

We sincerely thank the reviewer for the careful assessment and constructive comments. We understand that the main concerns relate to three aspects: (1) the mathematical basis and diagnostic role of the two-segment approximation, (2) the justification and sensitivity of the weighting and scoring choices, and (3) the distinction between in-sample fitting and reliable estimation of TCEV parameters and upper quantiles. In the revision, we will clarify the theoretical basis of the method, add focused sensitivity and uncertainty analyses, replace the current simulation comparison with a TCEV-based experiment, and revise overly broad claims regarding predictive performance and consistent superiority over other methods.

Main

1. Section 2.3.2: The authors argue that Equation (3) implies that under a double-log transformation, the TCEV is described by two linear segments. However, this statement is not mathematically correct because this transformation gives:

$$-\ln[-\ln(F_{TCEV}(x))] = -\ln[\exp(-\alpha_1 \cdot (x - \beta_1)) + \exp(-\alpha_2 \cdot (x - \beta_2))],$$

and the right-hand side does not simplify to two linear segments. Therefore, while the argument in Equation (3) is correct, we require strong assumptions for the right-hand side to approximately have the form of two linear segments. Specifically, we require the two Gumbel distributions to be sufficiently different. The authors need to carefully explore this aspect and provide a more rigorous analysis under which conditions this two-line segment representation can be assumed. Plotting some of the estimated models in Table 1, the necessary assumption does not seem to hold for all the applications. For instance, when I use the estimates from Case 1 in Table 1 using the equation above, I obtain a curve that looks close to linear, and it is noticeably different from the estimate in Figure 4 the authors obtain with their method. This aspect also results in follow-on issues related to Equations (4) and (5). Since the linear segments in Equation (4) are at best an approximation, it is not clear how the Gumbel parameters can be directly recovered. This probably requires some choice on how segments should be defined, but this aspect is not clearly explained. While I understand that these methods are mainly used to identify a starting point for the parameter values, this aspect is important, also because for the subsequent decision on whether a TCEV model should be considered.

Re: We agree that the statement in the original manuscript that the double-log-transformed TCEV distribution is represented by two linear segments is not mathematically exact. Equation (3) is exact for a single Gumbel distribution, but its direct extension to the TCEV distribution requires additional assumptions.

This expression shows that $(-\ln[-\ln(F_{TCEV}(x))])$ approaches the dominant linear component only when $(d(x))$ is sufficiently large. Near the point at which the two components make similar contributions, the transformed TCEV curve exhibits a smooth transition rather than an abrupt breakpoint. An approximate two-segment representation therefore requires the two Gumbel components to be sufficiently separated in their slopes and locations, the transition between the dominant components to occur within or close to the observed data range, and a sufficient number of observations to be available on both sides of the transition.

We will revise section 2.3.2 to replace the original exact statement with this approximate component-dominance interpretation and provide the corresponding derivation and conditions. Equations (4) and (5) will also be revised to clarify their purpose. The slopes and intercepts of the fitted segments do not provide exact estimates of the four TCEV parameters. Instead, they are used only to construct initial parameter values for the subsequent nonlinear optimization of the complete TCEV function. The final TCEV parameters are estimated using all observations and are not obtained directly from the two-line approximation.

The practical breakpoint-estimation procedure will also be described explicitly. After sorting the transformed empirical plotting positions according to x , all candidate breakpoints satisfying the specified minimum number of observations in both segments are examined. For each candidate, two linear regressions are fitted, and the breakpoint producing the lowest combined residual error is selected. The revised manuscript will report the candidate range, minimum segment size, residual criterion, and the treatment of cases in which several candidate breakpoints produce similar errors.

We agree that the existence of an RMSE-minimizing breakpoint does not, by itself, provide sufficient evidence of a two-component structure. We will therefore compare the two-segment approximation with a single-line representation using a criterion that accounts for their different model complexities and examine the stability of the breakpoint and segment slopes under resampling. The slope ratio will consequently be treated as preliminary diagnostic information rather than as a stand-alone criterion for TCEV

applicability.

