~~An Uncertainty Quantification Framework for Simulation-based Flood Frequency Analysis~~Residual Error Modelling for Process-Based Flood Frequency Analysis

Jonathan Romero-Cuellar[1], James R. Craig[1], Bryan A. Tolson[1], and Rezgar Arabzadeh[1]

[1]Department of Civil and Environmental Engineering, University of Waterloo, Waterloo, N2L 3G1, Canada

Correspondence to: Jonathan Romero-Cuellar (jromeroc@uwaterloo.ca)

Abstract. ~~Flood frequency analysis (FFA) is essential for flood risk management and infrastructure design, yet flood quantile estimates remain highly uncertain when based on limited streamflow records. In data-scarce settings, conventional statistical FFA is strongly constrained by sampling uncertainty: short annual maximum flow records often yield broad credible intervals for high-return-period floods, making the resulting estimates difficult to use in practical flood-risk decisions. To address this limitation, we introduce a simulation-based uncertainty quantification framework—UQ-flood—that integrates a process-based hydrologic model, a stochastic weather generator, and a residual error model (REM). Designed for annual maxima, the REM accounts for model bias and residual variability, enabling the generation of probabilistic streamflow ensembles tailored to extreme event analysis. We apply UQ-flood to three Canadian watersheds with long streamflow records and contrasting hydroclimatic conditions. We compare its performance against traditional statistical FFA using the Generalized Extreme Value distribution and Bayesian inference. UQ-flood yields flood quantile estimates consistent with long-record statistical methods but with substantially narrower uncertainty bounds. Under short-record conditions (e.g., 30 years), UQ-flood produced quantile estimates that remained close to long-record empirical/statistical benchmarks, whereas statistical FFA produced much wider intervals. Additional experiments reveal that omitting the REM introduces systematic bias in flood magnitude estimates. UQ-flood avoids parametric assumptions about flow distributions, circumvents hydrologic model biases, and is adaptable to data-limited conditions. By explicitly propagating uncertainty from hydrologic simulation to flood quantiles, UQ-flood offers a practical alternative for robust flood risk management, including applications in infrastructure design and floodplain mapping. We recommend integrating residual error models into continuous simulation frameworks to improve bias correction and uncertainty quantification in flood risk estimation.~~ Flood frequency analysis (FFA) is essential for flood risk management and infrastructure design, yet estimates of flood quantiles remain strongly influenced by the length and representativeness of available streamflow record. We introduce UQ-flood, a process-based flood-frequency framework that integrates a stochastic weather generator, a process-based hydrologic model, and a residual error model developed at the annual-maximum scale. The framework explicitly addresses the presence of errors in simulated annual flood peaks generated by models calibrated to continuous streamflow data. We apply UQ-flood to three Canadian watersheds with contrasting hydroclimatic conditions using 30 years of meteorological and streamflow observations. Long-record empirical annual-maximum estimates provide the primary benchmark, while Bayesian Generalized Extreme Value estimates based on short and long records provide secondary statistical benchmarks for the central flood-quantile

estimates. Across the three watersheds, the central UQ-flood estimates remained within the confidence intervals for the long-record discharge quantiles, although their relative proximity to the empirical estimates varied by watershed and return period. Process-based simulations without residual-error correction systematically underestimated flood magnitudes, whereas applying the residual error model shifted the estimated flood-frequency curves toward the long-record observational benchmarks. The framework also propagates residual predictive variability directly at the annual-maximum scale. These results demonstrate that annual-maximum residual error modelling can correct for systematic discrepancies in process-based flood estimates and support central flood-quantile estimation even at sites with short data records.

1 Introduction

Floods are among the most prevalent natural disasters, with both their intensity and frequency increasing due to global change (Milly et al., 2002). In this context, flood frequency analysis (FFA) plays a fundamental role in engineering hydrology, as it is essential for accurately estimating flood risks. These estimations are critical for designing flood protection infrastructure, planning land use, ensuring dam and industrial safety, conducting risk assessments, and making informed water resources management decisions. Industry-standard, at-site FFA typically involves fitting a statistical distribution to a sample of observed flood data, often based on annual maximum flows or peaks over threshold. These distribution models estimate the probability of exceedance for extreme flood peaks and are used to extrapolate high flood quantiles beyond observed data, often following regulatory guidelines (Hamed and Ramachandra Rao, 2019).

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., 2012a; Merz and Thieken, 2005a)~~.~~. 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., 2017a; Cameron et al., 2000a; 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.

~~Despite its central role in engineering practice, uncertainty quantification in FFA is often neglected, insufficiently characterised, or not transparently communicated~~ (Hall and Solomatine, 2008; Merz and Thieken, 2005b, 2009; Vidrio-

Sahagún et al., 2025). ~~. This is critical, as flood frequency estimates are affected by multiple sources of uncertainty — including sampling uncertainty, model structure uncertainty, and parameter uncertainty. Sampling uncertainty arises from the limited length of observed discharge records, which makes it difficult to reliably characterise the tail of the distribution, especially for rare events like the 100-year flood~~ (Lucas et al., 2023)~~. Structural uncertainty stems from the subjective selection of probability distributions. Goodness-of-fit tests often lack discriminatory power with small samples, allowing different distributions to fit the data equally well while producing vastly different flood quantiles~~ (Cunnane, 1985; Laio et al., 2009)~~. Parameter uncertainty further contributes to the problem, with flood estimates varying substantially depending on the parameter estimation method used~~ (Hu et al., 2020)~~.

Statistical FFA based on limited streamflow records, for example fewer than 30 years of annual maximum flows, can yield highly uncertain and often impractical estimates for rare events. In such cases, uncertainty intervals may be statistically defensible but so wide that they provide limited actionable guidance for infrastructure design, floodplain mapping, or emergency planning. This limitation highlights a key gap: the need for accessible and reliable uncertainty quantification approaches that remain useful under data-scarce conditions (Pappenberger and Beven, 2006).

Several strategies have been developed to reduce the limitations imposed by short flood records. A first option is to extend the flood record in time by incorporating historical information, documentary evidence, or paleoflood indicators (Benito et al., 2004). These approaches can improve the estimation of rare flood quantiles by adding information about events that occurred before systematic gauging began. However, they also require careful treatment of additional uncertainty sources, including event dating, magnitude reconstruction, threshold interpretation, and differences between historical, paleoflood, and instrumental data quality. Thus, while historical and paleoflood information can increase the effective sample size, it can also introduce new epistemic uncertainties that must be explicitly characterised.

A second option is to borrow information across space through regional flood frequency analysis, where data from hydrologically similar catchments are pooled to increase the effective flood sample (Cunnane, 1988). Regional approaches can reduce sampling uncertainty relative to single-site analysis, particularly in basins with short records. However, their reliability depends on the assumption that the selected catchments form a homogeneous region. Defining such regions is partly subjective and depends on choices regarding similarity criteria, catchment descriptors, regional boundaries, and pooling rules. As a result, different regionalization decisions can lead to different pooled samples and, consequently, different flood quantile estimates.

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, 2002a; Blazkova and Beven, 2009a; Haberlandt and Radtke, 2014a; Perez et al., 2019a; Sivapalan et al., 2005a; Yu et al., 2019a). 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. ~~.Previous research has established the concept and potential of model-based, or continuous simulation, FFA frameworks~~ (Arnaud and Lavabre, 2002; Blazkova and Beven, 2002; Eagleson, 1972; Franchini et al., 1996; Sivapalan et al., 2005)(Arnaud and Lavabre, 2002; Blazkova and Beven, 2002; Eagleson, 1972; Franchini et al., 1996; Sivapalan et al., 2005)~~. In these frameworks, process-based hydrologic models are driven by synthetic meteorological inputs to generate long discharge time series for flood frequency estimation~~ (Blazkova and Beven, 2009)(Blazkova and Beven, 2009)~~. Continuous simulation has been successfully applied to derive flood frequency distributions in a range of hydrologic settings (Haberlandt & Radtke~~ (2014)(2014)~~, Yu et al.~~ (2019)(2019)~~, Perez et al.~~ (2019)(2019)~~)~~.

Model-based approaches are particularly well-suited for this task because they capture complex interactions among flood drivers—such as rainfall, snowpack, and soil moisture—and their influence on runoff production (Sivapalan and Samuel, 2009)(Sivapalan and Samuel, 2009). In contrast to design storm methods, model-based FFA avoids the assumption of equal return periods for rainfall and discharge. It also enables the integration of land use change (e.g., urbanization), management interventions, and future climate scenarios (Lamb et al., 2016)(Lamb et al., 2016). Another key advantage of this approach is its ability to generate large synthetic samples, which reduces sampling uncertainty—a common limitation in statistical methods relying on short observational records. Common methods to generate large prediction samples for this framework include stochastic weather generators using multivariate non-Gaussian time series (e.g., Yu et al. (2023)(2023)), stochastic storm transpositions (Andersen et al., 2024; Wright et al., 2020)(Andersen et al., 2024; Wright et al., 2020), and reforecast ensemble pooling (Brunner and Slater, 2022)(Brunner and Slater, 2022)(Brunner and Slater, 2022)(Brunner and Slater, 2022).

However, continuous simulation does not automatically solve the uncertainty problem in FFA. A key limitation is that hydrologic models are commonly calibrated to reproduce daily streamflow dynamics or general hydrograph behaviour, whereas FFA depends on accurately representing annual maximum flows. This creates a mismatch between the calibration target and the design variable of interest. A model that performs well for daily streamflow may still exhibit systematic bias or inadequate variability in simulated flood peaks (Cameron et al., 2001; Brunner et al., 2021; Grimaldi et al., 2019; Spellman et al., 2018). If this model error is ignored, continuous simulation can produce biased flood frequency estimates, even when very long synthetic streamflow records are generated.

Several methodological frameworks have been proposed to quantify the uncertainty of flood frequency estimates, yet none have achieved widespread adoption in practice. One class of approaches combines continuous hydrologic model simulation with stochastic rainfall generation and rainfall–runoff modelling, often using methods like Generalized Likelihood Uncertainty Estimation (GLUE) to quantify uncertainty (Cameron et al., 2000b)(Cameron et al., 2000b)(Cameron et al., 2000b)(Cameron et al., 2000b). While GLUE is useful for exploring parameter and model structure uncertainty, it has been criticized for relying on subjective likelihood measures and for its tendency to overestimate uncertainty (Stedinger et al., 2008)(Stedinger et al., 2008)(Stedinger et al., 2008)(Stedinger et al., 2008). Formal Bayesian inference methods offer a more robust probabilistic treatment of uncertainty, but they are often computationally demanding and require careful specification of prior distributions, limiting their accessibility for routine applications (Costa and Fernandes, 2017)(Costa and Fernandes, 2017)(Costa and Fernandes, 2017)(Costa and Fernandes, 2017). Other approaches, including resampling methods, hierarchical Monte Carlo frameworks, variance decomposition, and ensemble pooling, have also been used to reduce or attribute uncertainty in flood estimates (Apel et al., 2004; Arnaud et al., 2017; Bomers et al., 2019; Burn, 2003; Mirzaei et al., 2015; Newman et al., 2021; Brunner and Slater, 2022). Despite this diversity of methods, many remain constrained by subjectivity, computational complexity, high data requirements, or limited treatment of hydrologic model error at the flood-peak scale. A more accessible, process-consistent framework for uncertainty quantification in FFA—particularly one that performs well under limited data availability—therefore remains a pressing need.

Given these limitations, ~~simulation~~process-based FFA requires an explicit treatment of hydrologic model error at the scale most relevant for design floods: the annual maximum flow. A key limitation of model-based flood estimation frameworks is the frequent neglect of model error and uncertainty in hydrologic predictions. To address this, we apply the Residual Error Model (REM) approach—a strategy designed to quantify and propagate residual uncertainty by embedding error model samples into calibrated deterministic simulations (Vogel, 2017)(Vogel, 2017)(Vogel, 2017)(Vogel, 2017)(Vogel, 2017)(Vogel, 2017). REM transforms deterministic outputs from process-based hydrological models into streamflow ensembles, offering a robust means of uncertainty quantification that is particularly valuable for hydrologic risk management. The REM framework consists of two key components:

1. A deterministic component, which represents changes in meteorological forcings, land use, and their non-linear interactions.

