Response to reviewer 1

An Uncertainty Quantification Framework for Simulation-based Flood Frequency Analysis

We are grateful to the reviewer for carefully reading our manuscript and for the constructive critique of Section 3.6 and Figures 2–3. We agree that our presentation can be improved to align with standard practice in flood frequency analysis and to avoid any possible misinterpretation. Below we detail the changes we will make (added text, new figures, and additional analyses), and how these address each comment. Please see below, the notation, for a point-by-point response to the reviewers' comments and concerns. All page numbers refer to the revised manuscript file with tracked changes.

Notation:

Bold font: referee's comment.

Black font: authors' comments.

Comment 1 — Section 3.6 (methodology and terminology)

Reviewer's comment:

"I do not agree with the methodology presented by the authors. On page 10, line 296, they write To account for uncertainty in plotting positions used in flood frequency analysis, we applied a probabilistic approach based on fitting parametric distributions to annual maximum flow (AMF) data. This method enables the estimation of confidence intervals around plotting positions, which are critical for interpreting the reliability of estimated quantiles".

Serinaldi (2009) compared several methods for computing confidence intervals for extreme quantiles (and NOT for order statistics), and recommended a method using fractional order statistics. He is thus able to compute a confidence interval for extreme values exceeding the highest value in the sample (see figures 2 and 4 of Serinaldi, 2009)."

Response:

We appreciate the reviewer's careful reading and agree that our wording in Section 3.6 caused confusion. Our intention was not to propose confidence intervals "around plotting positions".

Rather, we implemented an analytical method to compute confidence intervals for discharge quantiles (i.e., vertical uncertainty), which is the quantity of direct interest for design and risk assessment.

In addressing this comment, we realized our derivation of this analytical method was equivalent to that of Baâ et al. (2001) and hence now properly cite them for this analytical method. We selected Baâ et al. (2001) because their method (i) provides a closed-form solution that avoids the computational burden and implementation choices of resampling; (ii) focuses on quantifying the uncertainty of the discharge quantile conditional on the available information and the chosen statistical representation; and (iii) is straightforward and transparent to apply in our multi-basin comparison. In the original text, our shorthand reference to "plotting-position confidence intervals" was imprecise—we will correct this throughout.

Regarding Serinaldi (2009) and fractional order statistics, we fully acknowledge their value for constructing confidence intervals for extreme quantiles, including beyond the sample maximum. However, in this paper our primary objective is to compare the UQ-flood simulation framework against a GEV fit in terms of uncertainty in discharge quantiles relevant for design (e.g., the regulated 100-year event). Adding a second, conceptually different CI construction (fractional order statistics) would duplicate comparisons to our model-based predictive distributions, complicate the presentation, and move us away from the paper's central goal of contrasting model-based quantile uncertainty. We therefore propose not to implement fractional order statistics here; instead, we will:

1. explicitly clarify in Section 3.6 that our intervals are quantile CIs (vertical), not order-statistic AEP bands (horizontal); and
2. add a short paragraph in the Discussion acknowledging fractional order statistics as a valuable alternative for extreme-quantile CIs and citing Serinaldi (2009) as a recommended approach for future sensitivity analyses.

Finally, we agree with the reviewer's broader point that uncertainty in the frequency (AEP) of a given observed discharge is conceptually distinct from uncertainty in the discharge associated with a given return period. Because engineering decisions typically require estimates of discharge from a fixed return period rather than estimates of the return period given a discharge, our analysis prioritizes uncertainty in discharge (vertical intervals). We will make this distinction more explicit in both the text and the captions to prevent misinterpretation.

Manuscript changes:

- Section title: Rename Section 3.6 from "Plotting Position Confidence Interval" to "Confidence Intervals for Discharge Quantiles".
- Replace any mention of "confidence intervals around plotting positions" with precise language: "confidence intervals for discharge quantiles (vertical intervals)".

