Response to reviewers

An Uncertainty Quantification Framework for Simulation-based Flood Frequency Analysis

We thank the reviewers for their thoughtful and constructive comments, which have significantly improved the quality of this manuscript. Below we address each comment in detail. These changes are highlighted within the manuscript. 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.

Reviewer 1 (Dr. Ricardo Mantilla)

Comment 1 — HBV calibration length in the 30-vs-108-year comparison

Reviewer request:

Clarify whether the underlying HBV model was calibrated with only 30 years of data when comparing 30 vs. 108 years.

Response:

Yes—in both experiments the HBV model was calibrated using only a 30-year subset of daily streamflow, and that single 30-year calibration was then held fixed while we contrasted short-record (30-yr) and long-record (108-yr) estimation settings for AMAX frequency analysis. We will state this explicitly in the Methods to avoid ambiguity.

Specifically, we added the following text:

"Across all experiments, HBV calibration used only a 30-year daily streamflow subset. The same calibrated parameters were used when comparing 30-year vs. 108-year estimation settings."

The revision can be found on page 12, lines 321 – 323.

Comment 2 — Sufficiency of the AR residual-error component

Reviewer request:

Was the AR component added based on diagnostics or subjectively?

Response:

We thank the reviewer for raising this important point. The structure of the residual error model (REM) was determined based on diagnostic analysis of the raw residuals (defined as observation minus simulation), rather than subjective choice.

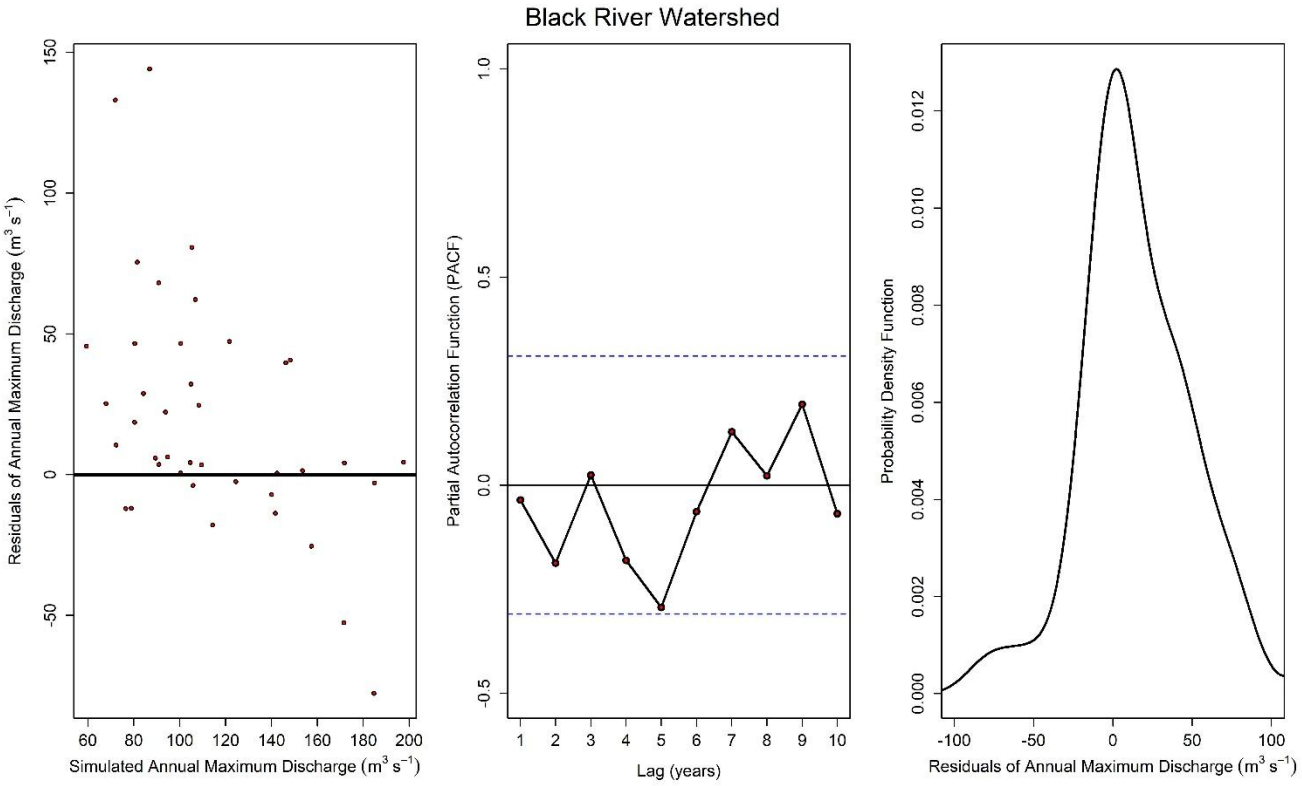
Specifically, we performed the following diagnostic assessments:

- Examined the variance behavior using time-series plots and segment-wise summaries to evaluate potential heteroscedasticity;
- Computed sample partial autocorrelation functions (PACF) of the annual maxima (AMAX) residuals to assess serial dependence;
- Evaluated the residual distribution using histograms and density plots.

