Reply on RC1

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

August 26, 2021

Dear Rasmus Benestad,

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 and a table.

1 Answers

1. Due to the small samples of the rainfall from clouds which rain gauges with a diameter of centimetres represent, I wonder if it's justifiable to combine data from stations with a distance below 250 m into a single station data set. It's probably OK for the statistical properties, but maybe not so for the raw data - especially on short time intervals (too much random sampling noise?). At least, there ought to be a test of the robustness of the results to this assumption.

Answer: Thank you for this remark. We value transparency in data processing which is why we described carefully how station data was obtained. We agree that the grouping of stations has to be treated carefully. In most cases, data with different temporal resolutions are obtained from different devices. For more transparency, we present the distance between devices of the same station in Tab. 1 (in this document). It shows that the average distance between stations is much lower than 250m. We argue that the grouping should only be relevant for maxima of $d \geq 24$h (which can be derived from minutely, hourly or daily data) and on this time scale, the spatial distance should have less impact on the results for most stations.
To verify the robustness of this approach, we investigate the influence of the grouping on the resulting estimated IDF curves. For this purpose, we use for the modeling of maxima with $d \geq 24$h (1) the maxima originating from the time series with daily resolution and (2), if available, the maxima originating from the higher resolution time series to estimate the IDF curves. In both cases the time range covers all years in which minutely data is available, so that the number of used data points is the same. Exemplarily, the two IDF curves for Bevertalsperre and Buchenhofen are shown in Fig. 1. For Buchenhofen, the differences are sufficiently small. For Bever the intensities derived from daily data are smaller, which was expected, since daily sums from daily data come from a fixed time frame (e.g. 00:00 to 23:59) whereas daily sums from minutely data come from a 24-hour window that is shifted minute-

by-minute and so, larger sums are captured. When measurements for daily precipitation sums exist from different devices with different temporal resolutions, we decided to keep minutely data in our study because of the flexible time window. For most other stations, IDF curves for both cases do not differ significantly. For few stations, there are differences, especially when only short records are available. We will add these remarks to the discussion.
A major advantage of this procedure is that the full potential of the duration-dependent GEV model can be used when grouping daily and minutely data. This way, the model can profit from both long daily records and short records with high temporal resolution at the same time and information can be used and transferred efficiently.
Additionally, we will correct a wrong number. The total number of used stations is 115. After the grouping process, 92 stations remain for our analysis.

2. Perhaps it would make the flow of the text better if there was some connection between the return value (probability) and the quantile? I thought the part on Quantile Skill Index (e.q. 13) wasn't as easy to follow as the preceding sections (how does it link to the preceding discussion on the GEV and the estimation of the parameters?).

Answer: Thank you for pointing out that the part about the Quantile Skill Index was not clear. Sect. 2.4 explains how the model is trained to get the best parameter estimate. Sect. 2.5 describes the verification to evaluate the model's performance and to compare it to other models. The words "quantile" and "return value" are used synonymous. In the manuscript we refer (non-exceedance) "probabilities" p to the corresponding "return periods" T since they can be easily converted with $T = 1/(1-p)$. We will rephrase the introduction in this subsection and link it to the estimation of GEV parameters.

3. The description of the bootstrapping was a bit difficult to follow - perhaps explain it more carefully or add an illustration?

Answer: Thank you for this remark. We will improve the description and add a reference to Davison & Hinkley (1997) where this procedure is described in detail.

4. Very brief catches (e.g. minutes) of rainfall with rain gauges are expected to be subject to a large degree of sampling uncertainty, aren't they? (depending on the number of rain drops falling onto the cross section representing the measurement). Maybe this also can explain some discrepancies at the extreme short end of the scale?

Answer: Thank you for this remark. Minutely measurements might indeed be less accurate when only a small number of rain drops is recorded. However, for events that are identified as annual maxima we expect the rain amount to be large enough that a higher sampling uncertainty compared to larger measurement accumulation sums can be neglected. We will add this note to the discussion.

5. One reason why annual maxima of different durations do not follow the same scaling process could be that different rain-producing meteorological phenomena have different temporal and spatial scales. If the rainfall can be considered as a 'by-product' of different processes and conditions (e.g. convection, weather fronts, cyclones, and derechoes), then different statistics may perhaps show the true situation? But I'm still struggling to understand what the skill estimates really say.

Answer: Thank you for this comment. We agree that maxima for different durations might origin from processes on different spatio-temporal scales. The aim of this new model is to be more flexible and cover extremes from different processes. Especially, it is able to combine the estimation of extremes from both short, mainly convective, processes and processes on longer time scales like derechos. It would be interesting to investigate different statistics for different durations, but then we would need to have a reliable theory that justifies which statistic should be used for which duration. The quantile score (QS) is a proper scoring rule (Gneiting, 2011) and is obtained for each duration individually. Therefore, it can evaluate the model performance, independent of underlying physical processes. Its purpose is to asses how well the modeled quantiles represent the observed maxima. Skill estimates describe how well the model performs in terms of QS, compared to a reference, especially to the commonly used models IDF or $\mathrm{IDF_c}$ or $\mathrm{IDF_m}$ (for model description, see Tab. E1 in the manuscript).
We will improve Sect. 2.5 about the verification accordingly to describe the purpose of the skill estimates.

