Interactive Discussion: Final response

From Worst-Case Scenarios to Extreme Value Statistics: Local Counterfactuals in Flood Frequency Analysis

Paul Voit, Felix Fauer, and Maik Heistermann
NHESS Discussions, doi:10.5194/egusphere-2025-4951

RC: ***Reviewer Comment***, AR: *Author Response*, □ Manuscript text

Dear Editor,

thank you for your efforts to coordinate the peer review process, for your editorial remarks and your quick response.

Please find our responses to your comments below as well as the responses to the review of the three referees.

Kind regards,
Paul Voit, Felix Fauer, and Maik Heistermann

1. Comments and responses to the Referee 1

RC: ***The paper would benefit from clearer definitions and a distinct conceptual separation between several key ideas that are currently addressed somewhat implicitly. In particular, the notion of "local counterfactuals" should be more rigorously defined early in the manuscript, including its hydrological interpretation and how it differs from related concepts such as storm transposition, spatial counterfactuals, and regionalisation. While the background section is strong, readers unfamiliar with these concepts may find it difficult to immediately understand what is novel versus what is adapted from existing approaches.***

AR: We suggest to add the following description of existing concepts to the introduction.

- **Regionalization**: Data from hydrologically similar catchments are incorporated into the estimation of distribution parameters to enhance the robustness of extreme value analysis (EVA) (e.g., Gaume et al., 2010; Guse et al., 2010; Nguyen et al., 2014; Halbert et al., 2016).
- **Probable maximum precipitation** (PMP): rainfall events from a "meteorological homogeneous" transposition domain are included in the analysis to increase the robustness (Fuller, 1914; District and Morgan, 1916) and to estimate the PMP (Hansen, 1987; WMO, 2009). Instead of exceedance probabilities this method only yields upper and lower bounds of precipitation. PMP can be used to estimate the upper bounds of a probable maximum flood (PMF), if used as forcing of a hydrological model. While PMP is widely applied in North America and Australia for designing high-risk infrastructure (e.g., dams and nuclear power plants), it is not prominently used in Europe. However, in recent years various studies regarding flood risk management have proposed and investigated different concepts of storm transposition, referring to the idea as "spatial counterfactuals" (Montanari et al., 2023; Merz et al., 2024; Voit and Heistermann, 2024; Vorogushyn et al., 2024; Thompson et al., 2025).
- **Stochastic storm transposition**: Building on the PMP/PMF concept, historical HPEs from a transposition domain are sampled using a Poisson distribution and randomly assigned (uniform distribution) within the domain, potentially affecting the catchment of interest (CoI). For flood frequency analysis, the resulting runoff in the CoI is simulated (e.g., Wright et al., 2014). This approach allows for the calculation of occurrence probabilities. For a detailed description see Wright et al. (2017). Globally, stochastic storm transposition (SST) remains rarely applied in practice (Wright et al., 2020) but it will form the core of the U.S. Federal Emergency Management Agency's "Future of Flood Risk Data" initiative, aimed at remapping the nation's floodplains (Abbasian et al., 2025).
- **Stochastic weather generators** are statistical models that simulate sequences of weather variables, such as temperature and precipitation, by randomly generating data based on observed patterns. They can be used to generate very long time series of meteorological forcings for a hydrological model (e.g. Falter et al., 2015; Apel et al., 2016).

RC: ***The selection of a 30 km radius and ten neighboring catchments seems reasonable but is largely based on empirical judgment. The manuscript should offer a clearer rationale for these choices, whether based on meteorological homogeneity, hydrological similarity scales, or sensitivity analysis.***

AR: We agree that the choice and combination of similarity metrics is a pragmatic one – an expert guess, if you will. The only way to actually *assess* the validity of our similarity metrics is to compare them against others in a kind of benchmark experiment: considering our KDTree-approach as a filter, the question behind such an experiment would be which filter could provide the best results in terms of improvement of our performance metric (i.e. the QSS). In fact, we did this when we analysed the sensitivity of the QSS to different neighborhood radii around our CoI. But while it would be highly interesting to expand such an analysis to the similarity metrics, we think that this is beyond the scope of the present study in which we rather aim to introduce a *framework* and provide proof-of-concept. We will, however, expand our discussion of the limitations of similarity metrics, and also outline perspectives for future research to assess the validity of similarity metrics.

RC: ***The criteria used to define catchment similarity via the KDTree deserve further explanation.***

AR: Yes, and your comment is in line with the other referees. The catchment similarity is a critical point. We

added more information (l. 123 ff) to further describe the process.

> 1. For each CoI, we identified the ten most similar catchments located entirely within a 30 km buffer around the CoI. We based similarity mostly on descriptors of topography, land use and soil which should i) strongly govern the formation and concentration of surface runoff and ii) ensure that potential orographic effects could occur both in the CoI and the NCs. Following descriptors were chosen:
>
>   - Peak [m³/s], time to peak [s] and standard deviation [m³/s] of the unit hydrograph: The unit hydrograph is derived directly from the DEM, similar hydrographs imply, to a certain degree, similar topography.
>   - Total catchment area including upstream basins.
>   - Curve number (soil moisture class 2): The curve number represents soils and land use in our model. A similar curve number would lead to a similar runoff generation in our model.
>   - Mean and standard elevation of the DEM and mean slope. With this descriptor we try to avoid sampling rainfall events from catchments which are e.g. situated at a substantially different elevation. If the CoI was e.g. close to a mountain range, rainfall events should not be sampled from this mountainous area, because they might not be representative for the rainfall events occuring in the CoI.
>   - Unit Peak Discharge: The peak of the unit hydrograph divided by the catchment area is yet another descriptor of the hydrological character of the catchment.
>
>   We used the KDTree-algorithm from the Python library "SciKit-Learn" and scaled all catchment descriptors with the "StandardScaler" from this library to ensure that none of this descriptors dominates the decision for similarity. However, we acknowledge that some descriptors are correlated.

RC: ***Applying an uncalibrated SCS-CN and GIUH-based lumped model across thousands of catchments introduces significant uncertainty. The model mainly captures fast runoff generation and overlooks Hortonian runoff, slow flow components, and spatial variability in precipitation. This is especially important for catchments with winter flood regimes, where soil saturation and slower processes may prevail. The authors should discuss how these simplifications could impact both annual maxima and GEV tail behaviour.***

AR: The referee is right. Uncertainties introduced by the hydrological model were already discussed in section 5 of the preprint, and we will expand this discussion based on the referee's comment. Here, we would like to maintain that for extreme events, we assume slow flow components to be negligible at the scale of small catchments. It is correct that high flows may occur during winter subject to saturated soils and low evapotranspiration or in spring subject to snow melt events, and that these high flows may constitute some of the annual maxima and hence govern return levels for small return periods. The tail behaviour of small catchments, however, is clearly dominated by convective heavy rainfall that occur in the summer. As for the effect of Hortonian surface runoff (infiltration excess), the referee is correct that this process is inadequately represented in the SCS-CN framework. We assume that this would lead to an underestimation of runoff generation for extreme events at very short durations, or, in other words, that a model that represents infiltration excess would lead to heavier tails of the GEV distribution.