2. A stochastic component, which accounts for hydrologic uncertainty not captured by the deterministic model alone (Shabestanipour et al., 2023)(Shabestanipour et al., 2023)(Shabestanipour et al., 2023)(Shabestanipour et al., 2023)(Shabestanipour et al., 2023)(Shabestanipour et al., 2023).

This dual structure enables REM to estimate predictive uncertainty—integrating data, parametric, and structural uncertainties—by comparing model simulations with observed streamflow records (Montanari and Koutsoyiannis, 2012)(Montanari and Koutsoyiannis, 2012)(Montanari and Koutsoyiannis, 2012)(Montanari and Koutsoyiannis, 2012)(Montanari and Koutsoyiannis, 2012)(Montanari and Koutsoyiannis, 2012). Various error modelling approaches have been proposed to represent this uncertainty (Jiang et al., 2019; Koutsoyiannis and Montanari, 2022; Li et al., 2016; McInerney et al., 2017, 2020; Shabestanipour et al., 2023; Todini, 2008)(Jiang et al., 2019; Koutsoyiannis and Montanari, 2022; Li et al., 2016; McInerney et al., 2017, 2020; Shabestanipour et al., 2023; Todini, 2008)(Jiang et al., 2019; Koutsoyiannis and Montanari, 2022; Li et al., 2016; McInerney et al., 2017, 2020; Shabestanipour et al., 2023; Todini, 2008)(Jiang et al., 2019; Koutsoyiannis and Montanari, 2022; Li et al., 2016; McInerney et al., 2017, 2020; Shabestanipour et al., 2023; Todini, 2008)(Jiang et al., 2019; Koutsoyiannis and Montanari, 2022; Li et al., 2016; McInerney et al., 2017, 2020; Shabestanipour et al., 2023; Todini, 2008).

In this study, we apply a REM that is additive in the transformed space, using the Box-Cox transformation, ~~a first-order autoregressive (AR(1)) model,~~ Gaussian innovations, and a flow-dependent error mean—following the formulation in Hunter et al., (2021)(2021)(2021)(2021)(2021)(2021). This error model is particularly attractive due to its ability to consolidate multiple sources of uncertainty into a single error term, enabling transparent, efficient, and straightforward generation of streamflow predictive distributions (Evin et al., 2014; Romero-Cuellar et al., 2024a)(Evin et al., 2014; Romero-Cuellar et al., 2024a)(Evin et al., 2014; Romero-Cuellar et al., 2024a)(Evin et al., 2014; Romero-Cuellar et al., 2024a)(Evin et al., 2014; Romero-Cuellar et al., 2024a).

Here, we present UQ-flood, an uncertainty quantification framework for ~~simulation~~process-based flood frequency analysis that integrates a process-based hydrologic model, a stochastic weather generator, and a residual error model developed at the annual maxima scale. The framework is designed to generate large synthetic flood samples while correcting systematic biases and propagating residual uncertainty in simulated flood peaks. By modelling the relationship between observed and simulated annual maximum flows directly, UQ-flood addresses the mismatch between daily streamflow calibration and flood-quantile estimation. This allows uncertainty from hydrologic simulation to be propagated into flood frequency curves and associated quantile estimates. The novelty of this study lies not in process-based continuous simulation alone, but in applying a residual error model directly at the annual-maximum scale and embedding it within a stochastic simulation framework for at-site FFA. This differs from daily-flow post-processing and from parameter-uncertainty approaches because it targets the design variable itself: annual maximum discharge.

Unlike traditional statistical FFA methods that rely solely on observed flood samples and assumed parametric distributions, UQ-flood is informed by hydrologic process representation and by the full information contained in the available meteorological and discharge records. To our knowledge, this is the first study to apply the REM approach within the context of at-site FFA for uncertainty quantification. In contrast to conventional approaches that can produce very wide uncertainty intervals under short-record conditions, UQ-flood offers a process-consistent and practical alternative for data-scarce environments.

2 Study Watersheds and Data

To evaluate the proposed UQ-flood framework for at-site application across diverse hydrologic conditions, we selected three Canadian case study watersheds based on two main criteria: (1) the availability of long daily streamflow records—each with more than 80 years of data—to enable validation of flood frequency estimates, and (2) minimal anthropogenic regulation, as the hydrologic models used in this study do not simulate human interventions such as reservoir operations or flow diversions. In addition to these criteria, we deliberately chose three watersheds that represent contrasting physiographic and climatic regions. The Lahave River in Nova Scotia was selected to represent coastal catchments influenced by maritime climate conditions; the Moyie River in British Columbia serves as a representative mountain watershed with a snowmelt-dominated regime; and the Black River in Ontario captures characteristics of the Great Lakes region, including wetland and lake-dominated hydrology.

2.1 Black River Watershed

The Black River near Washago is located in central Ontario, Canada, and drains a watershed of approximately 1,506 km^2 upstream of the Water Survey of Canada gauge 02EC002, which has over 100 years of streamflow records. The watershed is characterised by a humid continental climate, with cold, snowy winters and warm, humid summers. The climate is strongly influenced by the Great Lakes, which moderate seasonal temperatures and enhance precipitation (Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000). Hydrologically, the Black River is classified as a mixed-regime system, with a dominant nival (snowmelt) component; the nival fraction is approximately 0.7 (Burn and Whitfield, 2023)(Burn and Whitfield, 2023)(Burn and Whitfield, 2023)(Burn and Whitfield, 2023)(Burn and Whitfield, 2023)(Burn and Whitfield, 2023)(Burn and Whitfield, 2023)(Burn and Whitfield, 2023)(Burn and Whitfield, 2023)(Burn and Whitfield, 2023)(Burn and Whitfield, 2023)(Burn and Whitfield, 2023). The primary flood-generating mechanisms are spring snowmelt and rain-on-snow events, which have historically produced the highest flows (Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle

et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016). The watershed contains a high proportion of wetlands, which play a key role in sustaining groundwater recharge, and maintaining baseflow during dry periods.

2.2 Moyie River Watershed

The Moyie River is in southeastern British Columbia within the South BC Mountains region and drains an area of approximately 1,480 km² upstream of the Water Survey of Canada gauge 08NH006, which provides a 90-year observational record. The watershed lies within a complex mountainous terrain characterised by steep slopes, dense forest cover, and variable elevations. The climate is strongly influenced by elevation and orographic effects, resulting in a continental to subalpine climate with distinct seasonal patterns. Winters are cold with substantial snowfall, while summers are warm and relatively dry (Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000). The hydrologic regime is snowmelt-dominated, with peak flows typically occurring in late spring to early summer (May–June). Rain-on-snow events in late winter or early spring also contribute to flood risk, particularly when rapid warming occurs on an established snowpack (Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016).

2.3 Lahave River Watershed

The Lahave River at West Northfield is in Nova Scotia and drains an area of approximately 1,250 km² upstream of the Water Survey of Canada gauge 01EF001, providing 108 years of continuous daily flow records. The watershed is situated within a coastal, lowland physiographic region and is characterised by rolling hills, mixed forest, and numerous lakes and wetlands. The region experiences a humid maritime climate, with relatively mild winters, cool summers, and high precipitation distributed evenly throughout the year. This climate is strongly influenced by proximity to the Atlantic Ocean, which moderates seasonal temperature fluctuations and contributes to frequent rainfall. The area is also regularly affected by storms, including tropical cyclones and hurricanes, due to its exposure to Atlantic storm tracks (Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000)(Whitfield and Cannon, 2000). Hydrologically, the Lahave River exhibits a rainfall-dominated regime, with flood events occurring year-round, but most frequently in the fall and winter when intense rainfall coincides with saturated soil conditions. Rain-on-snow events may also contribute to elevated flows during transitional seasons; however, snow accumulation is generally lower than in inland or higher-elevation watersheds

(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016)(Buttle et al., 2016).

~~During a review of the annual maximum flow series, we identified an extreme event in 1956 that appeared as a statistical outlier in the historical record. Further investigation using provincial archives confirmed that this event was caused by an ice jam flood, a process not represented in the HBV-EC model used here. Because the hydrologic model does not simulate ice-related hydraulic dynamics, inclusion of this event would have introduced bias into the flood frequency estimates. Consequently, the 1956 event was excluded from model calibration and analysis to preserve the consistency and physical relevance of the simulation-based framework.~~

Figure 1 shows the location of three study watersheds.

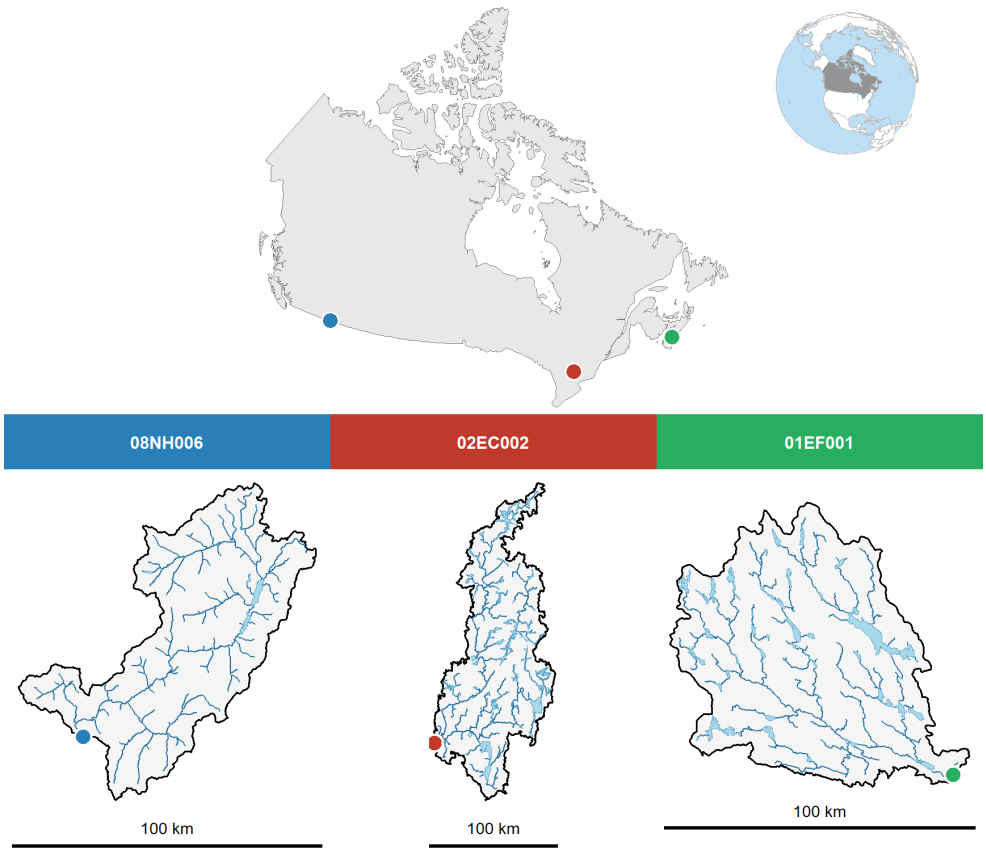


Figure 1: Geographic location of the three hydrological case studies analysed in this study. The figure shows the spatial distribution and watershed boundaries of the selected study areas.

275

3 Methodology

The UQ-flood framework proposed in this study integrates three key components: a process-based hydrological model, a residual error model (REM), and a stochastic weather generator (SWG). These components are used in conjunction to

quantify the model uncertainty in flood estimates. The framework consists of two main phases: 1) model setup and parameter estimation; and 2) long-term continuous simulation, as schematically illustrated in Fig. 2. The key steps are outlined as follows:

Phase 1: Model Setup and Parameter Estimation

1. Build a process-based hydrological model. The first step involves developing a hydrological model capable of simulating river discharge based on the physical processes governing the watershed. This model serves as the backbone for simulating runoff and streamflow under given climatic conditions.
2. Develop a residual error model (REM) based on annual maxima. The residual error model is developed to characterise and quantify uncertainties in hydrological predictions, specifically for extreme events. It is constructed using the annual maximum flow series, by analysing the differences between observed and simulated annual maxima. This approach captures the statistical properties of model errors—including their magnitude and variability at the scale most relevant for flood frequency analysis.
3. Set up a stochastic weather generator (SWG). The SWG is configured to generate synthetic weather data, including precipitation and maximum and minimum temperature, based on historical observations. This tool accounts for the variability in climatic conditions by preserving key statistical properties of precipitation and temperature such as marginal distributions, autocorrelation structures, and intermittency patterns.