- Add two sentences distinguishing order-statistic AEP dispersion (horizontal) from quantile CIs (vertical) and state that we focus on the latter.
- Properly cite and briefly summarize the Ba et al. (2001) analytical method and its rationale for this study's objectives.
- Add a sentence in the Discussion acknowledging Serinaldi (2009) and explaining why fractional order statistics are not implemented in the present comparison.

Reference:

Baâ, K.M., Díaz-Delgado, C. & Caârsteanu, A. Confidence Intervals of Quantiles in Hydrology Computed by an Analytical Method. Natural Hazards 24, 1–12 (2001). https://doi.org/10.1023/A:1011101700032

Serinaldi, F. (2009). Assessing the applicability of fractional order statistics for computing confidence intervals for extreme quantiles. Journal of Hydrology, 376(3–4), 528–541. https://doi.org/10.1016/j.jhydrol.2009.07.065

Comment 2 — Standard presentation with the experimental (empirical) distribution

Reviewer's comment:

"The standard method for analysing flood distribution involves plotting the experimental distribution as a reference, using observed annual maximum values and representing plotting probability position based on order statistics, as explained by Cunnane (1978) https://doi.org/10.1016/0022-1694(78)90017-3). This experimental distribution is then compared with a fitted distribution (Log Normal, Gamma, Gumbel, GEV…) or a simulated distribution. Using the SCHADEX simulation method, Paquet et al. (2013) compare the distribution of simulated daily or peak discharges against the observed CDF (Figures 11 and 13). Using the SHYPRE simulation method, Arnaud and Lavabre (2002) compare observed and simulated AMAX peak discharge (Figure 4). Using the SHYREG simulation method, Arnaud et al. (2017) present the median value and confidence interval of flood quantiles, estimated using the SHYREG method or Gumbel or GEV distributions, as well as the experimental distribution (Figure 4)."

Response:

We thank the reviewer for this valuable reminder and agree entirely. The standard practice is to use the experimental (empirical) AMAX distribution, plotted with plotting positions based on

order statistics, as the reference, and then overlay fitted or simulated distributions for comparison (e.g., Cunnane, 1978; implementations in SCHADEX, SHYPRE, and SHYREG). In our initial figures we emphasized a small set of design return periods (e.g., 2, 5, 10, 20, 50, 100 years) for engineering context, which regrettably obscured the full empirical distribution and may have limited the reader's ability to appraise model fit across the entire range.

We will therefore revise Figures 2 and 3 to follow the reviewer's recommendation and align with the cited literature. Specifically, we will:

1. Plot the full experimental AMAX distribution using plotting positions for all observed years (all sample points shown).
2. Overlay the fitted GEV distribution with 95% confidence bands (Bayesian approach), and
3. Overlay the UQ-flood simulated predictive distribution with its 95% simulation envelope.
4. Clearly distinguish AMF_sim (simulation chain without residual-error model, using black dots) from UQ-flood (simulation convolved with the residual-error model), so the reader can assess both simulation skill and predictive uncertainty.

These changes will bring our figures in line with standard presentations in simulation-based flood frequency analysis (e.g., Paquet et al., 2013; Arnaud & Lavabre, 2002; Arnaud et al., 2017) and the foundational plotting-position guidance of Cunnane (1978).

We also note here that our approach is explicitly non-parametric: we intentionally avoid assuming a fitted distribution.

Comment 3 — Interpreting Figures 2–3; AMFpp vs. AMFsim vs. UQ-flood

Reviewer's comment:

"*In this paper, Figures 2 and 3 are misleading. One would expect having to see a good agreement between AMFpp (in red) with AMFsim (in blue), which is not the case. The term "plotting-position confidence interval" is inappropriate, as one would expect a confidence interval for a plotting position to provide a probability interval. If the authors wish to highlight that a plotting position involves uncertainty, as expressed by order statistics, they should present a horizontal interval expressed as annual exceedance probability. This is linked to the fact that the maximum value of a AMAX sample over N years may have a return period greater than N years (it sometimes happens that a 100-year flood occurs during a 20-years observation period). Attached is an example of a confidence interval for a plotting position. It was computed from the simulation of 1000 samples of 20, 50 or 100 years of record. It can be seen that the length of the confidence interval is greater with the largest values of the sample, and that is greater when the sample size is shorter."*