RC: ***The conclusions would benefit from clearer guidance on when the proposed method may be unsuitable.***

AR: Thank you, this was also mentioned by another referee. We added following sentence in line 313:

In regions with high orographic gradients or highly heterogeneous rainfall patterns the proper size of the TD might have to be reduced or optimized in benchmark experiments similar to the one carried out in this study.

RC: ***The manuscript frequently refers to "worst-case scenarios", but this term is not clearly defined.***

AR: The manuscript mentions the term "worst-case flood" (or scenario) two times: in the title and in the conclusions. But the referee is completely right that it lacks clarity, an issue that was also noted by the other referees.

In fact, we think that the term "worst-case" is not required in the context of our study. Based also on other comments, we changed the manuscript title to "Considering rainfall events from a neighborhood improves local flood frequency analysis". We will remove the sentence in which the term occurred in the manuscript (ll. 305 ff. of the preprint) as it does not essentially relate to the subject of our study.

RC: ***The appropriateness of using annual maxima with maximum likelihood estimation for such short samples could be briefly discussed in relation to alternative approaches such as POT or L-moments.***

AR: We agree. We have started using the ML method in our first publications but have also realized that L-Moments is considered to be the more stable method. We also agree in regard to POT but this method is also not widely used by practitioners due to its more complex application, e.g. for the definition of the threshold. We are currently involved in a study in which we compare extreme rainfall events in southern India which are evaluated by an index based either on annual maxima or POT. Based on the results of that study, we might move to the POT method in the future to increase the robustness of our assessment. With regard to this manuscript, we suggest to add a brief discussion to the section "Limitations" in which we should point out that the use of a POT approach together with the Generalized Pareto distribution (GPD) might be preferable in case of limited time series lengths, and certainly compatible with the framework presented in our manuscript.

2. Comments and responses to the Referee 2

Main comments:

RC: ***From a comprehensive analysis of 13000 catchments, I was hoping to see a map with the regional performance of this method, and the areas that might be problematic.***

AR: We did not include a map because we could not distinguish any spatial patterns. We include here the map for the QSS for q=0.99 and q=0.995 (100-yr and 200-yr flood) for the 30km-buffer in Figure 1 and 2. Still, we think that these figures would not really add to the paper, so we would rather not include them in the revised version.

RC: ***What is the effect of the catchment area on method performance and the size of the transposition domain?***

AR: This is a very interesting and thoughtful comment. We originally had expected that small catchments would benefit more from our counterfactual approach (in terms of QSS) since the likelihood to be "hit" by small convective rainfall events decreases with catchment size. Instead, we do see a larger improvement of the QSS for larger basins in Figure 4 (but note that there are only few basins larger than 40 km² in our model setup, see Fig. 3). This might be caused by the generally larger amount of counterfactual peaks for the larger basins, because they often consist of several subbasins (and we move the rainfall event to the centroid of

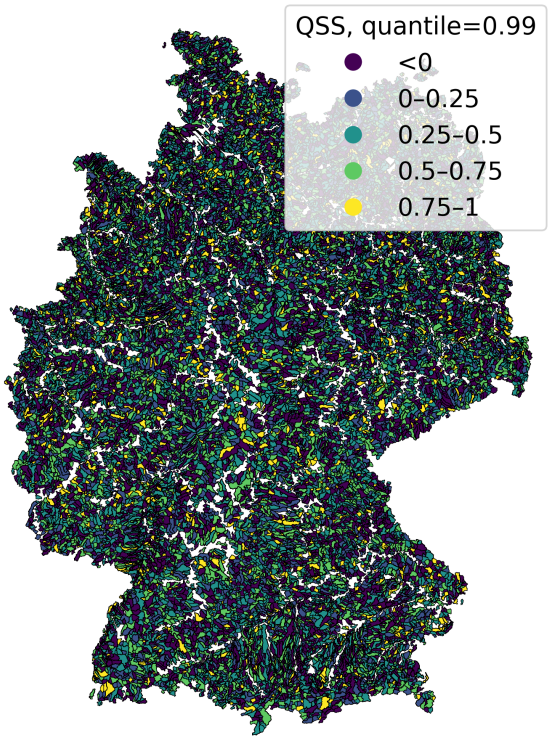

Figure 1: Quantile skill score for German subbasins for the 30-km buffer and the 0.99 quantile. Catchments with an upstream area > 750 km² are excluded (white).

every subbasin of the CoI). The larger amount of counterfactual peaks might simply result in a more robust GEV fit (see Table 1).

RC: ***The model you applied seems to have multiple limitations, and I am wondering what's the effect on your results. You choose to apply a CN model over lumped (small) basins. This implies that the response is driven by the cumulated value of precipitation and its distribution in time, and not by its intensity. Hortonian runoff is not simulated. The spatial variability of precipitation is lost - might be ok if your subbasins are very small.***

AR: The subbasins were chosen to be small (see Fig. 3 in the attempt to compensate for the lumped nature of the model. We hence think that the resulting uncertainty is rather low. Surely, the SCS-CN approach introduces considerable uncertainty (as would any hydrological model under extreme rainfall conditions). Namely, its inability to represent overland flow from infiltration excess might be a relevant source of uncertainty. As a consequence of the referee's comment (and other referee comments), we will expand the section "Limitations" to discuss more comprehensively the uncertainties to be expected from our hydrological model. At the same time, we would like to emphasize that we essentially present an analysis *framework* and recommend, for practical applications by e.g. agencies, to use a hydrological that has proven valid in the region of application.

RC: ***You represent only "quick runoff", while many of those catchments have annual maxima in winter, when***

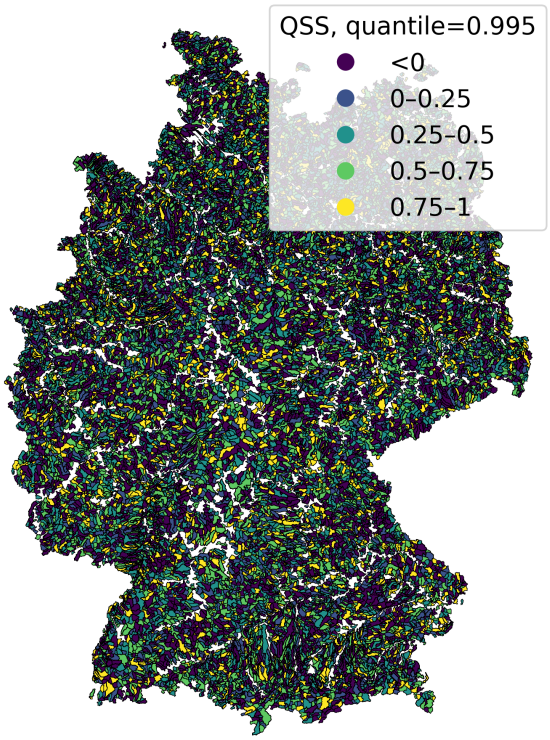

Figure 2: Quantile skill score for German subbasins for the 30-km buffer and the 0.995 quantile. Catchments with an upstream area > 750 km² are excluded (white).

soils are wet and slow runoff has a very high contribution on peaks.