We also appreciate the reviewer's specific observation concerning Case 1. Figure 4 shows a piecewise regression fitted to the transformed empirical plotting positions, whereas the curve calculated from the parameters in Table 1 represents the exact transformed fitted TCEV model. Although these are not identical quantities, their overall shapes should be reasonably consistent if the two-segment approximation is informative. We will re-examine this comparison for Case 1 and the other cases. If the exact transformed curve is close to a single line or the improvement from using two segments is limited, the corresponding case will not be interpreted as providing strong evidence of two clearly separated components.

Accordingly, the revised manuscript will present the piecewise linear analysis as an approximate procedure for generating initial parameter values and providing preliminary diagnostic information. It will no longer be described as an exact analytical representation of the transformed TCEV distribution or as sufficient evidence of TCEV applicability.

2. Slope-Ratio: The slope-ratio criterion of $R < 1.5$ feels very arbitrary to me, and the authors provide little justification for this choice. Should we not also consider the difference in β_1 and β_2 when deciding whether a simpler distribution may be used?

Re: We agree that the threshold ($R=1.5$) adopted in the original manuscript is an empirical screening value rather than a theoretically unique cutoff. It was introduced to require a noticeable difference between the slopes of the two fitted segments, but the original manuscript did not provide sufficient justification for this choice or discuss its uncertainty.

We also agree that the slope ratio alone is insufficient to determine whether a two-component TCEV distribution is preferable to a simpler distribution. Differences between the two location parameters, β_1 and β_2 , also provide relevant information about the separation of the ordinary- and extraordinary-event components. However, because the magnitude of $|\beta_1 - \beta_2|$ depends on the units and scales of the two components, we will consider this difference relative to the corresponding component scales and use it as supplementary evidence rather than introducing another unvalidated fixed threshold.

In the revised manuscript, $R=1.5$ will be described as a preliminary practical

criterion rather than a definitive or universally applicable threshold. We will supplement it by considering the separation of the location parameters, the improvement of the two-segment fit relative to a single-segment fit, and the stability of the estimated breakpoint and slope ratio. We will also conduct a sensitivity analysis using alternative values around ($R=1.5$) to examine whether the classification of the cases changes materially. Cases for which the evidence of a two-component structure is weak or unstable will be identified explicitly and interpreted cautiously.

3. Weighting and model fitting: The authors propose three weighting schemes, but it's not clear why these were selected in particular. Have the authors considered other schemes that just did not perform well in practice? A more serious issue for me is the overall fitting process. Some of the methods in Section 2.2 consider likelihood methods which allow us to account for the uncertainty in the observations, which tends to be very high for extreme observations. The approach by the authors considers the data as fixed as they look for the best match in terms of the empirical CDF. How is overfitting in such a situation avoided? I cannot find a discussion on this aspect in the manuscript.

Re: We agree that the original manuscript presented the mathematical definitions and normalization of the three weighting functions but did not sufficiently explain why these particular forms were considered. We will therefore revise section 2.3.3 to clarify the rationale for their selection and the effect of normalization. We will also clarify that these three functions are representative choices rather than an exhaustive set of possible weighting functions. We will also include an unweighted fit as a reference to assess whether the main conclusions depend on the use of upper-tail weighting.

We also appreciate the reviewer's concern regarding overfitting. The proposed method minimizes the discrepancy between the fitted and empirical CDFs, and the partitioned scoring procedure selects a weighting scheme using the same observed sample. The use of predefined weighting functions and the same four-parameter TCEV model constrains the flexibility of the procedure, but does not eliminate the possibility that the selected solution is excessively influenced by a small number of upper-tail observations. Weight normalization makes the WRMSE objectives comparable, but is not, by itself, a safeguard against overfitting. We will clarify this limitation in the revised manuscript.

To assess this issue more directly, we plan to extend the validation through

repeated-sample simulation experiments. Published TCEV parameter estimates available for several of the hydrological cases will be used as hydrologically relevant generating values. We recognize that these published values are themselves estimates from finite observed records and therefore cannot be regarded as the unknown true parameters of the corresponding hydrological populations. However, once used to define the simulated TCEV distributions, they are known generating parameters by construction.

