

Response to Reviewers

Manuscript title: An Uncertainty Quantification Framework for Process-based Flood Frequency Analysis

Manuscript number: egusphere-2025-5784

We sincerely thank reviewers for their careful assessment of our manuscript and for their thoughtful and constructive comments. We have revised the manuscript thoroughly in response to the issues raised during the interactive discussion. The comments have helped us better frame the manuscript's contribution and more clearly articulate the implications of our findings.

Below, we provide a point-by-point response to all comments. The original reviewer comments are reproduced in *italic*, followed by our responses in **bold** and a description of the corresponding changes made in the revised manuscript. Page and line numbers in our responses refer to the clean revised manuscript unless otherwise specified. Selected revised or newly added manuscript text is reproduced in boxes for ease of reference.

A marked-up version of the manuscript showing all revisions is also provided with the resubmission.

Response to Reviewer 1

General evaluation:

This manuscript proposes UQ-flood, a framework that combines a process-based hydrologic model, a stochastic weather generator, and a residual error model applied at the annual-maximum scale, and applies it to three Canadian watersheds. The paper is clearly written and well organised, and the central idea is a good one: correcting hydrologic model error at the same scale as the design variable, rather than assuming that a model calibrated on daily flows will reproduce flood peaks without bias. The experiments without the residual error model make that case convincingly, and the finding that the framework's central estimates stay close to long-record benchmarks using only 30 years of data is genuinely useful. My main concern is how the uncertainty bands are constructed and what the comparison against GEV is really showing. I think this is addressable within the existing results, but it does affect the paper's headline claim and should be resolved before publication.

Authors' response:

We sincerely thank the reviewer for the careful, constructive, and technically insightful evaluation of our manuscript. We appreciate the reviewer's positive assessment of the central idea, particularly the application of a residual error model at the annual-maximum scale and the finding that the central UQ-flood estimates remain close to long-record benchmarks when the framework is developed using only 30 years of observations. We agree that the interpretation of the uncertainty intervals is a central issue and that our original manuscript did not distinguish clearly enough between the uncertainty components represented by the UQ-flood and GEV intervals.

The reviewer's comments have helped us identify places where the manuscript overstated the implications of the narrow UQ-flood intervals. In the revised manuscript, we will distinguish more carefully among error model uncertainty, statistical sampling uncertainty, GEV parameter uncertainty, and uncertainty sources that are not propagated through the current framework. We will also revise the manuscript so that the principal contribution is framed around:

- 1. The stability of the central flood-quantile estimates under short-record conditions.**
- 2. The correction of systematic biases in deterministic continuous simulation.**
- 3. The explicit propagation of residual error at the annual-maximum scale.**

We will no longer include language that interprets a narrower UQ-flood interval, by itself, as evidence of universally greater precision or more complete uncertainty characterization.

Major comment 1: Interpretation of the UQ-flood uncertainty intervals

My main concern is with what the narrow UQ-flood uncertainty bands actually represent. As I understand the method, each realization produces 10,000 synthetic years of annual maxima, and flood quantiles are read off empirically using the Weibull plotting position. The reported band is the spread across the 1,000 realizations at a given return period.

The difficulty is that this spread depends heavily on a number the authors chose freely: the length of the synthetic series. With 10,000 years in each realization, every realization contains roughly 100 events at or above the 100-year level, so each one estimates the 100-year quantile very stably. The realizations therefore land close together, and the band is narrow. Had the authors used 200,000 years per realization, the band would have been narrower still; had they used 200 years, it would have been much wider. Nothing about the watershed, the data, or the physics would have changed in either case — only the amount of computing done.

This matters because the paper's headline claim is about width. The Abstract reports "substantially narrower uncertainty bounds," and the Black River example contrasts a $27 \text{ m}^3 \text{ s}^{-1}$ range from UQ-flood with $531 \text{ m}^3 \text{ s}^{-1}$ from GEV. But the GEV interval is wide for a real reason: only 30 annual maxima were observed, and that is genuinely all the information available. The UQ-flood interval is narrow largely because the authors simulated a very long series. These two numbers are not measuring the same thing, so placing them side by side as evidence of improved precision is not, in my view, supported.

It is worth stressing that the long synthetic series does not add information. Everything entering the framework still comes from the same 30 years of record used to fit the weather generator, calibrate HBV-EC, and infer the residual error model. Simulating more years reduces the noise in reading quantiles off the model, but it cannot tell us anything more about the real watershed than those 30 years contain. Presenting that reduction as reduced uncertainty risks giving the reader an unduly confident picture — a real concern for a paper aimed at design and floodplain-mapping practice, where a narrow band would justify smaller safety margins.

The manuscript comes close to acknowledging this in the result section, where the authors note that the two intervals represent different decompositions of uncertainty. But the text then goes on to interpret the narrower band as improved precision, which is the interpretation that caveat should have ruled out.

Authors' response:

We thank the reviewer for this important observation. We agree that the UQ-flood and GEV intervals represent different uncertainty components and should not be compared as equivalent measures of precision. We have therefore removed the direct comparisons of interval width and the associated claims of improved precision. Long-record observations are now the primary benchmark, while GEV is retained as a secondary benchmark for the central flood-quantile estimates.

Following the reviewer's suggestion, we have added the sensitivity analysis to the revised manuscript as Fig. 4 and included the following subsection in the Results (Page 18-19, lines: 476 - 497):

4.1.2 Simulation-length sensitivity analysis

A sensitivity analysis was conducted to examine whether the length of the synthetic streamflow sequence systematically controls the width of the UQ-flood interval. Simulated series of 100, 200, 500, 800, 1000, 2000 and 3,000 years were evaluated for the 50- and 100-year flood quantiles. For each series length, 1,000 residual-error realizations were generated, with each realization containing the specified number of annual maxima. The interval width, calculated as the difference between the 97.5th and 2.5th percentiles of the flood-quantile estimates across the ensemble. Figure 4 shows that the UQ-flood interval width is sensitive to the synthetic-series length, particularly for shorter simulations.

