**HESS-2024-159: Constructing a geography of heavy-tailed flood distributions: insights from common streamflow dynamics**

We thank the Editor and the two Reviewers for providing comments and suggestions. We have incorporated their comments in the revised version of the manuscript and answered each point below. **The Editor's and Reviewers' comments are in black font with gray shading while our replies are in blue font. Text in the original manuscript is reported in red, with $L_o$ refers to the line number. Text in the tracked-revised version is presented in dark-blue, with $L_r$ refers to the line number.**

**Editor**

We have received a detailed list of comments from two reviewers. After reading them carefully, and also from my own personal reading of the manuscript, I believe there are some issues that need to be carefully solved whether you decide to revise your manuscript.

Thank you for the review. We have addressed all the comments of the reviewers and made appropriate modifications in the manuscript, as outlined below.

**Response to Reviewer 1**

The authors provide empirical analyses on the pattern and drivers of the tail heaviness. They adopt hydrograph recession exponent as the indicator of watersheds with or without propensity for heavy tails of flood peak distributions. The contrasts in the tail heaviness across watersheds, climate regions, seasons, shed light on the importance of characterizing catchment storage in dictating flood regimes. The analyses are interesting and robust. A major of mine is that the manuscript is lengthy that obscure the new wisdom obtained. I would suggest the authors to further refine and remove unnecessary details.

Thank you for the summarized review and positive feedback. We have streamlined the details, particularly addressing the specific comments below, to better highlight the key findings of this work as suggested by the reviewer.

*Specific comments:*

1) The part that emphasize the utility of hydrograph recession exponents in characterizing tail heaviness is lengthy, and needs to be shortened. …

We have shortened the sections that describe the utility of hydrograph recession exponents in characterizing tail heaviness, as suggested by the reviewer. In particular, we shortened section 3.1 by making larger use of references to a previous publication where this approach was first introduced (Wang et al., 2023). See details, particularly, at the $\mathrm{L}_r$ 185-223:

Within the PHEV framework, we obtain consistent expressions for the probability distributions of various flow metrics, including daily streamflow (Botter et al., 2009), ordinary peak flows (local flow peaks resulting from streamflow-producing rainfall events), and floods (flow maxima within a specified timeframe) (Basso et al., 2016). ~~The mathematical expressions for these are presented as Eqs. (1) to (3), respectively:~~

~~$p(q) = C_1 \cdot q^{-a}\left(e^{\frac{-1}{\alpha K(2-a)} \cdot q^{2-a}}\right)\left(e^{\frac{\lambda}{K(1-a)} \cdot q^{1-a}}\right),$ (1)~~

~~$p_j(q) = C_2 \cdot q^{1-a} \cdot e^{-\frac{q^{2-a}}{\alpha K(2-a)}} \cdot e^{\frac{q^{1-a}}{K(1-a)}},$ (2)~~

~~$p_M(q) = p_j(q) \cdot \lambda\tau \cdot e^{-\lambda\tau \cdot D_j(q)},$ (3)~~

~~where $D_j(q) = \int_q^\infty p_j(q)\, dq$, $\tau[day]$ is the duration of the considered time frame, $C_1$ and $C_2$ are normalization constants.~~

By taking the limit of these distributions, insights into the tail behavior of these theoretical flow distributions are obtained (Basso et al., 2015; Wang et al., 2023)~~(see Eqs. (4) to (6), set $\boldsymbol{C_3 = \lambda\tau C_2}$)~~. In particular, Wang et al. (2023) showed that the tail of the distribution is exclusively governed by a power law function (indicating heavy tails) when the hydrograph recession exponent exceeds two, signifying discernible nonlinearity of catchment responses. Conversely, the tail appears as nonheavy when the recession exponent is below two, suggesting linearity of catchment responses (notice that recession exponents are found to be above one in most river basins; Biswal and Kumar, 2014; Tashie et al., 2020b). Eq. (1) provides the mathematical expressions for the case of floods as an example (Note that similar conclusions are drawn for the theoretical distributions of daily streamflow and ordinary peak flows). As a result, the hydrograph recession exponent has been proposed as a suitable indicator of heavy-tailed flood behavior, based on the analysis of common discharge dynamics. For further detailed information, please refer to Wang et al (2023).

~~$\lim_{q\to+\infty} p(q) = \lim_{q\to+\infty}\left\{C_1 \cdot q^{-a}\left(e^{\frac{-1}{\alpha K(2-a)}\cdot q^{2-a}}\right)\left(e^{\frac{\lambda}{K(1-a)}\cdot q^{1-a}}\right)\right\},$ (4)~~

~~$\lim_{q\to+\infty} p_J(q) = \lim_{q\to+\infty}\left\{C_2 \cdot q^{1-a}\left(e^{\frac{-1}{\alpha K(2-a)}\cdot q^{2-a}}\right)\right\},$ (5)~~

$$\lim_{q\to+\infty} p_M(q) = \lim_{q\to+\infty}\left\{C_1 \cdot q^{1-a}\left(e^{\frac{-1}{\alpha K(2-a)}\cdot q^{2-a}}\right)\right\}, \tag{1}$$

Here, $q$ represents the discharge, $p_M(q)$ denotes the mathematical expression of the probability distribution of flow maxima (i.e., floods) in PHEV, and *C* is a normalization constant.