AR: For extreme events, we assume slow flow components to be negligible at the scale of small catchments. However, it is correct that high flows may occur during winter subject to saturated soils and low evapotranspiration or in spring subject to snow melt events, and that these high flows may constitute some of the annual maxima and hence affect return levels for small return periods. As already mentioned above, we will expand the discussion of model uncertainties in the section "Limitations" based on this and other referee comments.

RC: ***L121: In SST one of the most critical points is the definition of a similar transposition domain. With your method, you seem to transfer this to catchment similarity. Can you give more information on how the catchment similarity criteria are mixed, and how much your approach is sensitive to this choice?***

AR: Surely, the 30 km radius is the prime filter to make sure we sample storms from an atmospheric environment that is governed by similar mechanisms as the CoI. In that sense, we maintain the concept of a transposition domain, in analogy to SST. Furthermore, by sampling storms that caused annual maxima in similar catchments, we ensure that the sampled storms have spatio-temporal characteristics that make them impact-relevant for the CoI (e.g. similar size or similar unit hydrographs) and that could also occur over the CoI given the potential for orographic effects (e.g. similar elevation in the catchment). Based on these considerations, we aim to create counterfactuals that are representative for our CoI. In the revised manuscript, we will explain

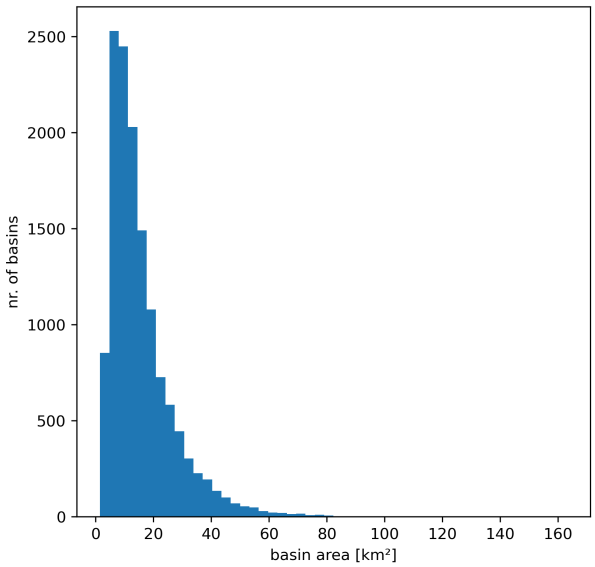

Figure 3: Catchment area distribution of the 13,452 headwater catchments.

Table 1: Number of basins for the "30km-buffer" transposition domain and mean number of counterfactual peaks for each size class.

size class	nr. of basins	mean nr. of counterfactual peaks
0-20 km²	5077	238
21-40 km²	1625	378
41-60 km²	732	564
>61 km²	1888	1240

in more detail the corresponding similarity criteria and how they are combined (”mixed”) by means of a KDTree-analysis: We used the KDTree-algorithm from the Python library "SciKit-Learn" and scaled all catchment descriptors with the "StandardScaler" from this library to ensure that none of this descriptors dominates the decision for similarity. However, we acknowledge that many descriptors are correlated.

We will address to your second question (”how much [is] your approach sensitive to this choice?”) in the following comment.

RC: ***How sensitive is your method to not finding similar catchments in the transposition domain?***

AR: We would like to include in this answer the question from the previous comment (”how much [is] your approach sensitive to [the choice and combination of similarity metric]?”). Admittedly, the choice and combination of similarity metrics is a pragmatic one – an expert guess, if you will. But while we think that our assumptions are plausible and well in line with ”hydrological common sense”, the only way to actually *assess* the validity of our similarity metrics is to compare them against others in a kind of benchmark experiment:

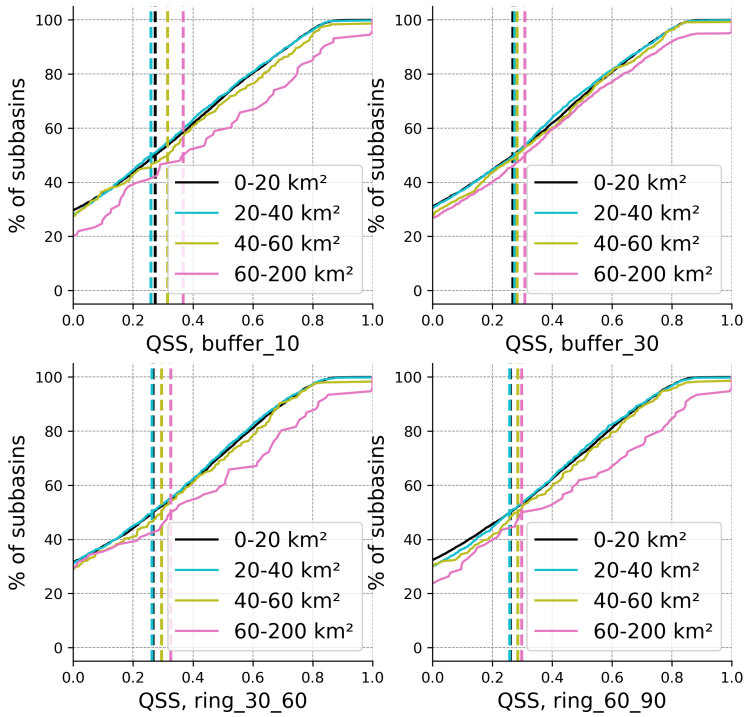

Figure 4: Quantile skill score for German subbasins for all four transposition domains and the 0.995 quantile.

considering our KDTree-approach as a filter, the question behind such an experiment would be which filter could provide the best results in terms of improvement of our performance metric (i.e. the QSS). In fact, we did this when we analysed the sensitivity of the QSS to different neighborhood radii around our CoI. But while it would be highly interesting to expand such an analysis to the similarity metrics, we think that this is beyond the scope of the present study in which we rather aim to introduce a *framework* and provide proof-of-concept. We will, however, expand our discussion of the limitations of similarity metrics, and also outline perspectives for future research to assess the validity of similarity metrics.