For the AMAX residuals at the study sites, these diagnostics indicated negligible serial dependence. Consequently, the inclusion of an autoregressive (AR) component is not required to adequately represent dependence in this application.

Nevertheless, we retained a flexible AR(1) structure within the REM framework to ensure applicability to other sites or use cases where residual autocorrelation may be non-negligible. In our case, the estimated AR coefficient was close to zero and did not have a material effect on the results.

An illustrative example from the Black River case study is provided:



We have clarified this point in the manuscript by adding the following text:

"The selection of the REM structure was guided by diagnostic analyses of the statistical properties of the raw residuals, including variance behaviour, temporal dependence, and distributional characteristics"

The revision can be found on page 12, lines 341 – 342.

Comment 3 — Parameter-uncertainty treatment (bootstrapping/ensembles)

Reviewer request:

Note that unmodeled parameter uncertainty could widen bands; suggest bootstrapping multiple, similarly performing parameterizations.

Response:

We thank the reviewer for this insightful comment. We agree that explicit propagation of parameter uncertainty through multiple behavioral parameterizations could widen predictive uncertainty bands.

In the present study, however, our uncertainty quantification is based on a residual error model (REM) used as a post-processor, which is intended to represent predictive uncertainty in an aggregate sense. Within this framework, discrepancies between simulations and observations are modeled directly, thereby implicitly capturing the combined effects of forcing uncertainty, parameter uncertainty, and structural model error. This approach is consistent with established post-processing paradigms in hydrology, including hydrologic uncertainty processors and related Bayesian or statistical error-modeling frameworks (Li et al., 2016; McInerney et al., 2017).

Under this modeling philosophy, explicit bootstrapping of parameter ensembles is not required for the scope of the present paper, as parameter effects are indirectly reflected in the predictive distribution inferred from simulation–observation residuals.

The revision can be found in Section 3.2 (page 12, lines 332 – 335)

"REM characterises residual uncertainty as the discrepancy between simulations and observations. This residual term implicitly reflects the combined effects of forcing uncertainty, parameter uncertainty, and structural model error, without separating their individual contributions. This formulation is consistent with established error-modelling frameworks in hydrology (Li et al., 2016; McInerney et al., 2017)"

Comment 4 — Rating-curve (stage–discharge) error and streamflow-estimate uncertainty

Reviewer request: Address uncertainty in discharge due to rating-curve error; compare to Velásquez & Krajewski (2024).

Response:

We thank the reviewer for highlighting this important source of uncertainty. We agree that rating-curve and discharge measurement errors can materially affect flood estimates and associated uncertainty bounds.

We now explicitly reference Velásquez & Krajewski (2024), who demonstrate—using Bulletin 17C flood frequency analysis combined with Monte Carlo experiments—that rating-curve uncertainty can substantially alter estimated flood quantiles, in some cases exceeding 50% and approaching 100% for long return periods. Their results clearly show that observational uncertainty can be of comparable or greater magnitude than hydrologic model uncertainty.

In the present study, uncertainty bands are derived from a residual error model (REM) applied as a post-processor to simulated discharge and therefore do not explicitly propagate rating-curve uncertainty. As a result, the reported predictive intervals should be interpreted as conditional on the discharge estimates used in model calibration and evaluation.

We have clarified this limitation in the Discussion and identify explicit treatment of rating-curve uncertainty—e.g., through coupling a stage–discharge error model with the REM or integrating observational uncertainty into the post-processing framework—as a priority direction for future work.

Specifically, we added the following text (page 27, lines 681 – 682):

"This study does not account for observational uncertainty associated with streamflow measurements. However, rating-curve uncertainty can materially alter estimated flood magnitudes, especially for long return periods (Velásquez & Krajewski, 2024)"

Comment 5 — Detecting different flood-generation mechanisms (e.g., snowmelt vs. rainfall) and conditional REM

Reviewer request: Comment on whether different flood mechanisms can be detected and whether a conditional REM may be needed.

Response:

We thank the reviewer for this excellent suggestion and agree that identifying distinct flood-generation mechanisms represents a promising direction for advancing residual error modeling. In principle, different flood mechanisms (e.g., rainfall-driven events, snowmelt, rain-on-snow, or seasonal circulation regimes) could be detected through event classification and used to define mechanism-conditional REM parameters.

Such an extension would align closely with the framework proposed by Fisher et al. (2025), who demonstrated that conditioning flood frequency models on process-based or climate-state information can improve inference and interpretation. Translating this idea to the REM context would require (i) explicit classification of flood events or hydroclimatic states and (ii) estimation of REM parameters conditional on those mechanisms (e.g., season, storm type, or large-scale weather patterns).