The sensitivity analysis shows that the UQ-flood interval width decreases substantially as the synthetic-series length increases from 100 to approximately 1,000–2,000 years. For both the 50- and 100-year flood quantiles, the interval width declines from approximately $2.9 \text{ m}^3 \text{ s}^{-1}$ at 100 years to approximately $1.7\text{--}1.8 \text{ m}^3 \text{ s}^{-1}$ at 1,000–2,000 years. Beyond 1,000 years, the changes are small, with a slight increase at 3,000 years. The nearly overlapping curves indicate that simulation length affects the 50- and 100-year flood intervals similarly. Overall, the results suggest that shorter synthetic sequences introduce greater finite-sample variability, whereas the flood-quantile intervals become relatively stable at sequence lengths of approximately 1,000 years or more. This stabilization reflects reduced numerical sampling variability conditional on the fitted UQ-flood modelling chain and should not be interpreted as adding observational information or reducing all sources of uncertainty about the watershed.

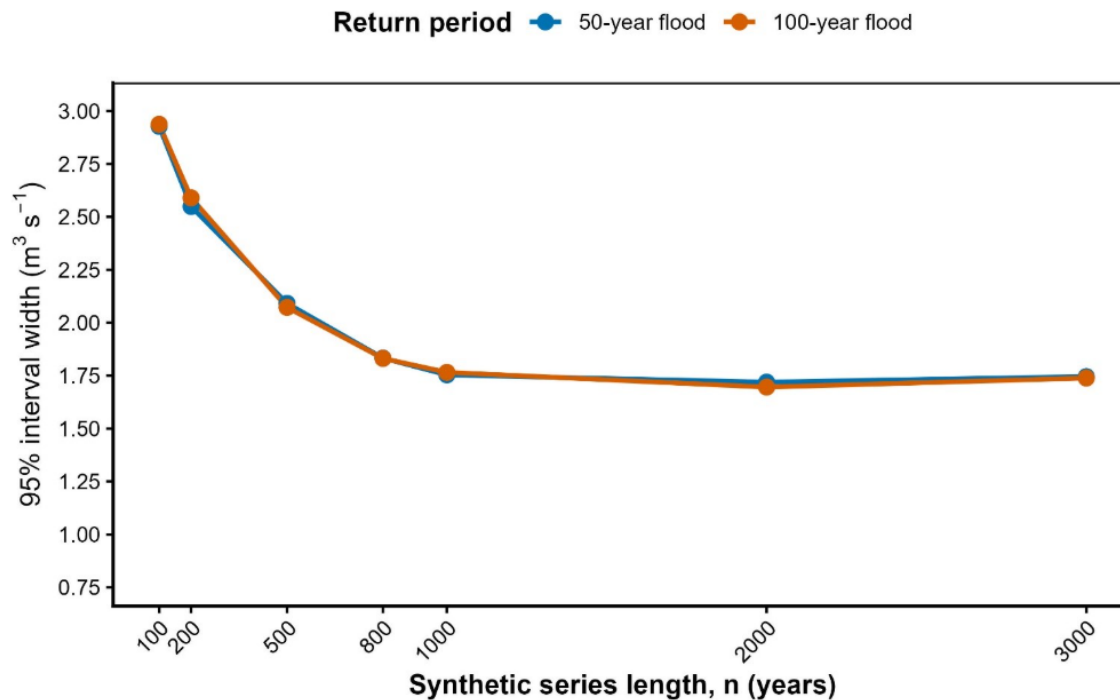


Figure 4. Sensitivity of the UQ-flood 95% interval width to simulated-series length for the 50- and 100-year flood quantiles in the Black River watershed. Interval width is defined as the difference between the 97.5th and 2.5th quantiles across the predictive ensemble.

Suggested revision 1: Simulation-length sensitivity analysis

Repeat the Black River analysis with synthetic series of, say, 200, 500, 1,000, and 10,000 years, and plot the width of the resulting band at the 50- and 100-year return periods against series length. If the width falls steadily as the series gets longer, that is direct evidence that the reported width is set by the simulation design rather than by anything hydrological. This would be a valuable figure either way, and it is cheap to produce with the existing runs.

Authors' response:

We sincerely thank the reviewer for this excellent suggestion. We conducted the requested sensitivity analysis for the Black River using synthetic annual-maximum series of different lengths and evaluated the resulting interval widths for the 50- and 100-year flood estimates. The results and corresponding manuscript revisions are described in our response to Major Comment 1, and the analysis has been added to the manuscript as Fig. 4.

We also identified an error in the original description of the experimental configuration. The reference to 10,000 simulated years was a typographical error. For all three case studies, the stochastic weather generator and hydrologic model produced a sequence of 1,000 annual maxima. For each annual maximum, the residual error model generated 1,000 realizations, resulting in 1,000 REM-corrected annual-maximum sequences, each containing 1,000 years. We corrected this description throughout the manuscript.

Suggested revision 2: Sampling-aware 30-year block analysis

For the short-record comparisons (Fig. 3 bottom; Fig. 4 C and D), divide the 10,000-year synthetic series into non-overlapping 30-year blocks, estimate the flood quantiles from each block separately, and use the spread across blocks as the uncertainty band. Each block then carries the same sampling limitation as the observed 30-year record, and the comparison against GEV becomes like-for-like. I would expect this band to be considerably wider than $27 \text{ m}^3 \text{ s}^{-1}$ — but it would be a defensible number, and the paper's argument does not depend on the band being extremely narrow. It is enough that UQ-flood stays close to the long-record benchmark while GEV drifts away from it, which the results already show clearly.

Authors' response:

We thank the reviewer for this thoughtful suggestion. We agree that a common 30-year sampling constraint could facilitate a more direct comparison between UQ-flood and the short-record GEV analysis. However, we decided not to implement the proposed block analysis because the revised manuscript no longer compares the widths of these intervals as equivalent measures of uncertainty.

The UQ-flood interval represents predictive variability across residual-error realizations, conditional on the fitted modelling chain, whereas the Bayesian GEV interval primarily represents sampling and parameter uncertainty associated with fitting the selected distribution to the observed annual maxima. Applying the same block length would therefore not make the two uncertainty decompositions equivalent.