Phase 2: Long-term Continuous Simulation

4. Generate long-term forcings using the SWG, which produces synthetic time series of daily meteorological variables (e.g., precipitation and max/min daily temperature) for an extended period, spanning greater than 1,000 years. This step ensures sufficient data for a robust non-parametric flood frequency analysis.
5. Simulate long-term continuous discharges using the hydrological model. The synthetic weather data generated by the SWG is used to drive the process-based hydrological model, simulating long-term streamflow for the watershed of interest.
6. Generate corresponding error realisations using the REM. The REM is applied to introduce stochastic error realisations into the simulated streamflow. This step adds probabilistic uncertainty to the hydrological model predictions, reflecting the inherent errors in model output.
7. Apply flood estimate uncertainty analysis. The resulting synthetic annual maxima, enhanced by uncertainty bounds, are ranked and used to compute flood frequency without assumptions or need of a specific parametric distribution. The annual exceedance probabilities are calculated using the Weibull plotting position formula, and the multi-realisation flood frequency curves illustrate the relationship between flood magnitude and return period, along with the associated uncertainties.

Figure 2 provides a schematic representation of the UQ-flood framework. The figure is divided into two phases: the Model Setup and Parameter Estimation phase (Steps 1–3) and the Long-Term Continuous Simulation phase (Steps 4–7). These steps collectively allow the computation of low-exceedance-probability flow magnitudes that explicitly account for weather

variability and uncertainty in hydrological modelling. This workflow ensures that uncertainties at all stages—model predictions, weather variability, and statistical error—are propagated into the final flood frequency analysis.

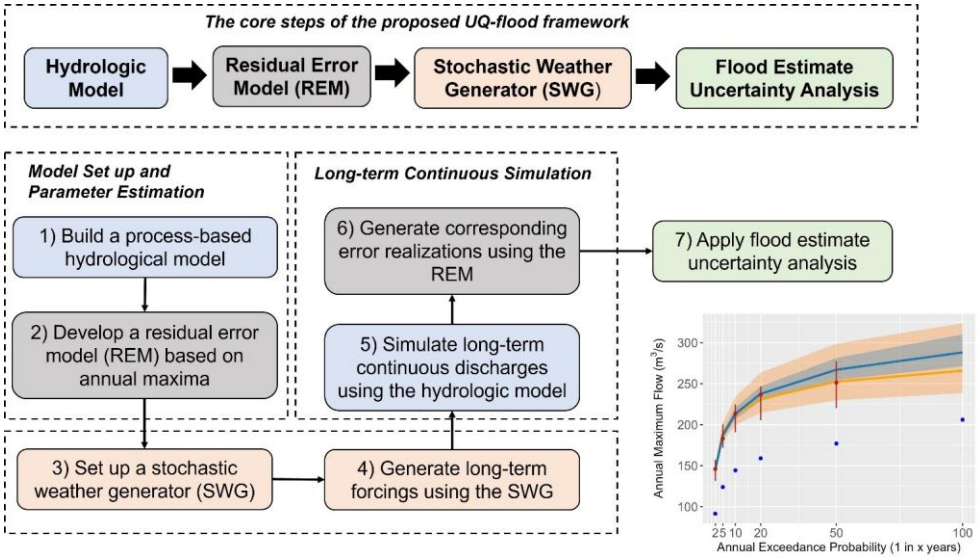


Figure 2: Conceptual diagram of the proposed UQ-flood (Residual Error Model - Flood Frequency Analysis) framework.

3.1 Hydrologic Model Development

The Hydrologiska Byråns Vattenbalansavdelning – Environment Canada (HBV-EC) model (Hamilton et al., 2000)(Hamilton et al., 2000)(Hamilton et al., 2000)(Hamilton et al., 2000)(Hamilton et al., 2000)(Hamilton et al., 2000)(Hamilton et al., 2000)(Hamilton et al., 2000)(Hamilton et al., 2000)(Hamilton et al., 2000)(Hamilton et al., 2000) was used to simulate watershed streamflow. A lumped representation of HBV-EC was implemented within the Raven Hydrological Modelling Framework version 3.8 (Craig et al., 2020)(Craig et al., 2020)(Craig et al., 2020)(Craig et al., 2020)(Craig et al., 2020)(Craig et al., 2020)(Craig et al., 2020)(Craig et al., 2020)(Craig et al., 2020)(Craig et al., 2020)(Craig et al., 2020). The HBV-EC model conceptualises catchments as a soil moisture zone atop two stacked buckets, representing upper and lower storage layers. Its structure includes three primary routines: a snow routine, soil routine, and runoff generation routine (Hamilton et al., 2000)(Hamilton et al., 2000)(Hamilton et al., 2000)(Hamilton et al., 2000)(Hamilton et al., 2000)(Hamilton et al., 2000)(Hamilton et al., 2000)(Hamilton et al., 2000)(Hamilton et al., 2000)(Hamilton et al., 2000)(Hamilton et al., 2000). The snow routine uses a classic degree-day approach to estimate snowmelt rates. The model partitions snowmelt and precipitation into

evapotranspiration and runoff based on water and energy balance principles. Runoff is divided into surface and subsurface flow, which is further partitioned between the upper and lower buckets according to soil moisture dynamics (Seibert and Vis, 2012)(Seibert and Vis, 2012)(Seibert and Vis, 2012)(Seibert and Vis, 2012)(Seibert and Vis, 2012)(Seibert and Vis, 2012)(Seibert and Vis, 2012)(Seibert and Vis, 2012)(Seibert and Vis, 2012)(Seibert and Vis, 2012)(Seibert and Vis, 2012). A lumped version of HBV-EC was used to simulate the average temporal dynamic of catchment behaviour without requiring detailed spatial information.

HBV-EC was selected due to its demonstrated strong performance in Canadian landscape (Beck et al., 2016; Stahl et al., 2008; Zégre et al., 2010)(Beck et al., 2016; Stahl et al., 2008; Zégre et al., 2010)(Beck et al., 2016; Stahl et al., 2008; Zégre et al., 2010)(Beck et al., 2016; Stahl et al., 2008; Zégre et al., 2010)(Beck et al., 2016; Stahl et al., 2008; Zégre et al., 2010)(Beck et al., 2016; Stahl et al., 2008; Zégre et al., 2010)(Beck et al., 2016; Stahl et al., 2008; Zégre et al., 2010)(Beck et al., 2016; Stahl et al., 2008; Zégre et al., 2010)(Beck et al., 2016; Stahl et al., 2008; Zégre et al., 2010)(Beck et al., 2016; Stahl et al., 2008; Zégre et al., 2010), its widespread use for flood simulations (Fischer and Schumann, 2021; Sikorska et al., 2015)(Fischer and Schumann, 2021; Sikorska et al., 2015)(Fischer and Schumann, 2021; Sikorska et al., 2015)(Fischer and Schumann, 2021; Sikorska et al., 2015)(Fischer and Schumann, 2021; Sikorska et al., 2015), its parsimony, ease of implementation, and its computational efficiency (Seibert and Bergström, 2022)(Seibert and Bergström, 2022)(Seibert and Bergström, 2022)(Seibert and Bergström, 2022)(Seibert and Bergström, 2022).

The model was run using daily time series of measured discharge, precipitation, maximum and minimum temperature, which were sourced from Environment Canada. The HBV-EC models were calibrated to a truncated discharge dataset using a combination of manual and automated approaches. Manual calibration followed a stepwise process: (1) parameters controlling the long-term water balance were adjusted to achieve volumetric flow simulations; (2) parameters governing snowpack process were calibrated to accurately simulate the freshet; (3) parameters influencing soil moisture dynamics were refined to improve recession and baseflow representation. Once an acceptable parameter set was identified through manual calibration, the Dynamically Dimensioned Search (DDS) algorithm (Tolson and Shoemaker, 2008)(Tolson and Shoemaker, 2008)(Tolson and Shoemaker, 2008)(Tolson and Shoemaker, 2008)(Tolson and Shoemaker, 2008)(Tolson and Shoemaker, 2008) was employed for further optimisation. The objective function minimised the daily discharge Kling-Gupta Efficiency (Gupta et al., 2009)(Gupta et al., 2009)(Gupta et al., 2009)(Gupta et al., 2009)(Gupta et al., 2009)(Gupta et al., 2009) often the last 30 years to ensure the best fit between simulated and observed hydrographs. The final set of parameters was evaluated against the long-term water balance, hydrographs, flow duration curves, and peak flows. 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.

3.2 Residual Error Modelling

The REM approach is a tool for propagating uncertainty into hydrological model predictions. It achieves this by incorporating samples of model error into deterministic predictions (Farmer and Vogel, 2016)(Farmer and Vogel, 2016)(Farmer and Vogel, 2016)(Farmer and Vogel, 2016)(Farmer and Vogel, 2016). REM combines deterministic predictions from process-based hydrologic models with stochastic components that characterise the uncertainties in model predictions (Montanari and Koutsoyiannis, 2012)(Montanari and Koutsoyiannis, 2012)(Montanari and Koutsoyiannis, 2012)(Montanari and Koutsoyiannis, 2012)(Montanari and Koutsoyiannis, 2012)(Montanari and Koutsoyiannis, 2012). This integration is accomplished by leveraging the statistical properties of the observed residual error, defined as the difference between observed and simulated streamflow (Vogel, 2017)(Vogel, 2017)(Vogel, 2017)(Vogel, 2017)(Vogel, 2017)(Vogel, 2017). By effectively characterising and modelling these residuals, the REM approach enables the generation of probabilistic streamflow estimates, accounting for uncertainties inherent in deterministic model outputs (Honti et al., 2013)(Honti et al., 2013)(Honti et al., 2013)(Honti et al., 2013)(Honti et al., 2013)(Honti et al., 2013). 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).

In this study, the residual error model is developed for the annual maxima, aligning with the requirements of flood frequency analysis. While much of the literature has focused on REM applications at the daily streamflow—where model errors are often non-normal, heteroscedastic, and autocorrelated —our approach focuses on modelling these properties for annual maxima. The statistical properties of annual maxima residuals require careful treatment to avoid biased estimates of extremes.

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. To address the heteroscedasticity—where errors tend to increase with larger streamflow—we apply a Box-Cox transformation (Box and Cox, 1964)(Box and Cox, 1964)(Box and Cox, 1964)(Box and Cox, 1964)(Box and Cox, 1964)(Box and Cox, 1964) with a power parameter fixed at 0.2, following the recommendations of McInerney et al. (2017)(2017)(2017)(2017)(2017)(2017). ~~To account for autocorrelation in the error time series—where errors exhibit similar sign and magnitude over consecutive time steps—we fit a first-order autoregressive AR(1) model (Evin et al., 2014; McInerney et al., 2017)(Evin et al., 2014; McInerney et al., 2017)(Evin et al., 2014; McInerney et al., 2017)(Evin et al., 2014; McInerney et al., 2017)(Evin et al., 2014; McInerney et al., 2017).~~ The innovation, or random component of the REM, follows a Gaussian distribution with a flow-dependent residual mean (Li et al., 2016)(Li et al., 2016)(Li et al., 2016)(Li et al., 2016)(Li et al., 2016)(Li et al., 2016). The parameters of REM were inferred using the method-of-moments (Hazelton, 2011)(Hazelton, 2011)(Hazelton, 2011)(Hazelton, 2011)(Hazelton, 2011)(Hazelton, 2011). For the detailed REM formulation, including the

model equations, readers are referred to Hunter et al. (2021)(2021)(2021)(2021)(2021)(2021), and Romero-Cuellar et al. (2024b)(2024b)(2024b)(2024b)(2024b)(2024b).

Formatted: Spanish (Spain)

Formatted: Spanish (Spain)

3.3 Stochastic Weather Generator (SWG)

In this study, we employed CoSMoS (Complete Stochastic Modelling Solution) (Papalexiou, 2018)(Papalexiou, 2018)(Papalexiou, 2018)(Papalexiou, 2018)(Papalexiou, 2018)(Papalexiou, 2018) as the stochastic weather generator (SWG). CoSMoS is an R package designed for the stochastic simulation of univariate, multivariate, or random field non-Gaussian time series. It implements a flexible framework that unified, extends, and improves the theory of stochastic modelling for hydroclimatic process, preserving key statistical characteristics such as marginal distributions, autocorrelation structures, and intermittency (Papalexiou, 2018)(Papalexiou, 2018)(Papalexiou, 2018)(Papalexiou, 2018)(Papalexiou, 2018)(Papalexiou, 2018). The core assumption underlying CoSMoS is that any process can be generated by transforming a parent Gaussian process.

