## Interactive Discussion: Author Response to Referee #1

# 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 Referee,

thank you for taking the time and effort to review this manuscript. Surely, incorporating your suggestions will improve the readability and overall quality of the manuscript.

Please find our responses to your comments below. These should be considered as preliminary (part of the interactive discussion) since the actual implementation of changes depends on the editorial decision.

Thanks again for your efforts!

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

**RC:** ***The term "local counterfactual" should be defined more firmly in the opening section so that readers unfamiliar with the concept can immediately grasp its hydrological meaning.***

AR: We agree and try to address the comment by the following changes:

In the sentence after l. 62 of the preprint:

> [...] Voit and Heistermann (2024a) introduced the concept of "local counterfactuals": they selected HPEs that had caused high runoff peaks in basins from a close (i.e. "local") neighborhood around the CoI (more specifically, a 20 km radius), transposed these events to the CoI and used it to force a rainfall-runoff model that would than return the counterfactual flood peak. The approach was based on the assumption that if an HPE were sampled from a local neighborhood, it would be more representative for HPEs that are "typical" for the CoI. Even with this local TD, local counterfactuals produced flood peaks comparable to a 200-year return level flood.

**RC:** ***The introduction would benefit from a clearer explanation of why a 30-km neighborhood and ten neighboring catchments were selected as the basis for local counterfactual generation.***

AR: While the size of the transposition domain (30 km neighborhood) will most likely always remain a matter of discussion, it would be generally preferable to select more than 10 neighboring catchments and thus, create more counterfactuals. This decision was mainly based on computational limitations because every additional neighboring catchment results in additional 23 counterfactuals which need to be modelled. We will clarify this by adding the following sentence in line 85:

> A 30 km radius neighborhood (transposition domain) can be still be considered as local and small, compared to the domain sizes in other studies (e.g. Voit and Heistermann, 2024; Abbasian et al., 2025) while the number of 10 neighboring catchments was chosen mainly to contain the computational load.

**RC:** ***The background section is strong, but it would be helpful to distinguish more explicitly between catchment similarity and storm similarity, as the manuscript presently assumes these are equivalent.***

AR: We entirely agree, and we will attempt, in the revised manuscript, to clarify the notion of similarity already in the introduction. Most importantly, we will try to better explain that the use of catchment similarity metrics is, to a considerable extent, *also* motivated by the aim to identify similar storms. How is that? 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. Yet, sampling storms that caused annual maxima in similar catchments ensures 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 mean and standard deviation of 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 extend the corresponding explanations on the design of the local counterfactuals, and we will also discuss in further depth the limitations that we face in creating such representative counterfactuals.

**RC:** ***The use of an uncalibrated SCS-CN and GIUH model across more than 13,000 catchments introduces considerable uncertainty, and the authors should include either a brief validation example or a reference to previous calibration results.***

AR: We agree that our hydrological model introduces considerable uncertainty, as would any hydrological model under extreme rainfall-runoff conditions. However, our modelling approach is well established in the flash flood community (Marchi et al., 2010; Borga et al., 2007; Ruiz-Villanueva et al., 2012; Tarolli et al., 2013). Furthermore, Voit and Heistermann (2024) could show for the Ahr flood in 2021 in Western Germany that our model was able to reproduce the reconstructed flood hydrograph at gauge Altenahr (Roggenkamp and Herget, 2022; Mohr et al., 2023) very well (Fig. 1). For that reason, we think that the presentation of additional validation results is not required within the present manuscript. This is also because we only compare model results within each catchment to one each other. Within such a comparison, any systematic model errors should tend to cancel out when used for GEV fitting. Furthermore, our study should be considered as a proof-of-concept. In the section ”Limitations”, we also recommended, for practical applications in the context of risk management, to rather use a model that has proven valid for the application region.

However, we suggest to expand (also in response to other referee comments) the section on ’Limitations” further with regard to the uncertainties of the hydrological model and potential implications for our analysis.