RC: ***One of your conclusions is that "the improvement declines when the radius of the transposition domain is extended beyond 30 km. I think this is based on Figure 2 alone. In this figure, performance seems to generally decline with larger radius, but I don't see a performance decline after 30 km2. I'd argue that the performance seems mostly independent from the size of the transposition domain (while it might be more sensitive to the measure of catchment similarity).***

AR: For the low return periods we can observe clear differences between the curves in Fig. 2 up to a QSS of 0.2. Generally speaking, CDF curves that are more shifted to the right indicate a better QSS. We hence do not quite agree with the referee that there is no performance decline after 30 km. However, we agree that the difference is becoming very small with increasing return periods.

Based on our results, we cannot refute the referee's hypothesis that the measure of catchment similarity is more important than the radius of the transposition domain. As pointed out in our response to the referee's previous comment, this would require an additional comprehensive benchmark experiment in which to investigate the

sensitivity of the QSS to competing definitions of similarity. As already pointed out above, this is beyond the scope of the present study which rather aims to introduce a framework and provide proof-of-concept. We will, however, expand our discussion of the limitations of similarity metrics, and also outline future research to assess the validity of similarity metrics.

RC: ***Transposing the HPE to the centroid of the catchment might result in biases (e.g. if the storm and the NC have an orientation, or if spatial distribution over the catchment is important - see Zhou, 2021).***

AR: We agree and we have read the publication of Zhou et al. (2021). We would argue, that the spatial distribution and storm direction is increasingly important for larger catchments, e.g. where various tributaries flow together. Nevertheless, the elongation or orientation of a catchment could be included in the similarity criteria. As pointed out above, a systematic study in which different definitions/implementations of similarity are systematically benchmarked would be certainly worthwhile. We suggest to expand the section "Limitations".

Other comments:

RC: ***"Worst-Case scenarios" is in the title and the conclusions, but what do you define as a "worst case"? Is taking the strongest HPE within 10km the worst that can happen to a basin? Sometimes compound scenarios or very high (500y?) return periods are considered.***

AR: We thank the referee for the comment. In fact, we think that the term "worst-case" is not required in the context of our study. Based also on other comments, we changed the manuscript title to "Considering rainfall events from a neighborhood to improve local flood frequency analysis". We will remove the sentence in which the term occurred in the manuscript's conclusions (ll. 305 ff. of the preprint) as it does not essentially relate to the subject and results of our study.

RC: ***As you say, in your analysis FFA is strongly limited by the relatively short data availability (23 years). You apply a GEV over annual maxima using maximum likelihood, but usually for small data samples, POT and maybe L-moments estimation might be more appropriate.***

AR: We agree. We have started using the ML method in our first publications but have also realized that L-Moments is considered to be the more stable method. We also agree in regard to POT but this method is also not widely used by practitioners due to its more complex application, e.g. for the definition of the threshold. We are currently involved in a study in which we compare extreme rainfall events in southern India which are evaluated by an index based either on annual maxima or POT. Based on the results of that study, we might move to the POT method in the future to increase the robustness of our assessment. With regard to this manuscript, we suggest to add a brief discussion to the section "Limitations" in which we should point out that the use of a POT approach together with the Generalized Pareto distribution (GPD) might be preferable in case of limited time series lengths, and certainly compatible with the framework presented in our manuscript.

RC: ***You use QS to evaluate particularly high return periods (200 years) over a short data record (23 years). If the quantile q is higher than every observations, is the best QS simply the closest to your observation? Is it correct to say that it's a better estimate?***

AR: Yes, exactly. Thank you for this remark. Indeed, if all observations are lower than the quantile, then the QS might reward the model that calculates the lowest quantile. Therefore, the QS is more reliable for the lower return periods, which are shown in Figs. 2 and 4. However, since the number of CoI is very large, some of them might experience observations above the 200-yr return period. Still, we suggest to add after line 184:

> Note that for very high return periods the QS might become unreliable, since only few or no observations are higher than the evaluated quantile. Then, the QS might just reward the model that predicts the lowest quantile.

Furthermore, we suggest to mention the limitation of the QS in section 5, ”Limitations”:

> In that context, the QSS has to be interpreted with care, specifically for very high return periods such as 100 or 200 years. In essence, the evaluation of the QSS for unseen quantiles is challenging because observations that exceed high quantiles are rare. Unfortunately, this limitation is difficult to overcome and applies to all scores known to us.

RC: ***L26: I'm not sure why you describe flash floods here. You didn't specifically analyze effect on flash floods, and your approach is general.***

AR: We are using the term "flash floods" because we specifically look at small and medium sized catchments which are specifically prone to flash floods. It is also these catchments for which the lack of long observational time series particularly evident. Furthermore, the overall methodological setup with a rather small transposition domain is geared towards small to medium sized catchments.

RC: ***L31: It's subjective, but I would not call a 750km2 basin "small". Maybe small + medium?***

AR: Yes, we agree. We suggest to change the sentence to: *...flash-flood–prone basins generally small to medium sized (<1000 km²).*

RC: ***L98: Isn't CORINE updated every 6 years?***

AR: Yes, at least according to the homepage. However, the most recent update is still the 2018 version (https://land.copernicus.eu/en/products/corine-land-cover).

RC: ***L106: you refer to your other paper for the model application, but I think it would be useful to add some more information on the model setup and characteristics that are important for your results.***

AR: We suggest to change the section "Modelling surface runoff" as following:

The hydrological model (Voit, 2024) was specifically tailored to simulate flash flood events in small- to medium-sized basins. A detailed model description is provided in Voit and Heistermann (2024). During flash floods, surface runoff dominates (Marchi et al., 2010; Grimaldi et al., 2010), while evaporation and groundwater dynamics are negligible. Accordingly, the model comprises two modules. First, effective rainfall is estimated for each catchment and timestep (hourly) using the SCS-CN method (U.S. Department of Agriculture-Soil Conservation Service, 1972), which is widely applied in flash flood modeling (Gaume et al., 2004; Borga et al., 2007; Emmanuel et al., 2017). Since flash flood events predominantly occur during the summer months, we slightly adjusted the CN values for agricultural areas to account for the effects of summer crops (based on Seibert et al., 2020). A single CN value for each subbasin was then derived using an area-weighted average.

Second, the geomorphological instantaneous unit hydrograph (GIUH), derived from the DEM, represents the concentration of quick runoff from effective rainfall. The flow velocities were computed with the method of Maidment Maidment et al. (1996). This approach accounts for the increase in hydraulic radius with rising flow volumes, as described by Manning's equation, thereby capturing the downstream acceleration of flow without requiring the estimation of roughness coefficients for individual grid cells. In addition, it removes the need to distinguish between hillslope and channel grid cells within the catchment. The method assumes a velocity field that is invariant in both time and discharge, enabling the convolution of GIUHs to simulate the catchment response to the effective rainfall of an HPE. When two subcatchments converge, the hydrograph of the upstream basin is superimposed on that of the downstream basin with an appropriate time lag. This delay is defined by the travel time from the downstream basin's inlet to its outlet.