Instead, we removed all interval-width comparisons and now use long-record observations as the primary benchmark, with GEV retained as a secondary benchmark for the central estimates. Under this revised framing, the proposed block analysis would address a separate question concerning short-sample variability within the fitted UQ-flood system.

Moreover, a 30-year block cannot provide direct empirical estimates of the 50- and 100-year quantiles using Weibull plotting positions. Estimating these quantiles would require an additional parametric or extrapolation model, moving the analysis away from the current empirical framework. We therefore believe that the block analysis would add methodological complexity without strengthening the revised central contribution.

Therefore, while we appreciate the rationale for the reviewer's proposed experiment, we respectfully decided not to include it in the revised manuscript. Once direct comparisons of UQ-flood and GEV interval widths are removed, the 30-year block analysis no longer addresses the principal evaluation question of this study. It would instead introduce a separate sampling-uncertainty experiment and require an additional extrapolation model for estimating 50- and 100-year quantiles from 30-year blocks. We believe that the revised benchmarking strategy, together with the new simulation-length sensitivity analysis, provides a clearer and more internally consistent evaluation of UQ-flood.

Suggested revision 3: Clearly distinguish interval contents and revise performance claims

Wherever interval widths are compared, state plainly what each interval covers and what it leaves out, and avoid describing the difference as greater precision or better performance without that qualification. The claim the results genuinely support is that UQ-flood's central estimates remain close to long-record benchmarks when only 30 years are available — a strong and useful finding on its own.

Authors' response:

We thank the reviewer for this important recommendation. We agree that the UQ-flood and GEV intervals represent different uncertainty components and should not be compared as equivalent measures of overall precision or performance. We have revised the Abstract, Methods, Results, Discussion, and Conclusions to emphasize the findings supported by the analysis: UQ-flood's central estimates remain close to long-record benchmarks when the framework is developed using 30 years of observations, and the residual error model corrects

systematic biases in deterministic process-based flood estimates. We also clarify that UQ-flood intervals remain conditional on the selected modelling chain and should not be interpreted as comprehensive uncertainty bounds.

Minor comment 1: Statement that uncertainty quantification is often neglected

[Line 35] I disagree with the claim that uncertainty quantification in FFA is "often neglected, insufficiently characterised, or not transparently communicated." Standard practice guidance says otherwise: Bulletin 17B and its successor Bulletin 17C both provide confidence limits for flood quantiles, and reporting a confidence interval alongside each return-period estimate is routine in practice, not exceptional. I would ask the authors to soften or reframe this statement. There is a defensible version of the argument — for example, that confidence limits in current practice address sampling uncertainty for a fitted distribution but not model-structure or hydrologic-model error — and that framing would actually strengthen the motivation for the paper rather than overstating the gap.

Authors' response:

We agree that the original wording was too broad and could incorrectly imply that uncertainty intervals are generally absent from conventional flood-frequency practice. That was not our intended meaning.

We have revised the Introduction and Discussion to recognize that conventional statistical FFA commonly reports confidence or credible intervals associated with flood-quantile estimates. We will then identify the more specific gap addressed by the study: conventional at-site intervals generally focus on uncertainty associated with finite flood samples and fitted statistical-distribution parameters, conditional on the selected model. They do not necessarily represent hydrologic-model structural error, meteorological simulation uncertainty, or the mismatch between daily-flow calibration and annual-maximum prediction.

Proposed revised manuscript text in the Introduction (page 2, lines:32-43):

In this study, process-based flood-frequency analysis refers to the generation of synthetic meteorological sequences, their transformation into streamflow through a hydrologic model, and the subsequent estimation of flood quantiles from the resulting annual-maximum series. Uncertainty quantification is an established component of statistical flood-frequency analysis, and conventional guidance includes procedures for reporting confidence limits for estimated flood quantiles (England et al., 2018). These intervals commonly characterize sampling and parameter uncertainty conditional on the selected probability model, although statistical frameworks can also incorporate distributional model uncertainty explicitly (Liang et al., 2012; Merz and Thielen, 2005). In process-based flood-
--

frequency analysis, additional uncertainty arises from the broader hydrometeorological modelling chain, including meteorological inputs and weather-generator parameters, hydrologic-model structure and parameters, antecedent conditions, and errors in simulated flood peaks (Arnaud et al., 2017; Cameron et al., 2000; Newman et al., 2021). These components are not automatically represented by conventional at-site statistical confidence intervals and therefore require explicit evaluation and propagation within the process-based simulation framework.

We have made the same distinction in the Discussion and remove the statement that traditional statistical approaches “often provide point estimates without a formal treatment of uncertainty,” which is also too general in its current form.

The revised text on page 22, lines 544–558, reads:

5.1 Placing the UQ-flood Framework in Context

Uncertainty is an established consideration in flood-frequency analysis, and conventional statistical methods provide confidence or credible intervals for estimated flood quantiles (England et al., 2018). These intervals are valuable for characterizing sampling and parameter uncertainty conditional on the observed flood sample and selected probability model, but they do not necessarily represent uncertainty arising from a process-based modelling chain. In process-based FFA, flood estimates are additionally affected by the representation of meteorological variability, hydrologic-model structure and parameters, antecedent conditions, and errors in simulated flood peaks; the relative importance of these components may also vary with return period and watershed characteristics (Arnaud et al., 2017; Cameron et al., 2000; Newman et al., 2021). Process-based simulation therefore does not eliminate uncertainty associated with short observational records but redistributes it across a broader set of modelling assumptions and components. This distinction is important when interpreting the present results: the UQ-flood and Bayesian GEV intervals characterize different conditional uncertainty components and should not be compared directly as equivalent measures of overall precision. Accordingly, the GEV analysis is retained as a secondary statistical benchmark, while the primary evaluation focuses on whether the central UQ-flood estimates remain consistent with the long-record observational benchmarks and whether the residual error model corrects systematic biases in the process-based flood estimates.

