Dear Reviewers, dear Editors,

Thanks a lot for the thorough and constructive reviews with many useful comments! We have tried to improve the original manuscript accordingly, following the reviewers' suggestions as much as possible. Please find our responses to all comments below (in *italic*). Line numbers in our responses refer to the revised version of the manuscript.

RC1: 'Comment on hess-2023-186', Anonymous Referee #1, 28 Aug 2023

The manuscript proposes a derived flood frequency analysis based on the combination of a weather stochastic generator and a relatively complex lumped rainfall-runoff model, namely the HBV model (Parajka, 2007), with 15 parameters. Various stochastic rainfall series are generated according to a GEV random distribution, using three different mean daily precipitation values and varied shape parameter values. For the RR model, based on previous publications and on a sensitivity analysis, focused on the activation of the "very fast runoff" component of the HBV model which controls the magnitude of the generated extreme events, some parameters are fixed and others are varied over reasonable ranges, to cover a large spectrum or possible rainfall-runoff dynamics. The analysis of the relation between these dynamics, the statistical characteristics of the generated rainfall series and the shape of the resulting generated flood frequency distributions is the central focus of the proposed paper.

The manuscript is potentially interesting, but contains several weaknesses and approximations that lead to misinterpretations and erroneous conclusions, that have absolutely to be corrected before it can be published.

We thank the reviewer for the thorough review and comments. We agree with some of the points, find them valuable and made changes accordingly. However, in some cases, we believe that some of the aspects that the reviewer sees as shortcomings are only evaluated that way based on a different set of assumptions than are valid for our work. In the following we address each of the points raised to clarify where and why we see things differently, and also to state which points we adopted and what changes we made to the manuscript.

1) The proposed analysis is based on the determination of a return period over which both distributions of (a) the rainfall depth over a duration equal to the time of concentration of the watershed Ptc and of (b) the peak discharges Q, can be considered to be parallel (i.e. the discharge distribution is entirely controlled by the rainfall distribution because the very fast component dominates the generated runoff in the RR model). The authors argue that "for impervious catchments, the curves of Pct and Q are assumed to run in parallel" (Page 8, L. 185). This interpretation key is false! And this is highly problematic for a manuscript aiming at drawing general conclusions on the shapes of flood peak distributions. In fact, for impervious catchments, the curves of Pct and Q are not only parallel, but superimposed (simple consequence of the rational formula, see also Gaume 2006). This error by the authors is due to the fact that they surprisingly neglect the importance of the deep percolation parameter Cperc in the HBV model, yet clearly visible in the sensitivity analysis (Figure 4). If the deep percolation is not set equal to zero, a significant deep percolation will remain in the model, even for large and rare rainfall events and the RR model can therefore not be considered to represent the behavior of an impervious catchment, unlike what is said by the authors (Fc=Luz=1 mm, line 185). I strongly suspect that the distance between the distributions of Pct and Q is controlled by the parameter Cperc. The influence of the parameter Cperc should therefore be considered and analyzed in the manuscript. And, at least, runs where this percolation is set to 0 must be considered in the analysis as reference runs for asymptotically impervious watersheds. It must be clearly stated that "for impervious catchments, the curves of Pct and Q are superimposed" (this is a straightforward statement…). I let the authors and the readers consider if it is realistic to think that a deep percolation flux,

feeding a very slow runoff component, can remain significant for extreme events. In many cases, this would appear as an unrealistic assumption.

- *We thank the reviewer for raising this point and bringing it to our attention that we set the criteria for what can be considered an impervious catchment too wide. Indeed, we should only consider the model runs to represent impervious conditions where – in addition to FC and Luz – also the percolation parameter Cperc has a value very close to 0. In our simulations the minimum value of Cperc is 0.00042 mm/h. We decided against setting any model parameter to 0 in any of the model runs as we wanted to ensure that the same processes are acting in all model runs, just with different intensity or relative importance. To reflect that neither FC nor Luz nor Cperc are exactly 0, we changed the wording from "impervious catchments" to "close to impervious catchments". This now refers to all model runs for which FC = Luz = 1 mm and Cperc = 0.00042 mm/h. We changed this definition in the manuscript (l.192). In doing so, the number of model runs which are considered to have close to impervious conditions decreases from 105 to 21. When repeating the analysis for finding the duration of P which best represents the concentration time of the catchment with this stricter definition of close to impervious catchment conditions, the results stay qualitatively the same – the concentration time is still considered to be 6h. Figure B1 has been updated accordingly.*
- *We find that the distance between the distributions of Pct and Q does indeed change with the parameter Cperc, as suspected by the reviewer. However, this change is marginal. With regards to the influence of catchment characteristics on the return period beyond which the distributions of Pct and Q run in parallel, we analysed the effect of Cperc and found that it does not show a notable influence (l. 263, Fig. B2).*
- *Regarding the statement that "for impervious catchments, the curves of Pct and Q are not only parallel, but superimposed" we must disagree, at least for our model set-up. Even for impervious catchments, evapotranspiration remains active and therefore Q will always be lower than Pct (see also newly added Figure 5). In the HBV model, a part of the rainfall leaves the system as evapotranspiration and will never become runoff. This is true also during rainfall events, as the actual evapotranspiration in each time step is linked only to the potential evapotranspiration and the state of the soil moisture storage, but not to the incoming rainfall. In fact, when checking this for one of the model runs with close to impervious conditions, the difference between P and Q corresponds to the actual evapotranspiration. Pct and Q would be superimposed for impervious catchments when evapotranspiration is neglected or when only effective rainfall is considered. We therefore stick to the assumption that the curves of Pct and Q run in parallel (l. 204) instead of stating that they must be superimposed as requested by the reviewer. For the analysis where we consider close to impervious conditions, i.e. finding the duration of P which best represents the concentration time of the catchment, the distance between the curves of Pct and Q is not of relevance but only their slopes. To address this aspect in the manuscript, we added (l.196): "One might expect that for impervious catchments, the curves of Q and Pct do not only run in parallel, but are identical. This is not the case for our model set-up for the following reason: In the TUW model, evapotranspiration is active even during extreme rainfall events and so Q is always lower than Pct by at least the amount of actual evapotranspiration taking place. This is a shortcoming of the model, but does not affect our findings as for our analyses only the slopes of the curves and not their distance to each other is relevant."*
- *With regards to the last point, we agree with the reviewer that it seems unrealistic that a deep percolation flux can remain significant for extreme events. However, we do not see that we would make such an assumption. The highest value of Cperc adopted in any model run is 0.25 mm/h which is one to two orders of magnitude lower than the hourly rainfall during extreme events. We argue therefore that the percolation rate is close to negligible compared to the influx of rainfall during extreme events, rather than remaining significant. As already stated in l. 334, the percolation rate seems to act on a longer timescale than is relevant for the generation of extreme runoff.*

2) Second, the authors should absolutely use rigorously the term "tail behavior". In the same manuscript, they conclude both: 1) that the distributions of Pct and Q are asymptotically parallel

and they define a threshold return period based on this property and 2) that the distributions of Q have heavier tails than the distributions of Pct, based on the estimated shape parameters (figure 7 in particular). Both statements are clearly incompatible. The reason is that the generated distributions of Q are obviously not of the GEV type and show a more or less rapid transition phase (see figure 3). It is possible to calibrate a shape parameter of the best suited GEV distribution, based on the generated discharge series, but, since the Q distributions differ clearly from the GEV distribution, the estimated shape parameter value, if it may encapsulate some information about the "transition phase" of the distribution of Q (i.e. its sharpness), it does by no means inform about the tail of this distribution of Q. The authors should correct their misleading interpretations and conclusions on the tails of the distribution of generated peak discharges and should not consider the estimated shape parameters as characterizing the tail of the discharge distributions. This is really an essential point.