The model's lightweight design allows the computation of large numbers of counterfactual scenarios. As it does not account for channel hydraulics or engineered structures, the analysis is restricted to headwater catchments smaller than 750 km². Because of the lumped nature of the model it is crucial that the catchments are small enough to account for the spatial variability of rainfall. In our analysis, this corresponds to 13,452 sub-catchments with an mean area of 15.7 km² and a maximum headwater catchment size of 163 km².

RC: ***L106: Please clarify: if I understand well, you apply a lumped CN model over basins with a median size of 15.7 km2. These basins are also combining into larger basins. I was confused how sometimes you talk about "upstream catchments" "transposition to each catchment in the CoI".***

AR: Yes, this is correct. We hope that this gets more clearer now with the suggested extended model description. If we look for similar NCs, we have to look at the total catchment area (including upstream basins) of both CoI and NCs. To clarify this, we suggest to change some parts in section 3.2:

For each CoI, we identified the ten most similar catchments located entirely within a 30 km buffer around the CoI. Similarity was quantified using a KDTree (SciKit-Learn) based on the following scaled (Scikit-Learn StandardScaler) catchment attributes: GIUH time to peak, GIUH standard deviation, GIUH unit peak discharge, mean slope, mean elevation, elevation standard deviation, *total catchment area (including the upstream subbasins)*, and mean curve number

and in line 132:

> If the CoI consists of various subbasin, we additionally transpose the HPEs to the centroid of every upstream subbasin.

RC: ***L134: do you apply the HPE multiple times by transposing the same HPE to the centroid of each subcatchment? If so, do you think it's generates realistic precipitation fields?***

AR: Yes, this is what we are doing. We extract the HPE with a large spatial buffer so the whole HPE will always cover the CoI including all its upstream subbasins. The shifting distance between these subbasins is just a few kilometers due to the generally small size of the subbasins. We consider all these scenarios as realistic counterfactuals.

RC: ***L157: do you mean that you disregard shape below 0 (0 is ok) and above 0.5?***

AR: Thank you for spotting this. It is of course the opposite to what is written in the manuscript. We checked our scripts and the correct version should be:

> For this reason we disregard catchments where one of the previous GEV distributions has a < 0 or shape ≥ 0.5.

RC: ***L195: GEV CoI is fitted over 22 years?***

AR: Yes, for the cross validation we only use 22 years. We repeat this 23 times and the take the average.

RC: ***L217: I don't see the supplement, also online.***

AR: Indeed, it is not there. We apologize and are not sure about the cause for the missing supplement. We will make sure to upload it with the revised manuscript. For the time being, we put the information from the supplement here:

The figure in the supplement (Figure 5 here in the response letter) shows the results for all TDs and for four different return periods (20, 50, 100 and 200 years). According to Fauer et al. (2021) negative values of the QSS cannot be easily interpreted which is why we show only QSS >= 0. The inclusion of the data from the CoI improves the quantile estimation only marginally compared to GEV_{NCs} (Fig. 2).

RC: ***L220: are the HPE over larger domains less typical than (comment authors "for the"?) CoI or just less correlated? How much do the HPE of the CoI overlap with the floods over NC?***

AR: We chose two ring sized TPs to ensure that we actually only sample HPEs from further away. The hypothesis is, that these HPEs are less representative for the CoI, as pointed out in the manuscript. We did not check whether or not the HPE that caused the annual maximum over the CoI also caused the annual maximum over the NC.

RC: ***L254: aren't NCs analyzed as a set? So with 230 values.***

AR: Sorry, this sentence is misleading. It hopefully becomes clearer if we write "23 for each NC". The total then is 230.

> First, although the counterfactual dataset exhibits some higher peaks, these peaks occur jointly with the entire set of annual maxima from this NC (23 values *for each NC*).

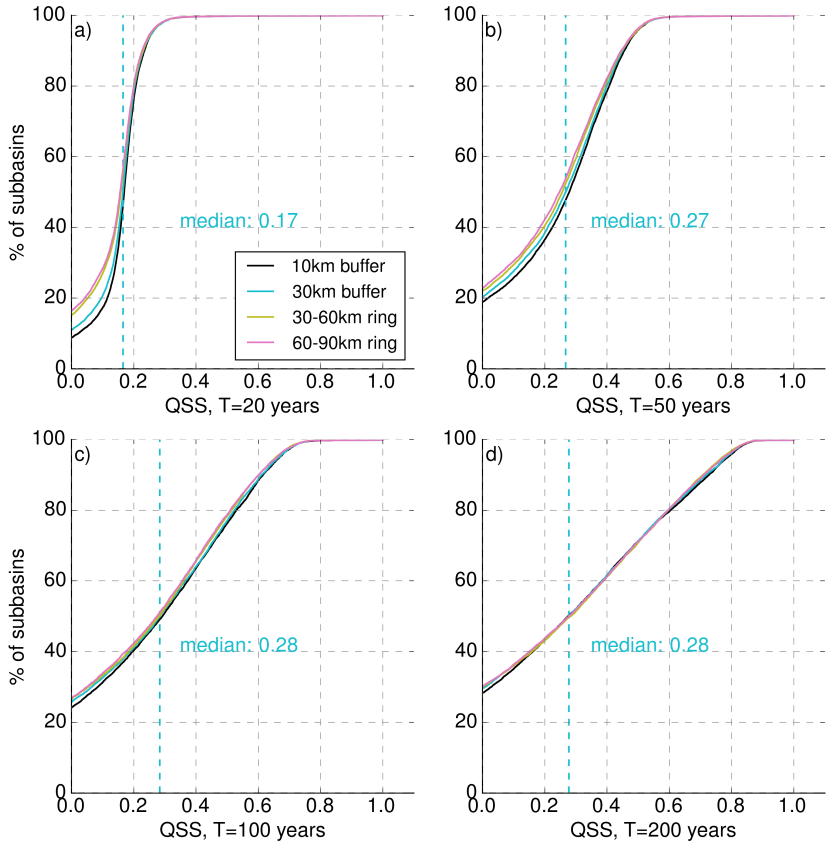

Figure 5: Cumulative distributions showing the quantile skill scores for GEV_{NCs} in reference to GEV_{all}, for all subbasins and for four different transposition domains (10-km buffer: yellow, 30-km buffer:blue, 30-60-km ring: green, 60-90-km ring: orange. Subplots a)-d) show different quantiles that relate to the a) 20-year, b) 50-year, c) 100-year and d) 200-year flood. A quantile score > 0 indicates the superiority of the GEV_{NCs}. The median QSS of the 30-km buffer is indicated with the vertical blue dashed line

RC: ***L258, L266, L267: I can't find those numbers reflected in figure 4. Am I reading it wrong?***

AR: Thanks again for your level of attention. We accidentally compiled an older figure into the PDF. This figure (Fig. 6 here in the response letter) is the correct one and will be implemented in the revised manuscript.