For this study, we inferred the parameters of the generalized gamma distribution, autocorrelation structure, and dry-period distribution of daily precipitation using 30 years of gauge-based observations, with statistics estimated separately for each calendar month. Similarly, for maximum and minimum temperatures, we estimated the monthly marginal distributions and autocorrelation structures. Using the inferred parameters, CoSMoS was used to generate long-term (e.g., 10,000 synthetic years) daily time series of precipitation, maximum temperature, and minimum temperature. These synthetic meteorological forcings were then used to drive the hydrologic model, enabling the simulation of annual maximum flows and the estimation of flood frequency distributions.

To adapt CoSMoS for flood frequency purposes, we implemented two key modifications: (1) a bootstrap-based smoothing strategy across monthly boundaries to address abrupt discontinuities in statistical properties, and (2) the introduction of physical constraints on extreme precipitation to prevent unrealistically long-tailed behaviour in the generalized gamma distribution. These constraints help avoid infinite or implausible values in simulated extremes. Full details of these modifications are provided in the Supplements.

3.4 ~~Simulation~~Process-based Flood Frequency Analysis

Following calibration of the hydrologic model, fitting of the stochastic weather generator (SWG), and inference of the residual error model (REM) parameters, the process-based flood-frequency analysis proceeds in two stages. First, the SWG generates 1,000 years of synthetic meteorological forcings, which are transformed into streamflow using the calibrated hydrologic model. The annual maximum flow is then extracted from each synthetic year, producing a deterministic sequence of 1,000 simulated annual maxima.

Second, the REM is applied to each simulated annual maximum to generate 1,000 residual-error realizations. This procedure produces 1,000 REM-corrected annual-maximum sequences, each containing 1,000 years of flood values. For each corrected

sequence, the annual maxima are ranked and assigned annual exceedance probabilities using the Weibull plotting-position formula. Flood quantiles are then estimated for the selected return periods, resulting in an ensemble of 1,000 flood-frequency curves.

The central UQ-flood estimate at each return period is obtained from the median of the ensemble, while the 95% predictive interval is defined by the 2.5th and 97.5th percentiles. This interval represents variability across the residual-error realizations, conditional on the synthetic meteorological sequence, calibrated hydrologic model, fitted SWG, and selected REM formulation.

~~Following the calibration of hydrologic model parameters, inference of residual error model (REM) parameters, and fitting of the stochastic weather generator (SWG), the methodology proceeds with simulation-based flood frequency analysis. For each realisation of the REM, a synthetic sequence of 10,000 years of annual maximum streamflow is generated using climate inputs from the SWG and the calibrated hydrologic model. This yields one estimate of the flood quantile distribution. The process is repeated 1,000 times, each with a different REM realisation, resulting in 1,000 samples of flood quantile estimates. These samples are then used to construct multi-realisation flood frequency curves, illustrating the relationship between flood magnitude and annual exceedance probability, along with associated uncertainty intervals. Return periods are computed using the Weibull plotting position formula~~.

3.5 At-site Statistical Flood Frequency Analysis

A Bayesian at-site flood-frequency analysis based on the Generalized Extreme Value (GEV) distribution was used as a secondary statistical benchmark for the central flood-quantile estimates. Long-record empirical annual-maximum-flow estimates remained the primary observational benchmark. The GEV distribution was selected because it is grounded in extreme-value theory and is widely used to model annual maximum flows in hydrologic applications (Coles, 2001a; Jenkinson, 1955a; Vogel and Wilson, 1996a; Zaghloul et al., 2020a)~~.~~

The GEV distribution was fitted to the observed annual maxima under two data configurations: the full available record and the same 30-year record used to develop UQ-flood. Bayesian inference was used to characterize uncertainty in the GEV parameters and corresponding flood quantiles (Liang et al., 2012a; Parkes and Demeritt, 2016a). Markov chain Monte Carlo sampling was applied to draw from the joint posterior distribution of the GEV location, scale, and shape parameters. Posterior mean flood quantiles were used as the central estimates, and the 2.5th and 97.5th percentiles of the posterior flood-quantile distributions defined the 95% credible intervals.

~~To provide a benchmark for evaluating the proposed simulation-based approach, we applied an at-site statistical flood frequency analysis (FFA) using the Generalized Extreme Value (GEV) distribution (Jenkinson, 1955)(Jenkinson, 1955)(Jenkinson, 1955)(Jenkinson, 1955). The GEV distribution is the standard model for fitting annual maximum flow~~

~~(AMF) data and is widely used in hydrologic design due to its foundation in extreme value theory and its ability to represent light, moderate, and heavy tails~~ (Coles, 2001; Vogel and Wilson, 1996; Zaghloul et al., 2020)(Coles, 2001; Vogel and Wilson, 1996; Zaghloul et al., 2020)(Coles, 2001; Vogel and Wilson, 1996; Zaghloul et al., 2020)(Coles, 2001; Vogel and Wilson, 1996; Zaghloul et al., 2020)~~.~~

~~In this study, we fit the GEV distribution to observed annual maxima using a Bayesian inference framework to quantify parameter uncertainty and generate credible intervals for flood quantiles~~ (Parkes and Demeritt, 2016)(Parkes and Demeritt, 2016)~~. This approach offers a probabilistic characterisation of uncertainty, in contrast to classical methods such as maximum likelihood estimation~~ (Liang et al., 2012b)(Liang et al., 2012b)~~. Specifically, we implemented Markov Chain Monte Carlo (MCMC) sampling to draw from the joint posterior distribution of the GEV parameters (location, scale, and shape). Posterior mean estimates were used for point estimation, whereas the posterior predictive distributions were used to estimate flood quantiles and their 95% credible intervals.~~

The analysis was ~~performed~~ implemented using the extRemes R package (Gilleland and Katz, 2016)(Gilleland and Katz, 2016)~~.~~ 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 diagnostics did not indicate important concerns with the fitted GEV models. The complete diagnostic results are presented in Figs. S2–S7 of the Supplement. ~~which provides a Bayesian GEV fitting module with built-in MCMC capabilities. Convergence diagnostics—including trace plots, effective sample size, and Gelman-Rubin statistics—were used to assess sampling reliability and ensure robustness of the posterior estimates.~~

Central flood-quantile estimates were evaluated at ~~This statistical approach serves as a benchmark method against which the performance of the REM-based flood frequency estimates was compared, particularly in terms of uncertainty characterisation.~~

~~We compare simulated to observed flood quantiles corresponding to~~ return periods of $T = 2,5,10,20,50,100$ years. For each return period, the interval width was calculated as:~~by computing the uncertainty range (UR), defined as the difference between the 2.5th and 97.5th percentiles of the ensemble outcomes (q). The uncertainty range is calculated at each computational time step t in the hydrographs, as expressed in the following Eq. (1)~~:

$$UR_{Tt} = q_{Tt}^{97.5} - q_{Tt}^{2.5}, t = 1, \ldots, T \quad (1)$$

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.

The GEV and UQ-flood intervals represent different conditional uncertainty components. The Bayesian GEV credible intervals primarily characterize sampling and parameter uncertainty conditional on the observed annual-maximum sample and selected statistical model. In contrast, the UQ-flood intervals represent predictive variability across residual-error realizations conditional on the calibrated hydrologic model, fitted stochastic weather generator, and selected REM formulation. Consequently, their widths were not compared as equivalent measures of precision or performance. Instead, the GEV results were used to provide statistical context for evaluating the central UQ-flood estimates against the long-record observational benchmarks.

Formatted: Justified

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.

3.7 Numerical Experiment Configuration

To evaluate the proposed UQ-flood framework, we designed a numerical experiment focused on two questions: (1) whether the central flood-quantile estimates obtained using 30 years of observations are consistent with long-record observational

benchmarks, and (2) how explicitly accounting for hydrologic-model error through the residual error model affects process-based flood estimates.

Long-record empirical annual-maximum-flow estimates were used as the primary benchmark. A Bayesian GEV analysis was retained as a secondary statistical benchmark for the central flood-quantile estimates. Because the UQ-flood and GEV intervals characterize different uncertainty components, their widths were not compared as equivalent measures of precision or performance. ~~To evaluate the performance and added value of the proposed simulation-based flood frequency analysis (UQ-flood), we designed a numerical experiment that compares it against a benchmark statistical FFA approach based on the Generalized Extreme Value (GEV) distribution, as described in Sect. 3.5. The objective of this experiment is to assess the ability of each method to produce robust flood quantile estimates and quantify associated uncertainty under different data availability conditions and modelling assumptions~~.

The evaluation consists of two main components:

1. Evaluation under different observational record lengths ~~Comparison under Varying Record Lengths:~~

UQ-flood was developed using 30 years of observed meteorological and streamflow data. Its central flood-quantile estimates were evaluated against:

- Empirical annual-maximum-flow estimates derived from the full available observational record.
- Bayesian GEV estimates fitted to the full observational record.
- Bayesian GEV estimates fitted to the same 30-year annual-maximum series used to develop UQ-flood.

This experimental design evaluates whether UQ-flood estimates developed from a limited observational period remain consistent with the long-record empirical benchmark. The long- and short-record GEV analyses provide statistical context for interpreting the effects of observational record length on the central flood-quantile estimates. The comparison does not imply that the UQ-flood and GEV uncertainty intervals represent the same uncertainty components. ~~We compared the UQ-flood method—applied using only 30 years of observed meteorological and hydrological data—to the statistical GEV-based FFA under two observational record lengths:~~

- ~~A long-record scenario using the full historical record (≥100 years), and~~
- ~~A short-record scenario using the same 30 years of observed annual maximum flows.~~

~~This setup allows us to isolate and evaluate the impact of sampling uncertainty on statistical FFA, while keeping the simulation-based method fixed in its data requirements. By comparing these scenarios, we assess whether the UQ-flood can provide reliable flood estimates even when observational records are limited, where statistical methods tend to suffer from high uncertainty~~.

2. Evaluation of ~~the Continuous~~process-based ~~Simulation~~ simulation ~~Approach Without~~ without residual ~~Error~~ error ~~Modelling~~modelling:

Formatted: List Paragraph, Bulleted + Level: 1 + Aligned at: 0.63 cm + Indent at: 1.27 cm

To assess the contribution of the REM, we conducted an additional experiment in which flood-frequency estimates were derived directly from annual maximum flows generated by the hydrologic model and stochastic meteorological inputs, without residual-error correction. This configuration is referred to as the process-based simulation without REM (det-SIM).

Comparison of these uncorrected estimates with the UQ-flood estimates and long-record observational benchmark was used to identify systematic discrepancies in the deterministic process-based simulations and evaluate the extent to which the REM corrects these discrepancies at the annual-maximum scale. ~~To evaluate the role of the residual error model (REM) in improving simulation based flood estimates, we conducted an additional experiment in which the continuous simulation method was applied without the error model correction. In this case, flood frequency estimates were derived directly from the annual maximum flows produced by the hydrologic model driven by stochastic meteorological inputs (referred to as AMF_{sim}). This configuration helps assess the systematic bias and underestimation issues that may arise when simulation-based FFA is used without explicitly accounting for model errors.~~

Together, these experiments evaluate the consistency of the central UQ-flood estimates with long-record observations and isolate the contribution of residual-error modelling to process-based flood-frequency estimation. ~~Together, these experiments provide a comprehensive assessment of the REM framework's performance relative to traditional methods and explore the consequences of omitting explicit error modelling in continuous simulation-based FFA~~.

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.

