Reply on RC2

Felix S. Fauer, Jana Ulrich, Oscar E. Jurado, Henning W. Rust

August 27, 2021

Dear Reviewer,

We would like to thank you for your constructive comments. In the following, we will address them point by point and support them with figures.

1 Answers

1. Line # 29: could be written as Methods

Answer: Thank you for this suggestion. We will change this sentence.

2. In lines # 65-70: A few aspects of nonstationarity could be discussed here, discuss briefly the added value of this study to address nonstationarity as compared to the methods discussed in the literature (Cheng and AghaKouchak 2014; Ganguli and Coulibaly 2017).

Answer: Thank you for suggesting these relevant papers. We will include non-stationarity in the discussion because we agree that this is an important concept that should be considered in IDF curves. However, this study is not supposed to add value in non-stationarity, since stationarity is an assumption of our model. We plan to include non-stationarity in future studies.

3. Section 2.1: Lines 90 – 95: This is not very clear - why 8 stations were merged into a single station leaving 92 overall stations out of 100 stations? Which physiographic or hydrologic similarity measures were adopted for regionalization?

Answer: Thank you for pointing out that this is not clear. We group the stations because the model can use its full potential, when long daily records are combined with high resolution (minute) records which are often much shorter. Typically, two different measuring frequencies are obtained from different devices and in some cases, those devices are not positioned at the same site. However, in most cases those stations are very near to each other or even at the same position. Therefore, whenever two stations have a distance of less than 250m, those respective stations are merged into one station. The only measure for this procedure is the distance. For more details about the distance of the stations and a test of robustness,

see also the answer to comment 1 in RC1 (https://hess.copernicus.org/preprints/hess-2021-334#AC1).
We will describe that more clearly in the manuscript. Also, we will correct a mistake in the manuscript. The total number of stations is 115. The number of grouped stations is 92.

4. It has been shown in the literature that the Generalized Maximum likelihood method, in general, does not provide a credible estimate of the shape parameter, yielding an abrupt estimate of shape estimate (Martins and Stedinger 2000). Have your values lie within the credible limits of shape parameter range as shown in the literature, boxplots showing the range of shape parameters for different duration could help to identify this issue

Answer: Thank you for this remark. Since in our model we assume the shape parameter to be constant over duration, we cannot present the range of the shape parameter depending on duration. However, we present the range of the shape parameter at all used stations in Fig. 1 to investigate whether it lies within credible intervals. Almost all parameters lie within $-0.3 < \xi < 0.6$. Only for one station (Dabringhausen), an unrealistically large shape parameter was estimated which could be explained by the scarce data availability at this station (15 years).
Martins and Stedinger (2000) report that maximum likelihood estimation (MLE) in small samples can lead to unrealistic shape parameters and that for large sample sizes the RMSE of both methods become similar (Figs. 4 and 5 in their study). We argue that due to the duration-dependent GEV in our study the number of data points available for estimation is multiplied by the number of duration steps and so, sample size should be sufficiently large for using MLE.

5. In skill score index in lines 190-200: what M and R represent, If R denotes empirical distribution, which empirical plotting position formula was used to estimate it? Typically Gringorten's plotting position formula is in use to characterize extremes.

Answer: R does not denote the empirical distribution but another IDF model. The quantile skill score (QSS) compares the quantile score (QS_M) of a new model M with the quantile score (QS_R) of a reference model R. The difference between models and references are only characterized by the features (curvature, multiscaling, flattening) that are used/not used in this model or reference. So, different combinations will be compared, e.g. a model using curvature and multiscaling ($\mathrm{IDF}_{\mathtt{cm}}$ as model M) vs. a model using only curvature ($\mathrm{IDF}_{\mathtt{c}}$ as reference model R). Table E1 in the manuscript lists which feature combinations are used as model M and reference R. The quantile score (QS) is a propper scoring function (Gneiting, 2011) for comparing modeled quantiles to all observations. We do not calculate a difference between model quantiles and empirical quantiles. Empirical distributions are only used for visualization ("+" in Figures 5 and 7 in the manuscript).
Thank you for pointing out that this part was not clearly described. We will improve Section 2.5 accordingly.

6. Fig. 7: Could you please show the difference in return levels in an inset diagram with vs without flattening? How much is the percentage difference between the two statistics in order to qualify as significant?

Answer: We created additional panel plots that show the difference in intensity and also the ratio of both estimates between the two models (Fig. 2 in this document). In comparison with Fig. 5 from the manuscript the differences are expected to overlap with the confidence intervals, suggesting non-significance. However, these confidence intervals are mainly depending on data size and would become smaller with more data available. Nevertheless, the purpose of this visualization was to demonstrate that allowing for flattening of IDF curves in long durations has an impact on the shape of IDF curves in short durations. These figures only show two selected stations, exemplarily.
In our study, the verification of model performance for all stations was done with the Quantile Skill Index (QSI) where we could show that certain models improve IDF estimation in certain duration-regimes, depending on the use-case. Here, a QSI<0.05 is considered as irrelevant (see white regions in Figs. 3 and 4 in the manuscript).

7. Between lines 360-363: Any discussion on copula-based IDF estimation that claims to preserve the inherent non-linearity between intensity vs duration?

Answer: Thank you for this remark. Bezak et al. (2016) used copula-based IDF curves and reported that IDF curves might be sensitive to the choice of method. This is important to consider when deciding on the appropriate way to create IDF curves. We will include this reference in the manuscript.