For each generating parameter set and selected sample size, independent TCEV samples will be generated repeatedly and analysed using the unweighted fit and the three weighting schemes. Parameter estimates will be compared with their known generating values using bias and RMSE. We will also evaluate errors in selected upper quantiles or return levels and examine the variability of the weighting scheme selected across repeated samples.

4. Section 2.3.4 The authors select amongst the three weighting schemes by considering the fit across three evaluation areas, $[0.8, 1.0]$, $[0.6, 0.8)$ and $[0, 1]$ which again feels quite arbitrary. Have the authors explored sensitivity of their results in later sections to the choice of weighting given to these three areas?

Re: We agree that the original allocation of 60, 30, and 10 points to the intervals $[0.8, 1.0]$, $[0.6, 0.8)$, and $[0, 1.0]$, respectively, reflects the intended hydrological priority given to the upper tail, but is not a uniquely determined statistical choice. The original manuscript did not sufficiently justify this allocation or evaluate the sensitivity of the results to it.

In the revised manuscript, we will explain that the $([0.8, 1.0])$ interval was assigned the greatest influence because the study focuses on rare upper-tail events, the $([0.6, 0.8))$ interval was included to account for the transition toward the upper tail, and the full interval $([0, 1.0])$ was retained as a constraint against excessive deterioration in the overall fit. Nevertheless, we will no longer present the 60/30/10 allocation as a generally optimal or theoretically unique choice.

We plan to add a sensitivity analysis in which several predefined allocations are applied to the same three evaluation areas, including a more balanced allocation and alternatives placing weaker or stronger emphasis on the upper tail. For example, the analysis will consider allocations such as 40/30/30, 60/30/10, and 70/20/10 for the upper-tail, transition, and overall areas, respectively. For each allocation, we will examine whether the selected weighting scheme changes and quantify the resulting variation in the four TCEV

parameters and selected upper quantiles or return levels.

For the simulated datasets with known generating parameters, the alternative allocations will also be evaluated using parameter bias and RMSE, together with upper-quantile or return-level errors. These external evaluation measures are independent of the score used internally to select the weighting scheme. This will help determine whether the conclusions are robust to the scoring allocation rather than being a consequence of the originally selected 60/30/10 values. The sensitivity results and any cases in which the selected scheme or upper-quantile estimates change substantially will be reported and discussed explicitly.

5. Benchmark methods: It is unclear how the POT method was applied. Have the others only applied it to the annual maximum series, or the full data set. I am also missing a discussion on which threshold selection techniques were used to determine the 80th percentile for the analysis. Can the authors please add this information to Subsection 2.4? Finally, would it not be useful to compare against the other TCEV fitting methods too, instead of just considering methods that cannot handle data generated from two different processes?

Re: In the original analysis, the POT method was applied to the annual-maximum series because sub-annual or event-level records were not consistently available for all cases. The threshold was fixed at the empirical 80th percentile of each annual-maximum sample, and a generalized Pareto distribution was fitted to the resulting exceedances using maximum likelihood estimation. However, we agree that this is not a standard implementation of the POT framework, which is generally applied to a more complete event series rather than to data that have already been reduced to annual maxima. We also acknowledge that the 80th-percentile threshold was adopted as a fixed empirical choice and was not selected through a formal threshold-diagnostic procedure, such as a mean residual life plot or parameter-stability analysis. The original manuscript did not explain these limitations adequately.

In response to this comment and the reviewer's related comments on the simulation design, we plan to replace the current comparison in section 4.4 with a more direct simulation-based comparison of TCEV estimation procedures. The revised simulations will be generated from TCEV distributions with known parameter values, and the proposed SR-MWS method will be compared with existing TCEV procedures for which reliable implementations can be obtained

or reproduced. The comparison will focus on parameter bias and RMSE, upper-quantile or return-level errors, and numerical convergence across repeated samples. This design is more consistent with the purpose of the study and allows the methods to be evaluated under data generated by two underlying processes.

