Author's response

Dear Editor and dear Reviewers,

Thank you for your time and efforts to review this manuscript. Your suggestions have improved the study structurally, methodologically and in the findings, for which we cordially thank you. The revision has been carried out following the answers to the reviewers' comments (https://doi.org/10.5194/egusphere-2025-6419-AC1 (AC1) and https://doi.org/10.5194/egusphere-2025-6419-AC2 (AC2)). We will briefly list the major changes as overview and then address all comments point-by-point. Please also find the tracked-changes version of the manuscript.

We hope that the revised version of the manuscript and my answers can satisfyingly address your comments and suggestions.

Kind regards,
Marc Lennartz, Benjamin Poschlod and Bruno Merz

Major changes:

- Revision of the study aim
- Revision of the study novelty
- More detailed return level comparison between the sMEV and GEV
- Making the software publicly available that results are reproducible
- Additional analysis regarding the RRMSE of GEV and sMEV to show the consistency of spatial and temporal patterns across different global warming levels
- Inclusion of a deviance test comparing temperature and time dependent non-stationary distributions of the GEV and sMEV to their stationary counterparts
- More comprehensive description of the iterative test to find an appropriate left-censoring threshold

Point-by-point answers RC1:

(*Reviewer comments in blue and italic*, answers in black; line numbers correspond to the new manuscript version without tracked changes)

If the methodological objective is to say that the SMEV method is more robust than the GEV approach, there is nothing new here, as this has been demonstrated in previous work.

We agree with this statement and admit that the objectives of the manuscript have not been formulated clearly. We have revised the claim of novelty in the abstract as such:

Page 1, line 8ff: "This comparison is novel in combining a large spatial and temporal extent with a high spatial and temporal resolution. The results show that, in the current climate, the GEV method estimates on average higher return levels than the sMEV for hourly extremes.

However, this difference reverses with further global warming. The results also reveal that the sMEV outperforms the GEV in the sensitivity to small samples across almost all return periods regardless of the length of observational records. In addition, there is a north-south gradient in the return level difference, the uncertainty difference and crucially the adequacy of the sMEV left-censoring threshold. In addition, we investigate non-stationary versions of the sMEV and GEV under climate change. The chosen non-stationary sMEV shows more promise in describing the change in return levels compared to the non-stationary GEV."

Additionally, we have reformulated the objectives in the introduction to include the following:

Page 3, line 81ff: "This approach allows us to investigate differences between the GEV and sMEV for hourly return level estimates, as well as their sensitivity to small samples. The latter is particularly important when estimating magnitudes of large return periods based on short observational records. Regional patterns for return level and uncertainty differences are discussed in the context of varying climate characteristics in conjunction with a fixed left-censoring threshold of the sMEV approach. Additionally, the uncertainty differences between the GEV and sMEV are investigated under varying sample sizes. The long time series used in the analysis allows us to reveal trends in return level estimates under global warming."

On the other hand, I think it would be interesting to present future projections for hourly rainfall in Germany using these two approaches. If this objective is retained, a more comprehensive literature review is needed to explain what projections are currently available in Germany on hourly extremes and how the results of this study either support them or produce new results.

Thank you for the suggestions of viewing the results of this study from a different angle. We agree that comparing return level estimations to other studies would be interesting. We have added a set of figures in the supplement showing the following:

The spatial distribution of the 10-year return level for different regional warming levels of the GEV (Fig. S9) and sMEV (Fig. S10) distribution.

In the main article we have added the following figure:

Fig 5: Showing the spatial distribution of the temperature scaling rates for the GEV and sMEV 10-year return level.

Additionally, we have added the following context in the results section:

Page 10, line 284ff: "The spatial distribution of the temperature scaling rate is shown using the 10-year return level as an example. Both the GEV (Fig. 5(c)) and sMEV (Fig. 5(d)) exhibit a similar north-south gradient with higher scaling rates in the north of the study region and smaller values towards the south. An exception is the Alps region, which shows higher scaling rates compared to other locations at similar latitudes. Compared to the sMEV, the GEV exhibits slightly higher scaling rates in the northern sea regions and slightly lower values in the center of the study domain, as well as smaller values around the Alps region (Fig. S11)."