Minor comment 2: Recent continuous-simulation and process-based FFA literature

[Line 75] The references supporting continuous-simulation and process-based FFA are somewhat dated and miss a body of recent work in this area, which has been quite active. I would suggest the authors review and cite more recent contributions, including:

• Baer, J. A.: *Design Storms Underestimate Flood Hazard and Risk Derived From Stochastic Storm Transposition*, MS thesis, University of North Carolina at Chapel Hill, 2025.

• Perez, G., Coon, E. T., Rathore, S. S., and Le, P. V. V.: *Advancing process-based flood frequency analysis for assessing flood hazard and population flood exposure*, *Journal of Hydrology*, 2024.

The recent work of Wright and colleagues on stochastic storm transposition is also relevant here; the reference lists of the two papers above point to several further studies worth considering.

Authors' response:

We thank the reviewer for these recommendations and agree that the literature review should better reflect recent developments.

We have revised the Introduction to include the suggested recent studies after verifying their publication details and direct relevance. We will also expand the discussion of recent work by Wright and colleagues on stochastic storm transposition and clarify how UQ-flood differs from, and complements, these approaches.

The revised text on page 3, lines 67–85, reads:

A third option is to generate larger flood samples using process-based continuous simulation, in which stochastic weather generators, stochastic storm transposition, or other meteorological resampling approaches are coupled with hydrologic models to produce long synthetic streamflow records (Eagleson, 1972; Franchini et al., 1996). This approach is attractive because it can represent physical controls on flood generation, including interactions among rainfall, snowpack, soil moisture, antecedent storage, and runoff production, while reducing sampling uncertainty through large synthetic samples. Previous studies have established the conceptual and practical basis of process-based FFA across a range of hydrologic settings and modelling configurations (Arnaud and Lavabre, 2002; Blazkova and Beven, 2009; Haberlandt and Radtke, 2014; Perez et al., 2019; Sivapalan et al., 2005; Yu et al., 2019). More recent developments have expanded this approach by using stochastic storm transposition to preserve the spatial and temporal characteristics of observed storms. Comparisons with conventional design-storm methods show that simplifying the spatiotemporal variability of rainfall can misrepresent flood hazard and risk, with important implications for infrastructure design, floodplain mapping, land-use planning, and disaster-risk management (Baer et al., 2026). Process-based FFA has also advanced using spatially distributed and more physically detailed hydrologic models. For example, coupling stochastic storm transposition with an integrated surface–subsurface hydrologic model extends the analysis beyond peak-flow frequency to include spatially distributed flood extent and population exposure (Perez et al., 2024). Together, these developments demonstrate how process-based FFA can move beyond simplified design-event assumptions and more explicitly represent the interacting meteorological, surface, and subsurface processes controlling flood magnitude and spatial extent. Nevertheless, the resulting estimates remain conditional on the adequacy of the meteorological sampling method, hydrologic-model structure, parameterization, and representation of the dominant flood-generating processes.

Minor comment 3: “Simulation-based FFA” versus “process-based FFA”

[Line 110] I would reconsider the term "simulation-based FFA." "Simulation" can equally describe purely statistical simulation such as Monte Carlo resampling from a fitted distribution not clearly distinguish what the authors are doing. "Process-based FFA" is more precise and is the term used in much of the recent literature. Since the term appears in the framework throughout the manuscript, this is worth settling early.

Authors’ response:

We agree that “process-based FFA” more clearly distinguishes the proposed framework from purely statistical simulation and Monte Carlo resampling.

We will adopt “process-based flood-frequency analysis” as the principal term throughout the revised manuscript. Where necessary, we will use “continuous-simulation process-based FFA” to describe the specific implementation involving a stochastic weather generator and hydrologic model.

Proposed definition at first occurrence (Page 2, lines: 32 - 34)

In this study, process-based flood-frequency analysis refers to the generation of synthetic meteorological sequences and their transformation into streamflow through a hydrologic model, followed by estimation of flood quantiles from the resulting annual-maximum series.

We also changed the title of the manuscript:

Residual Error Modelling for Process-Based Flood Frequency Analysis

Response to Reviewer 2

General evaluation:

This manuscript presents a novel and interesting framework for better estimating flood frequencies than the traditional GEV fitting approach. The proposed framework also integrates a more robust uncertainty quantification process. The topic is timely and important, with potentially broad implications, particularly for flood frequency analysis when the available observational record is relatively short. However, there are several aspects that require further explanation and/or discussion and the interpretation of the results could be strengthened. Please find my detailed comments below.

Authors' response:

We sincerely thank the reviewer for the positive and constructive evaluation of our manuscript. We appreciate the reviewer's recognition of the significance and timeliness of the UQ-flood framework, particularly its treatment of hydrologic-model error at the annual-maximum scale and its potential relevance when observational records are limited. The reviewer's detailed comments helped us clarify the study's objectives, strengthen the methodological description, and improve the interpretation of the results.

Following the reviewer's comments, together with the related concerns raised by Reviewer 1, we have revised the manuscript's central framing. The submitted version occasionally presented UQ-flood as outperforming the GEV approach, particularly by comparing the widths of their uncertainty intervals. We recognize that this comparison was not wholly appropriate because the UQ-flood and Bayesian GEV intervals represent different conditional uncertainty components.

In the revised manuscript, we therefore do not position UQ-flood as a method that is categorically better than conventional GEV-based flood-frequency analysis. Instead:

- 1. Long-record flood observations provide the primary empirical benchmark for evaluating the central flood-quantile estimates.**
- 2. The GEV analysis is retained as a secondary statistical benchmark, not as a competing method whose uncertainty width is directly compared with that of UQ-flood.**
- 3. Direct comparisons of UQ-flood and GEV interval widths have been removed, together with claims that narrower UQ-flood intervals demonstrate greater overall precision, reliability, or performance.**
- 4. The interpretation now focuses on whether UQ-flood's central estimates remain consistent with the long-record observational benchmark, despite the framework being calibrated and inferred using only 30 years of observations.**
- 5. The role of the residual error model is evaluated by comparing UQ-flood with the process-based simulation without REM, thereby demonstrating how annual-maximum-scale error correction affects the estimated flood quantiles.**

Major comment 1:

While the proposed approach is designed for annual maxima, several studies (e.g., Fischer & Schumann (2016)) have argued that the peaks-over-threshold (POT) approach can make more efficient use of the available information and can provide more robust flood-frequency estimates than annual maxima, particularly for relatively short records. Given that one of the main motivations of UQ-flood is to improve flood frequency estimation under limited observational records, I think this issue deserves more discussion. Could you explain why the annual-maxima approach was selected and discuss how the proposed framework could potentially work in the POT approach?