While we view this as a compelling and conceptually consistent extension of our framework, implementing a mechanism-aware or conditional REM is beyond the scope of the present study,

which focuses on a controlled comparison between UQ-Flood and GEV-based uncertainty representations. We now explicitly highlight this opportunity as a priority direction for future work.

Specifically, we added the following text:

“An important extension of the proposed framework would be the development of mechanism-aware or conditional residual error models, in which REM parameters vary by flood-generation process or hydroclimatic state, consistent with process-based conditioning approaches such as those proposed by Fisher et al. (2025).”

The revision can be found on page 28, lines 702 – 704.

Minor comments — Authors’ response

- **Line 123**: Use km² (superscript).
- **Line 389**: Replace “providing” with “provides.”

Reviewer 2

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.

Comment 1 — Section 3.6 (methodology and terminology)

Reviewer's comment:

"The authors present a simulation framework known as the 'UQ-flood framework'. This framework is based on a stochastic weather generator, a hydrologic model, and a residual error model for annual maximum discharges (AMAX). It enables the simulation of long series of AMAX values, which are used to estimate extreme floods and the associated uncertainties. A case study involving three long time series in the United States (with 100, 90 and 108 years of record) shows that the flood quantiles estimated by the UQ-flood framework are less uncertain than those estimated by a GEV distribution.

I need detailed explanations regarding Section 3.6 and Figures 2 and 3.

Section 3.6.

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 thank the reviewer for the careful reading and for highlighting this important conceptual distinction. We agree that our wording in Section 3.6 was imprecise and may have led to confusion. Our intent was not to construct confidence intervals "around plotting positions" (i.e., uncertainty

in order statistics or exceedance probabilities). Rather, our objective was to quantify uncertainty in discharge quantiles for specified return periods—that is, vertical uncertainty, which is the quantity most relevant for flood-risk assessment and engineering design.

In revisiting this section, we recognized that the analytical approach we employed to estimate discharge-quantile confidence intervals is equivalent to the method presented by Baâ et al. (2001). We now explicitly cite Baâ et al. (2001) and clarify that their framework provides the basis for our analysis. We selected this approach because it:

(i) yields a closed-form solution, avoiding the computational cost and methodological choices inherent in resampling-based methods;

(ii) directly quantifies uncertainty in the discharge quantile conditional on the assumed statistical model and available data; and

(iii) is transparent and practical for application across multiple basins in a comparative study.

We acknowledge that our earlier reference to "plotting-position confidence intervals" was incorrect and we will correct this terminology throughout the manuscript.

Regarding Serinaldi (2009), we fully agree that fractional order statistics provide a rigorous and valuable approach for estimating confidence intervals for extreme quantiles, including values exceeding the observed sample maximum. However, the primary aim of this paper is to compare uncertainty in discharge quantiles derived from the UQ-Flood simulation framework with those obtained from a GEV-based flood frequency analysis, both of which produce model-based predictive distributions for discharge. Introducing a second, conceptually distinct method for quantile uncertainty estimation based on fractional order statistics would broaden the scope of the study, complicate the comparison, and divert attention from the central objective of contrasting simulation-based versus distribution-based uncertainty representations.

For this reason, we do not implement fractional order statistics in the present analysis. Instead, we will (i) clearly distinguish between uncertainty in exceedance probability (horizontal uncertainty) and uncertainty in discharge for a fixed return period (vertical uncertainty), and (ii) acknowledge fractional order statistics, citing Serinaldi (2009), as a robust alternative approach that could be explored in future sensitivity analyses.

Finally, we agree with the reviewer's broader conceptual point that uncertainty in the annual exceedance probability of a given discharge is fundamentally different from uncertainty in the discharge associated with a prescribed return period. Because engineering and regulatory applications typically require estimates of discharge for fixed design return periods, our analysis prioritizes uncertainty in discharge quantiles. We will make this distinction explicit in both the main text and the figure captions to avoid misinterpretation.

Specifically, we rewrite the section 3.6, the main changes are:

- 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 Baa 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.

The revision can be found on page 14, lines 402 – 439:

“3.6 Confidence Interval for Discharge Quantiles

This study quantifies uncertainty in discharge quantiles associated with specified annual exceedance probabilities (AEPs), rather than uncertainty in plotting positions or order statistics. The focus is therefore on vertical uncertainty—i.e., uncertainty in discharge for a fixed return period—which is the quantity most relevant for flood-risk assessment and engineering design.

To estimate confidence intervals for discharge quantiles, we adopt an analytical approach based on fitting parametric probability distributions to annual maximum flow (AMF) data. Several candidate distributions (e.g., Weibull, exponential, gamma, and lognormal) are fitted using maximum likelihood estimation. Model selection is guided by a combination of visual diagnostics (density plots, Q–Q plots, and cumulative distribution function comparisons) and statistical criteria, including the Akaike Information Criterion (AIC) and goodness-of-fit statistics. The selected distribution provides the basis for estimating uncertainty in discharge quantiles corresponding to target AEPs.