Response:

We appreciate the reviewer's careful reading and agree that our figures and legend did not clearly convey the role of the residual-error model and the intended interpretation of the curves. We address the points as follows:

1. "What is the difference between AMFpp (empirical) vs. AMF_sim (simulation chain)?":

- AMFpp is the experimental (empirical) AMAX distribution constructed from the observed annual maxima using plotting positions. It is the visual reference for model comparison.
- AMF_sim is the raw output of the weather-generator + hydrologic model chain (no residual-error model). While a well-calibrated simulation chain can reproduce the observed AMAX distribution reasonably well, discrepancies between simulated and observed flows can persist due to structural limits, parameter equifinality, input uncertainty, and calibration choices (e.g., daily-flow calibration prioritizing hydrograph volume/timing over peak magnitude), which are well-recognized sources of bias for high flows. Continuous simulation modeling (weather-generator + hydrologic model) demonstrates how simulation can match empirical distributions when the chain is tuned appropriately but also highlight the value of post-processing when systematic differences remain.

2. ***What is the difference between AMFsim (in blue) and UQ-flood (in blue)? I understand that AMFsim represents the output of the simulation chain, without an error model, whereas UQ-flood includes the error model. In general, the simulation chain allows results to be obtained that are consistent with the experimental distribution of AMAX values, provided that its components are properly calibrated (several examples in Paquet et al., 2013; Arnaud and Lavabre, 2002; Arnaud et al., 2017):***

- UQ-flood represents the predictive distribution obtained by integrating a residual-error model (REM) with AMF_sim. The REM is designed to correct systematic bias and inflate/deflate dispersion so that predictive intervals achieve realistic coverage for high flows—an approach consistent with established post-processing frameworks in hydrology (e.g., Hydrologic Uncertainty Processor, Bayesian post-processing, kernel dressing, and modern ML-based bias correction). These frameworks explicitly model errors from the simulation chain to improve predictions, reduce bias, and yield better uncertainty quantification.
- Empirical support and novelty in our context. Numerous studies show that post-processing/bias-correction reduces flow bias and improves predictive reliability when raw hydrologic simulations under- or over-shoot extremes—even after careful calibration (Hunter et al., 2021; McInerney et al., 2017, 2018, 2020; Romero-Cuellar et al., 2024).

These findings hold for both physics-based and large-domain models, and in both forecasting and simulation contexts. While residual-error model approaches are typically applied to daily streamflow, here we extend the residual-error model to extremes—specifically AMAX—within our continuous-simulation FFA framework, to align simulated peaks with observed flood behavior and to provide calibrated predictive intervals for design-relevant quantiles.

3. Why AMF_sim may not align with AMFpp even with "decent" calibration:

- Calibrating on daily flows can under-represent peak magnitudes (objective functions emphasize hydrograph fit more than tail behavior); precipitation/routing errors and rating-curve uncertainty also propagate into differences between simulated and observed peaks. These limitations are widely documented; hence, post-processing the simulation output is now common practice to achieve unbiased, reliable predictive distributions for decision-relevant quantiles.

4. What the revised figures will show (to avoid any misleading impression):

- AMFpp (empirical points): all observed AMAX with plotting positions (reference curve).
- AMF_sim (simulation chain only): shown distinctly (e.g., black dots) to emphasize it is the un-corrected simulation.
- UQ-flood (predictive distribution with REM): solid colored curve (e.g., navy) with a 95% predictive envelope to show bias correction and calibrated dispersion.
- GEV fit: fitted curve with 95% CI (method stated), to allow side-by-side comparison with UQ-flood and the empirical AMAX cloud, consistent with SCHADEX/SHYPRE/SHYREG figure conventions.
- We will make AMF_sim vs. UQ-flood differences explicit in the legend.

