Appendix: Response to Reviewers

Reviewer 1 - Marie-Claire ten Veldhuis

In this paper a large set of (52) metrics are investigated, that describe temporal loading (i.e. temporal evolution) of rainfall events. This first part constitutes a rather elaborate literature review that concludes with a summary table listing the various metrics and in how many studies the authors found them to be used. Then, these metrics are calculated for a large number (>100,000) rainfall events recorded at Danish rain gauges between 1975 and 2025. The metrics are evaluated with respect to their degree of overlap (redundancy) and robustness to changes in temporal resolution (aggregation) and choices made during data processing. The study aims to answer four research questions: RQ1: What key properties of rainfall event temporal loading are commonly measured, and why? RQ2: How sensitive are these metrics to the temporal resolution of the rainfall data? RQ3: How does de-dimensionalisation of rainfall events affect metric values? RQ4: Which metrics are strongly correlated, suggesting they may be redundant or are suitable for use in cross-comparison studies.

The study represents a diligent amount of work, but what does not become clear after reading through the results and conclusions, is a justification for why this work is needed. A summary and comparison of the many metrics is surely informative, but the question remains what new insights we can gain from it. What can we not do now that this evaluation of metrics will enable us to do? One of the reasons mentioned for conducting the study is the need to incorporate temporal loading characteristics in the development of design storms for flood modeling and flood risk assessment. Representing temporal loading of storms in a statistically representative way is indeed not straightforward and new ideas or insights could be really valuable for the field. It seems like a missed opportunity that a study of such a large number of rainfall events restricts itself to just descriptive metrics and does not provide any statistical analysis that could feed into, for instance, recommendations for calculation of return periods in flood risk analysis and development of design storms.

In its current version, the analyses seems more suitable for submission in the form of a technical note. This would require drastically shortening the content, some suggestions are provided in the following.

General reply - Review 1

> Thank you for your comments and review. We appreciate the recognition of the scale of the work, and acknowledge your critique that the original manuscript did not sufficiently make the case for why such a comprehensive evaluation of rainfall temporal loading metrics is needed, nor what new insight or capability it provides. In response to this overarching point, conveyed in Comments 1.5-1.7, we have revised the manuscript to make clearer that the main purpose of this work was never to propose new metrics, or to offer specific guidance on design storm development, rather to advance development of a conceptual and empirical framework that supports consistent interpretation and comparison of results across studies, and more deliberate selection of metrics in future analyses.

The study is intentionally structured around two linked components. The literature review establishes how rainfall temporal loading is conceptualised and quantified across different research communities, highlighting substantial variation in terminology and metric usage. This novel synthesis is a key scientific contribution of this work, and provides the necessary context for the large-scale empirical evaluation of 52 metrics across more than 200,000 rainfall events. This empirical analysis quantifies robustness to aggregation and calculation on dimensionless mass curves, and studies the relationships between metrics. Without the literature-based context, the empirical analysis would be difficult to interpret and its relevance for cross-study comparison would be limited.

We acknowledge the reviewer's suggestion that the analysis might be more suitable for a technical note, but we believe that such a format would require removal of the synthesis that underpins the empirical contribution. In particular, the ability to interpret metric redundancy and robustness in a generalisable way relies on first establishing how and why these metrics are used across the literature. For this reason, we maintain that the Research Article format remains appropriate, while recognising the need for greater concision and clarity.

To address these concerns, we have substantially shortened the literature review from 7.5 to 5.5 pages (approximately a 30% reduction), with the review methodology, Figure 3, and Table 1 moved to an appendix. The remaining text has been edited to reduce verbosity and improve focus. In addition, the Introduction and Discussion have been revised to more clearly motivate the need for a coherent metric categorisation and to explicitly state the intended contribution of the work. In particular, we now state at the end of the Introduction that the paper develops a conceptual categorisation of temporal loading aspects, based on the four research questions, to support cross-study interpretation and meaningful metric selection in future research.

We hope that these revisions clarify both the motivation for the study and the nature of its contribution: not the introduction of new metrics or direct design-storm prescriptions, but a framework for understanding what different rainfall temporal loading metrics measure, when they are comparable, and when they are not.

Question 1.1. *Introduction: Several statements are made here that would benefit from a cited reference. A couple of examples: - P2, L 39: "rainfall event temporal loading is often oversimplified or overlooked in impact modeling applications." Not sure that this still represents current practice? - P2, L 44: "symmetrical, centrally peaked intensity profiles are commonly used in flood modeling (..)". Many other approaches are used for flood modeling these days. Please place statement into context - P3, L58: "The relevance of each aspect varies by (. . .) and can result in misplaced emphasis and misinterpretation". Please provide a reference for this statement?*

Response 1.1.