Authors' response:

We thank the reviewer for this valuable comment. We agree that peaks-over-threshold, or POT, methods can use more information from an observed record than an annual-maximum approach. Practically, there is no reason we could not extend the UQ-flood method to use POT instead of annual maxima. We decided here to use the simpler option in part because the northern watersheds we were simulated tend to have a single annual maxima, and the POT and annual maxima results would not significantly vary. However, because of this comment we have now added a discussion of the potential extension to POT sampling in the discussion. The revised text on page 24 and 25, lines 631–635, reads:

The current implementation uses annual maxima, but UQ-flood is not structurally limited to this formulation. A peaks-over-threshold extension could use more high-flow information from short records but would require explicit treatment of threshold selection, event independence, declustering, exceedance frequency, and event magnitude. Future work should also use event-scale diagnostics to investigate how hydrologic-model performance, meteorological simulation, residual-error characteristics, and flood-generating mechanisms contribute to differences among watersheds.

Major comment 2:

Could you add figures on how good the fitted GEVs are, since they act as a benchmark for evaluating your approach?

Authors' response:

We thank the reviewer for this helpful suggestion. We agree that, because the GEV estimates are retained as a statistical benchmark, the quality of the fitted distributions should be documented more clearly.

We evaluated the GEV fits for all three watersheds under both the short-record and long-record configurations using quantile plots, probability plots, and return-level plots. These diagnostic plots did not reveal important concerns regarding the adequacy of the fitted GEV distributions for the benchmark analyses.

To improve transparency without interrupting the main narrative, we have added the complete set of diagnostic plots to the Supplement as Figs. [S2–S7]. We have also revised Sect. [3.5] of the manuscript to refer explicitly to these diagnostics and briefly summarize their interpretation.

The following text has been added to the manuscript (Page 12, lines: 343 - 345):

The adequacy of the fitted GEV distributions was evaluated using quantile plots, probability plots, and return-level plots for each watershed under both the short-record and long-record

configurations. The diagnostic plots did not indicate important concerns with the fitted GEV models. The complete diagnostic results are provided in Figs. [S2–S7] of the Supplement.

We emphasize that, following the revisions made in response to Reviewer 1, the GEV analysis is used as a secondary statistical benchmark for evaluating the central flood-quantile estimates. The widths of the GEV and UQ-flood intervals are no longer compared as equivalent measures of uncertainty.

Major comment 3:

Which 30 years of data did you use for calibrating the hydrologic model and inferring the parameters of the weather generator?

Authors' response:

We thank the reviewer for identifying this omission. We agree that the exact periods used for calibration and parameter inference should be reported to improve the transparency and reproducibility of the analysis.

Proposed manuscript text (Page 15, lines: 421 - 426):

For the Black River and Moyie River watersheds, the HBV-EC hydrologic model, CoSMoS stochastic weather generator, and residual error model were calibrated or inferred using observations from 1 January 1991 to 31 December 2020. For the Lahave River watershed, the corresponding period extended from 1 January 1992 to 31 December 2021. These periods represented the most recent complete 30-year observational records available for the respective watersheds. For each watershed, the same period was also used for the short-record GEV analysis, ensuring consistency across the modelling and statistical comparisons.

Major comment 4:

L339-343: I do not understand why there is a need to calculate the uncertainty range at each time step in the hydrographs, since normally the peak Q_s are used for flood frequency analysis.

Authors' response:

We thank the reviewer for identifying this confusing description. We agree that the sentence stating that the uncertainty range was calculated “at each computational time step t in the hydrographs” was incorrect. Rather, the uncertainty range was not calculated at individual time steps of the simulated hydrographs. In Eq. (1), the index t refers to each target return period considered in the flood-frequency analysis. Specifically, the uncertainty range was evaluated for return periods of 2, 5, 10, 20, 50, and 100 years using the corresponding ensemble of flood-quantile estimates.

To avoid confusion, we have removed the reference to computational time steps and revised the notation so that T explicitly represents the return period.

Revised manuscript text (Page 13, lines: 347 - 351)

Central flood-quantile estimates were evaluated at return periods of $T = 2, 5, 10, 20, 50, 100$ years. For each return period, the interval width was calculated as:

$$UR_T = q_T^{97.5} - q_T^{2.5} \quad (1)$$

Where $q_T^{97.5}$ and $q_T^{2.5}$ denote the upper and lower limits, respectively, of the 95% interval for the flood quantile associated with return period T .

Major comment 5:

Are those black dots from the simulation without the REM? Would it make sense to add confidence intervals as you have 10,000 years of annual maxima, e.g. by get 1,000-year samples 100 times?

Authors' response:

We thank the reviewer for this comment. Yes, the black points represent the annual maximum flows generated by the process-based hydrologic simulation without applying the residual error model. These points are derived from a synthetic sequence of 1,000 annual maxima and are included as a deterministic reference to illustrate the systematic bias that may occur when process-based flood-frequency analysis is conducted without explicitly correcting hydrologic-model error. This purpose is consistent with the numerical experiment described in the manuscript.

To make this interpretation unambiguous, we have changed the corresponding label in the figures and captions from “SIM” to “det-SIM.” We have also revised the main text to clarify that the black points are not intended to represent an uncertainty estimate. Rather, they show the central flood-frequency relationship obtained directly from the hydrologic model driven by the stochastic meteorological inputs before the REM correction is applied.