RC: ***L284: why would catchment biases cancel out?***

AR: We will revise the manuscript in order to make the statement more comprehensible: if the model should have a systematic error (bias) in a specific catchment, than this bias should affect the peak discharge of all events simulated for that catchment and hence reflect in all the different GEV distributions fitted for that specific catchment. Apart from that, we will expand, as already stated in response to various referee comments, the section "Limitations" in order to more comprehensively discuss hydrological model uncertainties.

3. Comments and responses to the Referee 3

RC: ***The paper would benefit from clearer definitions and a distinct conceptual separation between several key ideas that are currently addressed somewhat implicitly. In particular, the notion of "local counterfactuals" should be more rigorously defined early in the manuscript, including its hydrological interpretation and how it differs from related concepts such as storm transposition, spatial counterfactuals, and regionalization. While the background section is strong, readers unfamiliar with these concepts may find it difficult to immediately understand what is novel versus what is adapted from existing approaches.***

AR: We suggest to add the following description of existing concepts to the introduction.

- **Regionalization**: Data from hydrologically similar catchments are incorporated into the estimation of distribution parameters to enhance the robustness of extreme value analysis (EVA) (e.g., Gaume et al., 2010; Guse et al., 2010; Nguyen et al., 2014; Halbert et al., 2016).
- **Probable maximum precipitation** (PMP): rainfall events from a "meteorological homogeneous" transposition domain are included in the analysis to increase the robustness (Fuller, 1914; District and Morgan, 1916) and to estimate the PMP (Hansen, 1987; WMO, 2009). Instead of exceedance probabilities this method only yields upper and lower bounds of precipitation. PMP can be used to estimate the upper bounds of a probable maximum flood (PMF), if used as forcing of a hydrological model. While PMP is widely applied in North America and Australia for designing high-risk infrastructure (e.g., dams and nuclear power plants), it is not prominently used in Europe. However, in recent years various studies regarding flood risk management have proposed and investigated different concepts of storm transposition, referring to the idea as "spatial counterfactuals" (Montanari et al., 2023; Merz et al., 2024; Voit and Heistermann, 2024; Vorogushyn et al., 2024; Thompson et al., 2025).
- **Stochastic storm transposition**: Building on the PMP/PMF concept, historical HPEs from a transposition domain are sampled using a Poisson distribution and randomly assigned (uniform distribution) within the domain, potentially affecting the catchment of interest (CoI). For flood frequency analysis, the resulting runoff in the CoI is simulated (e.g., Wright et al., 2014). This approach allows for the calculation of occurrence probabilities. For a detailed description see Wright et al. (2017). Globally, stochastic storm transposition (SST) remains rarely applied in practice (Wright et al., 2020) but it will form the core of the U.S. Federal Emergency Management Agency's "Future of Flood Risk Data" initiative, aimed at remapping the nation's floodplains (Abbasian et al., 2025).
- **Stochastic weather generators** are statistical models that simulate sequences of weather variables, such as temperature and precipitation, by randomly generating data based on observed patterns. They can be used to generate very long time series of meteorological forcings for a hydrological model (e.g. Falter et al., 2015; Apel et al., 2016).

RC: ***The selection of a 30 km radius and ten neighboring catchments seems reasonable but is largely based on empirical judgment. The manuscript should offer a clearer rationale for these choices, whether based on meteorological homogeneity, hydrological similarity scales, or sensitivity analysis.***

AR: We agree that the choice and combination of similarity metrics is a pragmatic one – an expert guess, if you will. The only way to actually *assess* the validity of our similarity metrics is to compare them against others in a kind of benchmark experiment: considering our KDTree-approach as a filter, the question behind

such an experiment would be which filter could provide the best results in terms of improvement of our performance metric (i.e. the QSS). In fact, we did this when we analysed the sensitivity of the QSS to different neighborhood radii around our CoI. But while it would be highly interesting to expand such an analysis to the similarity metrics, we think that this is beyond the scope of the present study in which we rather aim to introduce a *framework* and provide proof-of-concept. We will, however, expand our discussion of the limitations of similarity metrics, and also outline perspectives for future research to assess the validity of similarity metrics.

RC: ***The criteria used to define catchment similarity via the KDTree deserve further explanation.***

AR: Yes, and your comment is in line with the other referees. The catchment similarity is a critical point. We added more information (l. 123 ff) to further describe the process.

> 1. For each CoI, we identified the ten most similar catchments located entirely within a 30 km buffer around the CoI. We based similarity mostly on descriptors of topography, land use and soil which should i) strongly govern the formation and concentration of surface runoff and ii) ensure that potential orographic effects could occur both in the CoI and the NCs. Following descriptors were chosen:
>
>   - Peak [m³/s], time to peak [s] and standard deviation [m³/s] of the unit hydrograph: The unit hydrograph is derived directly from the DEM, similar hydrographs imply, to a certain degree, similar topography.
>   - Total catchment area including upstream basins.
>   - Curve number (soil moisture class 2): The curve number represents soils and land use in our model. A similar curve number would lead to a similar runoff generation in our model.
>   - Mean and standard elevation of the DEM and mean slope. With this descriptor we try to avoid sampling rainfall events from catchments which are e.g. situated at a substantially different elevation. If the CoI was e.g. close to a mountain range, rainfall events should not be sampled from this mountainous area, because they might not be representative for the rainfall events occuring in the CoI.
>   - Unit Peak Discharge: The peak of the unit hydrograph divided by the catchment area is yet another descriptor of the hydrological character of the catchment.
>
>   We used the KDTree-algorithm from the Python library "SciKit-Learn" and scaled all catchment descriptors with the "StandardScaler" from this library to ensure that none of this descriptors dominates the decision for similarity. However, we acknowledge that some descriptors are correlated.

RC: ***Applying an uncalibrated SCS-CN and GIUH-based lumped model across thousands of catchments introduces significant uncertainty. The model mainly captures fast runoff generation and overlooks Hortonian runoff, slow flow components, and spatial variability in precipitation. This is especially important for catchments with winter flood regimes, where soil saturation and slower processes may prevail. The authors should discuss how these simplifications could impact both annual maxima and GEV tail behavior.***

AR: The referee is right. Uncertainties introduced by the hydrological model were already discussed in section 5 of the preprint, and we will expand this discussion based on the referee's comment. Here, we would like to maintain that for extreme events, we assume slow flow components to be negligible at the scale of small catchments. It is correct that high flows may occur during winter subject to saturated soils and

low evapotranspiration or in spring subject to snow melt events, and that these high flows may constitute some of the annual maxima and hence govern return levels for small return periods. The tail behavior of small catchments, however, is clearly dominated by convective heavy rainfall that occur in the summer. As for the effect of Hortonian surface runoff (infiltration excess), the referee is correct that this process is inadequately represented in the SCS-CN framework. We assume that this would lead to an underestimation of runoff generation for extreme events at very short durations, or, in other words, that a model that represents infiltration excess would lead to heavier tails of the GEV distribution.