We employ an event-based analysis for estimating hydrograph recession exponents, a method deemed more robust than cloud-based analysis (Biswal and Marani, 2010; Dralle et al., 2017; Jachens et al., 2020). Specifically, we estimate the parameters of empirical power laws for individual hydrograph recessions through a linear regression of the pairs of $dq/dt$ and $q$ in log-log scale. The constant time step (CTS) method has been commonly used in various studies to estimate the time derivatives $dq/dt$ (e.g., Biswal and Marani, 2010; Mutzner et al., 2013; Dralle et al., 2017; Tashie et al., 2020b; Basso et al., 2023)~~, while other methodologies (Wittenberg, 1999; Rupp and Selker, 2006; Roques et al., 2017) have been developed for specific conditions. For instance, the exponential time step (ETS) method was introduced to reduce estimation uncertainty mainly caused by the latter-stage (low-flow) recession (Roques et al., 2017). In this study, recession exponents have been calculated using both CTS and ETS methods, yielding similar conclusions regarding flood tail behavior (see supporting information Text S1 and Figure S1). Therefore, we chose to base the subsequent analyses on the results obtained from the more commonly utilized approach (i.e., the CTS). The supporting information, Figure S2, provides an example of this process.~~ We use the constant time step (CTS) method to estimate the time derivatives dq/dt (see an example in supporting information Figure S2), which has been commonly used in various studies (e.g., Biswal and Marani, 2010; Mutzner et al., 2013; Dralle et al., 2017; Tashie et al., 2020b; Basso et al., 2023). We acknowledge the availability of alternative methods, such as the exponential time step (ETS), and confirm that ETS yields results similar to CTS in this study (see supporting information Text S1 and Figure S1).

…One question might be, as far as can be seen from the dataset, the record lengths are quite adequate despite of variance, some of other tail heaviness indicators would be able to perform as well.

The dataset employed in this study spans 24 to 148 years. We acknowledge that other indicators could also be used; however, we are specifically interested in the recession exponent because it is a novel index that allows us to infer the propensity of rivers to experience extreme floods. This index enables us to identify potential risks even in the absence of recorded extreme floods, which is often not possible with other indicators. Additionally, the recession exponent is suggested to mitigate the bias often introduced by the variance in dataset lengths across cases (Smith et al., 2018; Wietzke et al., 2020; Wang et al., 2023).

We have improved the literature review on this topic in $L_o$ 72-80. We further supplement this with the following statement, $L_r$ 99-103: "Nonetheless, we acknowledge that other indicators could also be used; however, we are specifically interested in the recession exponent because it is a novel index that allows us to infer the propensity of rivers to experience extreme floods. Such an index enables us to identify potential risks even in the absence of recorded extreme floods, which is often not possible with other indicators. Its stability provides additional value to mitigate the bias often introduced by the variance in dataset lengths across cases."

2) What is the rationale of using five days as the minimum duration, considering the vast variance in drainage areas?

Event-based recession analyses of daily discharge data typically consider recessions which last for a minimum of 3 to 5 days (e.g., Shaw and Riha, 2012; Biswal and Marani, 2010) to minimize noise from short events (Ye et al., 2014) while ensuring a sufficient sample size (i.e., a sufficient number of analyzed recessions) to obtain representative values of recession parameters (Shaw, 2016). We acknowledge that the most suitable recession length may vary depending on the drainage area. Although drainage area is an important factor, our catchments are somehow mesoscale catchments. Hence, for the sake of simplicity, we applied a single minimum duration, i.e., 5 days, a common practice in other studies (Biswal and Marani 2010; Shaw and Riha, 2012; Dralle et al., 2017; Jachens et al., 2020; Tashie et al., 2020a).

We have modified the manuscript accordingly:

$L_o$ 212-213: “~~The minimum duration of a recession is set to five days (Dralle et al., 2017), denoting that recessions shorter than five days are not considered.”~~

$L_r$ 233-237: “To reduce noise from short events (Ye et al., 2014) and ensure sufficient sample sizes (i.e., a sufficient number of analyzed recessions) to obtain representative values of recession parameters (Shaw, 2016), it is common practice to set a minimum recession duration. Following previous studies, we did not vary this duration across basins and set it equal to 5 days (Biswal and Marani 2010; Shaw and Riha, 2012; Dralle et al., 2017; Jachens et al., 2020; Tashie et al., 2020a).”

3) “The upper tail is defined by an optimized lower boundary of the discharge, determined by selecting the best fit based on the KS statistic”. This is not quite clear. How the upper tail is defined and statistically modelled is important. Section 3.2 also needs to be concise and informative. Please reconstruct.

Thank you for your suggestions. We have rephrased the sentence to improve clarity as below:

$L_o$ 232-233: “~~The upper tail is defined by an optimized lower boundary of the discharge, determined by selecting the best fit based on the KS statistic~~”

$L_r$ 259-265: “Empirical data following a power-law distribution (if applicable) typically do so above a certain lower bound (Clauset et al., 2009). Therefore, we employ the approach proposed by Clauset et al. (2007) to determine an optimized lower boundary above which a power-law tail may emerge. This method selects the boundary for which the empirical probability distribution and the best-fit power-law model are most similar as evaluated by the Kolmogorov-Smirnov (KS) statistic, which is used to quantify the distance between these distributions. If the optimized lower boundary is higher than the true lower boundary, the reduced data sample leads to a poor match due to statistical fluctuations. If it is lower, the distributions differ fundamentally.”