We believe these clarifications and changes will make Figures 2–3 transparent and informative: the empirical AMAX remains the reference, AMF_sim shows the raw model behavior, and UQ-flood demonstrates how residual-error modeling corrects bias and calibrates uncertainty.

Comment 4 — "Plotting-position confidence interval" and horizontal AEP bands

Reviewer comment: ***The term "plotting-position confidence interval" is inappropriate, as one would expect a confidence interval for a plotting position to provide a probability interval. If the authors wish to highlight that a plotting position involves uncertainty, as expressed by order statistics, they should present a horizontal interval expressed as annual exceedance probability. This is linked to the fact that the maximum value of a AMAX sample over N years may have a return period greater than N years (it sometimes happens that a 100-year flood occurs during a 20-years observation period). Attached is an example of a confidence interval for a plotting position. It was computed from the simulation of 1000 samples of 20, 50 or 100 years of record. It can be seen that the length of the confidence interval is greater with the largest values of the sample, and that is greater when the sample size is shorter.***

Response:

Thank you for highlighting this important distinction. We have clarified the terminology in our response to Comment 1 (Section 3.6) to avoid any confusion between (i) uncertainty in order-statistic plotting positions (probability/AEP space) and (ii) uncertainty in discharge quantiles (magnitude space). In this revision we:

- Do not introduce horizontal AEP intervals on the empirical points.
- Remove the ambiguous phrase "plotting-position confidence interval" and explicitly state that our uncertainty displays are vertical confidence/prediction intervals for discharge quantiles (GEV and UQ-flood) determined as a byproduct of incomplete sampling from an (unknown) distribution, which are the decision-relevant quantities for design and risk assessment.
- Retain the full empirical AMAX distribution (with standard plotting positions) as the reference curve and compare it to the fitted GEV and UQ-flood predictive distributions and their vertical uncertainty bands.

This keeps the focus on the paper's primary objective—quantifying uncertainty in design-relevant flood magnitudes—while preventing graphical clutter or a mixing of two distinct types of uncertainty on a single axis. We recognize the reviewer's valid point about the sampling variability of plotting positions (e.g., the sample maximum possibly corresponding to a return period $T > N$); this clarification will be incorporated textually in Section 3.6 without adding horizontal AEP whiskers to the figures.