- *We thank the reviewer for this comment. In our study, we are considering both asymptotic and pre-asymptotic properties and are aware that this can lead to seemingly contradictory conclusions. Please, refer to the discussion in Merz et al. (2022) on this issue. While we addressed this in the manuscript (e.g. l. 36, l. 344, l. 388), the reviewer's comment shows that we did not make it sufficiently clear yet. When comparing the distribution curves of Pct and Q, we argue that the distributions run in parallel beyond a certain threshold return period. This can be considered an asymptotic property of the distributions. When fitting GEV distributions to Pct and Q, and using their shape parameters for characterizing the tail behaviour, this should be considered a pre-asymptotic property given that the time series for fitting the distribution is of limited length. As rightfully mentioned by the reviewer, the shape parameter of a GEV distribution fitted to a limited time series does not reflect the "true" tail behaviour of the underlying distribution. It can only be considered as an approximation of the true tail behaviour, and this approximation becomes increasingly uncertain the shorter the available time series is. We also agree with the reviewer that the GEV distribution is not the best fit for some Q series (see also the discussion in l. 388) – nevertheless, we fit GEV distributions as this is common practice in hydrology when analysing annual maxima of observed series. However, one needs to keep in mind that especially in these cases the GEV shape parameter does not reflect the true tail behaviour of the underlying distribution. Nevertheless, it can help with achieving more accurate estimations of the occurrence probabilities of extreme events in light of limited time series. To make the distinction between the true tail behaviour of the underlying distribution and the tail behaviour of the fitted GEV distribution clearer, we adopt the following terminology: when fitting a GEV distribution, we will refer to the tail behaviour as characterized by the GEV shape parameter as "apparent tail behaviour" in contrast to the true tail behaviour. In the manuscript, we added the following sentences in the methods section (l. 177): "It should be noted that the shape parameter of a GEV distribution fitted to a time series of limited length does not necessarily reflect the true tail behaviour of the underlying distribution but is only an approximation thereof. 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."*
- *When considering this distinction between asymptotic and pre-asymptotic properties, the two statements mentioned by the reviewer are compatible. Even when the distributions of Pct and Q are asymptotically parallel, their apparent tail behaviour as characterized by the GEV shape parameter can be quite different from one another.*

3) The Scale and position parameters of the Rainfall GEV random model are adjusted to fit three values of mean annual precipitations: i.e. MAP serves as a calibration parameter for the random generator. But it is suggested in the analysis and discussion of the manuscript, that some behaviors may be related to the MAP (see fig. 6 and the corresponding discussion for instance). This is clearly an over-interpretation. Are the results depending on the MAP per se or on the statistical characteristics of generated extreme rainfall events fixed according to the parameter MAP? The later statement is true, but the fact that the MAP by itself may be an explanatory factor is never

demonstrated: are a more frequent saturations of the soil and sub-soil during wet periods really observed and has it a decisive impact on the results? Other choices for the rainfall random generator could have produced similar distributions for the extreme daily rainfall amounts with varied MAP. Would then the discharge peak distributions depend on the MAP? I Wonder. Unless they provide evidence that the annual total rainfall amount has a real impact on the results, I suggest that the authors avoid ambiguity and unsupported conclusions and replace references to the MAP by reference to the average value of the generated rainfall amounts in the interpretation and conclusions.

- *Thanks for the comment. First, we need to clarify two things: 1) the stochastic weather generator is not a "rainfall GEV random model" – it uses an extended Pareto distribution and not a GEV distribution for generating rainfall time series (l. 119), and 2) the MAP does not serve as a calibration parameter in the way that the extended Pareto distribution is fitted to three different MAP values. Instead, rainfall time series are generated based on one MAP value and afterwards shifted to three different levels of MAP by multiplying the hourly rainfall depths with a factor (l. 147). Through this approach, two rainfall time series with the same tail behaviour only differ in their mean rainfall level, but not in other aspects such as the frequency of wet days. Because of this, MAP and the mean event rainfall depth, suggested as an alternative measure to MAP by the reviewer, are highly correlated. In our study, we adopt MAP as a simple measure for quantifying the general wetness of a catchment. We agree with the reviewer that we could also adopt the mean event rainfall depth as a measure. As mentioned above, the two are highly correlated in our set-up and so the results are qualitatively the same independent of which of the two is chosen. We decided to use MAP as it seems to be a more straightforward and widely-used representation of overall catchment wetness. Instead, the mean volume of precipitation events is not a common indicator and might be ambiguous to define and use especially with increasing catchment scale. This is especially true for defining the three wetness levels to which the rainfall series are shifted, and we prefer to stick to the same measure for representing catchment wetness throughout the manuscript. To avoid over-interpretation as pointed out by the reviewer, we added the following statement (l. 250): "MAP should be interpreted here as a measure of overall catchment wetness which controls the mean event rainfall depth."*

4) The authors conduct their analysis on generated series of different lengths – i.e. from 60 to 6000 years. Doing so, they mix two completely different issues and introduce confusion. The first question, the central question of the manuscript I suspect, is the relation between the distribution of Pct and Q, depending on the RR processes and on the rainfall statistics. To tackle this question, it is important to get the most accurate definition of both distributions and hence to work on long series (i.e 6000 years for instance). The second question, which is an interesting but different one, is related to the estimation of the characteristics of both distributions, once defined, in real-life applications, when the lengths of the available measured series are limited. Estimation uncertainties related to sampling may introduce significant variability and blur the general image. It is not uninteresting to mention this second question, to connect this theoretical work to real-life applications, but only once the answers to the first questions are settled. The two questions must clearly be separated in the manuscript to avoid any confusion.