**RC:** ***The criteria used to define "catchment similarity" deserve more explanation, especially regarding how the attributes were scaled and weighted in the KDTree analysis.***

AR: In our analysis, we aimed at sampling rainfall events that would have a strong impact in the CoI. Because we based the selection on the flood peaks in the NCs, we want to ensure that the catchments are hydrologically similar, as well as that the rainfall events are representative for the factual rainfall events in the CoI. For this reason, we based similarity mostly on descriptors of topography, land use and soil which should i) govern the

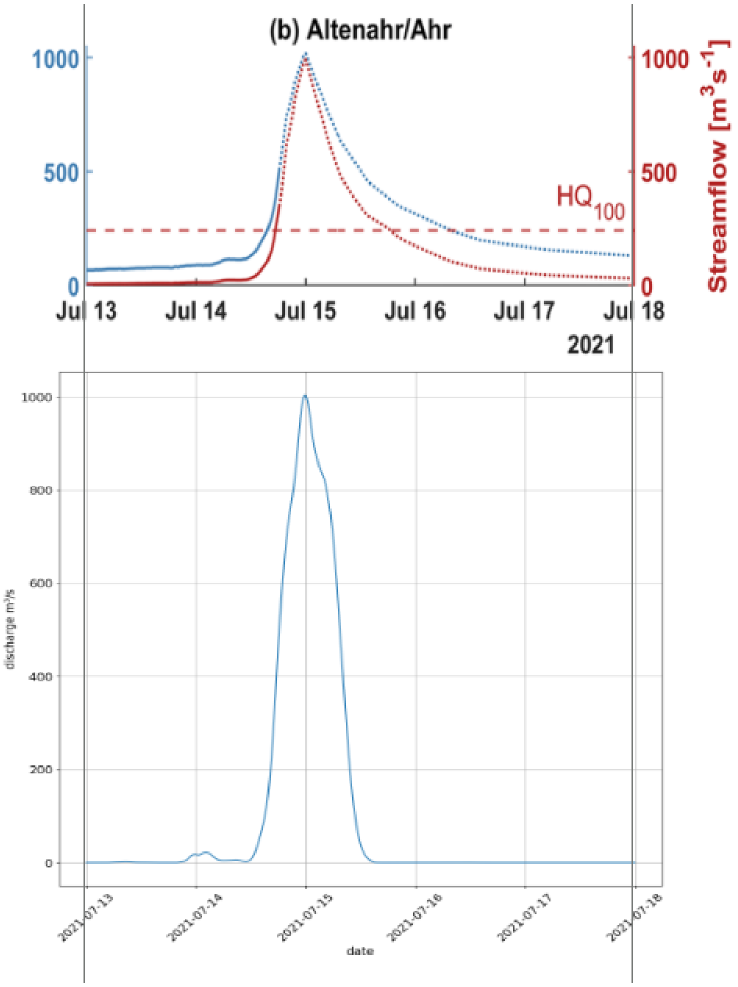

Figure 1: Top: Reconstruction by Roggenkamp and Herget (2022) published in Mohr et al. (2023) of the flood event in Altenahr, West-Germany, July 2021. Discharge is shown in red, water level in blue. Bottom: Modelled discharge for the same event with RADKLIM data with our model.

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.
- Upstream catchment area
- 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.

We suggest to change following sentence in section 3.2, line 123, to clarify the scaling:

> Similarity was quantified using a KDTree (SciKit-Learn) based on the following scaled (Scikit-Learn StandardScaler) catchment attributes:...