And the additional context in the discussion section:

Page 16, line 335: “Our results show a similar pattern to Rybka at al. (2023) which analyzed scaling rates of the 10-year return level for 3-hourly rainfall extremes. Their results are derived using the peak-over-threshold method and are constraint over Germany (Fig. 8 in Rybka at al. (2023)). Similar to our results they find higher scaling rates in the north of Germany, however our results show higher scaling rates on average, which is in part due to our study domain being larger.”

However, we note that gauging the reliability and plausibility of such projections is out of scope as they are governed by model uncertainty, scenario uncertainty, and internal climate variability beyond the differences in extreme value approaches.

[…] In fact, this is not really what is done here to evaluate the performance of one model compared to another; they must be compared to an observed reference. Furthermore, talking about performance is not very clear here, as it is not a very specific objective.

We agree that the phrase “evaluate performance” could be potentially misleading, by implying that the distributions are evaluated, based on comparisons between their respective quantile estimations and a common ground-truth. In fact, the return level estimations are not validated and only the sensitivity to small samples is used as a performance-related metric. However, this is a very important property, especially when estimating rare extremes, and studying this sensitivity has been done in many similar studies. We believe that the presented variability in sensitivity between the GEV and sMEV is not trivial and exhibits interesting patterns across return periods and across space. This has implication when deciding which distribution is most appropriate when estimating values of a specific return period across Germany.

A comparison to an observed reference, in turn, evaluates the capabilities of the climate model to reproduce the characteristics of hourly precipitation extremes. This has already been done by Rybka et al. (2023) as described in lines 106ff (in the initial manuscript).

For clarification, we have slightly extended this description so that the reader is well aware of the comparison COSMO vs. observational reference:

Page 11, line 290ff: “The sensitivity to small samples for varying return level estimates is quantified by the relative root-mean-square error (RRMSE). Here, the return levels derived by considering the whole 90-year time series serve as ground-truth.”

To our knowledge, this is the first application of the sMEV distribution to time series from a convection-permitting-mode => this is wrong, see these references below, among others: […] Generally speaking, as soon as I read an abstract "for the first time", I think to myself that the results of an article are potentially overstated.

Thank you for pointing out this inaccuracy. The relevant statement has been removed and paragraphs claiming novelty in this context have been reformulated accordingly (see answer to first comment).

Specific comments:

Page 4, line 95. It is somewhat surprising to read that the scenario is not suitable because it is too pessimistic. A little more context is needed here to explain why this scenario was selected in this context.

Here, the term “pessimistic” referred to recent literature suggesting that future emissions may remain below RCP8.5 levels (Hausfather, Zeke, and Glen P. Peters. "Emissions–the ‘business as usual’ story is misleading." *Nature, 2020*). The scenario was selected because it spans a wide range of regional warming levels, allowing analysis beyond weak warming conditions. However, we agree with the reviewer’s sentiment and reformulated the section to provide more clarity.

Page 4, line 105ff: “This scenario spans a wide range of regional warming levels, allowing analysis beyond weak warming conditions. It leads to a drastic temperature increase over the domain (Fig. 1(b)). However, recent literature suggests that future warming levels will likely stay below this scenario (Hausfather and Peters, 2020).”

Page 5, line 115. It is very good to cite the sources for the codes used. However, it would be interesting to specify what changes have been made and possibly produce the modified code.

Thank you for the suggestion. We now clearly specify that only selected sub-functions from the existing code (e.g. event separation and inversion routines) were used, while the greater framework of the analysis was developed independently.

Page 5, line 124ff: “The software for the statistical analysis, written in Matlab R2023b, is based on previous work by Marra (2020) and Marra and Peleg (2023). Certain smaller functions, like inversion or optimization routines, are used with minor modification. However, the greater framework of the analysis is entirely rewritten. For more information see the "Code availability" section.”

The full workflow, including preprocessing, main calculations, and plotting functions, will be made publicly available via GitLab. However, this process is currently subject to an GFZ-internal review. Thus, the repository may not yet be available at the time when this response is submitted, but will be available in the next weeks.

Page 5, line 116 The lmoment method does not focus solely on extreme values. Above all, the method allows for a more robust estimation of the parameters. This needs to be changed here.

We agree that the original phrasing was lacking important properties of the L-moments approach. We added an emphasis on robust parameter estimation rather than focusing solely on extremes.