- *We thank the reviewer for pointing out that we have not separated the two aspects clear enough. While the analysis of a threshold return period beyond which Pct and Q run in parallel is conducted only for 6000-year long series, the comparison of estimated GEV shape parameters between P and Q and the influence of the runoff generation on the apparent tail behaviour are addressed for different time series lengths. We restructured the respective parts of the results and discussion sections so that the results based on long time series are thoroughly presented and discussed before moving to shorter time series and the meaning of our findings for real-life applications with limited data availability.*

5) Except for figure 3, which corresponds to one specific case study, no illustration of the generated distributions is provided in the manuscript and the results are analyzed through aggregated values (estimated shape parameters, threshold return periods) … It is essential to provide such

illustrations. I would for instance suggest representing the distribution of Ptc along with the distributions of the various generated discharge series on the same graphic, to show the wide spectrum of generated peak discharge distributions and of their convergence speed towards the distribution of Ptc or a distribution parallel to the it, if the deep percolation is not asymptotically equal to 0 (see comment 1). This graphic could easily be repeated for all the settings of the stochastic rainfall model (combinations of Scale and Shape parameter values). Of course, the figure must be established for the longest available series. They may be repeated for shorter series to illustrate the effect of sampling variability and how it blurs the general sketch. Such a figure would present clearly the obtained results to the readers and illustrate one of the conclusions of Gaume (2006): i.e. concerning the shape of flood peak distributions, the range of possibilities is extremely large.

- *Thanks for the great suggestion. Such a figure will indeed enrich the manuscript as it clearly presents the "raw" results before aggregation. A new figure has been added to the results section (Fig.5): distributions of 6000-year long Pct series with 7 different tail behaviours and 3 different MAP levels are shown along with the distributions of all respective simulated Q series.*

6) The authors should remain prudent when extrapolating their results, especially the typical range of return periods of their defined threshold, to real-life. The rainfall stochastic model and the RR model remain simple approximation of real-life situations. Moreover, the RR model is implemented well beyond the range of events against which it could be calibrated and evaluated. Who knows really how watershed behave during extreme rainfall events, what type of thresholds, non-linearities or discontinuities may appear when extreme floods occur? Existing RR models, based on continuous and asymptotically converging reservoir models produce generally, to my experience, to smoothed evolutions of the RR relation. They may therefore help explore only part of the range of possibilities for the shape of flood peak distributions when used in derived flood frequency approaches… This limitation should be acknowledged in the manuscript.

- *We are aware of the limitations of extrapolating simulation results to the real world and state this in l. 416: "Furthermore, our findings are based on synthetic catchments and simulation runs. While such an approach has major advantages like the generation of long time series, results are not always directly transferable to the real world." In our case, the results are also limited in the way that we consider homogeneous catchment conditions and do not include spatial variability of rainfall or catchment characteristics (see l. 409). However, we agree with the reviewer that these limitations should be made more explicit. Following the reviewer's suggestion, we expanded this part as follows (l. 418): "In the adopted rainfall-runoff model only one nonlinearity in the runoff generation was considered, namely the activation of an additional very fast runoff component. However, in a real catchment multiple nonlinearities and process shifts might be present such as the onset of overland flow, the onset of subsurface stormflow, the activation of macropores or the temporary expansion of the river network. The model does not include all these processes explicitly and is therefore, as all models, a simplified representation of reality. Hence, the simulated flood peak distributions are also only representative for this simplified reality. Nevertheless, they can help us explore results which can be valuable for real-world applications."*

I join an annotated manuscript complementing this review. The discussion part, containing several questionable statements, should also be revised in-depth.

In the following, we address all comments made in the annotated manuscript one by one. Please note that here page and line numbers of the comments refer to the reviewed manuscript, i.e. the original submission, while line numbers in the responses refer to the revised version.

P.1 L.3: the occurrence of
In this sentence, we do not refer to an underestimation of the occurrence probability of rarely observed events, but rather an underestimation of their magnitude. Added: "the magnitude of"

P.1 L. 15: Can such an important conclusion be drawn based on simulation results only? I doubt. Be reasonable.
We removed this sentence.

P.2 L. 40: Should attempts based on historical or paleofloods or regional flood frequency analyses not be mentioned here? See for instance the attempts by Gaume et al (2010, http://dx.doi.org/10.1016/j.jhydrol.2010.01.008) for instance or the subsequent papers (https://doi.org/10.1016/j.jhydrol.2016.01.017, https://doi.org/10.1016/j.jhydrol.2013.09.058), not suffiicently considered in my opinion :=)
Figure 7 of the paper Gaume et al. (2010) shows a clear evidence of heavy observed tails, based on a regional set of ungauged extreme in the Cevennes-Vivarais region in the South of France, a region frequently affected by severe flash floods. This illustration could be mentioned here.
We added: "[…] or by including historical or paleoflood records (e.g. Stedinger & Cohn, 1986; Vorogushyn et al., 2021), or through regionalisation approaches (e.g. Merz & Blöschl, 2005; Gaume et al., 2010)."

P. 3 L.89: Yes and no : modelling approaches are flexible and help testing hypotheses, but models are not the reality. It can hardly been argued that models may realsitically represent extreme events on which they could not be calibrated or even tested...
To acknowledge this, we added: "It should be kept in mind though that models can only be a simplified representation of reality."

P.4 L.95: Already said before. Reorganise slightly the introduction to avoid repetitions...
We feel that summarising here what has been said before in more detail in the Introduction is crucial for establishing the research questions we want to address.

P.5 L.133: I do not understand why T and PET had to be averaged. This limits the possible variability of antecedent soil moisture conditions.
By using a "standard" year of T and PET, there is still variability of antecedent soil moisture conditions - but this variability is mainly linked to the P time series and the model parameter values set for the respective run (e.g. FC), i.e. by the factors that we are interested in and are analysing in our study. As stated, the aim was to minimize confounding effects that might blur the results.

P.5 L.138: This procedure, with exclusion of too large shape parameter values is unclear... GEV is fitted to 3 series ? Why is that necessary since the multiplication does only affect the position and sclae parameter of the GEV ? Why are the P series of 60 years with large estimated shape parameters excluded ? The shape parameter is also generated and known. Too large values can be exclude prior to the generation of the random 60 years series... Clarify and simplify this methodological part.
Thanks for this comment. We realized that we described the steps of GEV fitting and shifting to different mean levels in the wrong order, leading to some of the confusion expressed here. We first fit GEV distributions and exclude P time series based on their shape parameter, and then shift the remaining P series to three different levels. We changed the order of the sentences accordingly.
With regards to the exclusion of P series with high GEV shape parameters: the GEV shape parameter is not used for generating the data and so also not known beforehand. Instead, the generation of P data with the weather generator is based on the shape parameter of the extended Generalized Pareto distribution (extGP). We first fit the extGP distribution to daily data and also generate continuous daily time series, but then analyse GEV shape parameters based on annual maxima of hourly time series. We decided to generate P series with a rather wide range of extGP shape parameters, to ensure that when going from daily to hourly values and then to annual maxima, we would still cover a wide range of tail behaviours. In the end, we covered a range that was wider than what has been estimated for observed precipitation time series and so we excluded the ones well outside the observed range. Please also refer to our responses to the comments 1 and 3 by Reviewer #2.