**RC:** ***The assumption that storms producing high runoff in a nearby basin are hydrologically meaningful for the catchment of interest should be justified with either empirical evidence or literature support.***

AR: We are not entirely sure what the referee means by "hydrologically meaningful for the catchment of interest". We assume, however, the referee means that the selected storms should be "representative" for the kind of storms that cause flood peaks in the CoI. In that regard, we would like to refer to our above explanations on similarity. At the same time, the referee implies that all our assumptions on similarity and hence representativeness are only that: assumptions, or, more benevolent, "expert guess". That is correct. And 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 analysis: considering our KDTree-approach as a filter, the question would be which filter provides the best results in terms of improvement of our performance metric (QSS). That way, we could at least say which filter is superior over another one. In fact, we did this when we analysed the sensitivity of the QSS to different neighbourhood radii around our CoI. But while it would be highly interesting to expand such an analysis to other similarity metrics, we think that 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:** ***The manuscript should explain how independence among counterfactual annual maxima is ensured, given that neighboring catchments may experience correlated rainfall events.***

AR: Thank you for this remark. Indeed, we make the assumption that the counterfactual HPEs represent alternative variants of a given HPE in the CoI, that could have happened at another time within the CoI. With this approach we increase the sample size to improve the GEV parameter estimation. Also, we argue that events which cover two or more NCs at the same time, are allowed to have more influence on the GEV parameter estimation.

**RC:** ***Mixing factual and counterfactual peaks in a single GEV fit may violate standard assumptions, and this issue requires at least a clear justification in the methods section.***

AR: Thank you for this comment. We agree that further justification might improve the manuscript. Since the peaks of factual and counterfactual HPEs are determined with the same method, we argue that both can be pooled to fit a GEV. The discriminating characteristic between factual and counterfactual is that counterfactual peaks are derived from storm transposition.

We suggest to add in line 153:

> Since the peaks of factual and counterfactual HPEs are determined with the same method, both can be pooled to fit a GEV, given all assumptions above.

**RC:** ***Although the QSS results show improvements, the authors should comment on the fact that GEV$_{NCs}$ outperforms GEV$_{CoI}$ even without using any data from the catchment of interest, which may indicate over-smoothing or strong regional influences.***