4 Results

This section ~~presents~~ evaluates the UQ-flood framework across three case-study~~ies~~ watersheds. ~~to evaluate the performance of the proposed UQ-flood framework for flood estimates.~~

Section 4.1 ~~provides~~ presents ~~an~~ ~~in-depth~~detailed assessment ~~of~~ for the Black River watershed, including comparisons with the long-record observational benchmark, GEV estimates based on the full and 30-year records, and the process-based simulation without REM. Section 4.2 applies the same evaluation to the Moyie River and Lahave River watersheds to examine whether the central UQ-flood estimates remain consistent with the observational benchmarks under contrasting hydrologic conditions. ~~where we compare UQ-flood with the traditional statistical approach based on the Generalized Extreme Value (GEV) distribution using both long and short observational records.~~

~~To assess the generalisability of our findings across different hydrological conditions, Section 4.2 applies the same analysis to two additional watersheds—Moyie River (BC) and Lahave River (NS)—focusing on the consistency of the results rather than a detailed breakdown. Together, these case studies demonstrate the utility of the UQ-flood framework under diverse conditions.~~

~~The uncertainty intervals compared here do not represent identical uncertainty decompositions. The GEV intervals represent parameter/posterior uncertainty conditional on the selected statistical model and observed annual maxima sample, whereas UQ-flood intervals represent predictive uncertainty induced by residual-error realisations conditional on the calibrated hydrologic model and fitted weather generator. Therefore, comparisons of interval width are interpreted in terms of decision informativeness and consistency with long-record benchmarks, not as evidence that one method captures all uncertainty more completely.~~

4.1 Flood Frequency Estimates ~~in~~ for the Black River Watershed

4.1.1 Evaluation against Observational and Statistical Benchmarks~~UQ-flood Compared to At-site Statistical FFA~~

Figure 3 compares the UQ-flood estimates with the long-record empirical annual-maximum-flow estimates, Bayesian GEV estimates fitted to the long and short observational records, and the process-based simulation without REM. The upper panel presents the GEV analysis based on the full 108-year record, whereas the lower panel presents the GEV analysis based on the same 30-year period used to develop UQ-flood. The UQ-flood results are identical in both panels and are repeated to facilitate comparison with the two GEV configurations.

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.

The short-record GEV analysis is more sensitive to extrapolation at high return periods, with its central estimate diverging progressively from the long-record observational benchmark. This result illustrates the influence of the selected 30-year annual-maximum sample on the fitted GEV distribution. In contrast, UQ-flood incorporates information from the meteorological record, the hydrologic model, and the annual-maximum residual relationship. The comparison therefore evaluates whether this process-based modelling chain produces central flood-quantile estimates consistent with the long-record benchmark, rather than whether its uncertainty interval is narrower than the GEV interval.

The UQ-flood and GEV uncertainty intervals should be interpreted separately because they represent different conditional uncertainty components. The UQ-flood interval describes predictive variability across residual-error realizations, conditional on the calibrated hydrologic model, fitted weather generator, selected REM structure, and available calibration data. The Bayesian GEV interval primarily represents sampling and parameter uncertainty conditional on the observed annual-maximum sample and selected statistical model. Consequently, differences in interval width are not interpreted as evidence of greater precision or superior performance. The UQ-flood interval has the same width in the upper and lower panels; its apparently different visual width results solely from the broader y-axis scale in the lower panel.

Finally, the process-based simulation without REM systematically underestimates the long-record annual-maximum estimates. Applying the REM shifts the process-based flood-frequency curve toward the observational benchmark, demonstrating the importance of representing hydrologic-model discrepancies at the annual-maximum scale. Thus, the principal result for Black River is that the REM corrects systematic bias in the uncorrected process-based simulation and produces central flood-quantile estimates that remain consistent with the long-record observations.

~~The comparison of flood frequency estimates for the Black River watershed in Fig. 3 highlights notable differences between the traditional statistical approach using the GEV distribution and the proposed UQ-flood. The UQ-flood framework, which relies on only 30 years of precipitation and streamflow records, produces flood quantile estimates that align more closely with the observed annual maximum flows and their associated confidence interval for discharge quantiles, particularly for high return periods such as the 50- and 100-year floods. In contrast, the GEV-based estimates—derived from the full 108-year record—tend to underestimate extreme flows relative to the observational bounds (Fig. 3, top panel).~~

~~A key advantage of the UQ-flood framework lies in its ability to produce more constrained and informative uncertainty estimates. Across the full range of annual exceedance probabilities, UQ-flood yields narrower uncertainty bands compared to the GEV approach, demonstrating improved precision (Fig. 3, top panel). Moreover, the simulated annual maximum flows generated by UQ-flood closely follow the central tendency of the 108-years observed flood distribution, despite being calibrated with only 30 years of data. This close agreement indicates that the simulation-based approach, when corrected using an error model, effectively captures both the magnitude and variability of extreme hydrologic events.~~

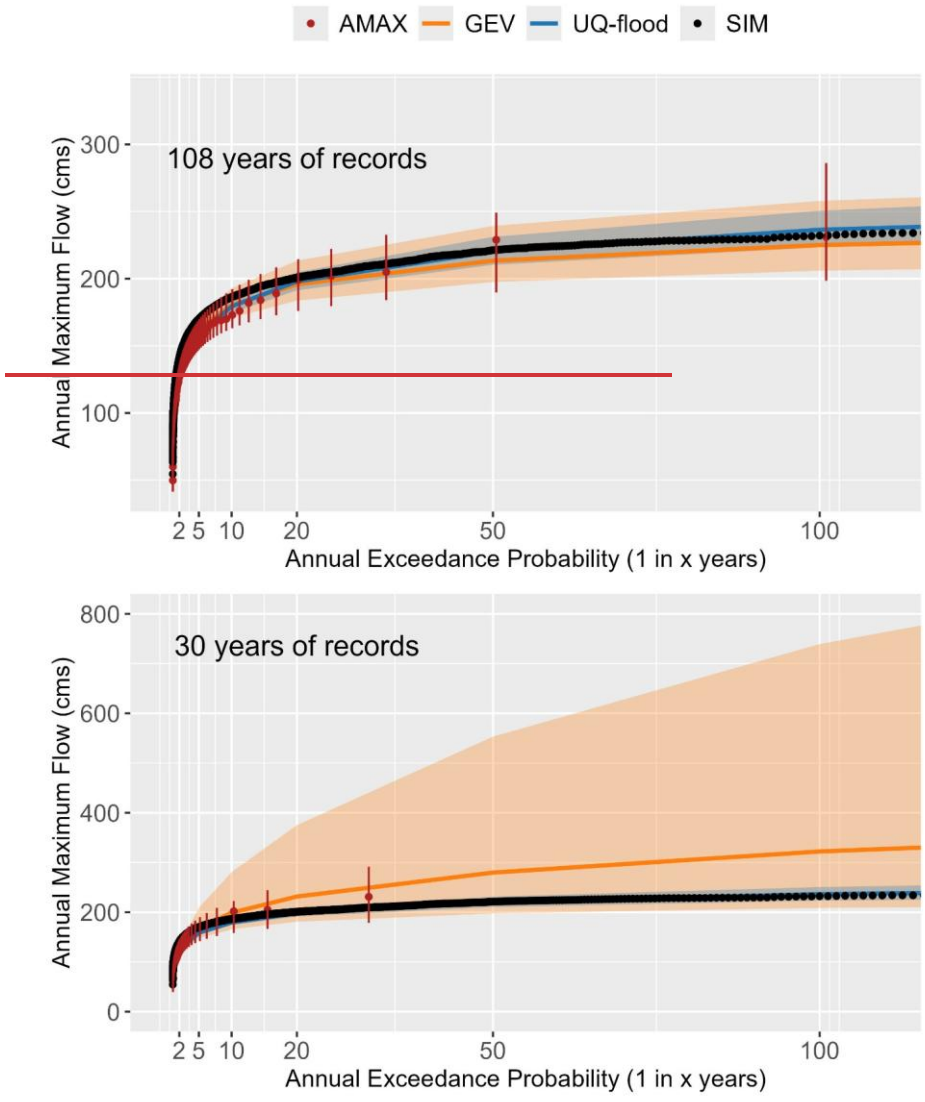
~~The bottom panel of Fig. 3 presents a critical and commonly encountered scenario in hydrology: applying at-site statistical FFA to short observational records (e.g., 30 years). In this case, the GEV-based estimates exhibit wide and often impractical uncertainty intervals. In contrast, the UQ-flood framework continues to provide narrow flood quantile estimates, despite relying on the same short record.~~

~~The differences in uncertainty ranges between the GEV and UQ-flood methods vary across return periods (Fig. 3). When using long observational records, the average difference in the width of the uncertainty intervals between GEV and UQ-flood is approximately 14 m^3s^{-1}. This difference increases substantially under short-record conditions, with the GEV method exhibiting an average uncertainty width approximately 198 m^3s^{-1} greater than that of UQ-flood. The orange GEV~~

~~uncertainty envelope is considerably larger for the 30-year record than for the 108-year record, widening dramatically at high return periods and reaching values several times larger than the estimated quantiles, which illustrates the profound sensitivity of GEV-based flood frequency estimates to limited data availability (Figure 3). The resulting uncertainty ranges undermine the reliability of statistical estimates, thereby limiting their usefulness for practical decision-making.~~

~~For example, consider the 100-year flood estimate for the Black River watershed. The observed 100-year flood is approximately 230 m^3s^{-1}. When applying the GEV distribution to short records, the estimated 100-year flood was 322 m^3s^{-1}, with a wide uncertainty range of 531 m^3s^{-1} — offering limited practical utility. In contrast, the UQ-flood framework estimated the 100-year flood as 236 m^3s^{-1}, with a much narrower uncertainty range of just 27 m^3s^{-1}. This result demonstrates the UQ-flood framework's ability to produce estimates that are not only closer to observed values but also more constrained, providing more reliable and actionable information for infrastructure design and flood risk planning—even when using only 30 years of data from a single site.~~

~~A critical consideration in interpreting flood frequency estimates is the treatment of uncertainty in confidence interval for discharge quantiles. Unlike traditional confidence intervals, which assess the proportion of observed data points falling within a given range, plotting positions are themselves derived from estimated exceedance probabilities—typically using formulas such as the Weibull method. These estimates are inherently uncertain, particularly when based on short observational records. Consequently, it is not statistically valid to evaluate uncertainty by simply counting the percentage of points within a 95% confidence interval on the flood frequency plot (Renard et al., 2013). To address this, the UQ-flood framework incorporates confidence interval for discharge quantiles (Sect. 3.6). This approach provides a more rigorous and informative representation of uncertainty in flood quantile estimation.~~


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

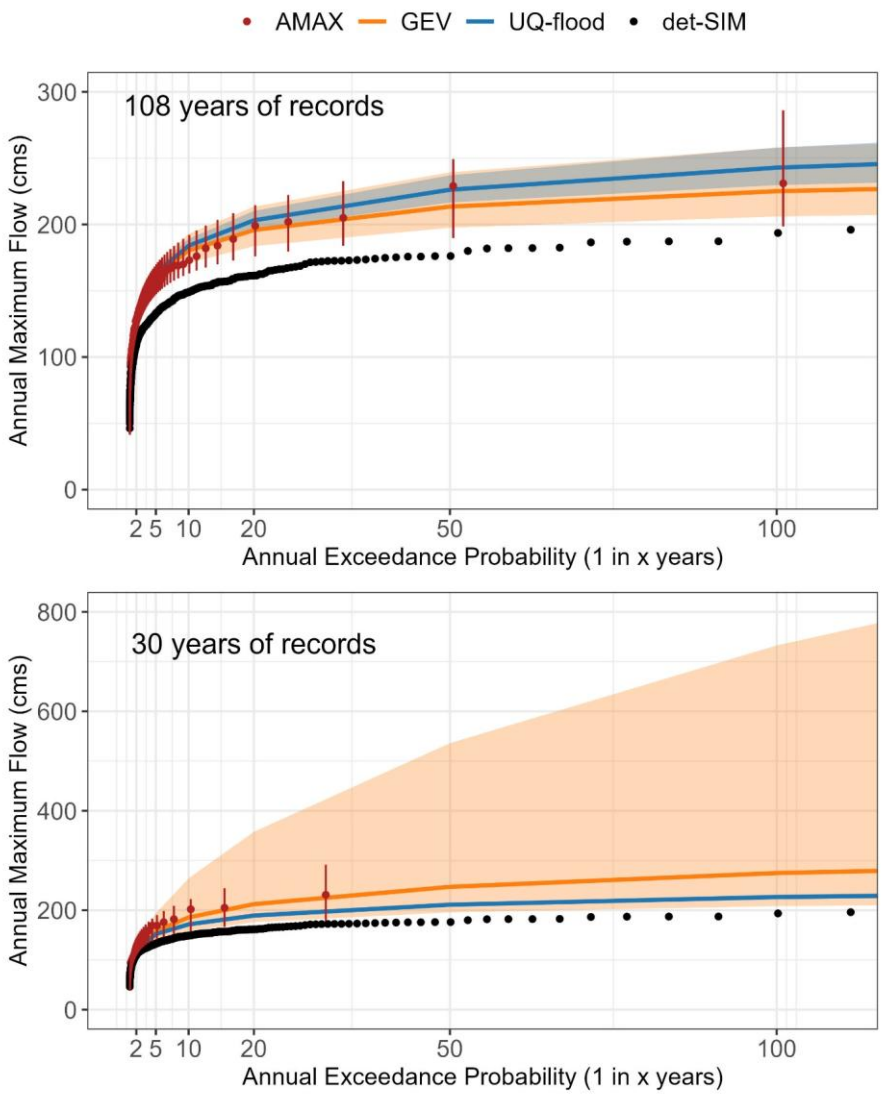

AMAX
GEV
UQ-flood
det-SIM
108 years of records
Annual Maximum Flow (cms)
300
200
100
0
2 5 10 20 50 100
Annual Exceedance Probability (1 in x years)
30 years of records
Annual Maximum Flow (cms)
800
600
400
200
0
2 5 10 20 50 100
Annual Exceedance Probability (1 in x years)

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.