We thank the reviewer for these comments and agree that several statements in the Introduction could benefit from clearer contextualisation and supporting references.

Regarding the statement that rainfall event temporal loading is often oversimplified or overlooked in impact modelling applications, our intention was not to suggest that this reflects all current practice, but rather that simplified temporal representations remain widely used in many design and applied modelling contexts. We have revised the text to make this scope clearer and to explicitly link this statement to commonly used design storm approaches, such as the FEH and Chicago profiles, which are discussed immediately thereafter.

Likewise, with respect to the comment that many other approaches are now used in flood modelling, we agree. Continuous simulation and event-based modelling using observed or stochastic rainfall are increasingly applied. We have amended the text to acknowledge these approaches, while noting that design storm methods remain prevalent in practice due to their simplicity and regulatory acceptance.

Finally, regarding the statement that the relevance of different aspects of temporal loading varies by application and can lead to misplaced emphasis, we agree that this is a general observation rather than a single result attributable to one study. We have clarified this point by briefly contrasting possible sensitivities across different application domains (e.g. flood modelling versus soil erosion).

Question 1.2. *Literature review: this review covers 7.5 pages summarizing metrics found in the literature that are then summarized in a nice overview table. The information density of this section is quite low (it's very wordy), and could easily be summarized in just the table with short descriptions of the metrics (Metric name| Metric description | References of studies where metric was used).*

Response 1.2.

We have carefully revised the manuscript to address this concern. The literature review has been substantially shortened (by approximately 30% / 2 pages), with the literature review methods and Figure 3 and Table 1 relocated to an appendix. The remaining sections have been edited throughout to reduce verbosity and improve readability. Additionally, we note that Tables B1 and C1-C3 in the appendices already summarise the information suggested as useful by the reviewer.

Question 1.3. *Methods: - L 275: A peculiar statement is made here that requires better justification: "For temporal loading, which is interested in what happens around the peak, the way in which the edges of the event are defined is particularly important, but in this research we do not investigate this further." If this aspect is so important, and central to the subject of study, should it not be particularly addressed?*

Response 1.3.

We thank the reviewer for highlighting this point and agree that the original wording could be interpreted as inconsistent. 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. 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 1.4. *Methods: L294: "Rainfall temporal loading metrics are implemented in Python based on the definitions provided in the original publications". It strikes me as odd that in a study that focuses on the calculation of metrics, no equations are provided of how metrics are calculated.*

Response 1.4.

> We thank the reviewer for this comment and agree that providing explicit equations improves the clarity and transparency of the methodology.
>
> In the original submission, we prioritised reproducible Python implementations because many of the metrics are inherently procedural, relying on ordered steps, thresholds, or windowing operations that are more clearly and precisely expressed in code than in equation form. This approach was intended to avoid ambiguity and support reproducible application by other researchers.
>
> That said, we recognise that including mathematical expressions aids readability and allows readers to more easily compare metrics conceptually. Based on a suggestion from reviewer 3, we have therefore revised the manuscript to include equations for the final recommended metrics (rather than all metrics), alongside references to the corresponding code implementation. This balances analytical clarity with reproducibility, while avoiding duplication for metrics whose definitions are inherently procedural.

Question 1.5. *Results and discussion: 4.1 Sensitivity of temporal aggregation: - the findings in this subsection on the effects of temporal aggregation are rather obvious, namely that values of peak intensities are particularly sensitive to temporal aggregation. It doesn't really seem worth analyzing and reporting?*

Response 1.5.

> We agree that it is intuitive that peak intensity metrics are sensitive to temporal aggregation. We certainly do not want to claim "surprise" where there isn't any. However, this subsection evaluates the sensitivity of a broad range of temporal loading metrics, many of which are more complex and for which the effects of aggregation are far less predictable. The results presented in Figures 5 and 6 provide quantitative evidence of how different metrics respond to aggregation, revealing substantial variability in sensitivity that cannot be inferred a priori.
>
> Even for metrics where sensitivity might be anticipated, we believe documenting the magnitude and consistency of this effect across events and metrics is important. Such evidence provides a reference against which methodological choices can be evaluated and supports more informed selection of metrics in future studies. In the revised manuscript, we will include a statement that explains that the strong sensitivity of peak intensity-related metrics to temporal aggregation is intuitive/expected. We will also highlight that it is

valuable to identify and contrast the magnitude of sensitivity and direction of change for specific metrics, especially the more complex ones.