We have revised the remaining of section 3.2 with the goal of shortening it. Below are the main modifications:

$L_o$ 221-231: “~~We used the Kolmogorov-Smirnov (KS) statistic ($\kappa$) to preliminarily assess empirical power law distribution reliability ($\kappa \in [0,\infty]$, with $\kappa=0$ denoting highest reliability) (Clauset et al., 2009; Klaus et al., 2011; Alstott et al., 2014). We acknowledge that there are alternative methods for evaluating goodness-of-fit. For example, the Anderson-Darling test, which is essentially a modified version of the Kolmogorov-Smirnov test, gives greater weight to the tails when assessing the distance between distributions. However, using such tests can be more conservative in determining the minimum threshold of the variable that delineates the tail range of the empirical distribution. This characteristic necessitates a much larger dataset to ensure a robust distribution fit for the tail; otherwise, the sample size of the identified tail may be insufficient for a reliable fit (Alstott et al., 2014; Clauset et al., 2009). In our case, we opted for the KS test because it allows for more appropriate~~

~~sample sizes based on our datasets (Hosking and Wallis, 1987) (see Figure S4 in the supporting information). To establish the reference point for plausible empirical power laws, we employed the framework introduced by Clauset et al. (2009).~~”

$L_o$ 233-246: “~~The goodness-of-fit test is conducted using the KS test with a significance level of p > 0.1. It's important to note that a higher *p*-value is considered more rigorous in this context, as the aim is to verify the null hypothesis rather than to reject it, as is often considered in other cases. Thus, p > 0.1 is a more stringent criterion than p > 0.05 in this scenario. All cases that meet the goodness-of-fit criteria are further evaluated for plausibility against common alternative distributions using the bootstrapping method. This involves generating 1000 sets of synthetic data from the optimized power law model, each with the same sample size as the fitted observations. In their experiments, the number of 1000 repetitions has proven sufficient to distinguish power laws from both exponential and lognormal distributions using the KS test with a p > 0.1 significance level. Once the empirical power law passes all these criteria, it is considered a dependable (plausible) representation of the empirical data distribution. We term such a case study a 'power-law-tailed case study,' while cases that don't meet these criteria are labeled as 'uncertain case studies' in subsequent analyses. The latter label acknowledges the awareness that we cannot definitively conclude whether these case studies are indeed not power-law-tailed or if their underlying distributions cannot be determined due to the high uncertainty caused by the small sample sizes of available observations.~~”

$L_r$ 278-288: “The goodness-of-fit is evaluated by means of the KS test. We used the KS test instead of alternatives like the Anderson-Darling test to ensure appropriate sample sizes for our datasets (Hosking and Wallis, 1987; Clauset et al., 2009; Klaus et al., 2011; Alstott et al., 2014) (see Figure S4 in the supporting information). The KS statistic ($\kappa$) is employed to preliminarily assess the reliability of empirical power law distributions ($\kappa \in [0,\infty]$, with $\kappa=0$ denoting the highest reliability). The test is applied with a significance level of p > 0.1, which is more stringent than the typical p > 0.05 since the goal here is to fail to reject the null hypothesis (i.e., to confirm the lack of evidence to conclude that a difference exists between the distributions). Cases meeting this criterion are further evaluated against alternative distributions (e.g., lognormal) by using a bootstrapping method, where 1000 synthetic datasets are generated from the optimized power law model (Clauset et al., 2009). Cases passing also this further criterion are deemed 'power-law-tailed.' Those that don't are labeled 'uncertain,' indicating that either these cases are not power-law-tailed, or their distribution cannot be determined due to high uncertainty from small sample sizes.”

$L_o$ 250-253: “~~The sample sizes of the identified upper tails for the distributions of daily streamflow range from 176 to 11260, with a median of 1280. For the distributions of ordinary peaks, the sample sizes range from 13 to 2240, with a median of 512. In the case of monthly maxima distributions, the sample sizes vary from 6 to 418, with a median of 132.~~

$L_r$ 287-288: “(the median sample sizes are 1280, 512, and 132 for daily streamflow, ordinary peaks, and monthly maxima, respectively)”

$L_o$ 259-263: “~~We acknowledge that the method of maximum likelihood used for parameter estimation could, in principle, be substituted with various established approaches based on the characteristics of the study data. For instance, Hosking and Wallis (1987) shown that the method of probability-weighted moments could offer better parameter estimation when working on less than 500 samples for generalized Pareto distributions, which are known to exhibit power law tails in asymptotic analysis.~~”

The following description is added for clarity (and in response to Reviewer 2's comments):
$L_r$ 302-305: “It's important to clarify that these sample sizes refer specifically to the tail of the empirical distributions. In other words, only the most extreme observations are analyzed to determine whether the empirical distributions exhibit power-law behavior in their tails. For the overview of the entire data series analyzed in this study, please refer to Section 2.”

4) Line 269, by “majority” you use “50 %” as the threshold?

Yes, we do. We have specified it in the revised manuscript ($L_r$ 314).

5) From Figure 1 and Figure 2, we can see there are many overlaps between heavy tails and nonheavy tails. This is especially evident in Figure 2 where we see the scatters are well mixed. These results make me wonder the utility of recession exponent (a=2) as the criteria. I would suggest the authors to explain and discuss the limitation.