P.6 L.141: Why are the shape parameters varied and the resulting series finally combined ? This is not logical. The simplest method would have consisted in generating 6000 years series with 3 contrasted shape parameter values.
This is again linked to the aspect that we are using extGP distributions to generate continuous daily data, while using GEV distributions and annual maxima in the analyses. For example, some time series with the lowest extGP shape parameter result in GEV distributions that have a higher shape parameter than the time series with the second lowest extGP shape parameter. By combining the different realisations of 60 years to long time series in the way we did, we ensure the strongest differences between the time series with regards to the aspect that we are interested in, i.e. the GEV shape parameter.
*To make it more clear when we are using which distribution, we added "extGP" or "GEV" in front of "shape parameter" in the respective parts of the methods section (l. 119-124, l. 142-147) (e.g.: "manipulation of the **extGP** upper tail shape parameter", "**GEV** shape parameters well outside the observed range").*

P.6 L.142: This last detail is not necessary...
We removed "i.e. it is doubled and later removed again."

P.7 L.158: Was it really necessary to add complexity running the model at hourly time scale to conduct this sensitivity analysis based on rainfallrunoff simulation. I am not convinced...
In the current study, we are assuming spatially homogeneous conditions of both the rainfall and the catchment characteristics. This assumption is only valid for small catchments, and running the model on a small catchment requires a temporal resolution that is higher than daily to capture all processes adequately.

P.7 L.166: The length of the series used for the fitting have an influence on the estimation uncertainties, but not on the tail behavior. I do not understand the usefulness of working with short simulated series. The authors try to adress two very different questions namely : 1) the resulting shape of the annual peak discharge distribution and 2) the evaluation of this shape based on a sample of limited size and the role of sampling variability. Be more accurate about the objectives of the poroposed approach. Moreover, the generated peak discharge series are not necessarily distributed according to a GEV law. Computing a corresponding GEV shape parameter reduces drastically the information content of the simulation results....
Please refer to our responses to comments 2 and 4 by Reviewer #1. Analysing time series of different lengths is important for evaluating the meaning of our findings for real-life applications with limited data availability. We have now separated the findings based on long time series and the ones based on shorter time series more clearly. Further, we added here: "It should be noted that the shape parameter of a GEV distribution fitted to a time series of limited length does not necessarily reflect the true tail behaviour of the underlying distribution but is only an approximation thereof. 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."

P.7 L. 171: This is not what I said in the 2006 paper. The equivalence of slope is the hypotehsis on which the gradex is based. I did demonstrate that this hypothesis was unfounded and that the distribution of Q converges asymptotically towards the distribution of P (i.e. when both are expressed in the same units) because generally the runoff coefficient tends towards 1 for extreme events.
This sentence is based on the conclusions (a) and (b) by Gaume (2006), where it is stated that "the distribution of [the peak discharge] Y will appear linear and with the same slope as the distribution of [the rainfall intensity] X […] on a log-log plot for 'extreme value' of type II distributions" and that the distribution "of the mean maximum rainfall intensity over a duration of the order of the time of concentration of a watershed should be considered and used as a guideline for any extrapolation of a flood peak distribution". We are sorry if we misunderstood or misinterpreted these conclusions. If we understand the reviewer's comment correctly, it points again to the question of whether the curves of P and Q run in parallel only or are even superimposed. This has been addressed in detail in our response to comment 1 by Reviewer #1. Here, we rephrased the sentence slightly so that the reference to Gaume (2006) only refers to the second part of the sentence, which now includes a direct citation.

This way, we hope that no-one feels misunderstood. The sentence now reads: "On such a plot, it is assumed that beyond the threshold return period the slope of the distribution of Q is the same as the slope of the distribution of P, given that P is considered "over a duration of the order of the time of concentration" (Gaume, 2006) of the catchment."

P. 7 L. 180: For an impervious watershed the distribution of Q is the same as the distribution of P over the time of concentration of the watershed (cite the rational formula). Of course again this holds is both are expressed in the same unit (mm/h for instance).
Please refer to our response to comment 1 by Reviewer #1 where the aspect of parallel vs. superimposed is discussed in detail. Here, we added: "One might expect that for impervious catchments, the curves of Q and Pct do not only run in parallel, but are identical. This is not the case for our model set-up for the following reason: In the TUW model, evapotranspiration is active even during extreme rainfall events and so Q is always lower than Pct by at least the amount of actual evapotranspiration taking place. This is a shortcoming of the model, but does not affect our findings as for our analyses only the slopes of the curves and not their distance to each other is relevant."

P.8 L.185: No, the curves should be superimposed and not only parallel. Please illustrate such curves in the manuscript.
Please refer to our responses to the comments 1 and 5 by Reviewer #1. A respective figure has been added (Fig. 5)

P.8 L.196: Again, I do not understand why the length of the time series is considered as important and what it should reveal...
As stated, the aim of analysing different time series length is to "compare results for very long time series and time series of typically observed lengths." This helps with transferring the findings to real-world applications where observed time series are usually limited in length. Please also refer to our response to comment 4 by Reviewer #1.

P.9 L.207: Rational method is not an asumption but a fact if the runoff coeffcient is supposed to be constant : case of impervious watersheds.
We are aware that this statement is not an assumption but a fact for very specific conditions (such as impervious watersheds). However, we are not interested in such very specific conditions but in the usual case, where this relation is not a fact but a (plausible) assumption.

P.9 L.212: The time of concentration clearly depends on the parameters of the RR model. It does not make sense to look for a constant value, independent of the model parameters. Likewise, the time of concentration depends on the dominant processes or flow path in the RR model : i.e. wether fast or slow runoff components dominate. 6 hours of TC for major floods on a 50 km2 watershed seems very high to me. I would rather guess realistic values - again for large floods - between 2 and 3 hours.
We agree that the time of concentration (tc) is not necessarily a constant value. In fact, Michailidi et al. (2018) found that "the time of concentration is a negative power function of the runoff intensity", i.e. tc decreases with increasing runoff intensity. However, due to the power function relation, differences in tc are larger between small and large events than they are between large and very large events. The larger the events, the more tc approaches a constant value. Since we are analysing annual maxima and are interested in extreme events, using a constant value for tc is a fair approximation.
We also agree that tc depends on whether fast or slow runoff components dominate. For estimating tc, we use impervious catchments where we know that the fast runoff component is active. We assume that this is also the case for the largest events in all other catchments, and so tc should be very similar between the largest events and impervious catchments. It is especially relevant to get tc right for these largest events because that is what we are interested in here and that is where tc becomes close to constant.
A lower value of tc might be estimated if also river routing was included, but for our current model set-up a tc of 6h is considered realistic.

P.10 L.224: In fact the increasing MAP results in an increasing rainfall event mean precipitation in the proposed Rainfall stochastic model...
Please refer to our response to comment 3 by Reviewer #1. We added: "MAP should be interpreted here as a measure of overall catchment wetness which controls the mean event rainfall depth."

P.10 L. 230: I do not agree with that statement. Why should this be unlikely ?
If rainfall gets large enough, every catchment will be saturated eventually, resulting in a direct transformation of rainfall to runoff. For some catchments this might be only for very high return periods, but from physical considerations this point will be reached eventually. We added here as explanation: "For all catchments a point is reached eventually where rainfall is so extreme that saturation occurs and rainfall translates directly to runoff."

P.10 L.235: I suspect that the distance between the two curves may be explained by Cperc. Could you verify this ? It would be an important information for an in-depth interpretation of the results.
Please refer to our response to comment 1 by Reviewer #1.