For the existing TCEV methods, we are currently examining the availability of the original implementations and the completeness of the published algorithm descriptions. Where a method is reconstructed from the literature, we will document its initialization, parameter constraints, stopping criteria, and other numerical settings, and verify the implementation against published case-study results as far as possible. If a method cannot be reproduced reliably because essential implementation details are unavailable, this limitation and the resulting scope of the comparison will be stated explicitly.

6. Simulation Setup: In Section 3.2, the authors that they also considered simulated data sets of size 30 years. Would one really consider the TCEV with four parameters for such a small data set due to the large uncertainty in the parameter estimates? Regarding my comment above on the benchmark methods, I am also not sure why the authors sampled from a Pareto distribution which is usually used to model POT and not annual maxima.

Re: The sample size of 30 was originally included to represent short hydrological records and was motivated by the shortest observed record in our case studies, which contains 31 annual maxima. We agree, however, that estimating four TCEV parameters from such a short record may result in substantial uncertainty and numerical instability. Therefore, the (N=30) setting will be retained only as a deliberately challenging small-sample scenario, rather than as a recommendation for routinely fitting the TCEV distribution to 30-year records.

In the revised simulation experiment, we will consider several fixed sample sizes, including N=30, 50, 100, and 200, so that the effect of record length can be evaluated explicitly. For each sample size, repeated samples will be generated from TCEV distributions with known parameters. We will report parameter bias and RMSE, errors in selected upper quantiles or return levels, confidence-interval coverage where available, and the frequency of convergence failures or invalid parameter estimates. These results will allow us to quantify the limitations of TCEV estimation for short records and to determine

whether the proposed procedure remains stable as the sample size decreases. The Discussion will also be revised to emphasize that reliable application of a four-parameter TCEV model requires sufficient evidence of a two-component structure and careful consideration of parameter uncertainty, particularly for short records.

We also agree that the rationale for generating annual-maximum samples from a Pareto distribution was not sufficiently justified. The Pareto-type framework is more commonly associated with threshold exceedances, whereas the present study focuses on annual maxima and the estimation of TCEV parameters. Following the reviewer's suggestion, we plan to replace the current multi-distribution simulation in section 3.2 and section 4.4 with simulations generated directly from TCEV distributions with known parameter settings. Consequently, the Pareto simulation and the associated comparison in the current section 4.4 will be removed. This revised design will provide a more direct and interpretable comparison of the TCEV estimation procedures.

7. Discussion on predictive performance: At several places the authors state that their method has a better predictive performance. However, their assessment is purely based on data fitting. Assessing predictive performance in extreme value analysis is generally tricky. For the cases with a larger number of data points, one may consider assessing predictive performance using cross-validation although such an approach may be debated. In any case, it would be good to add confidence intervals in Figure 9 to assess uncertainty in the predictions obtained by the different methods.

Re: We agree that the analyses presented in the original manuscript primarily assess in-sample goodness of fit and do not, by themselves, provide sufficient evidence of predictive performance. In particular, closer agreement between the fitted and empirical CDFs should be interpreted as improved representation of the observed sample rather than as direct evidence of more accurate prediction of unobserved extreme events.

We will therefore review the manuscript systematically and replace statements referring to “better predictive performance”, “prediction accuracy”, or similar claims with more precise terms such as “better goodness of fit”, “improved upper-tail representation”, or “lower fitting error”, wherever the results are based only on the observed data. Claims concerning estimation beyond the observed record will also be moderated and discussed in relation to the associated

sampling uncertainty.

We appreciate the reviewer's suggestion regarding cross-validation. Although cross-validation can provide useful information on out-of-sample performance when sufficiently large samples are available, its application to the annual-maximum records considered here is challenging. The upper tail is represented by only a small number of observations, and dividing the records into training and validation subsets would further reduce the number of extremes available for both fitting and evaluation. The results could therefore depend strongly on whether individual extreme observations are assigned to the training or validation subset. This is particularly problematic for identifying and estimating the extraordinary-event component of the TCEV distribution. For these reasons, we will not use cross-validation as the primary validation approach.