Ideally, the new index should be validated by comparing its predictions for both heavy-tailed and non-heavy-tailed case studies, as determined from empirical frequency distributions. However, from data we can only determine which case studies are power-law-tailed (black dots in Figure 2), but we cannot state which cases are certainly non-heavy-tailed. These cases are therefore labeled uncertain case studies (gray dots in Figure 2). The latter category encompasses case studies that either do not follow a power-law distribution or whose underlying distributions cannot be determined because of high uncertainty due to small sample sizes, which may cause the overlaps of these two groups. In fact, several years of data are still a small sample to reliably characterize the tail of empirical data distributions. Due to this limitation, the effectiveness of the recession exponent is assessed by comparing its predictions with the identified heavy-tailed cases only (black dots in Figure 2). Additionally, to rigorously validate the effectiveness of the recession exponent criteria, we also include Figure 1, which tests sensitivity of the index accuracy to the reliability level of the empirical power laws. The index accuracy is achieved by means of conditional probability. The results show that the accuracy improves as the reliability level of the empirical power laws increases. This suggests that, although the two groups overlap (in Figure 2), cases identified as heavy-tailed indeed have a recession exponent greater than 2, as expected from our index.

Notwithstanding the limitations inherent in identifying benchmarks for the new index based on data analyses, we acknowledge that misattribution can occur due to the recession exponent not always being able to properly distinguish between heavy and light tails, particularly when a is around the threshold value of 2. This issue is highlighted by the case studies in Norway, which we discuss in $L_o$ 356-363. We have clarified the matter discussed above at $L_r$ 286-288:

$L_r$ 286-288: "Cases passing all criteria are deemed 'power-law-tailed.' Those that don't are labeled 'uncertain,' indicating that either these cases are not power-law-tailed, or their distribution cannot be determined due to high uncertainty from small sample sizes."

Moreover, we have inserted additional discussion after $L_o$ 331:

$L_r$ 378-381: "We cannot conclude whether uncertain case studies (gray dots) represent cases that are indeed not power-law-tailed or if their underlying distributions cannot be determined due to the high uncertainty caused by small sample sizes. Therefore, we benchmark the recession exponent against the empirical power law exponent by focusing on the 'certain group,' i.e., power-law-tailed case studies (black dots)."

The following statement has also been inserted at $L_r$ 415:

$L_r$ 415: "However, we acknowledge that misattributions may occur, particularly when $a$ is around the threshold value."

6) Figure 4 and the text, please explain the rationale of using percentage. The absolute count of watersheds would matter, as can be seen that there is only one watershed in Bwk. This can be due to sampling uncertainties.

We recognize that the varying number of cases across climate types might introduce bias due to sample sensitivity (as we mentioned in $L_o$ 490-493). Nonetheless, the ratio (i.e., percentage) of heavy- to non-heavy-tailed case studies in each climate region is considered to be one of the most direct approaches to display the propensity towards a certain tail behavior in each region. We have revised the manuscript to emphasize this concern related to the employed dataset:

$L_o$ 490-493: "~~We acknowledge that these results are based on overarching conditions and do not encompass all climate types, and achieving an equal number of study sites across various climate regions might not always be feasible. Expanding the number of study sites could further enhance our understanding, especially for extreme cases.~~"

$L_r$ 554-558: “We acknowledge that these results are based on overarching climate conditions and do not encompass all climate types, and achieving an equal number of study sites across various climate regions might not always be feasible. We should be mindful of potential bias caused by sample sensitivity, particularly in regions with a limited number of cases (e.g., Csa, BSh, BWk in this study). Expanding the number of study sites in these climate regions could strengthen the current understanding.”

7) What do the authors mean by “catchment storage”? Please clarify.

Thank you for pointing this out. We have clarified this as follows:

$L_r$ 547-550: “We refer to catchment storage *sensu* Kirchner et al. (2009) and Botter et al. (2009), i.e., the varying amount of water contained in a catchment between dry and wet periods.. This capacity is dynamic and depends on various factors, such as soil moisture states, precipitation, and evapotranspiration (Merz and Blöschl, 2009; Zhou et al., 2022).”

8) Line 571, I would suggest the role of ET alone might not be that important. The ratio of ET to P worthwhile to be explored.

We agree that, also based on our analyses and findings, the temporal characteristics of rainfall and evapotranspiration collaboratively influence this seasonality, as discussed in detail at $L_o$ 539-548. We therefore revise $L_o$ 569-571 as follows:

$L_o$ 569-571: “~~Regions with pronounced temperature variations across seasons, particularly with higher temperature in summer, tend to display such dynamics and highlight the role of evapotranspiration in catchments in driving this seasonality.~~”

$L_r$ 634-638: “Regions with pronounced temperature variations across seasons, particularly with higher temperatures in summer, and characterized by relatively evenly distributed rainfall throughout the year tend to display such dynamics. This highlights the importance of both evapotranspiration and the temporal characteristics of rainfall in shaping flood tail behavior across seasons, aligning with previous studies (Guo et al., 2014; Basso et al., 2023).”

9) Line 605-606, the three references use indicators that quantify the heaviness of upper tails, while in this study, the authors are in fact addressing “propensity”.