Page 5, line 129ff: “All parameters are estimated using the method of L-moments, which allows for a more robust parameter estimation compared to the maximum likelihood estimation. In comparison, the L-moment method has reduced sensitivity to outliers (Poschlod, 2021) and places a greater emphasis on the tail of the distribution (Hosking, 1990).”

Page 5, line 117 There are also regional approaches that provide much more robust estimates for parameters of frequency models compared to pixel-by-pixel estimation. Even though such approaches are not used here, I think we should include this methodological warning.

Thank you for adding this important context, which have been added in the method section:

Page 5, line 133ff: "Note that other parameter estimation methods include regional information in local estimations. This information expansion serves to spatially smooth parameter estimates and, as a result, return level estimates as well (Hosking and Wallis, 1997; Blanchet and Lehning, 2010; Schellander et al., 2019). In turn, results are spatially less noisy and potentially more robust. However, these methods have the potential of hiding local extreme value characteristics. Additionally, for real world application such regional data may not always be available."

Page 5, line 120. I think some comments should be added about the difficulty of working with 200-year return periods calculated from 30-year time series, particularly for highly variable precipitation extremes. I think this is a highly questionable methodological choice given the methodology and data available.

We acknowledge the difficulty of estimating 200-year return periods from 30-year samples. However, this choice was made deliberately, as one advantage of the sMEV framework is robust estimation of extremes using short sample sizes. The large uncertainties present in such return period estimations are now discussed more explicitly in the limitation sector:

Page 19, line 455ff: "In general, estimating large return levels from few observational years is highly uncertain regardless of the extreme value method used. In reality, great care should be taken when estimating large return levels in such a way. The sensitivity to small samples may seem relativity small in case of the sMEV. However, the true range of uncertainty may be far greater, if the previously mentioned assumptions are not met. Thus, especially the results based on a 30-year sample size should be viewed with this limitation is mind (for instance Fig. 4 and Fig. 5)."

Page 5, line 123. This is my main issue with the methodology here. I don't see how repeating by bootstrap observational data over a long series will give more confidence in validating one model over another. I think the methodology here is not suitable for comparing the two distributions. To compare distributions, you need to use observational data and compare the fit to the observed data and calculate, for example, the errors between observed and simulated quantiles for the two distributions.

The reviewer is correct that the bootstrap approach is not suitable for validation against observations. We would like to clarify that the aim is instead to compare sensitivity to small samples. This is quantified using the relative root mean square error, applied to simulated climate model data. Please see our response to your earlier comment and the additional text in the revised manuscript (Page 11, line 290ff).

You can also look at the confidence interval produced by the two approaches with the two distributions.

If we understand the suggestion correctly, it seems to us, that the RRMSE values we present are not very different to calculating the confidence interval. Confidence intervals are typically calculated via bootstrapping the available years. They show the sensitivity of the return levels to sample variability. Similarly, via bootstrapping we show the return level sensitivity to the combined effect of sample variability and sample length.

In addition, as pointed out by the authors, this bootstrap approach does not take into account temporal dependence or possible trends at all. So it is difficult to see how this approach can be adapted to model comparison.

Indeed, effects of interannual autocorrelation and temporal trends are assumed to be negligible when comparing the sensitivity of small samples between the GEV and sMEV distribution. Justifying this assumption demands further investigation. We now compare the results of the RRMSE, if applied to the 3 time periods individually. These time periods are themselves not stationary, but they exhibit significantly less variability within the time period compared to the variability across time periods. These newly produced results reveal the same spatial patterns as well as similar relative differences between the GEV and sMEV for sample sizes below 30 years. Two figures (Fig. S13 and Fig. S14) have been added in the supplement of the revised article.

Page 6, line 147. The GEV parameters are associated with temperature as a covariate. However, there is no justification of the added value of this covariate. For example, a deviance test could be used to verify the added value of this covariate compared to a stationary model. Otherwise, the authors do not provide sufficient guarantees to verify that the principle of parsimony is satisfied.