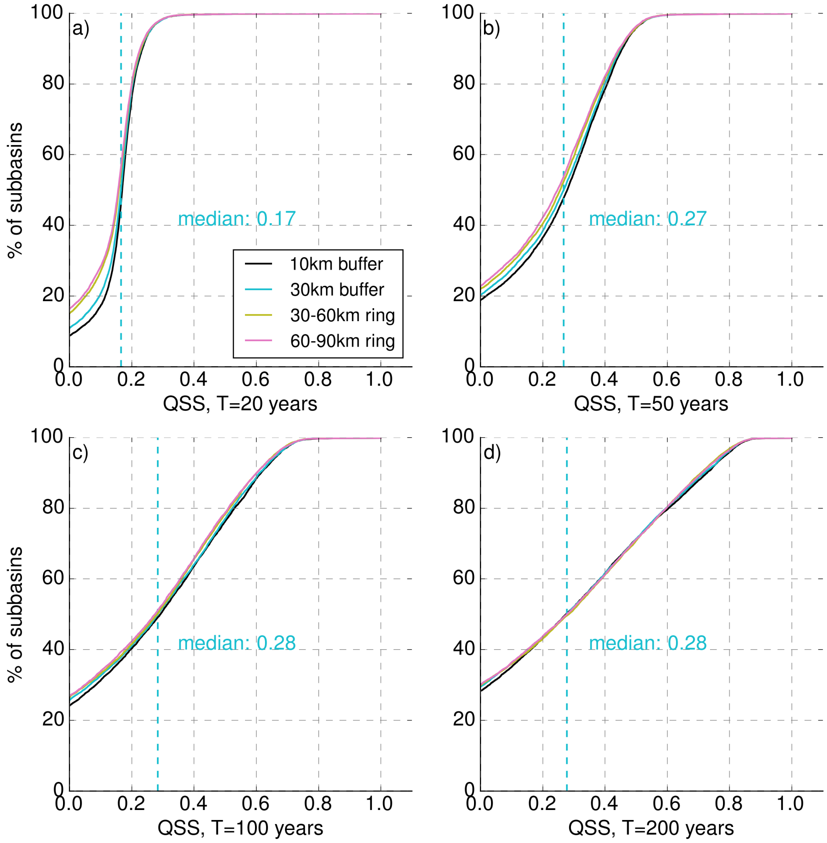

Figure 2: Cumulative distributions showing the quantile skill scores for GEV$_{\mathrm{NCs}}$ in reference to GEV$_{\mathrm{all}}$, for all subbasins and for four different transposition domains (10-km buffer: yellow, 30-km buffer:blue, 30-60-km ring: green, 60-90-km ring: orange. Subplots a)-d) show 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$_{\mathrm{NCs}}$. The median QSS of the 30-km buffer is indicated with the vertical blue dashed line

AR: Thank you for this comment. We believe that these results underline a strong regional influence and are a justification of our method. Apparently the flood peaks which we generated by sampling rainfall events from hydrological similar and nearby basins leads to counterfactual annual flood peaks that fit very well into the distribution of "observed" flood peaks. The uncertainty of the GEV fit decreases significantly when using 230 instead of 23 values. We already comment on the regional smoothing in section 4.3:

> Secondly, local counterfactuals also induce spatial smoothing (which is desired): each catchment is a CoI once, but serves as neighbor for many other CoIs. As a result, nearby and hydrologically similar catchments often share almost identical sets of peaks. When a counterfactual peak increases the return level estimate for one CoI, the peaks from that CoI will also enter the NC data pool once their roles are reversed. In this case, the inclusion of the peak can reduce the return level estimate for the neighboring catchment.

We also suggest to add another sentence in section 4.1 to explain the more robust fit of the GEV:

> These results serve as a proof of concept: for the majority of cases, we are able to better represent the quantiles in the data of the CoI by using a GEV distribution fitted exclusively to the counterfactual peaks (GEV$_{\text{NCs}}$). *Besides the fact, that the counterfactual peaks represent the distribution of CoI peaks well, the GEV$_{NCs}$ is also more robust because it is fitted to 230 values, instead of the 23 values used for GEV$_{CoI}$.* The improvement is more pronounced for higher quantiles (or return periods). In practice the GEV would be fitted to both factual *and* counterfactual peaks together (GEV$_{\text{all}}$), which only marginally increases the robustness of the return level estimates. The QSS for GEV$_{\text{all}}$ is shown in Figure S1 in the supplement.

As referee #2 pointed out, the supplement was not online. We apologize for that attach the figure here (Fig. 2 and will make sure, that the Supplement is properly uploaded with the revised version of the manuscript.

**RC:** ***The improvement of GEV$_{NCs}$ with increasing return period is convincingly shown, yet the manuscript should discuss why the lower tail benefits less from the counterfactual approach.***

AR: Thank you, we will try to make this clearer. The higher the return period, the more we need to extrapolate and the higher the uncertainty will be. When using only 23 years of data for extreme value statistics the uncertainty for the 200-yr return level is very high. With more data we do not extrapolate. In the case of the GEV$_{\text{NCs}}$ we already have 230 (counterfactual)-"annual" maxima. The uncertainty for the 200-yr return level is very low, as shown in the example in Figure 4. We will add two sentences in line 212 to clarify this further.

> We would like to take a closer look at the differences between the return periods. Increasing return periods lead to a decreasing fraction of catchments with positive QSS$_{\text{NCs}}$ values - obviously not desirable -, but also to a desirable increase of catchments with very high QSS values (for T=20 a, 0.2% of the catchments have a QSS > 0.5, while this fraction grows to 28% for T=200 a). Altogether, the median QSS continuously grows from a value of 0.16 for T=20 a to a value of 0.27 for T=200 a, suggesting that the value added by using GEV$_{\text{NCs}}$ increases with the return period. This is plausible, since return levels for low return periods can be estimated more robustly from short time series (for T=20 a, the estimation of a return level from an annual series of 23 years does not even imply extrapolation). *The uncertainty increases the more we extrapolate beyond the length of the annual series. Especially for high return periods the benefit of an increased data basis is visible in these results.*