7. Line # 433: few outlying events correspond to higher quantile (or at the tail of the distribution) leave the confidence intervals.

Answer: Thank you for this suggestion. We will change this sentence.

On behalf of all authors
Felix Fauer

2 References

Martins ES, Stedinger JR (2000): Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. Water Resources Research, 36,737–744, https://www.doi.org/10.1029/1999WR900330.

Nejc Bezak, Mojca Šraj, Matjaž Mikoš (2016): Copula-based IDF curves and empirical rainfall thresholds for flash floods and rainfall-induced landslides. Journal of Hydrology, 541,272-284, https://doi.org/10.1016/j.jhydrol.2016.02.058

Tilmann Gneiting (2011): Quantiles as optimal point forecasts, International Journal of Forecasting, 27, 197-207. https://doi.org/10.1016/j.ijforecast.2009.12.015

3 Figures and Tables

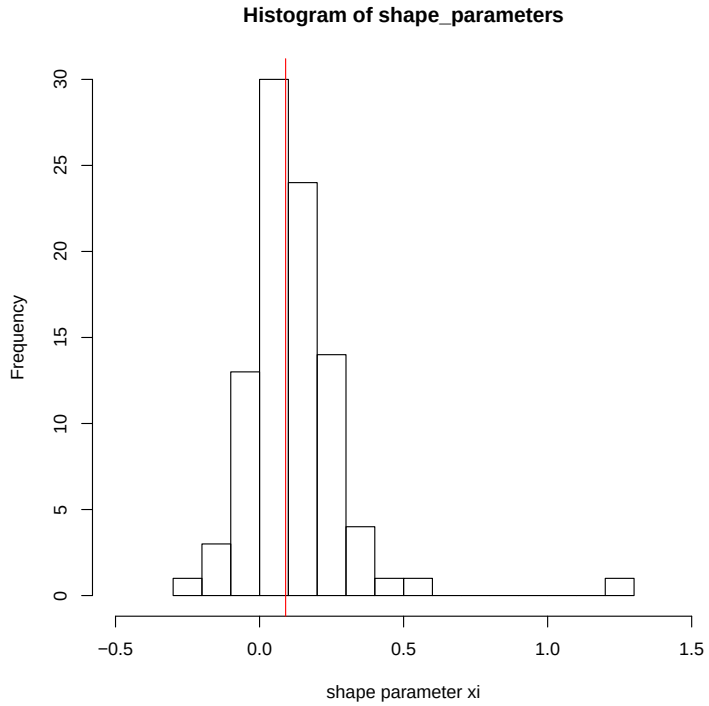

Figure 1: A histogram, showing the values of shape parameters from all stations. The shape parameter is constant across durations. The median is shown with a red line.

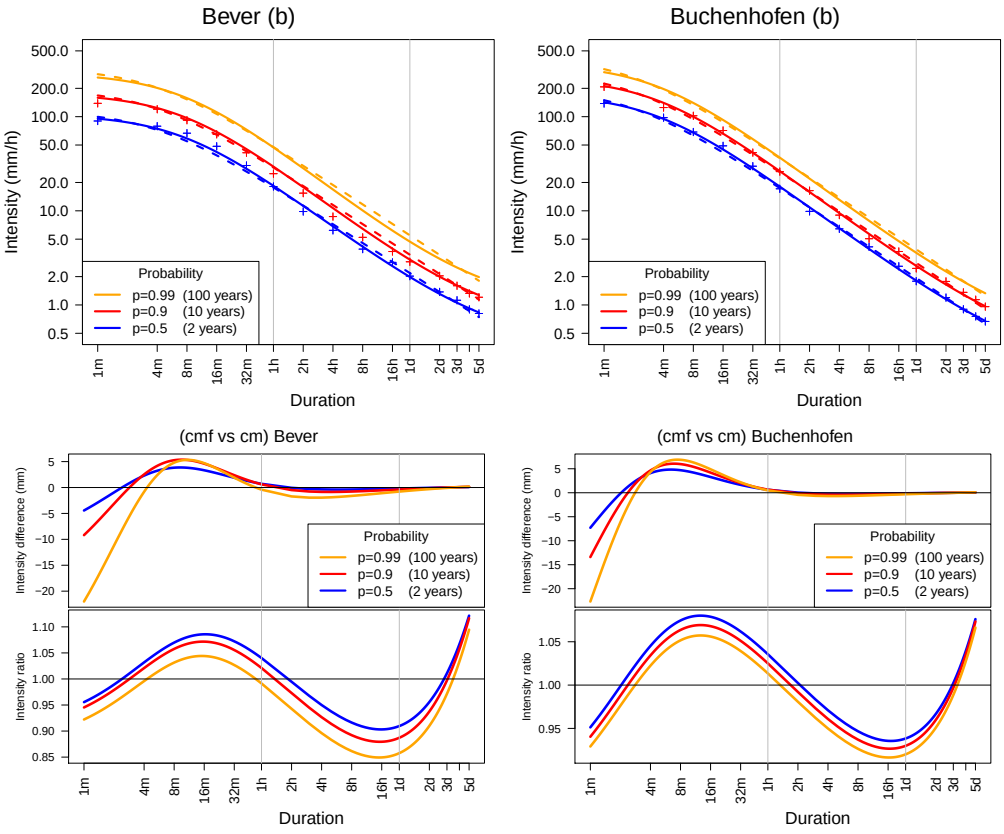

Figure 2: **Upper panels:** Fig. 7 from the manuscript. **Lower panels:** difference and ratio between the two models shown above.