RC: ***The conclusions would benefit from clearer guidance on when the proposed method may be unsuitable.***

AR: Thank you, this was also mentioned by another referee. We added following sentence:

> In regions with high orographic gradients or highly heterogeneous rainfall patterns the proper size of the TD might have to be reduced or optimized in benchmark experiments similar to the one carried out in this study.

RC: ***The manuscript frequently refers to "worst-case scenarios", but this term is not clearly defined.***

AR: The manuscript mentions the term "worst-case flood" (or scenario) two times: in the title and in the conclusions. But the referee is completely right that it lacks clarity, an issue that was also noted by the other referees.

In fact, we think that the term "worst-case" is not required in the context of our study. Based also on other comments, we changed the manuscript title to "Considering rainfall events from a neighborhood improves local flood frequency analysis". We will remove the sentence in which the term occurred in the manuscript (ll. 305 ff. of the preprint) as it does not essentially relate to the subject of our study.

RC: ***The appropriateness of using annual maxima with maximum likelihood estimation for such short samples could be briefly discussed in relation to alternative approaches such as POT or L-moments.***

AR: We agree. We have started using the ML method in our first publications but have also realized that L-Moments is considered to be the more stable method. We also agree in regard to POT but this method is also not widely used by practitioners due to its more complex application, e.g. for the definition of the threshold. We are currently involved in a study in which we compare extreme rainfall events in southern India which are evaluated by an index based either on annual maxima or POT. Based on the results of that study, we might move to the POT method in the future to increase the robustness of our assessment. With regard to this manuscript, we suggest to add a brief discussion to the section "Limitations" in which we should point out that the use of a POT approach together with the Generalized Pareto distribution (GPD) might be preferable in case of limited time series lengths, and certainly compatible with the framework presented in our manuscript.

References

Abbasian, M., Wright, D. B., Notaro, M., Vavrus, S., and Vimont, D. J.: Flood frequency sampling error: insights from regional analysis, stochastic storm transposition, and physics-based modeling, Journal of Hydrology, p. 133802, 2025.

Apel, H., Martínez Trepat, O., Hung, N. N., Chinh, D. T., Merz, B., and Dung, N. V.: Combined fluvial and pluvial urban flood hazard analysis: concept development and application to Can Tho city, Mekong

Delta, Vietnam, Natural Hazards and Earth System Sciences, 16, 941–961, 10.5194/egusphere-2025-495110.5194/nhess-16-941-2016, 2016.

Borga, M., Boscolo, P., Zanon, F., and Sangati, M.: Hydrometeorological analysis of the 29 August 2003 flash flood in the Eastern Italian Alps, Journal of hydrometeorology, 8, 1049–1067, 10.5194/egusphere-2025-495110.1175/JHM593.1, 2007.

District, M. C. and Morgan, A. E.: Exhibits to Accompany Report of the Chief Engineer, Arthur E. Morgan: Submitting a Plan for the Protection of the District from Flood Damage, Miami Conservancy District, 1916.

Emmanuel, I., Payrastre, O., Andrieu, H., and Zuber, F.: A method for assessing the influence of rainfall spatial variability on hydrograph modeling. First case study in the Cevennes Region, southern France, Journal of Hydrology, 555, 314–322, 10.5194/egusphere-2025-495110.1016/j.jhydrol.2017.10.011, 2017.

Falter, D., Schröter, K., Dung, N. V., Vorogushyn, S., Kreibich, H., Hundecha, Y., Apel, H., and Merz, B.: Spatially coherent flood risk assessment based on long-term continuous simulation with a coupled model chain, Journal of Hydrology, 524, 182–193, 10.5194/egusphere-2025-495110.1016/j.jhydrol.2015.02.021, 2015.

Fauer, F. S., Ulrich, J., Jurado, O. E., and Rust, H. W.: Flexible and consistent quantile estimation for intensity–duration–frequency curves, Hydrology and Earth System Sciences, 25, 6479–6494, 10.5194/egusphere-2025-4951/10.5194/hess-25-6479-2021, publisher: Copernicus GmbH, 2021.

Fuller, W. E.: Flood flows, Transactions of the American Society of Civil Engineers, 77, 564–617, 1914.

Gaume, E., Livet, M., Desbordes, M., and Villeneuve, J.-P.: Hydrological analysis of the river Aude, France, flash flood on 12 and 13 November 1999, Journal of hydrology, 286, 135–154, 10.5194/egusphere-2025-495110.1016/j.jhydrol.2003.09.015, 2004.

Gaume, E., Gaál, L., Viglione, A., Szolgay, J., Kohnová, S., and Blöschl, G.: Bayesian MCMC approach to regional flood frequency analyses involving extraordinary flood events at ungauged sites, Journal of hydrology, 394, 101–117, 2010.

Grimaldi, S., Petroselli, A., Alonso, G., and Nardi, F.: Flow time estimation with spatially variable hillslope velocity in ungauged basins, Advances in Water Resources, 33, 1216–1223, 10.5194/egusphere-2025-495110.1016/j.advwatres.2010.06.003, 2010.

Guse, B., Hofherr, T., and Merz, B.: Introducing empirical and probabilistic regional envelope curves into a mixed bounded distribution function, Hydrology and Earth System Sciences, 14, 2465–2478, 10.5194/egusphere-2025-495110.5194/hess-14-2465-2010, 2010.

Halbert, K., Nguyen, C. C., Payrastre, O., and Gaume, E.: Reducing uncertainty in flood frequency analyses: A comparison of local and regional approaches involving information on extreme historical floods, Journal of Hydrology, 541, 90–98, 10.5194/egusphere-2025-495110.1016/j.jhydrol.2016.01.017, 2016.

Hansen, E. M.: Probable maximum precipitation for design floods in the United States, Journal of Hydrology, 96, 267–278, 10.5194/egusphere-2025-495110.1016/0022-1694(87)90158-2, 1987.

Maidment, D., Olivera, F., Calver, A., Eatherall, A., and Fraczek, W.: Unit hydrograph derived from a spatially distributed velocity field, Hydrological processes, 10, 831–844, 10.5194/egusphere-2025-4951https://doi.org/10.1002/(SICI)1099-1085(199606)10:6<831::AID-HYP374>3.0.CO;2-N, 1996.

Marchi, L., Borga, M., Preciso, E., and Gaume, E.: Characterisation of selected extreme flash floods in Europe and implications for flood risk management, Journal of Hydrology, 394, 118–133, 10.5194/egusphere-2025-495110.1016/j.jhydrol.2010.07.017, 2010.