Instead, we plan to assess the recovery of TCEV parameters and upper quantiles through repeated-sample simulation experiments based on TCEV distributions with known generating parameters. For each parameter setting and sample size, independent samples will be generated repeatedly and analysed using the complete fitting and weighting-selection procedure. The estimated parameters and selected upper quantiles or return levels will be compared with their known generating values using bias and RMSE. This design will allow us to determine whether an improved fit to the empirical upper tail is accompanied by reliable recovery of the underlying distribution across independent samples, rather than merely reflecting closer agreement with sample-specific extremes.

We also agree that the uncertainty in the fitted results should be displayed in Figure 9. For the observed hydrological cases, we plan to use bootstrap resampling to construct pointwise 95% confidence intervals for the fitted quantile or return-level curves, with particular attention to the upper-tail region. The complete parameter-estimation procedure will be repeated for each bootstrap sample so that the intervals reflect the sampling variability of the estimated parameters and upper quantiles. Corresponding intervals will be calculated for comparison methods for which reliable refitting procedures are available. If a method is represented only by published point estimates and cannot be refitted to the bootstrap samples, statistically valid confidence intervals cannot be reconstructed from those estimates alone; this limitation will be stated explicitly.

The Methods section and the caption of Figure 9 will be revised to describe the bootstrap procedure, number of replicates, confidence level, and interpretation

of the intervals. The Discussion will distinguish clearly between in-sample goodness of fit and uncertainty in estimated upper quantiles.

8. Section 4.4: The authors compare their method against other methods not involving the TCEV. However, it's not clear to this referee whether Figures 11 and 12 are an average over all the distributions described in Section 3.2, or only for a certain distribution. There is also a lack of detail when it comes to the parameter settings considered. Overall, I would have preferred a comparison of the different TCEV methods on simulated data, as this would have allowed the authors to properly assess predictive performance of the fitted models. Would the authors be able to run such a comparison, in which case the current Section 4.4 may be replaced.

Re: We thank the reviewer for identifying the lack of clarity in the description of the simulation experiment and for suggesting a more direct comparison among TCEV estimation methods.

Figures 11 and 12 summarize the global and upper-tail RMSE values, respectively, obtained from all simulated datasets described in section 3.2. For each simulation case, the RMSE was calculated independently for each competing method. The resulting RMSE values across all simulation cases were then summarized using boxplots, while the scattered points indicate the RMSE obtained for individual simulation datasets.

We agree that simulations generated directly from known TCEV distributions would provide a more appropriate assessment of the proposed TCEV parameter-estimation method. We therefore plan to replace the current analysis in section 4.4 with a repeated-sample simulation experiment based on TCEV distributions with known generating parameters. Published TCEV parameter estimates available for several of the hydrological cases will be used to construct representative simulation scenarios with different degrees of separation between the ordinary and extraordinary event components. Although these published values are estimates rather than the unknown true parameters of the corresponding hydrological populations, they become known generating values when used to define the simulated distributions.

For each parameter scenario, independent samples will be generated at several fixed sample sizes representative of the observed records. Each simulated sample will be analysed using the proposed SR-MWS procedure and the existing TCEV estimation procedures for which reliable implementations

can be obtained or reproduced. For methods reconstructed from published descriptions, we will verify the implementations, as far as possible, against the parameter estimates reported for the published case studies and document the relevant algorithm settings, initial values, convergence criteria, and any remaining reproducibility limitations. If a method cannot be reproduced reliably because essential implementation details are unavailable, this limitation and the resulting scope of the comparison will be stated explicitly rather than assuming an implementation to be equivalent to the original method.

Because the generating parameters and distributions will be known, the ability of the methods to recover the generating parameters and quantiles will be evaluated independently of the empirical-CDF criterion used to fit SR-MWS. The revised analysis will report parameter bias and RMSE, bias and RMSE for selected upper quantiles or return levels, and numerical convergence or failure rates across repeated samples. The results will be presented separately by parameter scenario and sample size; any aggregated summaries will be clearly identified as secondary results. This design will allow us to assess whether the different TCEV estimation procedures recover the underlying distribution and upper-tail quantiles reliably, rather than evaluating them only by their agreement with the empirical CDF used for fitting.