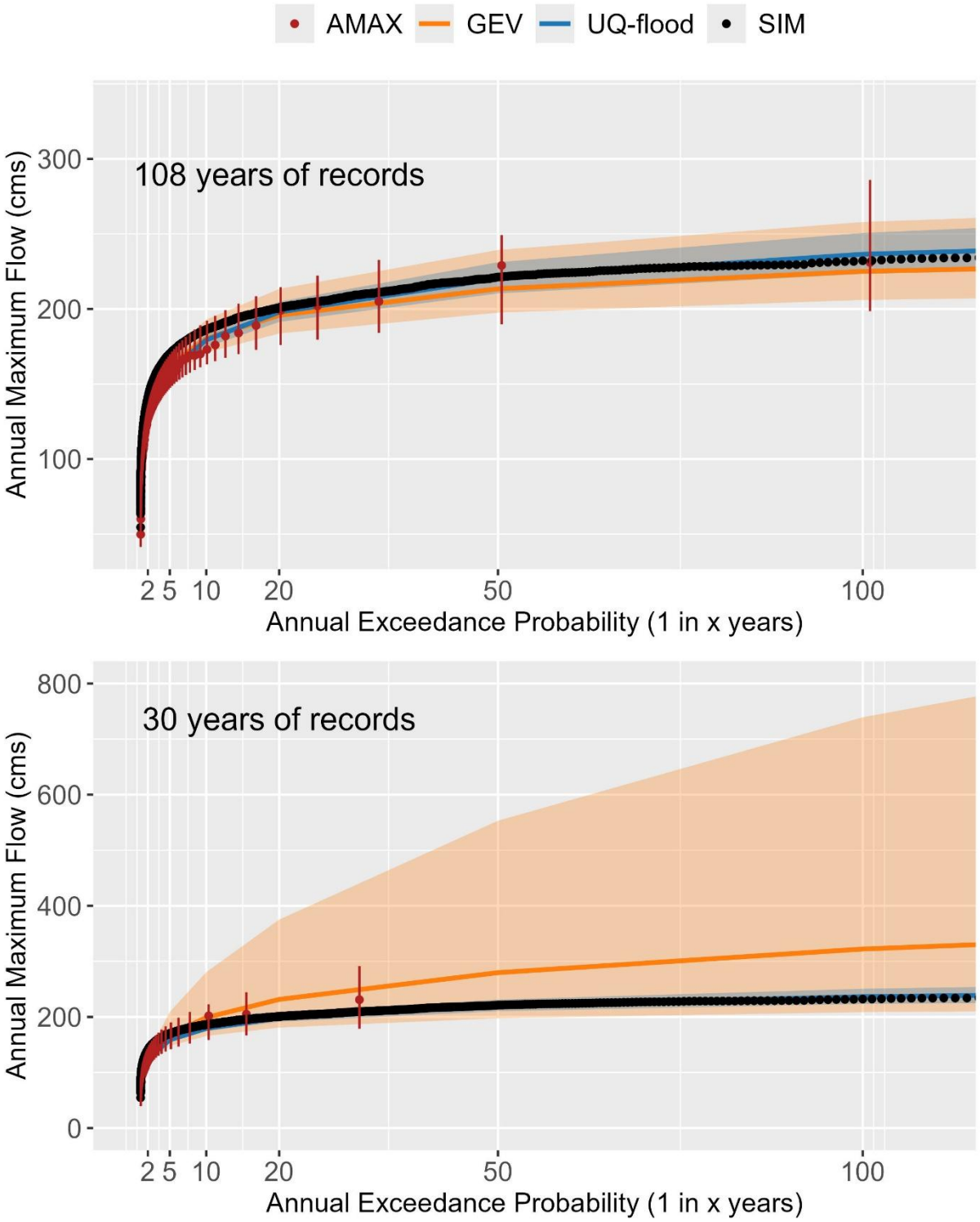
Comment 5 — Experimental (empirical) AMAX distribution

Reviewer comment: ***AMFpp (in red) is supposed to represent the plotting position of the AMAX values, but we have only 6 points. Authors should plot the experimental AMAX distribution of the three watersheds (with series length of 100, 90 and 108 years). We wish to compare this experimental distribution with the GEV distribution and the UQ-flood distribution.***

Response:

We thank the reviewer for this helpful suggestion. As noted in our response to Comment 2 — Standard presentation with the experimental (empirical) distribution, we agree and will revise the figures accordingly. Specifically, we will plot the full experimental AMAX distribution for each watershed—i.e., all observed AMAX values (100, 90, and 108 years, respectively) with their plotting positions—rather than the current subset of six points. We will then compare these empirical curves directly against both the GEV fit and the UQ-flood predictive distribution.

Here, there is an example of how the new figures will look like:


AMAX
GEV
UQ-flood
SIM
108 years of records
Annual Maximum Flow (cms)
300
200
100
2 5 10 20 50 100
Annual Exceedance Probability (1 in x years)
30 years of records
Annual Maximum Flow (cms)
800
600
400
200
0
2 5 10 20 50 100
Annual Exceedance Probability (1 in x years)

Comment 6 — GEV curve above the entire experimental distribution

Reviewer comment: ***In figure 2 (bottom) it is difficult to understand why the GEV distribution (in yellow) lies above the ENTIRE experimental distribution (in red).***

Response:

Thank you for raising this important point and for giving us the opportunity to clarify the behavior shown in Figure 2 (bottom). The bottom panel was constructed using a relatively short subsample of the record (approximately 30 years). For such small samples, estimation of the GEV parameters—particularly the shape parameter—is known to be highly uncertain and unstable.