Merz, B., Nguyen, V. D., Guse, B., Han, L., Guan, X., Rakovec, O., Samaniego, L., Ahrens, B., and Vorogushyn, S.: Spatial counterfactuals to explore disastrous flooding, Environmental Research Letters, 10.5194/egusphere-2025-495110.1088/1748-9326/ad22b9, 2024.

Montanari, A., Merz, B., and Blöschl, G.: HESS Opinions: The Sword of Damocles of the Impossible Flood, EGUsphere, 2023, 1–20, 10.5194/egusphere-2025-495110.5194/egusphere-2023-2420, 2023.

Nguyen, C. C., Gaume, E., and Payrastre, O.: Regional flood frequency analyses involving extraordinary flood events at ungauged sites: further developments and validations, Journal of Hydrology, 508, 385–396, 10.5194/egusphere-2025-495110.1016/j.jhydrol.2013.09.058, 2014.

Seibert, S. P., Auerswald, K., Seibert, S. P., and Auerswald, K.: Abflussentstehung–wie aus Niederschlag Abfluss wird, Hochwasserminderung im ländlichen Raum: Ein Handbuch zur quantitativen Planung, pp. 61–93, 10.5194/egusphere-2025-495110.1007/978-3-662-61033-6_4, 2020.

Thompson, V., Coumou, D., Beyerle, U., Ommer, J., Cloke, H. L., and Fischer, E.: Alternative rainfall storylines for the Western European July 2021 floods from ensemble boosting, Communications Earth & Environment, 6, 427, 2025.

U.S. Department of Agriculture-Soil Conservation Service: Estimation of Direct Runoff From Storm Rainfall, SCS National Engineering Handbook, Section 4, Hydrology. Chapter 10, 1972.

Voit, P.: A downward counterfactual analysis of flash floods in Germany – Code repository (v0.1), Zenodo [code], `https://doi.org/10.5281/zenodo.10473424`, last accessed: 15.08.2024, 2024.

Voit, P. and Heistermann, M.: A downward-counterfactual analysis of flash floods in Germany, Natural Hazards and Earth System Sciences Discussions, 2024, 1–23, 10.5194/egusphere-2025-495110.5194/nhess-2023-224, 2024.

Vorogushyn, S., Han, L., Apel, H., Nguyen, V. D., Guse, B., Guan, X., Rakovec, O., Najafi, H., Samaniego, L., and Merz, B.: It could have been much worse: spatial counterfactuals of the July 2021 flood in the Ahr valley, Germany, Natural Hazards and Earth System Sciences Discussions, 2024, 1–39, 10.5194/egusphere-2025-495110.5194/nhess-2024-97, 2024.

WMO: Manual on estimation of probable maximum precipitation (PMP), `https://library.wmo.int/viewer/35708/?offset=#page=1&viewer=picture&o=bookmarks&n=0&q=`, last accessed: 18 September 2024, 2009.

Wright, D. B., Smith, J. A., and Baeck, M. L.: Flood frequency analysis using radar rainfall fields and stochastic storm transposition, Water Resources Research, 50, 1592–1615, 2014.

Wright, D. B., Mantilla, R., and Peters-Lidard, C. D.: A remote sensing-based tool for assessing rainfall-driven hazards, Environmental modelling & software, 90, 34–54, 2017.

Wright, D. B., Yu, G., and England, J. F.: Six decades of rainfall and flood frequency analysis using stochastic storm transposition: Review, progress, and prospects, Journal of Hydrology, 585, 10.5194/egusphere-2025-495110.1016/j.jhydrol.2020.124816, 2020.

Zhou, Z., Smith, J. A., Baeck, M. L., Wright, D. B., Smith, B. K., and Liu, S.: The impact of the spatiotemporal structure of rainfall on flood frequency over a small urban watershed: an approach coupling stochastic storm transposition and hydrologic modeling, Hydrology and Earth System Sciences, 25, 4701–4717, 10.5194/egusphere-2025-495110.5194/hess-25-4701-2021, 2021.

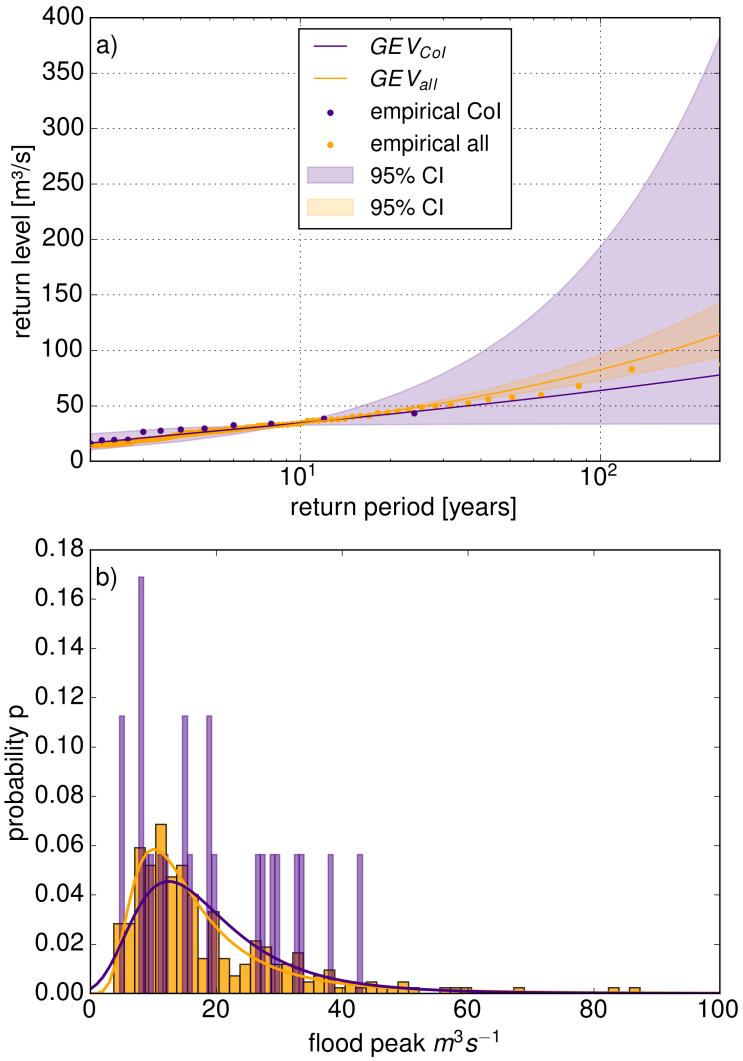

Figure 6: Comparison of two GEV_{CoI} and GEV_{all} for one exemplary basin. a) Return levels estimated by GEV_{all} (orange) are lower than by GEV_{CoI} (purple). The shaded areas mark the 95 % confidence interval estimated with boot strapping (n=500). The empirical return periods were estimated with the Weibull plotting position and are indicated with the semi-transparent dots. b) Density histogram of the annual maxima and fitted GEV distribution.