Thank you for pointing this out. We have improved the text as below:

$L_o$ 604-606: "~~These findings align with previous discussions on this matter (e.g., Merz and Blöschl, 2009; Villarini and Smith, 2010; Smith et al., 2018), which have suggested a relatively weak inverse correlation between catchment area and the occurrence of heavy-tailed flood behavior.~~"

$L_r$ 673-676: "These findings align with the results of previous studies (e.g., Merz and Blöschl, 2009; Villarini and Smith, 2010; Smith et al., 2018) which, by using different indices to quantify the heaviness of upper tails, have suggested a relatively weak inverse correlation between catchment area and the occurrence of heavy-tailed flood behavior."

*10) Line 649-650, this is obvious.*

This sentence ($L_o$ 649-650) serves as a contrast to the following one ($L_o$ 650-652). To address the reviewer's comment, we have revised it as follows:

$L_o$ 649-650: "~~Our findings first indicate that regions with relatively uniform~~ …"

$L_r$ 726: "Regions with relatively uniform …"

11) Section 5, I enjoy reading this section overall, but it can be further improved by explicitly highlighting what are found in this study, and what are proposed by previous studies, especially the review paper by Merz et al.

Thank you for the suggestion. We have enhanced the clarity of this section by adding or refining statements in Section 5. Particularly, we highlight the comparison between current understanding and the new findings contributed by this study in the revised version across hypotheses:

Hypothesis 2 ($L_o$717-729):

$L_r$ 806-808: "Therefore, this study provides evidence that the influence of flood generation processes is closely tied to the nonlinearity of hydrological behaviors. This finding enhances the understanding of these processes, supporting advancements in this area as suggested by Merz et al. (2022)."

Hypothesis 3 ($L_o$730-737):

$L_o$ 736-737: "~~Thus, our findings provide evidence that supports this hypothesis.~~"

$L_r$ 815-818: "Thus, this study addresses the knowledge gap by showing that a mix of flood event types can result in heavy-tailed flood behavior. It further suggests that this is especially critical for regions transitioning from snow-dominated flood generation processes to more mixed types, as observed in Northern Europe (Tarasova et al., 2023)."

Hypothesis 4 ($L_o$738-749):

$L_r$ 831-834: "In summary, this study proposes a quantification approach based on these acknowledged, robust drivers, using daily streamflow observations. This approach paves a broader path for exploring the relationship between flood tail behavior and other physioclimatic variables, enhancing our understanding of extreme hydrological events."

Hypothesis 5 ($L_o$750-763):

$L_r$ 848-850: "The interaction between evapotranspiration and the temporal characteristics of rainfall is suggested to be the underlying reason why drier catchments favor heavy-tailed floods, as observed in their seasonal flood tail behavior."

Hypothesis 6 ($L_o$764-774):

$L_o$ 772-774: "~~These findings underscore the importance of considering the dominant flood generation processes in each region and elucidate how catchment size interacts with flood tail behavior by influencing these dominant processes—either amplifying, reducing, or having no significant effect.~~"

$L_r$ 859-863: "These findings underscore the importance of considering the dominant flood generation processes specific to each region. To thoroughly address how catchment sizes affect flood tail behavior, it is important not only to focus on the size itself but also investigate how flood generation processes vary across different sizes within their study areas. This nuanced understanding can illuminate how catchment size interacts with flood dynamics—either amplifying, reducing, or exerting no significant effect on heavy-tailed flood behavior."

We sincerely appreciate the reviewer's valuable comments, which have certainly enhanced the quality and clarity of this manuscript.

**Response to Reviewer 2**

This paper investigates an interesting hypothesis, that the tail heaviness of flood distributions can be inferred from a recession analysis, through a large sample analysis with data from Europe and the United States. In hydrology, a heavy tail distribution of floods means that extreme floods are more likely to occur than would be predicted by distributions that have exponential asymptotic behaviour. Very large extremes may therefore happen with a probability that is not as small as one would expect if using, for example, a Gumbel distribution to model flood probabilities, and may result in huge damages due to their surprising nature. Identifying the properties of the distribution tails is a hard task which requires, usually, very long series of data and/or regional analyses when using statistical techniques. Linking tail heaviness to catchment behaviour and process understanding is therefore a very interesting avenue to follow in order to increase our confidence in the behaviour of the extremes. This is the aim of this paper and it is therefore of interest to the hydrology community. The paper is well-written with high-quality figures.

Thank you for summarizing the aims of this study.

1) The first aim of the paper is to validate the effectiveness of the method in identifying heavy tail flood behavior (line 95). However, the validation is based on analysing whether samples of (on average) 10 years of length (line 253), at the daily to monthly maxima timescales, can be represented in the tails by an empirical power law. My question is whether this "tail analysis" is representative of the extremes of practical interest? In hydrological practice, we typically focus on return periods of 100 or 200 years and sometimes more. It is unclear to me what are the return periods of interest in this paper. It is important to reflect on the range of return periods for which the analyses in this paper are meant because processes may emerge with increasing return periods which are not at work for less extreme floods. I am not convinced that the paper demonstrates that the method is relevant for extremes that are of interest in hydrology. Benchmarking the recession analysis on a statistical analysis on short samples does not constitute a proper validation of the method. In order to better evaluate the effectiveness of the method in identifying heavy tail flood behavior, an additional, and to me more convincing, benchmark should be the analysis of (many) long timeseries with methods usually adopted in flood frequency analysis (e.g., fit of the GEV shape parameter, better if using regional analysis).