4.1.2 Simulation-length sensitivity analysis

Formatted: Heading 3

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 $m^3 s^{-1}$ at 100 years to approximately 1.7–1.8 $m^3 s^{-1}$ 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.

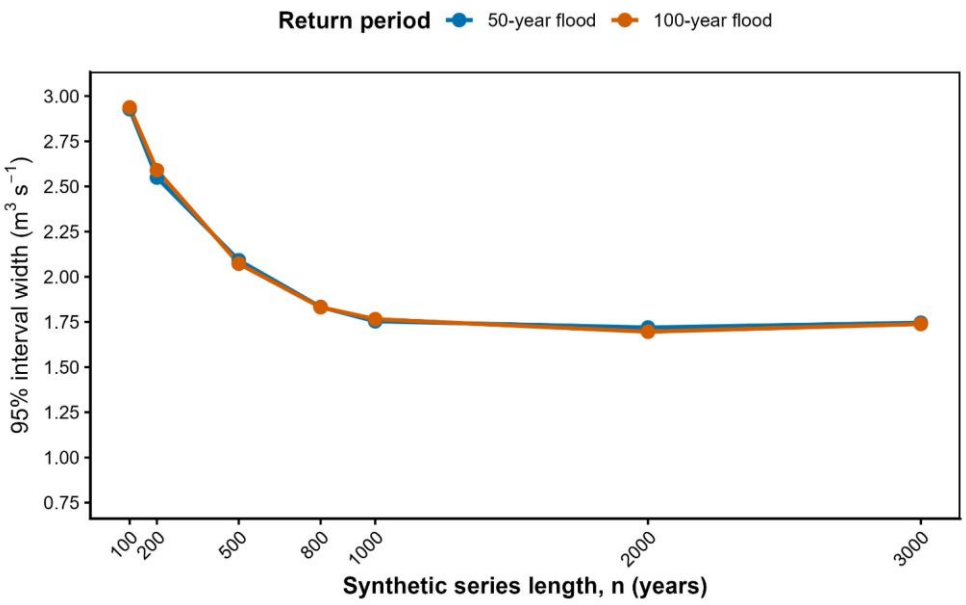


Figure 4: Sensitivity of the UQ-flood 95% interval width to simulated-series length for the 50- and 100-year flood quantiles in the Black River watershed. Interval width is defined as the difference between the 97.5th and 2.5th quantiles across the predictive ensemble.

Formatted: English (Canada)

4.2 Application to Additional Watersheds

The UQ-flood framework was further evaluated for the Moyie River and Lahave River watersheds, shown in the left and right panels of Fig. 5, respectively. As in the Black River analysis, the long-record empirical annual-maximum estimates provide the primary benchmark. Bayesian GEV estimates based on the full and 30-year records provide secondary statistical benchmarks, while the process-based simulation without REM is included to assess the contribution of residual-error correction. UQ-flood was developed using 30 years of meteorological and streamflow observations in both watersheds.

The central UQ-flood estimates remain within the confidence intervals for the long-record discharge quantiles across the evaluated return periods. However, neither UQ-flood nor the long-record GEV is uniformly closest to the empirical

estimates. For Lahave River, UQ-flood is closer to the empirical estimate at the 50-year return period, whereas the long-record GEV is closer at the 100-year return period. For Moyie River, the two approaches produce similar estimates at lower return periods, while the long-record GEV is closer to the empirical estimate at the 50-year return period. These results are notable because UQ-flood uses only 30 years of observations, whereas the long-record GEV uses the full available annual-maximum series. The corresponding numerical estimates are provided in Table S1.

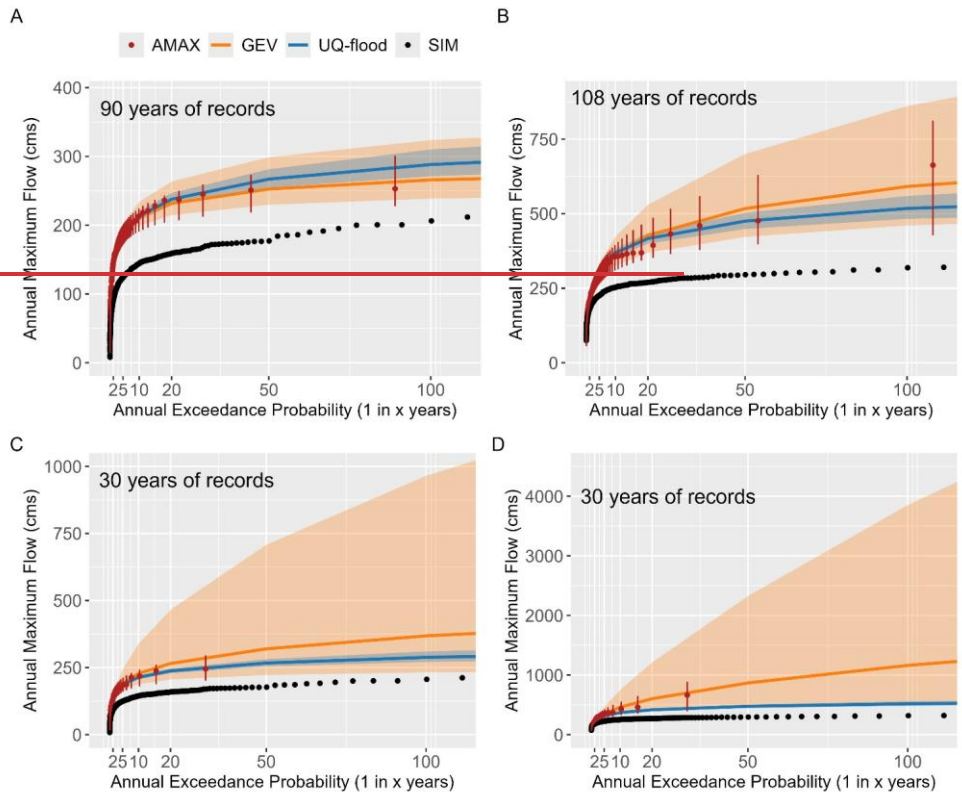
The UQ-flood and GEV uncertainty intervals represent different conditional uncertainty components and their widths are therefore not compared as equivalent measures of precision or performance. Instead, the evaluation focuses on whether the central estimates are consistent with the long-record observational benchmarks. The confidence intervals for the empirical discharge quantiles are particularly important at high return periods, where the observed plotting positions are affected by greater sampling uncertainty.

For both watersheds, the process-based simulations without REM underestimate the long-record empirical flood quantiles. Applying the REM shifts the process-based estimates toward the observational benchmarks, demonstrating the importance of representing hydrologic-model discrepancies at the annual-maximum scale. Overall, the additional case studies show that UQ-flood produces central estimates consistent with the observational confidence intervals and that the REM corrects systematic biases in the uncorrected process-based simulations. However, the relative proximity of UQ-flood and GEV to the empirical estimates varies by watershed and return period, and neither approach is uniformly closest across all cases.~~To assess the generalisability of the proposed framework, we applied the same analysis to two additional watersheds: the Moyie River (left panels in Fig. 4) and the Lahave River (right panels in Fig. 4). As in the Black River analysis, flood frequency estimates were generated using a confidence interval for discharge quantiles, traditional statistical approach based on the Generalized Extreme Value (GEV) distribution, and the simulation-based method incorporating a residual error model (UQ-flood). For the GEV method, both long-term and short-term observational records were used, while the UQ-flood framework consistently relied on only 30 years of data. The confidence interval for discharge quantiles results with sampling uncertainty use the long records.~~

~~The results for both watersheds are consistent with those observed in the Black River. In the top panels (A and B), where GEV is applied using long records, and in the bottom panels (C and D), where GEV uses only 30 years of data, UQ-flood consistently produces narrower uncertainty intervals while remaining consistent with long-record empirical benchmarks. These narrower bands are more informative and actionable for risk-based decision-making. In contrast, the GEV-based estimates—especially when derived from short records—exhibit substantially wider uncertainty bounds, limiting their practical utility in planning and design applications.~~

~~Across all panels and watersheds, the simulated annual maximum flows without error correction (AMF_{sim}, shown as black points) consistently underestimate flood magnitudes. This systematic bias illustrates a key limitation of using continuous simulation in isolation. The integration of a residual error model in the UQ-flood framework is essential to correct any model~~

~~bias and to propagate predictive uncertainty into flood quantile estimates. As a result, UQ-flood not only improves the reliability of simulation-based flood frequency analysis but also ensures that uncertainty estimates remain robust and interpretable, even in data-scarce environments~~.


A
AMAX
GEV
UQ-flood
SIM
90 years of records
Annual Maximum Flow (cms)
Annual Exceedance Probability (1 in x years)
B
108 years of records
C
30 years of records
D
30 years of records

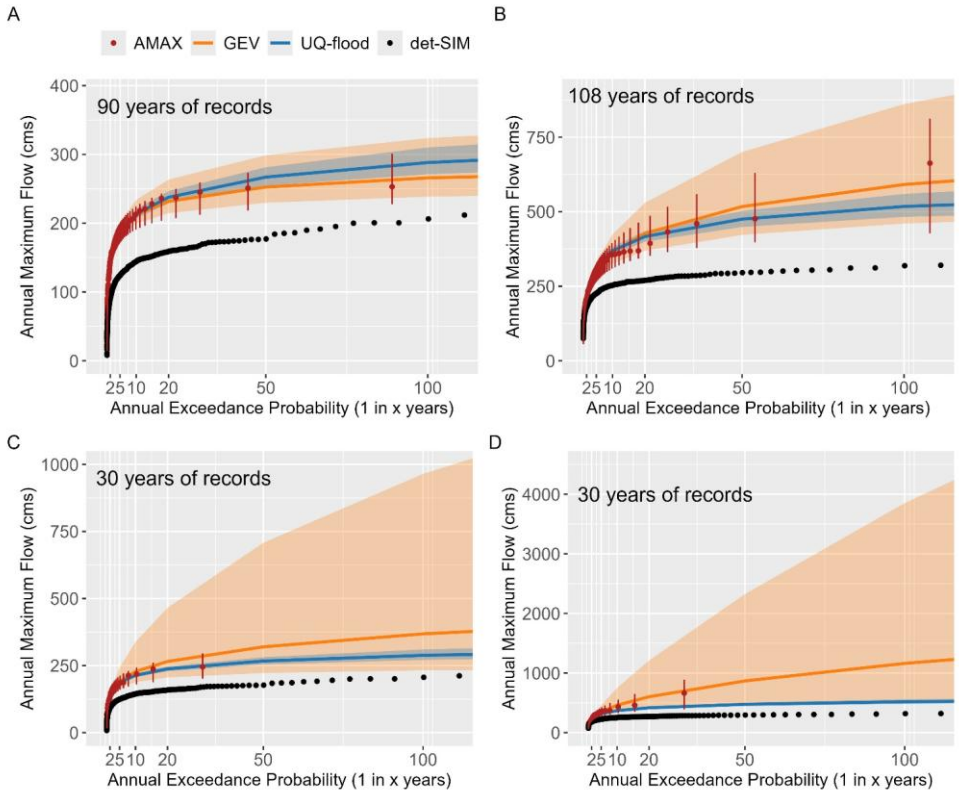


Figure 45: Flood frequency estimates and associated 95% uncertainty intervals for the Moyie River (left panels: A and C) and Lahave River (right panels: B and D) derived using the statistical GEV approach and the UQ-flood framework. The observed annual maximum discharge (AMAX; red) distribution, shown with confidence intervals for discharge quantiles, and the deterministic simulation output (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

5 Discussion

The findings of this study should be interpreted in the context of the short-sample problem in flood frequency analysis (FFA). As discussed in the Introduction, limited annual maximum flow records can lead to substantial uncertainty in rare flood quantile estimates, with confidence intervals that may be too wide to support practical design and planning decisions. Existing approaches address this limitation by incorporating additional information from historical or paleoflood records, leveraging spatial information through regional flood frequency analysis, or generating synthetic samples through continuous simulation. UQ-flood contributes to the latter approach by combining process-based hydrologic simulation with explicit bias correction and residual uncertainty propagation at the annual maximum scale. Rather than replacing historical-data or regional approaches, the framework offers a complementary, process-consistent alternative for situations in which local records are short and uncertainty intervals from conventional statistical FFA become impractically wide.~~The results of this study should be interpreted in the context of the broader short-sample problem in flood frequency analysis. As discussed in the Introduction, limited annual maximum flow records can produce uncertainty intervals for rare flood quantiles that are too wide to support practical design and planning decisions. Existing strategies address this problem by adding information in time through historical or paleoflood data, in space through regional flood frequency analysis, or synthetically through continuous simulation. UQ-flood contributes to the third pathway by using process-based simulation to generate large flood samples, while explicitly correcting hydrologic model bias and propagating residual uncertainty at the annual maxima scale. In this sense, the framework is not intended to replace historical data or regional approaches, but to provide a complementary, process-consistent alternative when local records are short and uncertainty bounds from conventional statistical FFA become impractical.~~

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)(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., 2017a; Cameron et al., 2000a; Newman et al., 2021)(Arnaud et al., 2017a; Cameron et al., 2000a; 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.

