: Indeed, it is easy to imagine cases where peak intensities are timed quite differently from the storm's centre of mass.*

Response 2.11.

Thank you. We agree, but also think that this distinction is often lost in the broader literature, when summary terms such as "front-loaded" are used in place of more specific descriptors. This can lead to misinterpretation, where metrics measuring different aspects are treated as equivalent because they are subsequently described using the same summary term. Highlighting this distinction is therefore one of the key motivations for our systematic metric evaluation.

Question 2.12. *Recommendations: The recommended metrics should be highlighted and listed within the text. This is probably the most important contribution of this manuscript, and it should be made more clearly.*

Response 2.12.

Thank you for this suggestion. We agree and have added the following sentence to the first paragraph of the recommendations: **"The recommended metrics are: the 4th with most rainfall and D_{50} (mass timing); the peak-position-ratio (peak timing); Gini coefficient (magnitude concentration); Temporal standard deviation (temporal concentration); and the event dry ratio (intermittency)."**

In response to reviewer 3's question 3.4, we further add the equations for the recommended metrics in the same section.

Question 2.13. *Summary and conclusions: This section is currently a bit long and repetitive, and some aspects would be better discussed elsewhere (for example, the sMAPE limitations could be moved to the Methods section)*

Response 2.13.

> The Summary and Conclusions section has been considerably shortened. As helpfully suggested here, the discussion of the sMAPE limitations has been moved to the Methods section. Additionally, unnecessary repetition of earlier details have now been omitted.

Question 2.14. *Appendix A: The exclusion of the term "hyetograph" is surprising given that a sizeable number of papers refer to rainfall distributions with this term*

Response 2.14.

> We thank the reviewer for this comment. The term 'hyetograph' was not explicitly included in the formal Boolean search because it is primarily used in the literature to describe a rainfall input to hydrological or hydraulic models, rather than as the object of study in investigations of temporal rainfall characteristics. In contrast, terms such as 'rainfall temporal profile', 'temporal loading', and 'intensity profile' are more likely to appear in papers that explicitly focus on the temporal structure of rainfall events, which was the aim of our review. After reading your concern, we tried conducting some ad-hoc searches using "hyetograph" to verify that no major studies were omitted. We were satisfied with the results, and therefore consider that the literature review remains comprehensive and representative of the field.