We appreciate the reviewer's suggestion to add confidence intervals around these points. After considering this option, however, we decided not to add an additional interval. Figures 3 and 5 already contain several observational, statistical, and process-based results, together with their associated intervals. Adding another uncertainty envelope would make the figures more visually complex and could obscure their main purpose, which is to demonstrate the effect of applying the REM to the process-based flood estimates.

We also propose the following revised caption wording (Page 18, lines: 471-475):

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 (without error model) (det-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.

Major comment 6:

In Fig.3, it is hard to see the estimate differences between different approaches for low return periods (<20 return years). It seems like the UQ-flood slightly overestimates the annual maximum flows for those return periods. Could you make a table (could be in the supplementary materials) listing the differences for a range of return periods e.g. 2, 5, 19, 20, 50, 100 years?

Authors' response:

We thank the reviewer for this helpful suggestion. We have added Table [S1] to the Supplement, reporting flood-quantile estimates for return periods of 2, 5, 10, 20, 50, and 100 years for all three study watersheds. The table includes:

- 1. the long-record observational benchmark.**
- 2. the UQ-flood central estimate.**
- 3. the GEV central estimate based on the full observational record.**
- 4. the GEV central estimate based on the 30-year record.**

Proposed manuscript text (Page 16, lines: 442 - 447):

The central UQ-flood estimates remain consistent with the long-record observational benchmark across the evaluated return periods, despite the framework being calibrated and inferred using only 30 years of meteorological and streamflow data. At the 50- and 100-year return periods, the UQ-flood estimates remain within the confidence intervals for the corresponding long-record discharge quantiles. The long-record GEV estimates also generally follow the empirical flood-frequency relationship, while differences among the central estimates vary with return period. Numerical estimates for selected return periods are provided in Table S1.

Major comment 7:

It seems to me that according to Fig. 4, the GEV estimates appear to be closer to the observed annual maxima (red dots) at high return periods (e.g. 45 and 90 years in Fig. 4A, and 109 years in Fig. 4B) than the UQ-flood estimates. In contrast, UQ-flood appears to provide closer estimates at lower return periods. This pattern seems different from the results for the Black River watershed shown in Fig. 3. Could the authors further investigate and discuss the reasons for these differences in model performance across watersheds? Could it be something related to the dominant flood generation mechanisms that are not well captured/modelled in the hydrologic model and/or the stochastic weather generator? The same applies to the REM, it seems to have little influence on

the estimates for the Black River watershed, while its influence is much larger for the other two watersheds. What might be the causes?

Authors' response:

We thank the reviewer for this careful interpretation of the figures. We agree that, based on the visual comparison in Fig. 4, the long-record GEV estimates appear closer to some observed annual maxima at the higher return periods than the central UQ-flood estimates.

However, proximity to individual empirical annual maxima is not the only criterion used to evaluate the results. The observed plotting positions are themselves affected by sampling uncertainty, particularly at the highest return periods. For this reason, the manuscript also presents confidence intervals for discharge quantiles, which explicitly represent sampling uncertainty in the long-record observational benchmark. The central UQ-flood estimates remain within these confidence intervals over the evaluated return periods. Therefore, we interpret UQ-flood estimates located within the observational confidence intervals as being consistent with the long-record benchmark.

We have revised the Results to make this evaluation criterion more explicit. We no longer state broadly that UQ-flood is consistently closer to the observed annual maxima than the GEV approach. Instead, we report whether the central estimates are consistent with the confidence intervals for the long-record discharge quantiles. This revised interpretation is also consistent with the manuscript's reoriented focus on central flood-quantile estimates rather than comparisons of uncertainty-band width.

We have also added the quantitative comparison requested in Major Comment 6 to the Supplement. Table [S1] reports the long-record observational benchmark, UQ-flood, long-record GEV and 30-year GEV at return periods of 2, 5, 10, 20, 50, and 100 years for all three watersheds. We now use this table to describe any watershed-specific differences quantitatively rather than relying only on the visual proximity of the curves. The numerical comparison shows that the apparent performance pattern is not uniform. UQ-flood is closest to the observational benchmark in some cases, including the Lahave 50-year flood, whereas the long-record GEV is closer in others, including the Moyie 50-year and Lahave 100-year floods.

The reviewer's observation that the REM appeared to have little influence on the Black River estimates also helped us identify an error in the submitted Black River figure. The black points in that figure inadvertently represented simulated annual maxima after bias correction rather than the intended process-based simulation without the REM. This made the effect of the REM appear smaller than it actually was. We have corrected the Black River figure so that the black points now represent the process-based simulation without REM, consistent with the numerical experiment and with the presentation for the other watersheds. The corrected figure shows the actual bias in the deterministic process-based simulation and

provides a clearer representation of the correction introduced by the REM. We have also replaced the label “SIM” with “det-SIM” in the figures and captions to avoid further ambiguity.

This is the new Fig. 3 for the black River Case study:

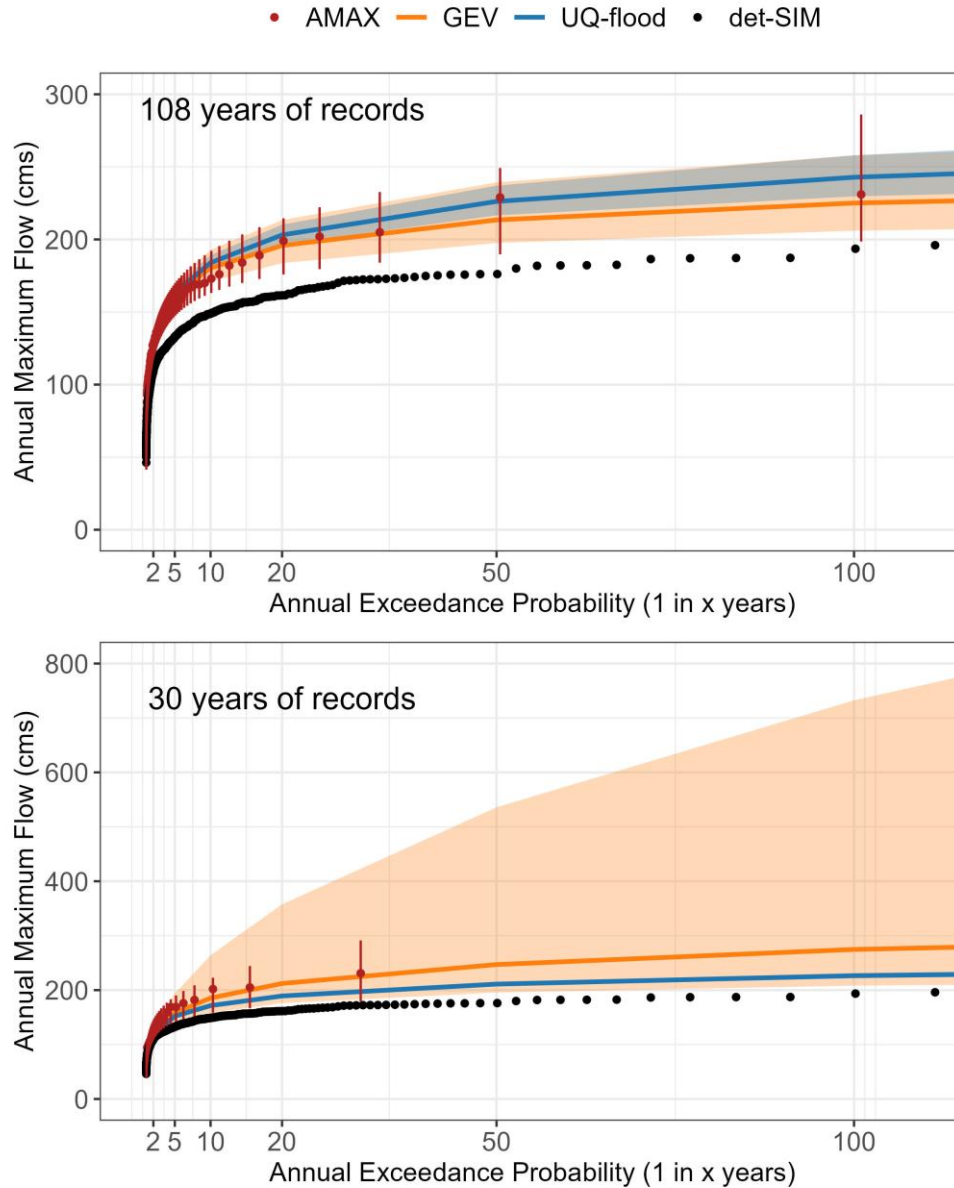


Figure 3: Flood frequency estimates and associated 95% uncertainty intervals for the Black River watershed derived using the statistical GEV approach and the UQ-flood framework.

The observed annual maximum discharge (AMAX; red) distribution, shown with confidence intervals for discharge quantiles, and the deterministic simulation output (without error model) (det-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.

The UQ-flood uncertainty-band width is identical in the upper and lower panels; the apparent difference is solely a visual effect caused by the broader y-axis scale used in the lower panel.

We appreciate the reviewer's suggestion that differences among watersheds could be associated with dominant flood-generation mechanisms or with limitations in the hydrologic model and stochastic weather generator. However, the current study does not include event-based process diagnostics that would allow us to attribute the observed differences to specific flood-generation mechanisms. We therefore avoid making a mechanism-based explanation that is not supported by the analysis.

Instead, we have revised the manuscript to state that differences among watersheds may reflect the combined influence of hydrologic-model performance, stochastic-weather-generator representation, and residual-error characteristics, but that the present results do not allow these influences to be separated or attributed conclusively. A formal investigation of mechanism-specific performance would require additional diagnostics and is outside the scope of the present study.

Proposed limitation statement (Page 24 -25, lines: 633 - 635)

Future work should also use event-scale diagnostics to investigate how hydrologic-model performance, meteorological simulation, residual-error characteristics, and flood-generating mechanisms contribute to differences among watersheds.

Table S1: Central flood-quantile estimates for return periods of 2, 5, 10, 20, 50, and 100 years for the Black River, LaHave River, and Moyie River watersheds. AMAX denotes the empirical flood-quantile estimate derived from the available long annual-maximum-flow record. UQ-flood denotes the central estimate from the proposed process-based framework developed using 30 years of observations. GEV-short denotes the central Bayesian GEV estimate based on the same 30-year observational period, and GEV-long denotes the central Bayesian GEV estimate based on the full available annual-maximum record. All discharge values are expressed in $\text{m}^3 \text{s}^{-1}$. NA indicates that an empirical Q100 estimate was not available for the Moyie River.

Watershed	Method	Q2	Q5	Q10	Q20	Q50	Q100
Black River	AMAX	128.5	155.4	176.6	194.0	223.1	230.7
	UQ-flood	128.7	163.3	183.8	202.9	226.1	242.7
	GEV-short	126.5	167.0	192.8	216.4	246.6	269.2
	GEV-long	130.1	162.1	180.3	195.8	213.4	225.1
Lahave River	AMAX	207.0	301.2	356.2	382.9	472.4	635.3
	UQ-flood	233.0	316.7	369.0	416.6	475.1	517.6
	GEV-short	225.5	348.8	460.4	602.6	867.3	1161.6
	GEV-long	215.3	301.3	363.3	427.1	517.3	591.3
Moyie River	AMAX	146.0	183.0	213.4	236.5	251.2	NA
	UQ-flood	144.3	187.2	213.4	237.7	267.0	288.1
	GEV-short	137.6	191.3	227.9	265.2	319.8	368.2
	GEV-long	146.5	189.4	212.6	231.6	252.5	265.8

Major comment 8:

I would be very careful with comparing the proposed approach with other approaches, as introduced in L534-542. The authors state that their approach has two main advantages over other models: 1) less subjectivity; and 2) being faster and simpler. However, the proposed framework still involves several subjective or user-defined choices, including the selection of the hydrological model and stochastic weather generator, as well as the use of a fixed $\lambda = 0.2$ for the Box–Cox transformation. The authors are encouraged to clarify what aspects of the proposed framework are considered less subjective and how these choices affect the results.