**RC:** ***The discussion should reflect that counterfactual extremes depend strongly on the selected time window and may not represent the full range of possible events.***

AR: We agree. The longer the record length, the higher the probability it will contain an event with an even larger magnitude (although we would be careful with the term ”full range of possible events” - even with very long time series, we will have difficulties in spanning that range). In that sense, an analysis that includes local counterfactuals shows exactly the same behavior as conventional flood frequency analysis within the CoI. We will expand the section "Limitations" accordingly.

**RC:** ***The authors appropriately highlight the short time series, but they omit discussion of potential non-stationarity in rainfall over the 2001–2023 period, which may influence GEV tail behavior.***

AR: We agree that non-stationarity of the extreme value distribution is not accounted for by our approach, and we will expand the discussion of "Limitation" in order to point this out. That being said, we would speculate (meaning that we cannot prove it) that the brevity of the time series underlying conventional GEV fitting is a more important source of uncertainty than the non-stationarity of the distribution.

**RC:** ***The conclusion section accurately summarizes the study, but it should offer clearer guidance on when the counterfactual method might be unsuitable—particularly in regions with strong orographic gradients or highly heterogeneous rainfall patterns.***

AR: Thank you for this suggestion. We will add following part to the "Conclusions":

> The selection of the TD affects the quality GEV estimation when local counterfactuals are employed. We showed that the QSS decreased when HPEs were sampled from a distance of more than 30 km away from the CoI. Still, the optimal definition of the TD will remain arbitrary and represents a subject for further research, as it represents an inherent trade-off: while an increasing distance allows us to sample from a larger variety of events and particularly from a larger choice of hydrologically similar catchments, an increasing distance will typically sample HPEs that are less representative for the meteorological processes that govern the CoI. At of now, the 30 km radius remains a rather pragmatic choice and a compromise between these two requirements. In regions with high orographic gradients or highly heterogeneous rainfall patterns the size of the TD might have to be reduced or optimized in benchmark experiments similar to the one carried out in this study.

## 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.

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.

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.

Mohr, S., Ehret, U., Kunz, M., Ludwig, P., Caldas-Alvarez, A., Daniell, J. E., Ehmele, F., Feldmann, H., Franca, M. J., Gattke, C., Hundhausen, M., Knippertz, P., Küpfer, K., Mühr, B., Pinto, J. G., Quinting, J., Schäfer, A. M., Scheibel, M., Seidel, F., and Wisotzky, C.: A multi-disciplinary analysis of the exceptional flood event of July 2021 in central Europe–Part 1: Event description and analysis, Natural Hazards and Earth System Sciences, 23, 525–551, 10.5194/egusphere-2025-495110.5194/nhess-23-525-2023, 2023.

Roggenkamp, T. and Herget, J.: Hochwasser der Ahr im Juli 2021–Abflussabschätzung und Einordnung, Hydrologie und Wasserbewirtschaftung, 66, 40–49, 2022.

Ruiz-Villanueva, V., Borga, M., Zoccatelli, D., Marchi, L., Gaume, E., and Ehret, U.: Extreme flood response to short-duration convective rainfall in South-West Germany, Hydrology and Earth System Sciences, 16, 1543–1559, 10.5194/egusphere-2025-495110.5194/hess-16-1543-2012, 2012.

Tarolli, M., Borga, M., Zoccatelli, D., Bernhofer, C., Jatho, N., and Janabi, F. a.: Rainfall space-time organization and orographic control on flash flood response: the Weisseritz event of August 13, 2002, Journal of Hydrologic Engineering, 18, 183–193, 10.5194/egusphere-2025-495110.1061/(ASCE)HE.1943-5584.0000569, 2013.

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.