We agree that the added value of temperature as a covariate should be formally tested. The revised supplement, therefore, includes a deviance test comparing stationary models to non-stationary models with temperature and time as a covariate (Fig. S15). More specifically, we estimate the Kullback–Leibler divergence using the corrected Akaike's information criterion, similar to Kim et al. (2017, http://dx.doi.org/10.1016/j.jhydrol.2017.02.005). This information criterion weighs added information against a penalty associated with higher model complexity. The analysis shows that the temperature and time dependent non-stationary distributions of both the GEV and sMEV add value and the principle of parsimony is satisfied. The deviance between the regional temperature and time dependent distributions compared to the stationary distribution is similar. However, the time dependent distribution shows a slightly better fit in both cases. We have added the following context in the method section:

Page 9, line244ff: "We have confirmed the added value of these non-stationary distributions applied to the whole dataset by calculating the corrected Akaike's information criterion, similar to Kim et al. (2017) (Fig. S15). Compared to stationary distributions the non-stationary distributions have a smaller value for the corrected Akaike's information criterion and thus add value when applied to this dataset."

Furthermore, it is rather difficult to understand why, in the context of trend detection, time is not used as a covariate, which is the most commonly used approach.

Regional temperature was chosen because of its more direct physical relationship (thermodynamic scaling) with precipitation extremes, compared to time. The regional domain of the temperature was chosen as in central Europe moisture from precipitation extremes are largely supplied by local/regional sources (Keune, Jessica, and D. G. Miralles. "A precipitation recycling network to assess freshwater vulnerability: Challenging the watershed convention." *Water Resources Research*). The assumption is that higher regional temperatures have a positive effect on the intensity of these extremes referring to the thermodynamic scaling of extreme precipitation.

Additionally, time as a covariate becomes problematic once the long-term global warming trend changes. We argue that regional temperature will outperform time as covariate, if the calibration period exhibits a significantly different global warming trend compared to the validation period.

Page 8, line 215 I don't understand the consistency here between using a GEV with a linear dependence on temperature, an SMEV model with an exponential dependence, and then explaining in this section that ultimately a linear dependence is used.

We would like to clarify that the ns-GEV and ns-sMEV formulations differ fundamentally: the ns-GEV includes a linear temperature dependence in the location parameter, while the ns-sMEV combines an exponential dependence in the scale parameter with a linear dependence in the rate of ordinary events. The relevant section has been rewritten to remove ambiguity.

Page 9. Line 240ff: “Based on Vidrio-Sahagún (2023) and Marra et al. (2024), this is a linear dependence of the GEV location parameter, while other parameters are kept constant. For the sMEV distribution the shape parameter is kept constant, while the scale parameter is exponentially dependent on temperature.”

Page 11, line 258. It is unclear how the root mean square error is calculated here. This concerns the observed data? or the simulation approach proposed in the methodology?

The RRMSE is calculated using only simulated data, as detailed in the methodology. The section is now rewritten to remove ambiguity.

Page 11, line 290ff: “The sensitivity to small samples for varying return level estimates is quantified by the relative root-mean-square error (RRMSE). Here, the return levels derived by considering the whole 90-year time series serve as ground-truth.”

Page 15, line 296. Could this result be influenced by the covariate temperature? I don't understand why the quantiles are not compared with the stationary models here.

These results are derived by the stationary models. At this point the covariate temperature is not yet included explicitly. Thus, it does not directly influence these quantiles of the distributions. The section has been reformulated adding this information.

Page 16, line 331f: “Comparing the return levels derived by the stationary models, we observe that for the past period the GEV generally provides higher return levels than the sMEV.”

Page 18, line 400 This sentence is incomprehensible. The authors write that they do not make scenarios about extreme precipitation events, yet that is exactly what they are doing. Explaining that this work is on the "challenges and opportunities of stationary and non-stationary distributions" does not mean much.

Thank you for pointing out this mistake. The sentence does not add any value and is removed. Please refer to the introduction for an updated overview of the objectives and novelties.

Page 19, line 431. ", which is why more complex versions are needed to represent strong change" = what does it mean?

Thank you highlighting this poor phrasing. We have rephrased the sentence as such:

Page 20, line 485ff: "However, neither of the chosen non-stationary distributions are able to fully capture the effect of climate change on extreme precipitation. It is likely that more degrees of freedom in the non-stationary parameters of the distributions are necessary to improve the ability of capturing the effect of climate change."

Point-by-point answers RC2:

(*Reviewer* comments in blue, answers in black; line numbers correspond to the new manuscript version without tracked changes)

In general, I don't believe that the application to modeling results represents a source of originality fundamental to the paper's value. In my opinion, the database to which the methodologies are applied is a tool rather than an objective, so it doesn't matter whether the application is the first, and therefore I would relax the various statements in this regard.