For a given AEP, the corresponding non-exceedance probability is defined as:

$$p = \left(1 - \frac{1}{AEP}\right)(N + 1) \quad (2)$$

Where N is the sample size and AEP is expressed in years. The uncertainty in this probability estimate due to finite sample size is modelled using the beta distribution, consistent with the analytical framework of Baâ et al. (2001). The lower and upper bounds of the 95% confidence interval are computed as:

$$x_{0.05} = qbeta(0.025, p, N - p + 1) \quad (3)$$

$$x_{0.95} = qbeta(0.975, p, N - p + 1)$$

These bounds are then transformed into flow quantiles using the cumulative distribution function (CDF) of the selected parametric model:

$$Q_{0.05} = F^{-1}(x_{0.05}) \quad (4)$$

$$Q_{0.95} = F^{-1}(x_{0.95})$$

where F^{-1} is the inverse CDF (quantile function) of the fitted distribution. This procedure yields confidence intervals for discharge quantiles that explicitly reflect sampling uncertainty conditional on the assumed statistical model. It provides a transparent, closed-form alternative to resampling-based methods and is well suited for comparative analyses across multiple basins. We emphasize that this approach differs conceptually from methods that quantify uncertainty in exceedance probabilities or order statistics (horizontal uncertainty), such as fractional order statistics (e.g., Serinaldi, 2009). While such approaches are valuable—particularly for estimating uncertainty beyond the observed sample maximum—the present study prioritizes uncertainty in discharge for fixed design return periods.”

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 important and well-founded comment and fully agree with the recommended presentation. The experimental (empirical) AMAX distribution, plotted using plotting positions based on order statistics, provides an essential reference for evaluating both

fitted and simulation-based flood distributions (e.g., Cunnane, 1978; Paquet et al., 2013; Arnaud & Lavabre, 2002; Arnaud et al., 2017).

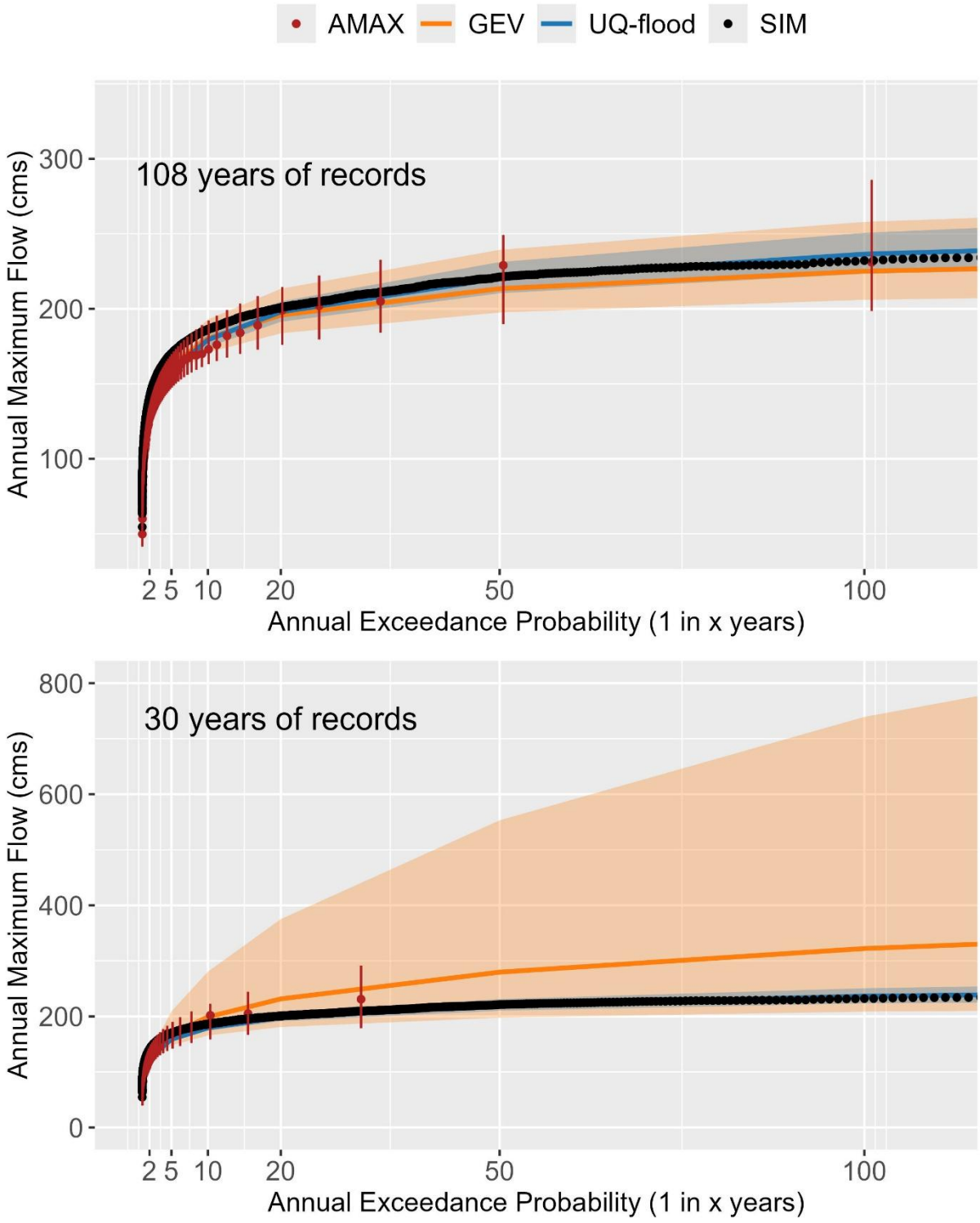
In the original submission, Figures 2 and 3 emphasized a limited set of design return periods (e.g., 2, 5, 10, 20, 50, and 100 years) to maintain an engineering-oriented focus and facilitate visualization. However, we agree that this presentation obscured the full empirical distribution and limited the reader's ability to visually assess model performance across the entire range of observed floods.