P.11 Fig.6: It is surprising to see that the parameters FC and Luz have almost the same impact on the computed threshold RP... Considering this fact, the analysis could certainly have been further simplified...
In retrospective we probably could have set FC to a constant value and only vary L_{UZ} between model runs. However, the sensitivity analysis identified FC as one of the relevant parameters and so we included it as a variable parameter in the analyses. Before conducting the analyses, it was not obvious that both storage parameters would show similar effects.

P.11 L.240: I am convinced that it is not directly the MAP that is controling the response of the RR model, but rather the average event rainfall amount which is proportional to the MAP. Please reformulate the analysis that could be misleading. It is not the storage/MAP ratio which is the explanatory factor but the rainfall event amount / storage ratio...
Please refer to our response to comment 3 by Reviewer #1.

P.11 L.250: Yes, but this is due to the complexity of the model used. Fc and Luz are close to zero, but still Cperc exists and the model combines a fast response and a slow flow component that remain non-negligeable for moderate rainfall events. To evaluate totally impervious situation, it would be necessary to set also Cperc to zero. This will solve two problems : the distributions of P and Q will be parallel but furthermore will be superimposed....
Please refer to our response to comment 1 by Reviewer #1. We now also set C_{PERC} to a value close to zero for impervious catchments.

P. 12 Fig.7: It is worth noting that Xsi calibrated on the discharge series appear most generally larger than the Xsi value of the rainfall distribution.
We agree and mention this in l.355: "[…] for time series longer than 200 years and shape parameters of P greater than 0.2, the shape parameter of P seems to be a kind of lower bound for the shape parameter of Q".

P.12 L. 255: The GEV shape parameter is still comprised between 0.18 and 0.38. It is the estimated value, based on a limited size series that varies between -0.31 and 0.56.
We added "estimated" in front of "GEV shape parameter".

P.12 L.257: the estimated GEV
Added.

P.12 L.260: As said before, what is evaluated for short tme series are the estimation uncertainties related to sampling variability. It is not directly related to the topic of the manuscript. The focus should be set on the results obtained for the 6000 year series.
Please refer to our response to comments 2 and 4 by Reviewer #1. We feel it is valuable to analyse time series of different lengths to make the findings relatable for observed time series of limited length.

We restructured the respective part of the results to separate more clearly between findings based on long time series and those based on shorter time series. We also added: "This is expected due to the larger estimation uncertainties for shorter time series."

P.13 L.268: Be careful in the interpretations. As is said earlier in the manuscript, the tail of the Q distribution is constrained by the distribution of P. The high estimated shape parameters are not characterizing the tails of the Q distribution but a transition phase beteween frequent floods controled by infiltration and rare floods dominated by fast runoff components... Be more accurate in your comments.
Please refer to our response to comment 2 by Reviewer #1 where this aspect is addressed in detail. Changed here to "stronger apparent heavy tail behaviour".

P.13 L.272: This threshold return period is related to the RR model used an its parameters. Do not try to extrapolate the conclusions to real life...
Please refer to our response to comment 6 by Reviewer #1.

P.13 L.283: I do not understand this eliptic sentence. Please explain what you mean with peculiarities in the rainfall...
Please see the next comment for a joint response.

P.13 L.286: Really unclear and not convincing. Illustrate the conducted tests and the "bump" or remove this explanation.
Based on this and the previous comment we feel that the respective paragraph led to more confusion than it helped clarifying things. We therefore removed the entire paragraph.

P.13 L.289: Why unlikely. On the contrary, Gaume et al (2006) concludes that the range of possibilities for the shape of the Q distribution is large.
We agree that the possibilities for the shape of the Q distribution are large, but this is mainly the case for low and medium return periods. Gaume (2006) states "that flood peak distributions may have a large variety of shapes depending on the dynamics of the rainfall-runoff process especially as far as the medium range return periods are concerned" and that there is a "distance between the asymptotic distribution and the low return period flood quantiles", while for "the quantile estimations of very large return period floods, the distribution of the maximum mean rainfall intensity over a duration of the order of the time of concentration of a watershed should be considered as the possible flood peak asymptotic distribution". Different Q distributions might converge differently towards the asymptotic distribution, but they will converge eventually. As also stated above in response to the comment on P.10 in L. 230, if P gets large enough, every catchment will be saturated eventually, resulting in a direct transformation of rainfall to runoff. We added a sentence stating that this might occur only for very high return periods in some catchments (see next comment).

P.13 L.292: Yes, but the return period of the saturating rainfall event may be extremely large...
Added: "For some catchments, the return period of such saturating rainfall events might be extremely large."

P.13 L.293: Yes of course, this is trivial.
We still deem it necessary to report this finding.

P.14 L.302: Again, even if compex the RR model used remains simplistic if compared to reality. Be prudent when extrapolating to real-world !!!
We do not state here that the results can be transferred to the real world, but that drawing conclusions as stated in the sentence before would assume the transferability. We also state in the next sentence that this needs to be tested, e.g. based on real-world observations. To clarify, we added: "[…], which would still need to be verified."

P.14 L.304: Yes but it has a major influence on the final result : distribution of Q parallel to the distribution of P rather than superimposed. It is really unfortunate that you missed this aspect.

Please refer to our response to comment 1 by Reviewer #1.

P.14 L.311: Some publications came to the opposite conclusions : high-elevation catchments have larger subsurface storage capacities due to the deep fracturation and weathering of the bedrock and to the importance of moraines (I am not able to find the paper I am thinking of...).
There might be high-elevation catchments with large subsurface storages. However, it is reasonable to assume that many low-lying catchments have thicker soils and with that larger soil moisture storage capacities than high-elevation catchments, and soil moisture is one of the relevant subsurface storages considered in our model set-up.

P.14 L.318: Confused ! Again do not say that the Shape parameter characterizes the tail of the distribution. Q and P tails are obviously identical – at least their shape parameters). Here Xsi is calibrated on the entire series which is also obviously not corresponding to a GEV distribution. The computed value reveal the existence of a transition phase in the simulated distributions and do not characterize the tail of the distribution.
Please refer to our response to comment 2 by Reviewer #1. Even when the tails of P and Q are asymptotically parallel, their GEV shape parameters estimated for time series of limited length (i.e. pre-asymptotic behaviour) can be different. In line with our response to comment 2 by Reviewer #1, we changed "tail behaviour" to "apparent tail behaviour" in this paragraph. We further changed the respective sentence as follows: "However, even when the tail of the rainfall distribution controls the tail of the flood peak distribution asymptotically, both distributions do not necessarily have the same shape parameter when estimated for a time series of limited length."

P.14 L.325: The results obtained for short time series introduces confusion rather than information and this is mainly because the authors confuse the value of the parameter of the underlying distribution and its estimated values based on limited size samples.
Please refer to our response to comments 2 and 4 by Reviewer #1. We restructured the respective parts to distinguish more clearly between results obtained for long time series and those for shorter time series, and also distinguish more clearly now between "true" and "apparent" tail behaviour.