The concern raised pertains to $L_o$ 253 ($L_r$ 301), where we mention that the average sample size used for fitting empirical power laws on monthly maximum streamflow is 132. We believe the reviewer interpreted this as representing roughly 10 years of observations, which is considered too short to represent extreme behavior. We would like to clarify that the average sample size of 132 refers only to the length of the identified 'tail' of the frequency distribution, not the length of the entire observation period. The full observation period is on average 62 years (ranging from 24 to 148 years) across the dataset used in this study (see $L_o$ 125 / $L_r$ 138). Empirical data, if they follow a power-law distribution, typically do so only for values above a certain threshold (i.e., the tail). Consequently, it is standard practice to first identify this threshold (i.e., where is the tail) before fitting a power law. The sample size of 132 hence refers to the most extreme monthly maxima above this threshold observed within an average 62-year long data series.

We regret any confusion caused by the previous wording and have added the following description to clarify this point in the revised manuscript:
$L_r$ 302-305: “It's important to clarify that these sample sizes refer specifically to the tail of the empirical distributions. In other words, only the most extreme observations are analyzed to determine whether the empirical distributions exhibit power-law behavior in their tails. For the overview of the entire data series analyzed in this study, please refer to Section 2.”

Nonetheless, We acknowledge that the length of the data series used in this study (62-years on average) does not allow to directly derive from data the magnitude of events with 100 or 200 years return period, as it occurs in most cases by using any other method. In fact, such an observation period aligns with what is normally available (e.g., Bertola et al., 2023) and used in flood frequency analysis for estimating (through extrapolation) 100- or 200-year floods (Zhao et al., 2021).

Several studies suggested that the shape parameter of the GEV may not be a reliable indicator of tail heaviness because it is highly sensitive to the length of observation series and the occurrence of outliers (Hu et al., 2020; Cai and Hames, 2010). However, to address the concerns of the reviewer and provide an additional benchmark for the method employed in this study, we also calculated the L-moment ratio diagrams, which have been shown to provide more robust results than the shape parameter of the GEV, particularly for evaluating highly skewed samples (Vogel and Fennessey, 1993). 97.8%, 100%, and 94.1% of the identified heavy-tailed case studies(based on empirical power law fitting of daily flows, ordinary peaks, and monthly maxima) exhibit greater L-skewness and L-kurtosis than the exponential distribution, thus indicating heavy-tailed behavior. These results, which confirm the results of the previous benchmarking through the use of L-moment ratio diagrams, are shown in Figure S5 and discussed at lines $L_r$ 77-80, 372-377.

$L_r$ 77-80: “Additional efforts to improve the reliability of tail heaviness estimates include the use of L-moments (Hosking et al., 1985), which ensure better upper tail estimation of GEV compared to maximum likelihood, and L-moment ratio diagrams (Vogel and Fennessey, 1993), which improve estimation in highly skewed samples.”

$L_r$ 372-377: “We also perform an L-moment analysis, a compelling method in order statistics used to quantitatively describe extremes and known for its robustness to stochastic sampling uncertainties (Hosking, 1990). This analysis serves to confirm the tail heaviness observed in the identified power-law-tailed case studies, in which these case studies show clearly heavier tails than exponential distributions (i.e., the widely accepted distinction of heavy- and nonheavy-tailed distributions; Merz et al., 2022) (see supporting information Figure S5)."

2) The second aim of the paper is the evaluation of the causes for differences in the recession coefficient and therefore, based on the hypothesis made here, of the tail heaviness of floods. The results are not that easy to interpret, since the method proposed uses a sharp threshold on the recession coefficient (a=2) to distinguish between heavy tail behaviour and (possibly) non-heavy tail behaviour. So only a binary distinction is made and differences between Germany and UK, for example, cannot be clearly identified. Since different degrees of tail heaviness exist, wouldn't it be more useful to link the recession coefficient to, for example, the exponent of the empirical power law b? The Authors show something like this in Figure 2 even though the relationship doesn't seem to be so strong. But wouldn't that be more useful in hydrological practice where, for instance, the estimation of the GEV shape parameter is of interest?

The identification of heavy-tailed floods through hydrograph recession analysis (employed in this study) uses a threshold of two on the recession exponent to distinguish heavy-tailed cases from non-heavy-tailed ones. The method further allows for evaluating the tail heaviness based on the specific exponent values, as noted by the reviewer. Such an approach resembles what done for other indices, such as the GEV shape parameter, where a threshold value of zero is used to differentiate between heavy-tailed and non-heavy-tailed distributions. Differently from the latter case, where the threshold of zero has a statistical meaning only, the threshold of two in the method adopted in this study has also a physical meaning, as it represents a degree of non-linearity of the catchment hydrologic response which cause a shift in the shape of the resulting streamflow and flood distributions (see Botter et al., 2009 and Kirchner et al., 2009 for details in this regard; see Basso et al., 2016, 2023 for how the shift in the shape of the streamflow distribution translates into a shift in the shape of the flood distribution).