We have therefore revised Figures 2 and 3 to follow standard practice and align with the cited literature. Specifically, the revised figures now:

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 SIM (simulation of hydrologic model plus weather generator 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.
5. Figure captions updated to clarify the role of the experimental distribution and the distinction between deterministic simulation and predictive uncertainty.

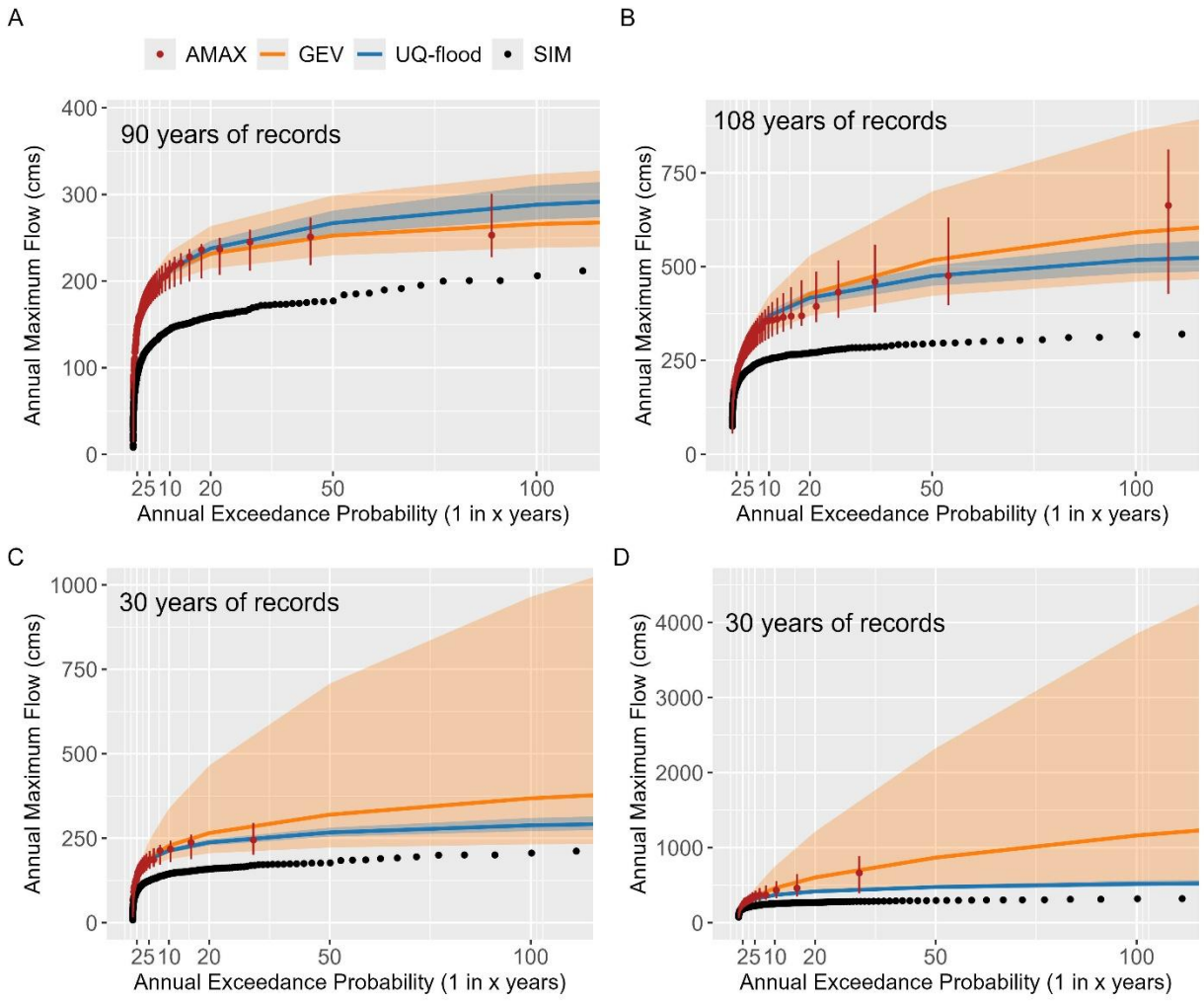
These modifications bring the figures into direct correspondence with established simulation-based flood-frequency presentations used in SCHADEX, SHYPRE, and SHYREG applications, and with the foundational guidance on plotting positions provided by Cunnane (1978).