With limited sample sizes, common estimation approaches (including maximum likelihood estimation and typical Bayesian specifications) can produce biased GEV parameter estimates, often resulting in inflated upper-tail quantiles relative to the empirical plotting positions. This behavior does not necessarily indicate an implementation error but rather reflects the inherent statistical limitations of fitting extreme value models to short records. Small samples can lead to a poorly behaved likelihood surface, yielding parameter estimates with high variance and bias. Martins and Stedinger (2000) demonstrate through simulation that the GEV shape parameter is especially sensitive to record length and can be substantially biased when sample sizes are small.

To verify that the elevated GEV curve is not due to a coding or methodological error, we repeated the analysis using an alternative parameter estimation technique based on L-moments. Table 1 compares the observed empirical quantiles with GEV estimates obtained using both Bayesian (MCMC) and L-moment methods. As shown, the L-moment-based GEV fit also lies above the empirical distribution across the range of observed values, consistent with the behavior seen in Figure 2 (bottom). This agreement between two independent estimation methods supports our interpretation that the discrepancy arises from small-sample effects rather than incorrect model implementation.

Table 1. Comparison of empirical and GEV-based annual maximum flow estimates for selected annual exceedance probabilities. Comparison of observed annual maximum flows (AMF_obs) with quantile estimates from the generalized extreme value (GEV) distribution fitted using Bayesian inference (AMF_Bay) and the L-moment method (AMF_Lmom) for a short-record subsample (~30 years). Across selected annual exceedance probabilities (AEP), both GEV estimation approaches produce quantile estimates that exceed the empirical values, illustrating the upward bias commonly associated with GEV parameter estimation—particularly the shape parameter—under limited sample sizes.

AEP	AMF_obs	AMF_Bay	AMF_Lmom
2	128.50	124.00	122.85
5	155.39	167.74	160.88
10	176.55	198.91	186.19
20	194.06	231.53	210.56
50	223.06	279.69	242.24
100	230.67	322.19	266.08

To make the comparison clearer, we will:

- State explicitly in the caption that the bottom panel is a 30-year short-record experiment, whereas the top panel uses the full record, to highlight the expected degradation of parametric fits under limited data.

Comment 7 — The vertical confidence interval

Reviewer comment: ***The vertical confidence interval (in red) on plotting position is misleading, as it can be interpreted as an uncertainty interval for the flood peak (which is estimated from the maximum stage and converted to discharge using a rating curve). It could be removed. We are more interested in comparing the confidence intervals of the GEV distribution and the UQ-flood distribution.***

Response:

Thank you for highlighting the potential for misinterpretation. We agree that the figures must make a clear distinction between uncertainties in probability space (plotting positions) and uncertainties in magnitude space (discharge quantiles). However, we will retain the vertical intervals because they are not intended as "plotting-position confidence intervals." Rather, they are the plotted results of an analytical method for confidence intervals of discharge quantiles computed directly from the AMAX sample, following Ba, Díaz-Delgado & Cârsteanu (2001). These intervals provide a non-parametric data-driven benchmark of quantile uncertainty attributable to sample size and tail behavior, without relying on normal-approximation heuristics or a fitted distribution and are complementary to the model-based intervals from the GEV fit and the UQ-flood predictive distribution.

To ensure clarity and avoid the specific confusion the reviewer notes, we will:

1. Explicitly re-label the vertical bars as "Confidence intervals for discharge quantiles" (Ba et al., 2001) in the captions and legend, and state that they are not measurement (rating-curve) uncertainty nor "plotting-position CIs."
2. Describe them in Section 3.6 as vertical (magnitude-space) intervals for specified return periods derived from the AMAX sample via the Ba et al. (2001) analytical formulation, to be used as a reference against which to compare the GEV and UQ-flood uncertainty bands.
3. Keep the focus on model comparisons by plotting, on the same axes, the GEV curve with its 95% CI and the UQ-flood predictive distribution with its 95% band, and by clearly distinguishing these model-based envelopes from the analytical quantile CIs (style and color).
4. Clarify the role of plotting positions in the captions: the empirical AMAX points use a standard plotting-position formula, but uncertainty in plotting positions (probability/AEP) is not displayed; instead, we present quantile CIs (vertical), which are the decision-relevant quantities for design.