This study emphasizes a binary distinction between heavy and non-heavy-tailed distributions - rather than discussing the degree of heaviness - for two reasons. First, a reliable identification of heavy-tailed distributions (i.e., even without any claim about their degree of heaviness) is per se a difficult task, as noted by the reviewer in the previous comment. Second, the identification itself holds significant hydrological importance, regardless of the degree of heaviness. In fact, the presence of a heavy tail alone can serve as a critical warning of a relatively high probability of extreme events. For the latter reason also other studies, even those using indices like the GEV shape parameter, often focus on distinguishing cases with heavy-tailed flood distributions from those without (e.g., Macdonald et al., 2022). In addition, there is a notable gap in conducting such investigations on an extended spatial scale. This gap is largely due to the fact that quantifying such behavior remains highly sensitive to the sample size (Wietzke et al., 2020; Hu et al., 2020), making reliable identification across different datasets challenging (Merz et al., 2022). The use of the empirical power law exponent b, as suggested by the reviewer, would similarly suffer from sensitivity to the sample size, as discussed at lines $L_r$ 291-301, 388-395.

Instead, the recession exponent used in this study has been tested and found to be more reliable to distinguish between heavy and non-heavy-tailed distributions than, e.g., the GEV shape parameter, especially in analyses with short data lengths, as shown in a previous work (see Hu et al., 2020 and Wang et al., 2023). This justifies the selection of such an index for investigations across a broader range of study areas. We have added the following statement to enhance the relevant discussion:

$L_r$ 721-725: "This study focuses a binary distinction between heavy and non-heavy-tailed distributions, rather than assessing the degree of heaviness, for two key reasons. First, identifying heavy-tailed distributions is inherently challenging. Second, the identification itself holds significant hydrological importance, regardless of the degree of heaviness. In fact, the presence of a heavy tail alone can serve as a critical warning of a relatively high probability of extreme events, making it a crucial issue also in studies using other indices (e.g., Macdonald et al., 2022)."

Besides, I think the spatial results obtained here should be compared to regional studies on flood frequency? One example is Macdonald et al. (2022) who identify the GEV shape parameter as a quantification of tail heaviness in Germany. Figure 3a here seems consistent with Figure 4 of Macdonald et al. (2022). What about the other regions? In the US there are maps of the regional skewness in the Bulletin 17b. These comparisons could strengthen the confidence in the effectiveness of the method, since they are based on longer timeseries and on regional analyses.

We agree that comparing our findings with previous regional studies would strengthen the conclusions of this work, as we described in $L_o$ 390-424. We thank the reviewer for suggesting further comparisons. Accordingly, we will enhance this section with the following modifications:

For Germany: $L_r$ 445-446: "This finding aligns with Macdonald et al. (2022), who used GEV shape parameters as an indicator of heavy-tailed behavior for gauges with more than 50 years of observations."

For the UK: $L_r$ 451-453: "According to our findings, heavy-tailed flood behavior is prevalent in the UK, with a prevalence of 77%, particularly in the eastern and southern coastal regions. This aligns with clues from historical events (European Environmental Agency, 2010) and clues from future flood risk assessments (Rudd et al., 2021)."

For the US: $L_r$ 473-477: “In particular, catchments on the eastern side of the Appalachian Mountains exhibit pronounced heavy-tailed flood behavior, while those on the western side mostly exhibit non-heavy-tailed behavior. This is consistent with several previous findings based on the skewness of annual maximum streamflow (Interagency Advisory Committee on Water Data, 1982), the GEV shape parameters (Villarini and Smith, 2010), and the upper tail ratio (Smith et al., 2018).”

We would like to highlight that the analyses of this study, which are based on shorter and more variable lengths of data (24-148 years) and on analyzing hydrograph recessions from ordinary flows rather than flood records, provide findings in agreement with studies that rely on longer data records (e.g., only gauges with more than 75 years in the work of Villarini and Smith (2010), and more than 50 years in Macdonald et al. (2022)). In our view, such an agreement not only confirms the effectiveness of this new approach but also highlights its advantages, as it allows for analyses across broader geographical areas where less data may be available, facilitating the investigation of diverse conditions.

Given these concerns, I am sorry I cannot recommend publication of this work in HESS.

We hope the responses provided above satisfactorily address the reviewer’s concerns. We would like to emphasize that this work not only aims at validating the effectiveness of the newly proposed approach (as demonstrated in greater detail in our previous work, Wang et al., 2023) but also at using this index to shed light on the relationships between heavy-tailed flood behavior and critical environmental factors (e.g., climate, catchment areas) that remain poorly understood.

## References

Basso, S., Merz, R., Tarasova, L., & Miniussi, A. (2023). Extreme flooding controlled by stream network organization and flow regime. Nature Geoscience, 16(April), 339–343. https://doi.org/10.1038/s41561-023-01155-w

Bertola, M., Blöschl, G., Bohac, M., Borga, M., Castellarin, A., Chirico, G. B., et al. (2023). Megafloods in Europe can be anticipated from observations in hydrologically similar catchments. Nature Geoscience, 16(11), 982–988. https://doi.org/10.1038/s41561-023-01300-5

Biswal, B., & Marani, M. (2010). Geomorphological origin of recession curves. *Geophysical Research Letters*, *37*(24), 1–5. https://doi.org/10.1029/2010GL045415

Botter, G., Porporato, A., Rodriguez-Iturbe, I., & Rinaldo, A. (2009). Nonlinear storage-discharge relations and catchment streamflow regimes. Water Resources Research, 45(10), 1–16. https://doi.org/10.1029/2008WR007658

Cai, Y., & Hames, D. (2010). Minimum sample size determination for generalized extreme value distribution. Communications in Statistics: Simulation and Computation, 40(1), 87–98. https://doi.org/10.1080/03610918.2010.530368