We note that while the UQ-Flood framework is fundamentally simulation-based and does not rely on fitting a parametric distribution to observed AMAX values, presenting the empirical distribution as a reference is essential for transparency and interpretability. The revised figures now clearly support this comparison without altering the underlying methodology or conclusions of the study.



“Figure 3: Flood frequency estimates and associated 95% uncertainty intervals for the Black River watershed derived using the statistical GEV approach and the UQ-flood framework. The observed

annual maximum discharge (AMAX; red) distribution, shown with confidence intervals for discharge quantiles, and the deterministic simulation output (SIM; black points) are included for reference. The top panel shows GEV results using long-term records (108 years); the bottom panel uses short records (30 years). The UQ-flood framework uses only the 30-year record in both cases."



"Figure 4: Flood frequency estimates and associated 95% uncertainty intervals for the Moyie River (left panels: A and C) and Lahave River (right panels: B and D) derived using the statistical GEV approach and the UQ-flood framework. The observed annual maximum discharge (AMAX; red) distribution, shown with confidence intervals for discharge quantiles, and the deterministic simulation output (SIM; black points) are included for reference. The top panel shows GEV results using long-term records (108 years); the bottom panel uses short records (30 years). The UQ-flood framework uses only the 30-year record in both cases."

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 thank the reviewer for this detailed and constructive comment. We agree that, in the original submission, the figures and legends did not sufficiently clarify the distinct roles of the empirical distribution, the raw simulation output, and the post-processed predictive distribution. We address each point below and describe how the revised figures resolve these issues.

1. Distinction between AMFpp (empirical) and AMFsim (simulation chain)

- AMFpp represents the observed (empirical) AMAX distribution derived from observed annual maxima and plotted using plotting positions based on order statistics. It serves as the primary reference for evaluating model behavior across the full probability range.
- AMFsim is the output of the continuous simulation chain (hydrologic model plus stochastic weather generator) without residual-error post-processing. While a well-calibrated simulation chain can often reproduce the empirical AMAX distribution reasonably well, systematic discrepancies may persist due to structural model limitations, parameter equifinality, forcing uncertainty, and calibration choices (e.g., daily-flow calibration objectives that prioritize hydrograph volume and timing over peak magnitudes). These limitations are well documented in continuous-simulation flood frequency analysis and motivate the use of post-processing (Cameron et al., 2001; Brunner et al., 2021; Grimaldi et al., 2019; Spellman et al., 2018).

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 convolving AMFsim with a residual-error model (REM). The REM is explicitly designed to correct systematic bias and to adjust dispersion such that predictive intervals achieve realistic coverage for high flows.
- This approach is consistent with established hydrologic post-processing frameworks, including hydrologic uncertainty processors. These frameworks explicitly model errors in the simulation chain to improve predictive reliability when raw simulations under- or over-estimate extremes.
- While residual-error modeling has most commonly been applied to daily streamflow, this study extends the approach to annual maxima, enabling bias correction and calibrated uncertainty quantification for design-relevant flood quantiles within a continuous-simulation framework. Empirical evidence supporting the effectiveness of post-processing for reducing bias and improving uncertainty coverage is well established (e.g., McInerney et al., 2017, 2018, 2020; Hunter et al., 2021; Romero-Cuellar et al., 2024).

3. Why AMFsim may not align with AMFpp despite "reasonable" calibration

Even with careful calibration, AMFsim may deviate from the empirical AMAX distribution because:

- calibration is often performed on daily flows, which can under-represent tail behavior;
- precipitation extremes, routing processes, and observational uncertainties propagate into peak-flow errors; and
- continuous-simulation models may exhibit systematic bias at the upper tail.

These factors motivate the use of post-processing to achieve unbiased and reliable predictive distributions for high-return-period events.

Clarification regarding plotting-position uncertainty is provided in Reviewer 2, Comment 1, while revisions to Figures 2–3 to avoid potential misinterpretation are described in Reviewer 2, Comment 2.