In addition, the hydrologic-weather generator approach is typically considered as a time consuming approach (e.g. Wing et al. 2020). Also, most hydrological models lack good-quality data for calibration and capturing extreme events (e.g. Wing et al. 2020). These considerations raise questions about the practical applicability and transferability of the proposed framework to catchments with limited data or substantially different hydrological characteristics. Given the

statement in L613–616 regarding the broader application of the approach, I suggest that the authors discuss these limitations more explicitly and clarify under what conditions the framework can be expected to be transferable.

Authors' response:

We thank the reviewer for this important comment. We agree that the original manuscript described the advantages of UQ-flood too broadly. The framework includes several consequential modelling choices, including selection and calibration of the hydrologic model, selection and fitting of the stochastic weather generator, specification of the residual error model, and selection of the Box–Cox transformation parameter. We have therefore removed the claim that UQ-flood is generally “less subjective” than other approaches.

We have also moderated the statements that UQ-flood is inherently faster and simpler. Our intended comparison was with fully joint Bayesian approaches (e.g., Costa and Fernandes (2017)) that simultaneously infer hydrologic-model, weather-generator, and residual-error parameters. UQ-flood uses a sequential modelling structure and a comparatively simple residual error formulation, which makes its implementation less complex than a fully joint inference framework. The revised manuscript now describes UQ-flood as less complex in its statistical implementation than fully joint Bayesian approaches and have removed unsupported statements about computational speed.

The Box–Cox parameter was fixed at $\lambda=0.2$ based on previous hydrological residual-error research rather than selected specifically to optimize the results for the three study watersheds. McInerney et al. 2017 evaluated several approaches to representing heteroscedastic streamflow residuals and identified fixed Box–Cox transformations with $\lambda=0.2$ and 0.50, together with the logarithmic transformation for perennial watersheds, among the Pareto-optimal schemes. That study also found that calibrating λ could sometimes overfit low flows at the expense of predictive precision at high flows. We selected $\lambda=0.2$ as an established residual-error formulation for streamflow, but we agree that fixing this value remains a modelling choice and that its suitability may vary among applications.

Regarding transferability, we agree that UQ-flood should not be assumed to perform reliably in any watershed without evaluating the adequacy of the underlying modelling components. The framework should be applied only where the hydrologic model can represent the dominant runoff and flood-generation processes and where the stochastic weather generator can reproduce the meteorological characteristics relevant to flood formation. Its reliability also depends on having sufficient and suitable data to calibrate the hydrologic model, infer weather-generator parameters, and characterize residual errors at the target flood scale. We now reduce the broader claims of transferability and have added a paragraph to the discussion.

Proposed revised manuscript text (Pages 23, lines:580-590):

UQ-flood uses a sequential modelling structure in which the hydrologic model, stochastic weather generator, and REM are developed separately. This structure is statistically less complex than fully joint Bayesian approaches that simultaneously infer parameters across the complete modelling chain. However, UQ-flood should not be interpreted as free from subjective or user-defined decisions. Hydrologic-model selection and calibration, weather-generator selection and configuration, REM specification, and transformation choice can all affect the resulting flood estimates. Moreover, long continuous simulations can remain computationally demanding.

The Box–Cox parameter was fixed at $\lambda=0.2$ based on previous hydrological residual-error studies identifying this value as a suitable option for representing heteroscedastic streamflow errors. Nevertheless, the fixed parameter represents a modelling assumption, and its influence was not evaluated in the present study. Future applications should assess whether the selected transformation and residual-error structure are appropriate for the watershed and flood regime under consideration.

In addition, for the generalization of the UQ-flood. The revised text on page 24, lines 626–629, reads:

Transferability depends on whether the hydrologic model and meteorological generator adequately represent the dominant flood-generating processes. The modular framework can accommodate alternative lumped or distributed models, but this flexibility does not guarantee reliable performance in different or ungauged watersheds. Applications to ungauged basins would require regionalization or parameter-transfer methods and case-specific evaluation.

We thank the reviewer for encouraging a more balanced description of the framework. We have removed claims that UQ-flood is generally less subjective or universally faster, clarified that its principal practical advantage is a less complex statistical structure relative to fully joint Bayesian inference, and explicitly acknowledged its computational and data requirements. We have also identified the fixed Box–Cox parameter as a modelling assumption and substantially moderated the claims concerning transferability and application to ungauged watersheds

Minor comment 1:

L38: Remove the extra ‘.’

Authors’ response:

Thank. It is done.

Minor comment 2

Are there any missing records in the daily observed streamflow? If so, how did you handle those missing data?

Authors' response:

We thank the reviewer for this question. We checked the daily streamflow records used for model calibration and parameter inference and found no missing values within the selected 30-year periods. Therefore, no gap-filling or interpolation was required.

Minor comment 3:

L193: The acronym HBV-EV has not been introduced before.

Authors' response:

Thank you. This has been corrected.

Minor comment 4:

In L205: Define the two phases more clearly, e.g. 1) model setup and parameter estimation; and 2) long-term continuous simulation.

Authors' response:

Thank you. This has been corrected.

Minor comment 5:

I suggest to move the first sentence (L253) to the previous paragraph. It makes more sense to first explain why HBV-EC was selected and then introduce how the model works..

Authors' response:

Thank you. This has been moved.

Minor comment 6:

L266: The mentioning on comparison 30-year versus 108-year estimations is somewhat confusing. Only the Lahave River watershed has 108 years of data, while the other two have shorter records.

Authors' response:

Thank you. We clarified that the Black and LaHave River records each contain 108 years of data, whereas the Moyie River record contains 90 years.

Minor comment 7:

L358: Remove the extra ‘.’

Authors’ response:

Thank you. This has been corrected.

Minor comment 8:

L383: Moyie River Watershed only has 90-year of data, right?

Authors’ response:

Yes, you are right.

Minor comment 9:

L398-410: Much of this information has already been introduced in the preceding sections. Consider shortening or removing the repetition.

Authors’ response:

Thank you. We shortened the section and removed the repeated information.

Minor comment 10:

L457: Remove the extra ‘.’

Authors’ response:

Thank you. This has been corrected.