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661–703. https://doi.org/10.1137/070710111

Clauset, A., Young, M., & Gleditsch, K. S. (2007). On the Frequency of Severe Terrorist Events. *Journal of Conflict Resolution*, *51*(1), 58–87. https://doi.org/10.1177/0022002706296157

Dralle, D. N., Karst, N. J., Charalampous, K., Veenstra, A., & Thompson, S. E. (2017). Event-scale power law recession analysis: Quantifying methodological uncertainty. *Hydrology and Earth System Sciences*, *21*(1), 65–81. https://doi.org/10.5194/hess-21-65-2017

European Environmental Agency. (2010). Mapping the impacts of natural hazards and technological accidents in Europe An overview of the last decade. Publications Office of the European Union. https://doi.org/10.2800/62638

Guo, J., Li, H.-Y., Leung, L. R., Guo, S., Liu, P., & Sivapalan, M. (2014). Links between flood frequency and annual water balance behaviors: A basis for similarity and regionalization. Water Resources Research, 50, 937–953. https://doi.org/http://dx.doi.org/10.1002/2013WR014374

Hu, L., Nikolopoulos, E. I., Marra, F., & Anagnostou, E. N. (2020). Sensitivity of flood frequency analysis to data record, statistical model, and parameter estimation methods: An evaluation over the contiguous United States. Journal of Flood Risk Management, 13(1), 1–13. https://doi.org/10.1111/jfr3.12580

Interagency Advisory Committee on Water Data. (1982). Guidelines for determining flood flow frequency: Bulletin 17B. https://water.usgs.gov/osw/bulletin17b/dl_flow.pdf

Kirchner, J. W. (2009). Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward. Water Resources Research, 45(2), 1–34. https://doi.org/10.1029/2008WR006912

Merz, B., Basso, S., Fischer, S., Lun, D., Blöschl, G., Merz, R., et al. (2022). Understanding heavy tails of flood peak distributions. Water Resources Research, 1–37. https://doi.org/10.1029/2021wr030506

Merz, R., & Blöschl, G. (2009). Process controls on the statistical flood moments - a data based analysis. *Hydrological Processes*, *23*(5), 675–696. https://doi.org/10.1002/hyp

Rudd, A. C., Kay, A. L., &Sayers, P. B. (2023). Climate change impacts on flood peaks in Britain for a range of global mean surface temperature changes. Journal of Flood Risk Management, 16(1), 1–15. https://doi.org/10.1111/jfr3.12863

Shaw, S. B. (2016). Investigating the linkage between streamflow recession rates and channel network contraction in a mesoscale catchment in New York state. *Hydrological Processes*, *30*(3), 479–492. https://doi.org/10.1002/hyp.10626

Shaw, S. B., & Riha, S. J. (2012). Examining individual recession events instead of a data cloud: Using a modified interpretation of dQ/dt-Q streamflow recession in glaciated watersheds to better inform models of low flow. *Journal of Hydrology*, *434–435*, 46–54. https://doi.org/10.1016/j.jhydrol.2012.02.034

Smith, J. A., Cox, A. A., Baeck, M. L., Yang, L., & Bates, P. (2018). Strange Floods: The Upper Tail of Flood Peaks in the United States. *Water Resources Research*, *54*(9), 6510–6542. https://doi.org/10.1029/2018WR022539

Villarini, G., & Smith, J. A. (2010). Flood peak distributions for the eastern United States. *Water Resources Research*, *46*(6), 1–17. https://doi.org/10.1029/2009WR008395

Vogel, R. M., & Fennesse, N. M. (1993). L moment diagrams should replace product moment diagrams. Water Resources Research, 29(6), 1745–1752. https://doi.org/10.1029/93WR00341

Wang, H., Merz, R., Yang, S., & Basso, S. (2023). Inferring heavy tails of flood distributions through hydrograph recession. Hydrol. Earth Syst. Sci, 27(24), 4369–4384. https://doi.org/10.5194/hess-27-4369-2023

Wietzke, L. M., Merz, B., Gerlitz, L., Kreibich, H., Guse, B., Castellarin, A., & Vorogushyn, S. (2020). Comparative analysis of scalar upper tail indicators. *Hydrological Sciences Journal*, *65*(10), 1625–1639. https://doi.org/10.1080/02626667.2020.1769104

Ye, S., Li, H. Y., Huang, M., Alebachew, M. A., Leng, G., Leung, L. R., et al. (2014). Regionalization of subsurface stormflow parameters of hydrologic models: Derivation from regional analysis of streamflow recession curves. *Journal of Hydrology*, *519*(PA), 670–682. https://doi.org/10.1016/j.jhydrol.2014.07.017

Zhao, G., Bates, P., Neal, J., & Pang, B. (2021). Design flood estimation for global river networks based on machine learning models. Hydrology and Earth System Sciences, 25(11), 5981–5999. https://doi.org/10.5194/hess-25-5981-2021

Zhou, X., Sheng, Z., Yang, Y., Han, S., Zhang, Q., Li, H., & Yang, Y. (2022). Catchment water storage dynamics and its role in modulating streamflow generation in spectral perspective: a case study in the headwater of Baiyang Lake, China. Hydrology and Earth System Sciences, (November). Retrieved from https://doi.org/10.5194/hess-2022-357