P.14 L.327: It is not an artefact, it is the result of sampling variability and its effect on estimation of parameters.
Changed to: "This is due to higher sampling uncertainty for short time series […]"

P.15 L.340: Too long comment to mention that the RR simulations may not be realistic, even for extreme events. An explanation that has been overseen by the authors is that the generated sample is much larger than the available measured sample and hence, the probability to observe large estimated shape values is higher in the simulated sample. If a comparison between simulated and observed values had to be conducted the distribution of both estimates had to be compared and not only the maximum values.
We agree that the larger sample size could potentially also be an explanation for estimating higher GEV shape parameters in our study than for observations, but this explanation does not apply here. As indicated by the reviewer, in such a case one would expect that the distributions of estimated shape parameters only differ in their maximum but not in their mean. We checked this and found that also the mean of the estimated GEV shape parameters is higher in our study than it is for the observed values. We will therefore stick with the explanations given in the manuscript. To address the reviewer's comment, we shortened and condensed the respective part.

P.15 L.352: Again, the tail of Qdistributios is not heavier !!!! Be accurate please !
Please refer to our response to comment 2 by Reviewer #1. Changed to "apparent heavy tail behaviour"

P.15 L.354: NO, no and no ! You do not show that Q tails are heavier ! Never !
Please refer to our response to comment 2 by Reviewer #1. Changed to "flood peak distributions with higher estimated GEV shape parameters and apparent heavy tail behaviour"

P.15 L.357: Why does the shift violate the assumption of IID ? Here you make a confusion between physical processes and statistical properties... Measurement errors have various possible sources, but series of measurement errors are considered and studied as iid Gaussian variables in statistics. Please remove this sentence.
It is true that a process shift does not automatically mean a violation of the assumption of IID values, but it does make it likely. We do not want to remove the sentence as we feel that it is important to discuss this aspect and also Reviewer #2 stated that this is a crucial point. To clarify, we changed the sentences as follows: "In this study, we fit one GEV distribution to the data even when we know that there is a process shift in the runoff generation. This might actually violate the assumption of IID values for distribution fitting, if the values below and above the threshold are not identically distributed."

P.15 L.358: I am not sure that this clumsy part of the discussion is really useful. PLease remove or reformulate. Mixture distribution is not a solution, but one conclusion can be that the distribution of Q may not ressemble any simple shape statistical distribution and that the common practice consisting in fitting such a distribution to short measured AM series is risky.
We feel it is adequate and important to mention mixture distributions in this context. When we say that fitting one GEV distribution might be inappropriate for some time series, we should also mention what might be an alternative to it. In line with the reviewer's suggestion, we added the following sentences: "Our results indicate though that this common practice can be problematic at times because a GEV distribution is not always a good fit, even when considering annual maxima. While the GEV distribution is the asymptotic distribution of independent block maxima (Fisher & Tippett, 1928), we usually consider time series of limited length and pre-asymptotic behaviour."

P.16 L.397: In fact, the mean annual precipitation controls here the mean event rainfall depth in the generator. Recall this to avoid misinterpretation of the sentence.
We added: "MAP reflects here the overall catchment wetness and controls the mean event rainfall depth."

P.16 L.400: This is of course a trivial result since sources of variabilities have been introduced in the RR model to generate the distributions of Q.
That there is some more variability might be trivial, but that we found "much larger" variability is not trivial. In our opinion, this is an important finding that should be reported in the conclusions.

P.17 L.407: Not heavier tail, but distributions of Q with "transition" phases.
Please refer to our response to comment 2 by Reviewer #1 where this aspect is addressed in detail. Changed here to "stronger apparent heavy tail behaviour".

RC3: 'Reply on AC1: Parallelism of Pct and Q, effect of Cperc, impervious catchments', Anonymous Referee #1, 07 Sep 2023

If I understand the authors' answer correctly, the parallelism is attributed to the fact that evapotranspiration remains active in the proposed model setting, even during heavy rainfall events. This, of course, must be clearly explained in the revised version of the manuscript. In fact, since the atmosphere is, by definition, saturated or close to saturation during rainfall events, especially during significant rainfall events, evapotranspiration is drastically reduced. Maintaining a high value of evapotranspiration during rainfall events is a shortcoming of the proposed model settings and is quite unrealistic, and this should also be acknowledged. I must insist on this point because it is key information: for impervious catchments and spatially uniform rainfall intensities, the distributions of Pct and Q should, at least asymptotically, not only be parallel but also superimposed if physically realistic conditions are considered for the RR relation !

Yes, this was understood correctly. In the HBV-like model that we are using, evapotranspiration can take place even during rainfall events. This is indeed a shortcoming of the model and is now acknowledged in the revised manuscript in the following way (l. 196): "One might expect that for impervious catchments, the curves of Q and Pct do not only run in parallel, but are identical. This is not the case for our model set-up for the following reason: In the TUW model, evapotranspiration is active even during extreme rainfall events and so Q is always lower than Pct by at least the amount of actual evapotranspiration taking place. This is a shortcoming of the model, but does not affect our findings as for our analyses only the slopes of the curves and not their distance to each other is relevant."

RC2: 'Comment on hess-2023-186', Anonymous Referee #2, 06 Sep 2023

The authors examine the relationship between upper tail behaviour of rainfall and flood peak distributions through analyses that are based on a stochastic weather simulator and a rainfall runoff model. Results are used to conclude that "runoff generation can strongly modulate the behaviour of flood distributions"… "threshold processes in runoff generation lead to heavier tails"… and that "for return periods that are mostly of interest to flood risk management, runoff generation is often a more pronounced control of flood heavy tails than precipitation". The modeling and analysis chain used in this study includes assumptions, approximations and subjective judgements that are not compatible with the strong conclusions that are drawn. The analyses are interesting and provide the foundation for a useful paper, with more modest scope and expanded treatment of uncertainties in the analyses.

We thank the reviewer for the very helpful review and comments. We have thoroughly checked our assumptions and conclusions, and it seems like some conclusions came across stronger than they were intended to. Thanks for bringing this to our attention. We have now addressed all the points raised by the reviewer and weakened some conclusions where deemed appropriate, for example by adding that the findings mainly hold for small, homogeneous catchments and Central European conditions. Below are our detailed responses to the reviewer's comments, along with the respective changes that we made in the manuscript.

Specific issues/questions with the modeling / analysis chain are enumerated below:

1) Is the stochastic weather generator suitable for representing the upper tail of rainfall? Do observations from Bamberg provide a suitable grounding for a general assessment of rainfall extremes? Why vary the GP shape parameters between 0.2 and 2.0? Do the 0.9/0.1 day/night PET scaling assumptions have an impact on results? The weather generator produces daily rainfall, but the authors note that shorter duration data are needed for modeling studies of a 50 km2 watershed. Is the "Method of Fragments" suitable for reproducing the climatology of sub-daily time scales? Each of these issues requires supporting arguments to justify the strong conclusions. The reliance on observations from a single site in Germany and the method used for producing hourly observations are of particular concern.