Comment 4 — 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 assess the robustness of this result, we re-estimated the GEV parameters using the L-moment method as an independent alternative to the Bayesian (MCMC) framework. Both approaches produced comparable flood quantile estimates and showed the same tendency for the fitted GEV curve to exceed the empirical distribution over the observed data range. The convergence of these independent methods provides strong evidence that the observed discrepancy arises from the limited sample size and the uncertainty of tail estimation, rather than from errors in model formulation or implementation.

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 5 — 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:

We thank the reviewer for highlighting this potential source of confusion and agree that the figures must clearly distinguish between uncertainty in probability space (plotting positions) and uncertainty in magnitude space (discharge quantiles). We also agree that uncertainty in flood-peak measurement (e.g., rating-curve uncertainty) should not be conflated with the uncertainty representations shown in Figures 2–3.

We would like to clarify, however, that the vertical intervals shown in red are not plotting-position confidence intervals, nor are they intended to represent measurement or rating-curve uncertainty. Rather, they correspond to confidence intervals for discharge quantiles derived analytically from the AMAX sample following Baâ et al. (2001). These intervals quantify uncertainty in discharge magnitude at fixed return periods attributable to finite sample size and tail behavior, without assuming a fitted parametric distribution.

We retain these vertical quantile confidence intervals because they provide a model-agnostic, data-driven benchmark against which the model-based uncertainty bands from the GEV fit and the UQ-Flood predictive distribution can be interpreted. In particular, they allow readers to assess whether the uncertainty implied by parametric or simulation-based approaches is over- or under-dispersed relative to what the observed AMAX sample alone would suggest—an issue that is especially relevant in the upper tail, where small-sample effects are pronounced and normal-approximation methods are known to understate uncertainty.