6. A new and relevant paper DOI: 10.1088/1748-9326/abd4ab suggests a simple formula for expressing IDF curves even for sites with limited data. This formula is based on more 'physical' parameters (wet-day mean precipitation and wet-day frequency), rather than the stronger reliance on the statistical/mathematical theory behind GEV. It would be interesting to compare the results presented here with this formula. It also fits in the comparison of different ways to parametrize IDF curves. At least, it could be included in the introduction and the discussion of different ways of calculating IDFs.

Answer: Thank you for suggesting this relevant paper. We will include it in the introduction and/or in the discussion where we are going to add a paragraph about non-stationarity, as suggested by another reviewer. We think that a reference to this paper would fit there as well.

7. Appendices: when describing what calculations and processing was done in this analysis, it's more elegant to use past tense rather than present tense (my subjective opinion). But mixing past and present tense makes the text inconsistent and a 'clumsy' read. Also check the references therein ('??').

Answer: Thank you for these two suggested improvements. We will improve both the tenses and the references in this Section.

On behalf of all authors
Felix Fauer

2 References

Davison, A., & Hinkley, D. (1997): Bootstrap Methods and their Application (Cambridge Series in Statistical and Probabilistic Mathematics), Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511802843

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

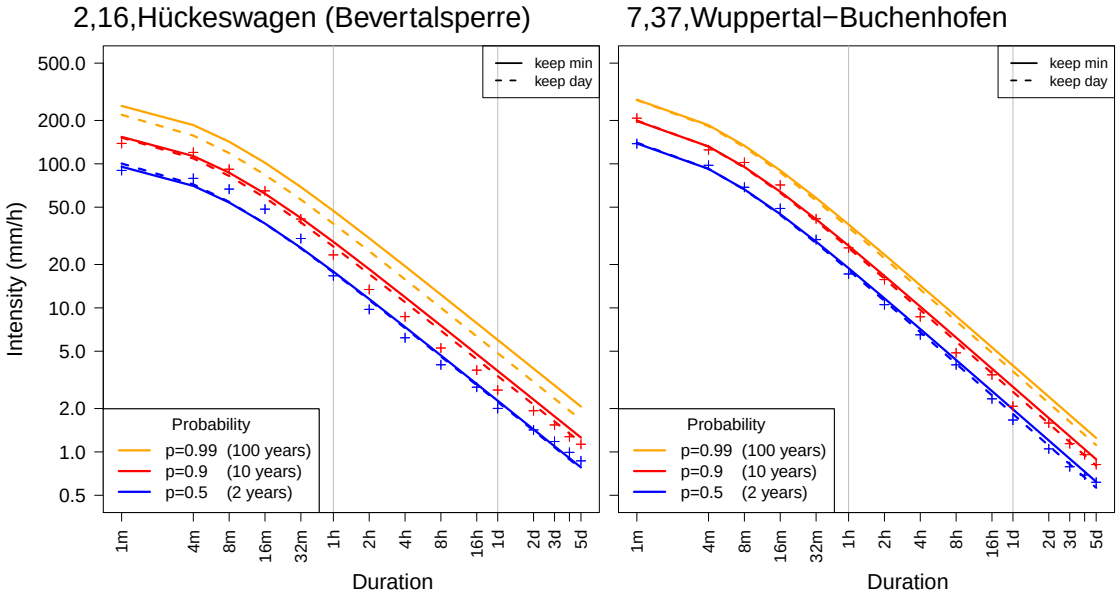

Figure 1: Comparison of IDF curves for two different ways of obtaining maxima for $d \geq 24$h. keep min: all maxima are derived from minutely data. keep day: maxima $d \geq 24$h are derived from daily data.

Table 1: Distance between devices of grouped stations. Two bold stations are chosen in the manuscript for detailed IDF curves.

Index	Number	Station name	Distance (m)
1	3	Breckerfeld-Wengeberg	90
2	16	**Hückeswagen (Bevertalsperre)**	101
3	18	Köln-Stammheim	49
4	30	Remscheid-Lennep	34
5	32	Solingen-Hohenscheid	22
6	35	Wipperfürth-Gardeweg	52
7	37	**Wuppertal-Buchenhofen**	217
8	50	Bochum	0
9	51	Dormagen-Zons	0
10	53	Köln-Bonn	0
11	54	Essen-Bredeney	0
12	64	Reichshof-Eckenhagen	0
13	65	Neunkirchen-Seelscheid-Krawinkel	0
14	66	Lüdenscheid	0
15	67	Meinerzhagen-Redlendorf	0
16	68	Overath-Böke	0
17	69	Gevelsberg-Oberbröking	0
18	72	Leverkusen	4
19	74	Neumühle	37
20	75	Schwelm	24