Dear Reviewers, dear Editors,

Thanks a lot for taking the time to look at our revised manuscript again and provide some more useful comments. We appreciate the time and effort you have put in. We have tried to improve the revised manuscript accordingly by addressing all of the reviewers' suggestions. Please find our responses to the comments below (in *italic*). Line numbers in our responses refer to the newly revised version of the manuscript.

Report #1

on the whole, the authors have taken my comments into consideration, even if I think they could have improved and clarified the content of their manuscript with more substantial changes in considering for instance the statistical distribution of the net rainfall (rainfall minus ETR) as a reference rather than the total rainfall since ETR is maintained in their model during rainfall events. The paper is almost ready for publication to my opinion, but an ambiguous and maybe misleading formulation has to be adjusted in the conclusion and may be introduction.

None of the generated peak discharge distributions exhibit heavy tails, but the article shows that there is likely to be a transitional phase between the distribution of frequent events and that of extreme events (see Figure 5). This behavior or shape does not characterize the tail of the peak discharge distribution but is limited to a specific range of return periods. The term tail should be avoided to describe these possible peculiarities of the shapes of peak discharge distributions. It should be more explicitely said that these distributions are likely to show a transition phase with a possible accelerated growth of discharge quantiles with the return period (this had already been stressed by Gaume 2006 if I remember well). Such a formulation would be much more consistent with the results presented in the manuscript...

We thank the reviewer for this comment. First of all, we feel the need to disagree with the statement that "none of the generated peak discharge distributions exhibit heavy tails". We understand that the comment's main focus is on distributions with a process shift, but leaving those aside for a moment, we can consider the flood peak distributions which do not show a transitional phase or step change. These distributions run in parallel to the respective rainfall distributions for all return periods, and some of the underlying rainfall distributions clearly show heavy tail behaviour (GEV shape parameters up to 0.376). Hence, the flood peak distributions closely following these rainfall distributions also exhibit heavy tails.

Regarding the distributions with what the reviewer calls a transitional phase, we agree that such a transitional phase or step change does not characterize the tail of the distribution. However, it does affect the shape parameter of the GEV distribution fitted to such data. This estimated GEV shape parameter is commonly used in hydrological practice to characterize apparent tail behaviour, which in turn is used for estimating magnitudes of rare floods and design floods. While we have tried to make this distinction between the actual *statistical tail of a distribution and the* apparent *tail behaviour of a fitted distribution clear throughout the manuscript, for example through using the terms "apparent tail behaviour" and "estimated GEV shape parameter", we appreciate the reviewer's suggestion for including a formulation regarding the transition phase in the conclusions. We will add the following sentence in l. 462: "Distributions with such a process shift tend to show a step change. While the step change itself does not characterize the tail of the distribution, it does result in a higher estimated value of the shape parameter of the fitted GEV distribution."*

Report #2

The authors have provided additional detail to the extensive methodological questions raised in reviewer comments, improving the paper significantly. The authors have provided expanded rationale for specific methods and assumptions in response to issues raised in my initial review. Many of these are difficult, open-ended problems for which the authors have provided helpful revisions. Two are noted below for additional consideration. The larger issues that need additional treatment concern not the individual modeling components, but the modeling and analysis chain used in the study. There are many links that are stitched together to yield hydrologic analyses. There are several formulations of model-chain questions that would provide useful ways of organizing additional discussion. The most direct question is "Why should one have confidence in the composite analysis system? A simpler formulation is "What are the model components/assumptions that are most questionable? And Why? The added discussion of nonlinearity in runoff generation is a start at addressing this question. The revisions that are recommended are significant, but can be accomplished in a reasonably short time, hence the recommendation of minor revisions.

We thank the reviewer for this comment and appreciate that our revisions are considered helpful. We also thank the reviewer for pointing out where the revisions have not been extensive enough. We will address those points below.