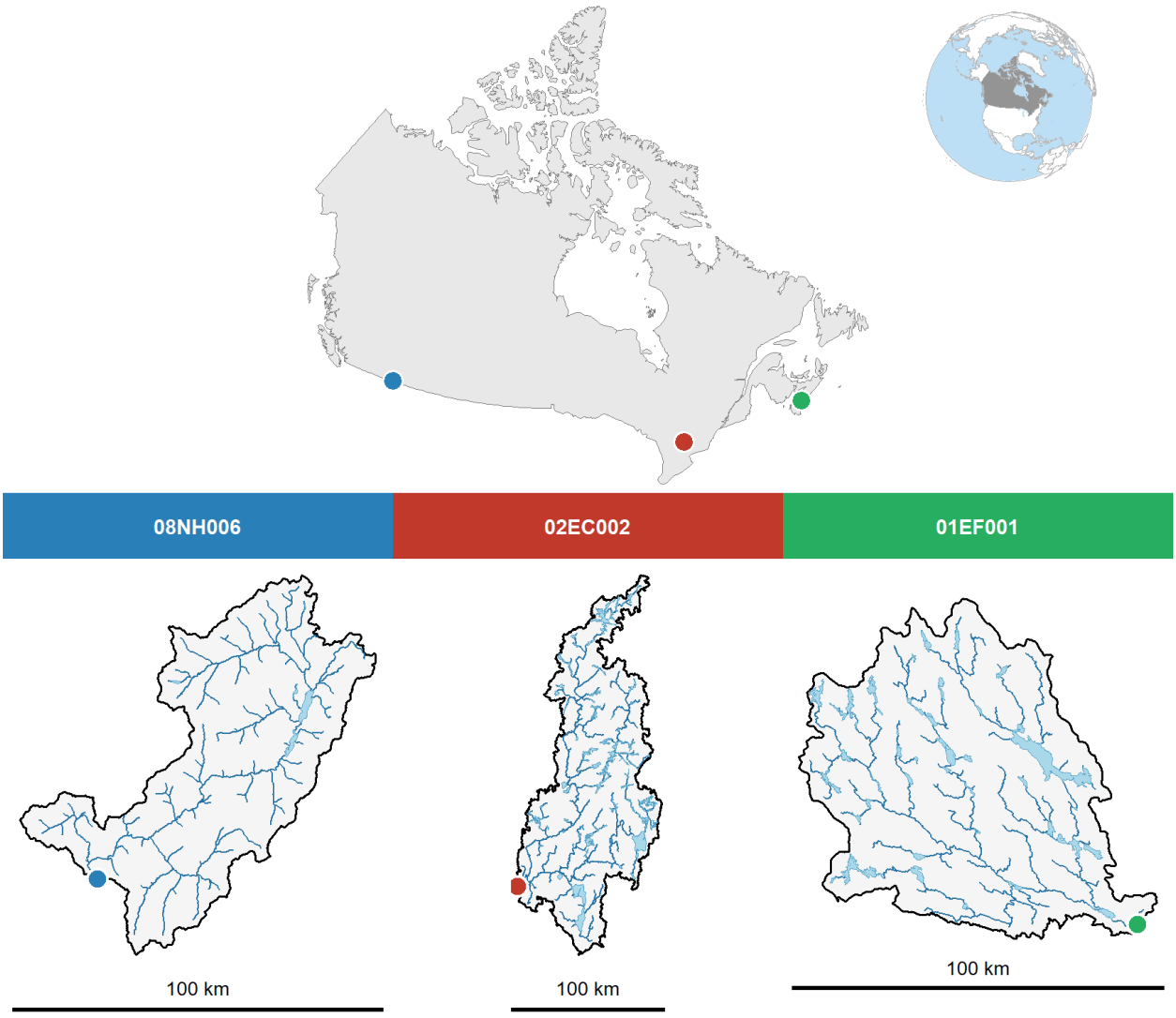
Minor comments — Authors' response

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

Action: We have added Paquet et al. (2013), thank you for the recommendation (Pag 3, line 85).

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

Action: We have included a map showing the locations of the three study watersheds (Pag 9, line 254)



“Figure 1. Geographic location of the three hydrological case studies analyzed in this study. The figure shows the spatial distribution and watershed boundaries of the selected study areas.”

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

Action: We have corrected the phrase to: “upstream of the Water Survey of Canada gauge”.

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

Action: We have added 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 have plotted 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). Please see the comments 2 of the reviewer 2.

We thank the editor for the opportunity to revise our manuscript. We have carefully revised the paper to improve clarity, structure, and the positioning of the contribution. The main modifications are summarized below:

Abstract:

The abstract has been revised to improve precision and clarity. It now more explicitly states the research problem, the proposed UQ-flood framework, and the key findings, with clearer emphasis on the contribution and relevance for flood frequency analysis under data-scarce conditions.

Introduction:

The introduction has been reorganized to provide a clearer and more up-to-date overview of the state of the art. We strengthened the literature review and improved the framing of the research gap. The contribution and novelty of the proposed approach are now more explicitly articulated.

Framing of the contribution:

The manuscript has been revised to better clarify the novelty of the method, particularly the integration of a residual error model (REM) at the annual maxima scale within a continuous simulation framework for flood frequency analysis.

Results:

The presentation of the results has been improved, and additional clarifications have been included to better support the interpretation of the comparisons between methods.

Discussion:

The Discussion section has been expanded to provide additional context and background. In particular, we have reinforced the positioning of the UQ-flood framework within existing approaches for flood frequency analysis. The manuscript now more clearly emphasizes that UQ-flood provides a complementary alternative for addressing short-record limitations, alongside approaches such as regional flood frequency analysis (RFFA) and the use of historical and paleoflood information.

Assumptions and limitations:

We have strengthened the discussion of the assumptions and limitations of the proposed framework, including aspects related to the hydrologic model, stochastic weather generator, and sources of uncertainty not explicitly considered.

Technical revisions:

We conducted a full editorial revision of the manuscript, including:

Standardization of spelling (consistent use of Canadian/British English),

Correction of minor grammatical issues,

Updating and correcting figure numbering and cross-references.

Overall, the manuscript has been revised to improve clarity, consistency, and transparency, and to better position the proposed framework as a process-based, complementary approach for flood frequency analysis under limited data availability.

We believe that these revisions have significantly improved the clarity and quality of the manuscript. We are grateful for the reviewers' thoughtful comments and hope that the revised version meets the journal's standards for publication.

Thank you for your consideration.

Sincerely,

The Authors

References

Bâa, 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

Brunner, M. I., Melsen, L. A., Wood, A. W., Rakovec, O., Mizukami, N., Knoben, W. J. M., and Clark, M. P.: Flood spatial coherence, triggers, and performance in hydrological simulations: large-sample evaluation of four streamflow-calibrated models, Hydrol. Earth Syst. Sci., 25, 105–119, https://doi.org/10.5194/hess-25-105-2021, 2021

Cameron, D., Beven, K., Tawn, J., and Naden, P.: Flood frequency estimation by continuous simulation (with likelihood based uncertainty estimation), Hydrol. Earth Syst. Sci., 4, 23–34, https://doi.org/10.5194/HESS-4-23-2000, 2000.

Coles, S.: An Introduction to Statistical Modeling of Extreme Values. Springer Series in Statistics, Springer-Verlag, Berlin, https://doi.org/10.1007/978-1-4471-3675-0, 2001.

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

Grimaldi, S., Schumann, G. J.-P., Shokri, A., Walker, J. P., & Pauwels, V. R. N.: Challenges, opportunities and pitfalls for global coupled hydrologic-hydraulic modeling of floods, Water Resour. Res., 55, 5277–5300, https://doi.org/10.1029/2018WR024289, 2019.

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

Spellman, P., Webster, V., Watkins, D.: Bias correcting instantaneous peak flows generated using a continuous, semi-distributed hydrologic model, J Flood Risk Management., 11:e12342. https://doi.org/10.1111/jfr3.12342, 2018.