Why retain these vertical quantile CIs?

They provide a model-agnostic, empirical baseline for the uncertainty in discharge magnitudes at design return periods, derived directly from the observed AMAX sample via an analytical formulation (Bâ et al., 2001). Presenting this baseline alongside the GEV and UQ-flood intervals helps readers judge whether model-based uncertainty is over- or under-dispersed relative to what the sample alone would suggest—particularly important in the upper tail, where normal-approximation methods are known to understate uncertainty and where small-sample effects are consequential.

Minor comments — Authors' response

- ***Page 3, line 70 (add SCHADEX citation and note its international applications).***

Action: We will add Paquet et al. (2013) and note that SCHADEX has been applied internationally, consistent with the description in the paper.

- ***Page 4, Section 2 (add a location map).***

Action: We will include a map showing the locations of the three study watersheds.

- ***Page 5, line 142 (wording fix).***

Action: We will correct the phrase to: "upstream of the Water Survey of Canada gauge".

- ***Page 13, line 387 (missing reference).***

Action: We will add the full reference for Renard et al. (2013), Water Resources Research, 49:825–843, doi:10.1002/wrcr.20087.

- ***Page 14, Figure 2 (plot all empirical AMAX points).***

Action: We will plot all plotting positions for the observed Annual Maximum Series in both panels (top: 108 values; bottom: 30 values), as requested, following recent presentation examples (e.g., Lucas et al., 2024, see Figs. 2 and 8).

References

Bâ, K. M., Díaz-Delgado, C., & Cârsteanu, A. (2001). Confidence intervals of quantiles in hydrology computed by an analytical method. *Natural Hazards*, *24*(1), 1–12. https://doi.org/10.1023/A:1011101700032

Cunnane, C. (1978). Unbiased plotting positions — A review. *Journal of Hydrology*, *37*(3–4), 205–222. https://doi.org/10.1016/0022-1694(78)90017-3

Hunter, J., Thyer, M., McInerney, D., & Kavetski, D. (2021). Achieving high-quality probabilistic predictions from hydrological models calibrated with a wide range of objective functions. *Journal of Hydrology*, *603*, 126578. https://doi.org/10.1016/J.JHYDROL.2021.126578

Martins, E. S., & Stedinger, J. R. (2000). Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research*, *36*(3), 737–744. https://doi.org/10.1029/1999WR900330

McInerney, D., Thyer, M., Kavetski, D., Bennett, B., Lerat, J., Gibbs, M., & Kuczera, G. (2018). A simplified approach to produce probabilistic hydrological model predictions. *Environmental Modelling & Software*, *109*, 306–314. https://doi.org/10.1016/J.ENVSOFT.2018.07.001

McInerney, D., Thyer, M., Kavetski, D., Laugesen, R., Tuteja, N., & Kuczera, G. (2020). Multi-temporal Hydrological Residual Error Modeling for Seamless Subseasonal Streamflow Forecasting. *Water Resources Research*, *56*(11), e2019WR026979. https://doi.org/10.1029/2019WR026979

McInerney, D., Thyer, M., Kavetski, D., Lerat, J., & Kuczera, G. (2017). Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. *Water Resources Research*, *53*(3), 2199–2239. https://doi.org/10.1002/2016WR019168

Romero-Cuellar, J., Arabzadeh, R., Craig, J. R., Tolson, B. A., & Mai, J. (2024). A multi-model evaluation of probabilistic streamflow predictions via residual error modelling. *Journal of Hydrology*, *635*, 131152. https://doi.org/10.1016/J.JHYDROL.2024.131152

Serinaldi, F. (2009). Assessing the applicability of fractional order statistics for computing confidence intervals for extreme quantiles. *Journal of Hydrology*, *376*(3–4), 528–541. https://doi.org/10.1016/J.JHYDROL.2009.07.065