Thanks for raising these questions. We address them one by one:

- *In the stochastic weather generator an extended Generalized Pareto (extGP) distribution is used to simulate precipitation. The Generalized Pareto (GP) distribution can be heavy-tailed and is suitable to capture extreme rainfall values, but might miss the lower bulk (Nguyen et al., 2021). Using an extended GP distribution "allows a smooth transition between bulk of the distribution and the heavy tails" (Nguyen et al., 2021). When evaluating the weather generator based on the extGP distribution for 528 stations in Central Europe, Nguyen et al. (2021) found that both the daily mean and the extreme (99.9th percentile) precipitation intensities are well captured. We are*

therefore convinced that the weather generator is suitable for our study. We added the following sentence in the methods section (l. 113): "The weather generator has been evaluated to capture both the daily mean and the extreme (99.9th percentile) precipitation intensities well for a large set of weather stations in Central Europe (Nguyen et al., 2021)."

- *The observations from Bamberg were chosen for setting up the weather generator due to the availability of long daily and hourly records at this station. First, we estimate the extGP distribution parameters using the station's daily data. Then, we manipulate the upper tail shape parameter with various reasonable values to create time series that illustrate different degrees of extreme frequency. For our study, it is of main importance to have a large set of different rainfall extremes, and this is achieved by setting the extGP upper tail shape parameter to a range of different values. We added (l.123): "Through this manipulation of the extGP upper tail shape parameter, time series with different degrees of extreme frequency are created, despite using observations from just one station as initial input."*
- *The shape parameters of the extGP distribution were not actually varied between 0.2 and 2.0, but the shape parameter of the distribution fitted to the data from Bamberg was multiplied with factors in this range. This way, the extGP shape parameters covered the range of shape parameters from distributions fitted to observations from 528 weather stations in Central Europe. These are the stations used by Nguyen et al. (2021) for calibrating and evaluating the weather generator. Factor 2.0 leads to shape parameters that are higher than most of the "observed" shape parameters. However, we wanted to create a wide range of tail behaviours and rather narrow it down afterwards based on GEV shape parameters as our analysis is based on annual maxima and not on daily data (see also our response to point 3). We added in the methods section (l. 121): "This way, the upper shape parameter covers the range of values that was found when fitting extGP distributions to observations from the large set of Central European weather stations analysed by Nguyen et al. (2021)."*
- *The fragments for the disaggregation of PET were assigned as 0.9 for day times and 0.1 for night times to represent that the largest share of evapotranspiration occurs during the day. To check the sensitivity, we set up a test case of 189 model runs over 60 years with different disaggregation of PET. We compared the original fragments of 0.9 for day times and 0.1 for night times to fragments of 0.5 each, meaning that PET was set to be constant for every 24 hours. This is considered to be an extreme and not very realistic disaggregation of PET from daily to hourly values and should only serve for the evaluation of the sensitivity. To analyse the sensitivity, the GEV shape parameters of the simulated discharge time series were estimated for both PET disaggregation schemes. The root mean squared error between the shape parameters was estimated to be 0.00921. Based on this we are confident that the disaggregation of PET has hardly any effect on the results – at least for small, homogeneous catchments – and that any realistic disaggregation scheme would lead to the same results as presented in the manuscript.*
- *The Method-of-Fragments (MOF) is a commonly used method when disaggregating rainfall time series (e.g. Carreau et al., 2019; Li et al., 2018; Lu et al., 2015; Sharma and Srikanthan, 2006; Westra et al., 2012). In a comparison of different rainfall disaggregation models, the nonparametric MOF was found to outperform point process-based and cascade models (Pui et al., 2021). It is able to better match the observed intensity-frequency relationship than the other models, and this was found to be particularly true for extreme rainfall characteristics (Pui et al., 2012). For more details on MOF and its performance, see also Guan et al. (2023). We added in the methods section (l. 127): "The MOF is a commonly used method for the disaggregation of rainfall (e.g. Carreau et al., 2019; Li et al., 2018; Lu et al., 2015; Westra et al., 2012), and has been found to outperform other disaggregation models, especially for extreme rainfall characteristics (Pui et al., 2012)."*

2) 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?

We thank the reviewer for this comment. We are aware that assuming homogeneous conditions throughout the catchment is an assumption that does not allow to include all processes which might affect tail behaviour. However, as stated in l. 414, including more processes and adding spatial variability of rainfall or runoff characteristics makes it difficult to isolate their effects. We strongly agree that spatial variability and the influence of the catchment size would be interesting aspects for future studies and suggest respective analyses (l. 412), but this is beyond the scope of the presented study. In fact, our follow-up study is devoted to the analysis of the effect of spatial variability of rainfall and runoff on tail heaviness. In the current set-up it would not be advisable to analyse different catchment sizes for various reasons: 1) for larger catchments, the assumption of homogeneous conditions is questionable, and 2) the current model set-up does not include river routing which would become increasingly important for large catchments. To acknowledge these limitations, we weakened some conclusions by stating more clearly that they are only valid for small, homogeneous catchments. For example, we added in l. 400: "[…] at least for small homogeneous catchments in Central Europe".

3) It is disconcerting that 300 of the 1000 precipitation series are discarded because of a large, estimated shape value. Having thresholds based on estimated shape parameters can contort inferred distributions in simulation studies like those used in this paper. How are results dependent on the 0.37 threshold for precipitation shape parameters? Was 0.37 chosen because it is a little larger than 0.33? The principal concern is that a subjective decision on the extreme nature of rainfall is an important component of the analysis chain that leads to the conclusion that flood peaks do not depend as strongly on rainfall as on runoff production.

We thank the reviewer for this comment. We understand that limiting the extremeness in a study on extremes can seem inappropriate or misleading in some cases, but we are confident that limiting the GEV shape parameter of precipitation (P) time series does not affect our conclusions. It would be a completely different story had we excluded individual extreme events instead of entire time series with too high GEV shape parameters. As mentioned in our response to point 1, we generated P series with a rather wide range of extGP shape parameters, to be able to potentially exclude some time series based on their GEV distributions. We wanted to ensure that when going from daily to hourly values and then to annual maxima, we would still cover a wide range of tail behaviours. As it turned out, we covered a range that was wider than what has been estimated for observed precipitation time series and so we excluded the ones well outside the observed range.

As seen in Fig. 7, we still cover a range of GEV shape parameters, even after excluding the P time series with the largest GEV shape parameters. It can also be seen that all P shape parameters result in a wide range of Q shape parameters, and that the minimum and maximum of the Q shape parameters seem to linearly increase with increasing P shape parameter for the time series of 6000 years. This relation would expand also for larger P shape parameters, so that adding the respective time series would not change the picture. However, we do not feel comfortable using P shape parameters well outside the observed range in this analysis. The cut-off of 0.37 was chosen because it is a little larger than the estimated maximum of 0.33 for observed P time series in Germany (Vorogushyn et al., 2023), but still seems to be a reasonable GEV shape parameter. As described above, using a different value would not qualitatively affect the results presented.