The two questions from the original review that would benefit most from additional discussion:

1) Is the rainfall-runoff model suitable for drawing strong conclusions about upper tails of flood peaks? The assumption of a homogeneous catchment over a 50 km2 scale removes the capability of assessing important processes that can contribute to flood peak response. Does spatial variability of rainfall contribute to tail behaviour at 50 km2 scale? How does this change for 10 km2 scale and 1000 km2 scale?

The authors' responses focus on future studies addressing these issues, the suggestion that 50 km2 catchments are homogeneous in Central Europe and the importance of open channel flow processes (tied to river routing) in larger catchments. The arguments and associated revisions are not persuasive.

We still agree with the reviewer that the aspects of spatial variability and catchment size would be very interesting to analyse in this context, however we can only reiterate that this is beyond the scope of our study. We decided for a simplified catchment representation in the model set-up, i.e. a small homogeneous catchment, to be able to derive clear statements on relations between model inputs and resulting flood peak distributions. Based on the implemented model set-up, we cannot make well-founded statements on the effects of rainfall variability and catchment size on flood peak tail behaviour. We address these limitations and the potential effects that these factors might have in the discussion (l. 410 onwards). In addition, we would like to point out that we are also not attempting to make conclusions about larger or spatially variable catchments, and that we state, both in the discussion and the conclusions, that the findings are valid only for small homogeneous catchments.

To address the reviewer's comment, we have expanded the discussion on spatial variability by adding the following sentence (l. 411): "Spatial variability in rainfall has been linked to heavy-tailed flood peak distributions, and it has been shown that this effect depends on the catchment size (Wang et al., 2023)." We further added (l. 419): "Based on this rationale, we decided for a simplified catchment representation without spatial variabilities. Expanding this set-up in future studies is however deemed very interesting and advisable."

Addressing the suitability of the model chain, we have added (l. 431): "In addition, the model components used have been shown to represent well real-world behaviour when calibrated with real-world data (e.g. Nguyen et al., 2021; Ceola et al., 2015; Parajka et al., 2007).

2) The authors note that "we fit one GEV distribution to the data even when we know that there is a process shift in the runoff generation, which actually violates the assumption of independent and identically distributed values for distribution fitting". This is a serious issue, which is tied to the broader issue raised in the introduction concerning the dependence of estimated upper tail behaviour of observed time series given the sensitivity to the "largest few events".

The authors conclude their response by noting that "when fitting GEV distributions to subsets of a time series of different lengths, the shape parameters may vary due to differences in the estimation uncertainties. To reflect this, we will use the terminology "apparent tail behaviour" when drawing conclusions based on the GEV shape parameter of a distribution fitted to a limited time series." Tail estimates are always subject to issues associated with bias and variance, even if the underlying assumptions hold. Resorting to "apparent tail behaviour" isn't helpful in addressing the issue. Here the key points to address are the consequences that follow from the fact that the underlying GEV assumption does not hold.

We thank the reviewer for addressing this important point again. We would like to stress that in this exercise, we mimic the situation of a hydrologist, who is not aware whether the IID assumption holds. This is a typical situation in practice of flood frequency analysis. If the IID assumption does not hold, we have to be aware that the still fitted GEV distribution might not represent the true underlying distribution well and that the fitted tail might not represent the actual tail. Nevertheless, the fitted distribution gives us a guess on the occurrence probability of extreme events, albeit with uncertainty. How large the error is, depends on a particular case and can only be quantified in dedicated synthetic experiments. As stated in l. 390, more accurate results might be achieved with a mixture distribution when a step change is apparent. However, in hydrological practice it is well established to fit GEV distributions to observed time series, and often it is not known whether the IID assumption really holds. As mentioned in l. 394, this common practice can be problematic at times, for the above-stated reasons. In the manuscript, we will add the following sentence (l. 391): "If we still fit one GEV distribution to the entire data, it does not represent the true underlying distribution and also not the true tail behaviour."