Question 1.6. *Results and discussion: 4.3 Metric redundancy: the same comment applies here: is it really a new insight that metrics related to time of the peak are related, and similar for mass distribution etc.*

Response 1.6.

It is indeed reasonable to expect that metrics which quantify similar aspects of temporal structure, such as the timing of peak intensity, may exhibit some degree of correlation. However, the assertion that redundancy within 'categories' is obvious presupposes the existence of those categories. One of the central contributions of this paper is the development of a coherent categorisation of temporal loading metrics, which enables these relationships, and their limitations, to be clearly identified and interpreted. The updated manuscript will articulate this more clearly.

While some metrics fall into categories quite obviously (e.g., time-to-peak being a peak timing metric), for other metrics this is much less clear (e.g., skew_p being a peak timing metric). Therefore, this manuscript also provides a systematic testing of assumptions about metric behaviour, something which we found to be lacking in the literature. For instance, our empirical analysis allowed us to classify skew_p in the peak timing category, and furthermore provided a clear assessment of the extent to which skew_p is correlated with time-to-peak. This is an invaluable output for anyone wishing to apply either of these metrics, and/or to compare their results to previous studies using either metric.

Additionally, Section 4.3 offers clear evidence that peak timing and mass timing metrics are not equivalent. While this distinction may appear intuitive, we argue that it is frequently obscured in the literature, where broad summary terms such as "front-loaded" are used interchangeably to describe results derived from both peak timing and mass timing metrics. This practice risks treating fundamentally different aspects of temporal structure as equivalent. Explicitly demonstrating and quantifying this distinction is therefore a key motivation for, and outcome of, the systematic metric evaluation presented here. Section 4.3 in the manuscript has been revised to makes this point more clearly.

Question 1.7. *5. Recommendations "A central contribution of this paper is the argument that analyses of rainfall temporal loading must begin with a clear definition of the aspects most relevant to the research question or application." This seems like a rather generic and common sense argument. Which comes back to the earlier comment that the new insight provided by this study is not very clear."*

Response 1.7.

We agree that, as a general principle, the idea that analyses of rainfall temporal loading should begin with a clear definition of the relevant aspects may appear self-evident. However, the motivation for this paper arises from the observation that, in practice, this step is often not made explicit in the literature. Instead, different metrics are frequently applied, compared, or summarised under common descriptive terms without a clear articulation of the specific aspect of temporal structure they are intended to represent.

The aim of this study is therefore not to argue that such clarity is desirable in principle, but to provide a structured framework through which it can be achieved in practice. By systematically reviewing existing metrics, grouping them according to the aspects of temporal loading they quantify, and empirically evaluating their behaviour, sensitivity, and redundancy, the paper offers a concrete basis for making these definitions explicit and defensible.

In light of this, we agree that the sentence quoted by the reviewer could be misinterpreted as positioning a common-sense principle as a central contribution. We have therefore revised the manuscript to clarify that the paper is *built* on the premise that analyses should begin with a clear definition of the relevant aspects of temporal loading, and that a central contribution of the study is the provision of a framework that enables such definitions to be made more consistently and transparently in future work. The revised text reads as follows:

"This paper adopts the premise that analyses of rainfall temporal loading should begin with a clear definition of the aspect(s) most relevant to the research question or application. Building on this premise, a central contribution of this work is the development of a conceptual framework that supports explicit, consistent, and reproducible metric selection. Rather than prescribing a single universal metric, we provide a structured approach to choosing metrics that are aligned with the intended interpretation."

Across the previous three comments (1.5, 1.6, 1.7) it's clear that we have not done enough to motivate the need and importance of establishing a coherent categorisation of metrics that can aid researchers in choosing proper metrics for their analyses. To correct for this problem generally, in the revised manuscript we add some new text on this objective to the end of the Introduction. After establishing the research gap, and the four research questions we add the following text: **"Based on the findings of the empirical analyses related to the four research questions, we will provide a new conceptual categorisation of the various aspects of temporal loading. This aims to provide a new, conceptual framework to assist researchers in interpreting and comparing results from existing research, and to aid the meaningful selection of metrics in future research.**

Question 1.8. *6. Summary and conclusions This section should preferably restrict itself to presenting the conclusions. A short summary is already provided in the abstract of the paper.*

Response 1.8.

> We agree that the Conclusions section was too wordy and lengthy. This section has now been considerably shortened. This includes movement of some of the sections of text to earlier parts of the manuscript (e.g., as suggested by Reviewer 2, the section on the limitations of sMAPE has been moved to the Methods section). Additionally, in the revised manuscript, unnecessary repetition of earlier details have been omitted.