Accordingly, section 3.2 will be revised to provide the complete simulation design, including the generating parameter sets, sample sizes, number of replicates, random-number generation procedure, and evaluation metrics. The current Figures 11 and 12 will be replaced by figures and tables reporting the results of the TCEV-based simulation comparison.

9. Discussion of results: The authors consider 10 Cases involving real-world data, but they do not discuss whether their conclusions agree with climatological knowledge. For instance, they argue that a TCEV is a suitable model for 8 out of 10 cases. I would like to see some comment on whether other studies also found that extremes at these sites are governed by two processes.

Re: We agree that the original manuscript focused mainly on the statistical fitting results and did not sufficiently relate the individual cases to previous hydrological or climatological knowledge.

Cases 1-6 and Case 9 were obtained from previously published studies. Cases 1-6 were taken from Campos-Aranda (2021), in which the TCEV distribution was applied to annual maximum flood records showing influential outlying

floods or mixed-population characteristics. Case 9 was obtained from De Luca and Napolitano (2023), where the TCEV distribution was also considered in the statistical analysis of extreme precipitation. These previous applications therefore provide a statistical basis for considering the TCEV model for these cases. Our purpose in reanalyzing them was primarily to examine whether SR-MWS could improve TCEV parameter estimation for the same datasets, rather than to claim solely from our slope-ratio diagnostic that two distinct physical processes had been demonstrated at these sites.

In contrast, Cases 7 and 8 were obtained from Clarke (2002), where the records were analysed using generalized linear models based on the Gumbel distribution, indicating that a simpler single-component model may be more appropriate. This is broadly consistent with our diagnostic results, which did not show a pronounced difference between the two fitted segments for these cases. We will therefore present Cases 7 and 8 as comparison cases illustrating the applicability boundary of the TCEV framework rather than as datasets that were simply excluded.

Case 10 is an observed precipitation record from the Alicante weather station and was included as a practical application of the proposed SR-MWS method to an independent dataset. To our knowledge, there is no previous TCEV analysis or explicit event-generating-process classification available for this station. We will therefore avoid inferring two distinct climatological processes solely from the dog-leg appearance of its probability plot.

In the revised manuscript, we will briefly clarify the source of each case and the statistical model used in the corresponding previous study. We will also clarify that the results of this study indicate statistical support for considering the TCEV distribution, but do not by themselves demonstrate the existence of two distinct hydrological or climatological processes. The interpretation of the individual cases and the corresponding conclusions will be revised accordingly.

Minor Comments

1. Line 41: The word “a” is missing before “Generalized Pareto Distribution (GPD)”.

Re: This wording will be corrected in the revised manuscript.

2. Line 106: It’s more common to write “log-likelihood function” than “logarithmic

maximum likelihood function”.

Re: This wording will be corrected in the revised manuscript.

3. Line 108: The text suggests that the second method is to maximize the “negative loglikelihood function”, but that does not align with the general statistical approach. If you were to use the negative log-likelihood function, you would try to minimize it. It also does not appear to me that methods 1 and 2 are really distinctly different, because they both consider the log-likelihood function.

Re: We agree that the description in the original manuscript contains an error regarding the direction of optimization. When the negative log-likelihood is used as the objective function, it should be minimized rather than maximized. Maximizing the log-likelihood and minimizing the negative log-likelihood are mathematically equivalent.

After re-examining the original reference, we also agree that TCEV-IML and TCEV-MOF are not fundamentally different in terms of their statistical estimation principle. Both are intended to obtain maximum-likelihood estimates of the TCEV parameters. Their distinction lies primarily in the numerical procedure: TCEV-IML iteratively solves the parameter-estimation equations derived from the log-likelihood using successive substitution, whereas TCEV-MOF directly minimizes the negative log-likelihood using the Rosenbrock algorithm.