~~Uncertainty is a fundamental aspect of flood frequency analysis (FFA), yet it has historically received limited attention in practical hydrology~~ (Blöschl et al., 2019; Merz et al., 2015; Pappenberger and Beven, 2006)(Blöschl et al., 2019; Merz et al., 2015; Pappenberger and Beven, 2006)~~. Traditional statistical methods—particularly those based on the Generalized Extreme Value (GEV) distribution—have long served as the foundation for regulatory and engineering applications~~ (Vogel and Wilson, 1996)(Vogel and Wilson, 1996)~~. However, these approaches can be sensitive to sample size, rely on parametric assumptions, and often provide point estimates without a formal treatment of uncertainty~~ (Vidrio-Sahagún et al., 2025)(Vidrio-Sahagún et al., 2025)~~. This can be especially limiting when working with short streamflow records, which are increasingly common in contemporary hydrologic design practice.~~

The UQ-flood framework presented in this study offers a complementary approach to existing methods by integrating process-based hydrologic simulation with a residual error model (REM) and a stochastic weather generator. This integration supports a shift towards process~~simulation~~-based uncertainty quantification that is consistent with hydrologic processes. Rather than replacing existing methods, UQ-flood aims to enhance uncertainty characterisation by propagating uncertainty from model inputs through to flood quantile estimates. In this context, incorporating error models into continuous simulation frameworks may help improve bias correction and support more transparent flood risk estimation.

5.2 ~~Advancing Uncertainty Quantification in FFA~~Annual-Maximum Residual Error Modelling in Process-Based FFA

A key contribution of UQ-flood is the integration of a process-based hydrologic model, a stochastic weather generator, and a residual error model developed specifically for annual maximum flows. Although these components have been used separately in previous flood-frequency studies, their integration allows systematic hydrologic-model discrepancies to be corrected at the same scale as the target design variable.

The annual-maximum formulation addresses an important mismatch in process-based FFA. Hydrologic models are commonly calibrated using daily streamflow metrics that emphasize overall hydrograph behaviour, water balance, and seasonal dynamics. However, satisfactory daily-flow performance does not guarantee unbiased simulation of annual flood peaks, which may be particularly sensitive to antecedent conditions, threshold processes, snowmelt–rainfall interactions, and upper-tail behaviour. This limitation was evident in the simulations without the REM, which systematically underestimated flood quantiles across the study watersheds. By modelling the residual relationship between observed and simulated annual maxima, the REM corrects these systematic discrepancies and represents residual predictive variability at the scale directly relevant to flood-frequency estimation.

The resulting UQ-flood intervals should not be interpreted as comprehensive measures of all uncertainty in the modelling chain. They describe predictive variability across residual-error realizations, conditional on the selected hydrologic model, calibrated parameter set, fitted weather generator, REM structure, and available observations. Uncertainty in alternative model structures, hydrologic-model parameters, weather-generator parameters, streamflow measurements, and the choice of residual-error formulation is not separately propagated. Similarly, the UQ-flood and Bayesian GEV intervals represent different conditional uncertainty components and should not be compared as equivalent measures of overall precision. The principal finding is instead that the central UQ-flood estimates remain consistent with the long-record observational benchmarks when the framework is developed using 30 years of data.

~~A key contribution of this study is the implementation of an integrated simulation-based framework that combines a process-based hydrologic model, a stochastic weather generator, and a residual error model (REM) to support uncertainty characterisation in flood estimates. To our knowledge, this is among the first applications of the REM approach within the context of at-site flood frequency analysis (FFA), allowing for direct quantification of uncertainty in flood estimates. The REM is applied at the annual maxima scale, which aligns with the requirements of engineering design and regulatory assessments. This scale-specific focus may offer a more targeted way to represent uncertainty compared to conventional daily-scale error modelling.~~

~~The REM contributes to the UQ-flood framework by accounting for systematic biases in hydrologic simulations and representing residual variability not captured by deterministic model structures. While components such as continuous simulation, stochastic weather generation, and error modelling have been explored individually in previous FFA studies, the integration of these elements into a unified framework is a distinguishing feature of UQ-flood (Fig. 2). This integration enables consistent propagation of uncertainty from meteorological inputs to flood quantiles.~~

~~A central motivation for applying the REM at the annual maxima scale is the mismatch between the calibration target of hydrologic models and the target variable in FFA. Hydrologic models are commonly calibrated using daily streamflow metrics that reward overall hydrograph performance, water balance, or seasonal dynamics. However, strong daily-flow performance does not guarantee unbiased representation of annual maximum flows, which are controlled by threshold processes, antecedent conditions, snowmelt-rainfall interactions, and tail behaviour. The experiments without the REM demonstrate this issue directly: deterministic continuous simulation tended to underestimate flood magnitudes across the study watersheds. By estimating residual errors between observed and simulated annual maxima, UQ-flood corrects the scale most relevant to design flood estimation rather than assuming that daily calibration is sufficient for flood-frequency applications.~~

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, and the present study did not formally compare computational requirements among alternative uncertainty frameworks. We therefore do not claim that UQ-flood is universally faster, simpler, or less subjective than other approaches.

The Box–Cox transformation parameter was fixed at λ=0.2, following previous hydrologic residual-error studies that identified this value as suitable for representing heteroscedastic streamflow errors. Nevertheless, the fixed value remains a modelling assumption, and its influence was not evaluated in this study. Future applications should assess whether the selected transformation and residual-error structure adequately represent the residual characteristics of the watershed and flood regime under consideration.

UQ-flood does not require fitting a parametric probability distribution directly to the synthetic annual-maximum series. Instead, flood quantiles are estimated empirically from the REM-corrected sequences using plotting positions. This avoids imposing a parametric form on the final synthetic flood sample, but it does not remove the assumptions embedded in the hydrologic model, weather generator, or REM. The framework should therefore be viewed as a complementary process-based approach rather than a replacement for conventional statistical FFA.

Overall, the contribution of UQ-flood lies in explicitly representing hydrologic-model error at the annual-maximum scale. The framework corrects systematic biases in deterministic process-based simulations and propagates residual predictive variability into flood-quantile estimates.

~~Compared to other uncertainty quantification approaches, UQ-flood seeks to balance methodological rigour with practical implementation. For instance, unlike GLUE-based methods, it does not rely on subjective likelihood thresholds or behavioural criteria~~ (Blazkova and Beven, 2009; Cameron et al., 2000b)(Blazkova and Beven, 2009; Cameron et al., 2000b)~~. Similarly, while full Bayesian approaches offer comprehensive uncertainty treatment, they can be computationally intensive. Costa and Fernandes~~ (2017)(2017)~~, for example, proposed a Bayesian method that jointly infers parameters of the weather generator, hydrologic model, and a skew exponential power error distribution using MCMC proposed a Bayesian indirect flood estimation method that jointly infers parameters of the weather generator, hydrologic model, and a complex skew exponential power error distribution using MCMC. While robust, such methods may be less accessible for operational use~~ (Costa and Fernandes, 2017)(Costa and Fernandes, 2017)~~. In contrast, UQ-flood applies a simplified error model, offering a faster and more accessible alternative without compromising statistical validity.~~

~~UQ-flood also offers an alternative to bootstrap-based single-site approaches, which depend on resampling the available data~~ (Arnaud et al., 2017b; Bomers et al., 2019; Burn, 2003; Mirzaei et al., 2015)(Arnaud et al., 2017b; Bomers et al., 2019; Burn, 2003; Mirzaei et al., 2015)~~. In small-sample contexts—such as 30 years of annual maxima—bootstrap methods may struggle to capture uncertainty adequately, especially when the sample is unrepresentative. Resampling from limited datasets can lead to narrow confidence intervals and potentially biased estimates~ (Hesterberg, 2015)(Hesterberg, 2015)~~. The commonly used percentile bootstrap interval is particularly sensitive to sample size. In contrast, UQ-flood appears to~~

~~perform reasonably well under data-scarce conditions (Fig. 3), providing interpretable quantile estimates even with limited records.~~

~~Importantly, UQ-flood does not require fitting a parametric distribution directly to the annual maximum flow sample. Instead, flood quantiles are estimated empirically from large synthetic ensembles generated through process-based simulation and residual error modelling. This flexibility may improve interpretability and help avoid tail sensitivity issues associated with parametric distribution fitting~~ (Macdonald et al., 2022; Merz et al., 2022)~~(Macdonald et al., 2022; Merz et al., 2022)~~.

~~Although this study does not explicitly address non-stationarity, the UQ-flood framework is well-suited to exploring such conditions. By perturbing the system with alternative land use scenarios, climate projections, or water management strategies, users can assess the influence of changing drivers on flood hazard while maintaining a consistent uncertainty quantification approach.~~

~~Overall, UQ-flood contributes to the broader movement towards physics-informed uncertainty quantification in flood risk assessment. By embedding uncertainty within the modelling chain – from climate inputs to flood quantiles – it offers a scalable and decision-relevant methodology that complements existing statistical approaches.~~

Formatted: Spanish (Spain)

5.3 Limitations and Future Directions

UQ-flood remains conditional on the choices and assumptions within its modelling chain. The UQ-flood interval represents residual predictive variability, whereas statistical FFA intervals primarily characterize sampling and parameter uncertainty conditional on a selected probability model. Because these intervals represent different uncertainty components, their widths should not be compared as equivalent measures of overall precision.

Flood estimates are influenced by hydrologic-model structure and calibration, stochastic weather-generator selection and configuration, and residual-error-model specification. In this study, CoSMoS initially produced implausibly long-tailed precipitation extremes in very long simulations, requiring a constraint to preserve physical plausibility. UQ-flood is not restricted to CoSMoS and could use other meteorological sampling approaches, including stochastic storm transposition (Wright et al., 2020), reforecast ensemble pooling (Brunner and Slater, 2022), or convection-resolving climate simulations (Dallan et al., 2024). However, all synthetic sequences represent the climate conditions reflected in the calibration period and should not be interpreted as future-climate projections.

A core assumption of UQ-flood is that process-based hydrologic models can derive more useful information than purely statistical extrapolation from short records. That is, expert knowledge about watershed processes—such as water balance, storage, and runoff generation—enables the model to constrain future flood behavior more effectively than sampling-based uncertainty alone (Su and Tung, 2013). In this sense, UQ-flood acts as a more intelligent flood sampler, embedding physical understanding directly into the generation of extreme event ensembles.

The present analysis does not propagate uncertainty in streamflow measurements, rating curves, hydrologic-model parameters, or weather-generator parameters. These uncertainties may be important for rare floods and should be incorporated in future applications. Moreover, UQ-flood requires sufficient meteorological and streamflow observations to calibrate the modelling components and characterize annual-maximum residuals.