We agree with your observation and appreciate the suggestion. Consequently, we have removed all claims of "first application" to avoid overstating the results. For a detailed overview please refer to the first response to Reviewer #1.

It seems to me that the variations in RL captured by the distributions for the different RPs are not very far from the RRMSE values for the same RPs. I would suggest, perhaps in the limitations section, mentioning this more explicitly when discussing the uncertainties of the distributions due to the assumptions.

Thank you for the hint. Indeed, the spatial variations in return levels, even between neighbouring grid cells, are of a similar magnitude to the relative root mean square error. This is important when putting the uncertainties between the GEV and sMEV into context. This has been added in the limitations section.

Page 19, line 459ff: "Moreover, small sample uncertainty and method uncertainty may have varying importance depending on the location and return level. For instance, when estimating a 100-year return level with 30 years of observations, the spatial average RRMSE of the GEV and sMEV is 34.6% and 12.5%, respectively, while the return level differences tend to be lower in the center of the study domain and regularly exceed 40% towards the extreme north or south (see Fig. 4(d))."

par. 2.1 study area: maybe adding a figure with the K-G climate regions would help readability (maybe in the supplement?)

We agree that a map of the climate regions would be very helpful. Especially, since climatological properties greatly influence the results of our analysis. Figure S8 now includes a map of the Köppen-Geiger climate regions based on Beck et al. (2023: https://doi.org/10.1038/s41597-023-02549-6) in the supplementary material.

par. 3 methods, row 130 "effects introduced by climate change... etc" is an important statement that should require some references

Indeed, effects of interannual autocorrelation and temporal trends are assumed to be negligible when comparing the sensitivity of small samples between the GEV and sMEV distribution. Justifying this assumption demands further investigation. We now compare the results of the RRMSE when applied to the 3 time periods individually. These time periods are themselves not stationary, but they exhibit significantly less variability within the time period compared to the variability across time periods. These newly produced results reveal the same spatial patterns as well as similar relative differences between the GEV and sMEV for sample sizes below 30 years. Fig. S13 and Fig. S14 have been added in the supplement of the revised article.

Par. 3.2 sMEV: row 180-189 I found it difficult to interpret what is written in the paragraph. It certainly refers to the method of Marra et al., but I would suggest expanding the section to include more details on the specific application of the method.

We recognize that the description of the iterative sMEV procedure was overly condensed. The methods are now explained in greater detail. The loop applied in the test is described in a more structured way, explaining every step and clearly stating the break condition.

Page 7, line 202ff: "This test iteratively evaluates the likelihood that the observed RX1h values arise from a specific left-censored sMEV distribution. Starting with a very low left-censoring threshold $p \in [0,1)$, the sMEV is fitted to all ordinary events greater than the p-quantile. Based on this sMEV distribution, $\bar{n}$ x N events are randomly sampled, with the maxima across $\bar{n}$ yielding N RX1h events. Repeating this random sampling many times yields the 90% confidence intervals of the nth largest RX1h values $\forall\ n \in \{1, \ldots, N\}$. Next, the proportion of actual RX1h values falling within this confidence interval is assessed. If the number of events outside the confidence interval is larger than 10%, the loop is repeated with an incrementally larger left-censoring threshold. Otherwise, this left-censoring threshold is chosen as the ideal value and the loop ends."

Par. 5.1 The first comment clearly refers to GEV, but it's not immediately clear from the text. I'd suggest rewriting it, immediately citing the context to which the comment refers. It would also be helpful to include, at row 290, a reference to the figure where the median shape parameter is represented (Fig. 3?).

Thank you for these suggestions for improving clarity. We have rewritten the opening of Paragraph 5.1 to explicitly state when we are referring to the GEV distribution.

Page 15, line 320ff: “For lower return periods the scale and location parameter of the GEV are both causing an increase of similar magnitude, while the change in the shape parameter can be neglected (Fig. S1). For higher return periods the location and shape parameter cause an opposite effect of similar magnitude, while the scale parameter causes a significantly larger increase. It is a noteworthy result that the median shape parameter of the GEV is decreasing, making the distribution less heavy-tailed (Fig. 3(a)).”