In the revised manuscript, the first paragraph of Sect. 2.2 will state, “The first two, denoted TCEV-IML and TCEV-MOF, are different numerical implementations of maximum-likelihood estimation, whereas the third, TCEV-MCS, combines maximum-likelihood estimation, statistical testing, and Monte Carlo simulation.” We have also replaced the incorrect expression “TCEV-MOF leverages the Rosenbrock algorithm to maximize the maximum likelihood function” with “TCEV-MOF directly minimizes the negative log-likelihood using the Rosenbrock algorithm,” and revised the surrounding text to ensure consistent terminology throughout the manuscript.

4. Line 144: I would suggest to write “where $\alpha > 0$ is the slope parameter and $(-\alpha\beta)$ is the intercept value of the line at the origin.”

Re: This wording will be corrected in the revised manuscript.

5. Table 1: Can the authors provide confidence intervals or standard errors for the standard errors please? This allows the reader to make a full comparison of the methods. Moreover, results for Cases 7 and 8 are also not presented, and no reason for their exclusion is provided, while this is done for Case 10.

Re: We understand the first part of the comment as requesting confidence intervals or standard errors for the TCEV parameter estimates reported in Table 1. We agree that information on parameter uncertainty would allow a more complete comparison among the estimation procedures.

The parameter values currently presented in Table 1 for Cases 1-6 and 9 were taken from previously published studies and transformed, where necessary, into the TCEV parameterization adopted in the present manuscript. They were not re-estimated by us using the original algorithms. The source studies generally reported point estimates but did not provide the corresponding standard errors or confidence intervals. Consequently, statistically valid uncertainty intervals cannot be reconstructed from the published point estimates alone. In the revised Table 1, we will identify these values explicitly as published reference estimates and indicate when uncertainty information was not reported in the original source.

Where the relevant TCEV procedures can be implemented and refitted reliably, we plan to supplement the comparison with bootstrap-based uncertainty estimates. For each bootstrap sample, the complete parameter-estimation procedure will be repeated, and the variability and percentile-based 95% confidence intervals of the four TCEV parameters will be reported. We will distinguish clearly between uncertainty measures calculated from our refitting analysis and information reported in the original publications. If a published method cannot be reproduced reliably because its code or essential implementation details are unavailable, we will not assign an uncertainty interval to its published point estimate and will state this limitation explicitly.

We also agree that the absence of Cases 7 and 8 from Table 1 was not adequately explained. These cases were included to evaluate the TCEV applicability diagnostic. Their slope ratios were below the adopted threshold, indicating that the required two-component structure was not sufficiently supported. Moreover, the source study analysed these datasets using a simpler Gumbel-based generalized linear model and did not provide TCEV parameter estimates. Cases 7 and 8 were therefore excluded from the subsequent TCEV

parameter comparison. We will explain this explicitly in the revised table caption and the corresponding text.

6. Figure 4: It seems that for some of the cases, only around 5 data points are used for Segment 2. How confident should we be in estimates for a Gumbel distribution based on only 5 data points given the high uncertainty in these observations? Further, the authors identified break points, but it's not clear how they are estimated in practice.

Re: We agree that a segment containing only approximately five observations yields highly uncertain estimates of its slope and intercept. In the proposed procedure, however, the two segment-specific linear approximations are used only to construct initial TCEV parameter values and provide preliminary diagnostic information on the possible two-component structure; the final four TCEV parameters are subsequently optimized using the complete dataset. Nevertheless, uncertainty in the upper segment may affect the estimated breakpoint and slope ratio, and we will clarify this limitation.

We will also provide a more detailed description of breakpoint estimation. Specifically, all candidate breakpoints satisfying the minimum number of observations in both segments will be evaluated. For each candidate, two linear regressions will be fitted, and the breakpoint producing the lowest combined RMSE will be selected. We will add a sensitivity and/or bootstrap analysis to examine the stability of the breakpoint and slope ratio when only a small number of upper-segment observations are available. The interpretation of cases with unstable breakpoints or slope ratios will be revised accordingly.

7. Figure 6: The authors comment on the differences in performance for the different weighting schemes, but one may argue that the results in Figure 6 are in line with Figure 2 which shows that the quadratic weighting assigns the most weight to the most extreme observations. As such, the estimation will be pushed more towards matching these values in the quadratic weighting scheme than the other two. Maybe it would be good for the authors to refer the reader back to Figure 2 in their interpretation.