With regards to the last sentence of this comment, we feel the need to clarify that we did not infer that flood peaks depend more on runoff generation than on rainfall – instead, we are showing that P becomes increasingly important the larger an event is, until eventually a threshold is reached beyond which the runoff generation has no effect.

We added in the methods section (l. 146): "Using a slightly different cut-off than 0.37 for excluding P time series with very high GEV shape parameters was not found to affect the findings."

4) Model simulations are grounded in subjective decisions concerning model parameters. Sensitivity analyses of model parameters are used to select a sub-set for numerical experiments. These

parameters are varied across a "reasonable range", with the other parameters set to fixed values based on previous studies in Austria. Strong arguments are needed to support the assumptions that the HBV model with the range of selected parameters captures the world of extreme flood response.

We thank the reviewer for this comment. As we are running the model simulations on a synthetic catchment, we could not calibrate the model against observations and so had to make some decisions regarding the model parameters. However, we aimed at being as little subjective about this as possible by basing our assumptions about the parameters on previous studies. The goal was not to capture the entire world of extreme flood responses, and we are aware that there might be other parameter combinations not covered in our set-up which also result in extreme floods. Instead, the main focus was to capture events with and without the fastest runoff component being active. The sensitivity analysis of the model parameters was set up accordingly. The parameter ranges that we used are based on the study by Parajka et al. (2007). The ranges cover a large span of values and are commonly used in studies using the TUWmodel (e.g. Merz et al., 2011; Ceola et al., 2015). For example, Ceola et al. (2015) adopted the same ranges of model parameters in a study where they calibrated the TUWmodel for catchments in Italy, Switzerland, Austria, Germany and Sweden – i.e. for catchments in different topographic settings and covering a range of meteorological conditions. We therefore believe that the ranges are wide enough to capture many extreme flood responses, even when not necessarily capturing all possible extremes. We added in the methods section (l. 164): "The same parameter ranges have been used by Ceola et al. (2015) for calibrating the TUWmodel for European catchments with different topographic and meteorological conditions, and are therefore deemed appropriate for capturing many different extreme flood responses."

5) How well does the upper subsurface storage in the HBV model represent threshold-dependent flood response, especially given the assumption of spatial homogeneity? Is the approach suitable for a broad range of basin scales? Choosing a storage-dependent variable for runoff production is a strong assumption that is likely violated in some settings, especially for the most extreme events. Incorporation of runoff processes that are more sensitive to rainfall rate could lead to markedly different conclusions. Infiltration excess runoff mechanisms, and their variants, play an important role for extreme floods in many settings, especially for arid/semi-arid regions (which are prominent settings for inferred heavy tails of flood frequency distributions).

We thank the reviewer for this important comment. The exceedance of the upper subsurface storage is of course not the only threshold process in the runoff generation that could act in a catchment. We selected it as one representation of threshold behaviour that could be reasonably implemented in the simulation model. In contrast, infiltration excess cannot be well represented in the model, as is the case for other threshold processes such as the onset of preferential flow through macropores or a temporary expansion of the river network. As also stated in response to point 2 raised by the reviewer, the approach and the findings should not be extrapolated to much larger basins, as the assumption of a spatially homogenous catchment storage does not hold for most large basins. For large catchments, it is unlikely that storage exceedance occurs simultaneously in the entire catchment (see also the discussion in l. 409 on this). In addition, for larger catchments river processes might become more important, and processes like network expansion or overland flow are not represented in the model. We agree with the reviewer that assessing the effects of other nonlinearities in the runoff generation would be highly interesting, especially in combination with different basin scales. However, this would require a different model set-up and, as discussed in l. 414, "in such a set-up, tail heaviness could be affected by a combination of catchment size, sub-basin response, spatial organization and river routing characteristics, making it difficult to isolate the effects of precipitation and runoff generation." In response to both this comment and comment 6 by reviewer #1, we expanded the discussion as follows: "In the adopted rainfall-runoff model only one nonlinearity in the runoff generation was considered, namely the activation of an additional very fast runoff component. However, in a real catchment multiple nonlinearities and process shifts might be present such as the onset of overland flow, the onset of subsurface stormflow, the activation of macropores or the temporary expansion of the river network. The model does not include all these processes explicitly and is therefore, as all models, a simplified representation of reality. Hence, the simulated flood peak

distributions are also only representative for this simplified reality. Nevertheless, they can help us explore results which can be valuable for real-world applications."

With regards to the reviewer's reference to arid/semi-arid regions we would like to emphasize that the model and its parametrization are set up for Central European conditions. Parameters for the rainfall-runoff model are based on Austrian catchments (Merz et al., 2011) and the weather generator was originally set up and evaluated for Central European stations (Nguyen et al., 2021). We added in the discussion (l. 427): "Nevertheless, the simulation model chain and its parametrization has been set up for Central European conditions and so the findings should not be directly transferred to other regions of the world where conditions are very different."

6) The procedure used to examine threshold return periods beyond which the flood peak distribution is governed by the rainfall distribution (evaluating slopes of rainfall and flood distributions on log-log-plots) is based on ad-hoc procedures with subjectively chosen parameters.

We do not agree that we have used an ad-hoc procedure to examine the threshold return periods. We have based the procedure on the following reasoning: For close to impervious catchments, the curves of Pct and Q are assumed to run in parallel. This would mean that the difference between their slopes is 0. However, due to some noise in the data, we observed that the slopes between the curves of Pct and Q are hardly ever exactly zero, even for the model runs on close to impervious catchments (l. 201). We therefore used the differences between the slopes of Pct and Q estimated for close to impervious catchments to evaluate what level of noise is to be expected, i.e. what amount of differences between the slopes we need to expect even for parallel curves. Based on this we defined a buffer around zero within which slope differences need to lie for the curves to be considered as parallel. We also tested the sensitivity of the threshold return period to the definition of the buffer to ensure that our results are reasonable (l. 206, l. 271). We found that both a narrower and a wider buffer result in some cases in unreasonable threshold return periods, while this was not the case for the buffer that we used (l. 273).

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".

We fit one GEV distribution even when we know that there might be a process shift in the runoff generation, as fitting GEV distributions to annual maximum series (AMS) is very common in hydrological practice. In hydrological practice, we often do not know whether there is a process shift acting or not, and so we simply fit one distribution to the observed AMS. Doing the same here, allows us to draw conclusions which can be of relevance for hydrological practice. However, it should be noted that when fitting one GEV distribution despite the presence of a process shift we cannot necessarily infer the tail behaviour of the true underlying distribution. Nevertheless, it can still be insightful with regards to the occurrence probability of extreme events. For more details on this aspect and the distinction between true and apparent tail behaviour, please also see our response to point 2 raised by reviewer #1. In the manuscript, we added the following sentences in the methods section (around l. 177): "It should be noted that the shape parameter of a GEV distribution fitted to a time series of limited length does not necessarily reflect the true tail behaviour of the underlying distribution but is only an approximation thereof. 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."

The authors have deployed a broad array of simulation and modeling tools to address an interesting and important problem. The assumptions and subjective decisions needed to implement this array of tools create the most serious obstacles to supporting the expansive conclusions that are reported.