Re: In the revised manuscript, we will expand the discussion of Figure 6 and refer readers back to Figure 2 when interpreting these results.

8. Figure 7: It seems the fit for the highest values is not always great. Could the authors add confidence intervals to the quantile-quantile plot so that the reader can assess the model fit for rigorously?

Re: In the revised manuscript, we plan to supplement Figure 7 with uncertainty intervals for the fitted values, particularly in the upper-tail region. We will investigate an appropriate resampling-based approach, such as bootstrap resampling, while taking into account the limited sample sizes and the uncertainty introduced by parameter estimation and weighting-scheme selection. The selected procedure, including the number of resamples and the interpretation of the resulting intervals, will be described in the Methods section and the figure caption. We will also revise the corresponding discussion so that the upper-tail performance is evaluated together with its estimated uncertainty rather than solely from the proximity of the fitted points to the 1:1 line.

9. Section 4.3: The authors state that their method outperforms the other methods, but Figure 8 suggests that this does not apply for Cases 5 and 6 in terms of the RMSE. One may also argue that this form of performance comparison is skewed towards the proposed method, because a part of the estimation procedure considers three of the intervals for $F(x)$ on which the methods are compared when selecting the weighting method, and the performance in Figure 9 is exactly the one other one which optimizes.

Re: We agree that the statement that SR-MWS consistently outperforms the other methods is too strong. As shown in Figure 8, SR-MWS does not achieve the lowest RMSE in all intervals for Cases 5 and 6. We will revise section 4.3 to describe these case-specific differences accurately and moderate the corresponding conclusions.

We also agree that the current comparison is partly influenced by the design of the SR-MWS procedure. Three of the $F(x)$ intervals used in Figures 8 and 9 are also used internally to select the weighting scheme. Consequently, the resulting RMSE-based score is not an independent validation metric and should not be used alone to demonstrate the superiority of the proposed method. In the revised manuscript, we will clearly distinguish the internal weighting-selection criterion from the external performance evaluation. Figures 8 and 9 will be interpreted as descriptions of in-sample fitting behavior rather than independent evidence of predictive superiority.

To provide a more independent comparison, we will supplement the analysis

with goodness-of-fit measures not directly optimized by SR-MWS and with repeated-sample simulations from TCEV distributions with known parameters. The simulation results will be evaluated using parameter bias and RMSE, upper-quantile or return-level errors, and convergence stability. We will also report the results of all three fixed weighting schemes and an unweighted reference, rather than relying solely on the scheme selected by the partitioned score. These revisions will allow us to evaluate whether the observed reduction in empirical-CDF error is accompanied by more reliable recovery of the underlying TCEV distribution.

10. Figure 9: Please check that the sentence “The diamond symbols represent observed” in the caption is not missing it at least one word.

Re: The caption will be revised as follows:

“The diamond symbols represent the empirical plotting positions of the observed data.”

11. Line 446: The authors state that the paper by Huser and Wadsworth (2022) mentions weighting, but this paper considers a very different framework to the one in this paper. Please elaborate on the connection.

Re: We originally cited Huser and Wadsworth (2022) because the paper emphasizes the need for flexible extreme-value models at finite sample levels, where the behavior of the extreme region may differ from that of the main body of the data. This general perspective is conceptually related to our motivation for emphasizing tail information and reducing the dominance of ordinary observations during parameter estimation.

In the revised manuscript, we will remove this citation from the interpretation of the weighting results. If the reference is retained, it will be used only as general background on flexibility in extreme-value modeling, with the difference between the two frameworks stated explicitly.

12. The authors should please check their references. For instance, the Coles (2001) reference in Line 532 is incorrect.

Re: The reference will be corrected in the revised manuscript as follows:

"Coles, S.: *An introduction to statistical modeling of extreme values*, Springer Series in Statistics, Springer, London, UK, 208 pp., <https://doi.org/10.1007/978-1-4471-3675-0>, 2001."

Following the reviewer's suggestion, we will also systematically check the entire reference list.