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.

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.

~~While the UQ-flood framework offers a robust approach for quantifying uncertainty in flood frequency analysis, several limitations and areas for future development should be acknowledged. First, caution is warranted when directly comparing the uncertainty ranges produced by UQ-flood and traditional statistical FFA. These methods quantify different sources of uncertainty. UQ-flood captures the predictive uncertainty of the hydrologic model, including uncertainties in inputs, structure, and parameters. In contrast, statistical FFA primarily characterises parametric uncertainty associated with fitting the GEV distribution. Despite these theoretical differences, both approaches provide valuable representations of uncertainty in flood quantile estimation and can be regarded as broadly comparable for practical decision-making (Abbasian et al., 2025)(Abbasian et al., 2025).~~

~~The accuracy and reliability of the resulting flood estimates are also inherently influenced by the choices made across the modelling chain and the methods used to represent uncertainty. Decisions related to hydrologic model structure, calibration strategy, weather generator configuration, and residual error modelling all affect the resulting flood quantiles. As highlighted by Kritidou et al., (2026), these modelling options play a critical role in determining the reliability of flood estimates and should therefore be carefully considered when applying simulation-based flood frequency analysis frameworks.A key enabler of the UQ-flood framework is the stochastic weather generator (SWG), which provides the synthetic meteorological inputs necessary for long-term continuous simulation. The SWG is essential for producing the large sample sizes required to support robust, nonparametric flood frequency estimation and effective uncertainty quantification. However, selecting and adjusting a reliable SWG is not a trivial task (Grimaldi et al., 2022)(Grimaldi et al., 2022). Currently, there are no standardised guidelines for choosing an appropriate SWG for FFA, and each available method carries distinct assumptions~~

and limitations (Fischer et al., 2025)(Fischer et al., 2025). Successful implementation may require advanced expertise in hydroclimatology and statistical modelling, as the choice of SWG can significantly influence the accuracy and realism of simulated flood extremes.

In this study, we used the CoSMoS weather generator (Papalexiou, 2018)(Papalexiou, 2018) due to its ability to simulate non-Gaussian time series. However, we identified and corrected a bias in the generation of extreme precipitation values, which initially exhibited unrealistically long-tailed behaviour when used to produce very long meteorological time series. This issue arises in part because SWGs tend to be fit to the full data distribution rather than specifically to the tail. To address this, we implemented a constraint on extreme precipitation to preserve physical plausibility. Importantly, UQ-flood is not tied to CoSMoS or any single weather generator. It can be informed by any meteorological sampling approaches, such as stochastic storm transposition (Wright et al., 2020)(Wright et al., 2020), reforecast ensemble pooling (Brunner and Slater, 2022)(Brunner and Slater, 2022), or convection-resolving climate models (Dallan et al., 2024)(Dallan et al., 2024)(Dallan et al., 2024)(Dallan et al., 2024). These alternatives may be particularly useful in regions where convective storms or surprise floods dominate the hydrologic regime. This flexibility ensures UQ-flood can be adapted to a wide range of climatic contexts and emerging flood risks. The synthetic sequences should be interpreted as stationary realisations of the climate represented by the calibration period, not as projections of future climate.

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. Similarly, uncertainty in the parameters of the weather generator used to simulate rainfall inputs is not considered. To some extent, we assume that rainfall records at each site are equal to or longer than the corresponding streamflow records, which may not hold in all cases. Including these additional sources of uncertainty could improve the robustness of future analyses and support more comprehensive flood risk assessments.

A core assumption of UQ-flood is that process-based hydrologic models can derive more useful information than purely statistical extrapolation from short records. That is, expert knowledge about watershed processes—such as water balance, storage, and runoff generation—enables the model to constrain future flood behavior more effectively than sampling-based uncertainty alone (Su and Tung, 2013)(Su and Tung, 2013)(Su and Tung, 2013)(Su and Tung, 2013). In this sense, UQ-flood acts as a more intelligent flood sampler, embedding physical understanding directly into the generation of extreme event ensembles.

Another strength of UQ-flood is its modular structure, which makes it readily extendable to different modeling environments. While this study employed a lumped, conceptual hydrologic model (HBV-EC), which was appropriate given the scale and data availability of the selected watersheds, the framework is equally compatible with physically based distributed models. Applying UQ-flood within a distributed modeling context could enhance the spatial realism of simulated flood responses, extend the approach to ungauged watersheds, and improve the representativeness of predictive ensembles (Perez et al., 2019)(Perez et al., 2019)(Perez et al., 2019)(Perez et al., 2019). However, the performance and added value of

~~UQ-flood in such models remain to be systematically evaluated, particularly in terms of how well the residual error model captures localized hydrologic variability and structural uncertainty across heterogeneous landscapes.~~

~~This study focused on three Canadian watersheds with long streamflow records to enable evaluation. While this provided validation of the UQ-flood framework under diverse hydroclimatic conditions, broader testing across data-scarce and climatically varied regions is desirable to establish its generalizability. Additionally, comparison with regional FFA methods is necessary to further assess its relative performance and applicability.~~

5.4 Implications for Design and Decision-Making

The findings across the three case studies highlight the potential practical value of UQ-flood for flood risk management and engineering design. For watersheds with long flow records, UQ-flood produced flood quantile estimates consistent with those obtained using conventional statistical flood frequency analysis (FFA), while yielding narrower uncertainty intervals. Under short-record conditions, such as a 30-year record, UQ-flood also provided robust and actionable flood estimates, whereas conventional statistical FFA produced substantially wider uncertainty intervals that may be impractical for design and planning purposes. These findings suggest that UQ-flood can help address the limitations of local data scarcity, a common challenge in operational flood risk assessment.

From a practical perspective, improved characterization of uncertainty can support more informed design and investment decisions. Uncertainty intervals that are excessively wide can complicate the selection of design floods, potentially contributing to overdesign and unnecessary costs or underdesign and increased flood risk. In applications such as infrastructure planning, floodplain mapping, and insurance risk assessment, more informative uncertainty estimates can support the prioritization of investments, the evaluation of alternative design options, and the development of climate-resilient infrastructure.

UQ-flood also provides a process-based framework for interpreting and communicating flood estimates. By combining hydrologic simulation, explicit bias correction, and residual uncertainty propagation, the framework links flood quantile estimates to modeled hydrologic processes and observed model performance. This provides a complementary perspective to conventional statistical FFA, in which estimates depend primarily on the selected probability distribution and the available sample of annual maximum flows. Nevertheless, the process-based nature of UQ-flood does not eliminate uncertainty associated with model structure, parameter estimation, or the representation of hydrologic processes; these sources of uncertainty should be considered when interpreting results and applying them to engineering decisions.

~~The findings from the three case studies highlight the practical value of UQ-flood for flood risk management. When applied to watersheds with long records, UQ-flood produced flood quantile estimates consistent with traditional statistical methods, but with significantly narrower uncertainty intervals. This has direct implications for design safety, regulatory compliance,~~

and investment decision-making. Critically, when applied under short-record conditions (e.g., 30 years), UQ-flood maintained robust and actionable flood estimates, while statistical FFA produced wide, often impractical uncertainty bounds. These results demonstrate the method's resilience to data scarcity, a common challenge in real-world applications.

From a practical standpoint, narrower uncertainty bounds enable engineers and planners to avoid both overdesign (leading to excessive costs) and under-design (resulting in heightened risk). In fields such as insurance, urban planning, and floodplain mapping, proper uncertainty quantification supports more accurate risk pricing, prioritization of investments, and the design of climate-resilient infrastructure.

Importantly, UQ-flood also enhances the interpretability and credibility of flood estimates. Unlike statistical methods that rely heavily on distributional assumptions, the UQ-flood framework is built on a process-based hydrologic model and observed model performance, making it more transparent and intuitive for engineers, scientists, and decision-makers to communicate and justify design choices.

Finally, the UQ-flood is conceptually aligned with the emerging design simulation approach framework aligns with the emerging concept of the design simulation approach (Fischer et al., 2025)(Fischer et al., 2025)(Fischer et al., 2025)(Fischer et al., 2025), which emphasizes process-based flood estimation to inform engineering design. Within this context, UQ-flood offers a framework for incorporating quantified uncertainty into decisions such as the selection of design floods and safety factors. Rather than relying exclusively on fixed safety margins, engineers and decision-makers can use uncertainty estimates to evaluate the robustness of alternative design choices against specified risk tolerances. The potential benefit is a more transparent and evidence-based balance between infrastructure costs and flood safety, particularly in watersheds where limited observational records constrain conventional flood frequency analysis. which advocates for simulation-driven flood estimates as the foundation for engineering design. In this context, UQ-flood offers a pathway to inform key design decisions—such as the selection of safety factors—based on quantifiable uncertainty. Rather than relying on fixed safety margins, decision-makers can assess how confident they are in their design flood estimates and adjust accordingly, thereby improving both cost-efficiency and risk robustness.

6 Conclusions

Flood-frequency estimates are affected by multiple sources of uncertainty, including limited observations, meteorological variability, hydrologic-model errors, parameter uncertainty, and model-structure choices. Transparent characterization of these uncertainties is important for the appropriate interpretation and use of flood estimates in engineering design and risk management.

This study introduced UQ-flood, a process-based flood-frequency framework that integrates a stochastic weather generator, a process-based hydrologic model, and a residual error model developed at the annual-maximum scale. The framework was designed to address the mismatch between hydrologic models commonly calibrated against daily streamflow and the annual

flood peaks used in flood-frequency analysis. The REM corrects systematic discrepancies in the simulated annual maxima and propagates residual predictive variability into the resulting flood-quantile estimates.

UQ-flood was evaluated in three Canadian watersheds with contrasting hydroclimatic and physiographic characteristics. Long-record empirical annual-maximum estimates were used as the primary benchmark, while Bayesian GEV estimates based on long and 30-year records provided secondary statistical benchmarks. The principal findings are as follows:

1. The central UQ-flood estimates remained consistent with the long-record observational benchmarks when the framework was developed using only 30 years of meteorological and streamflow observations. The relative proximity of UQ-flood and GEV to the empirical estimates varied among watersheds and return periods, and neither approach was uniformly closest in every case.

2. Process-based simulation without residual-error correction produced systematic biases in the estimated flood quantiles. Applying the REM shifted the simulated flood-frequency curves toward the long-record observational benchmarks, demonstrating the importance of accounting for hydrologic-model discrepancies at the annual-maximum scale.

3. The annual-maximum REM provided an explicit representation of residual predictive variability around the corrected process-based estimates. This scale-specific formulation targets the same variable used in engineering flood-frequency analysis rather than assuming that satisfactory daily-streamflow performance guarantees unbiased simulation of annual flood peaks.

Formatted: List Paragraph, Numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0.63 cm + Indent at: 1.27 cm

Formatted: List Paragraph, Numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0.63 cm + Indent at: 1.27 cm

Formatted: List Paragraph, Numbered + Level: 1 + Numbering Style: 1, 2, 3, … + Start at: 1 + Alignment: Left + Aligned at: 0.63 cm + Indent at: 1.27 cm

The UQ-flood and Bayesian GEV intervals represent different conditional uncertainty components. The UQ-flood interval describes predictive variability across residual-error realizations, conditional on the selected hydrologic model, calibrated parameter set, fitted weather generator, REM formulation, and available observations. The Bayesian GEV interval primarily represents sampling and parameter uncertainty conditional on the observed annual-maximum sample and selected statistical model. Their widths should therefore not be compared as equivalent measures of overall precision or uncertainty completeness.

The results do not imply that process-based simulation eliminates the information limitations associated with short observational records. Instead, UQ-flood uses the available meteorological and streamflow information through a process-based modelling chain and explicitly represents residual errors in the simulated design variable. Its estimates remain conditional on the adequacy of each component of that chain.

Overall, the results demonstrate that explicitly modelling residual errors at the annual-maximum scale can improve the consistency of process-based flood-frequency estimates with long-record observations. UQ-flood should therefore be viewed as a complementary framework for process-based FFA rather than as a replacement for conventional statistical methods.

Future work should evaluate the framework across a broader range of hydroclimatic regimes and examine uncertainty associated with hydrologic-model parameters, weather-generator parameters and structure, streamflow measurements, and alternative REM formulations. Further developments could also extend UQ-flood to peaks-over-threshold applications and investigate residual error models conditioned on flood-generating mechanisms.

~~Flood estimates are inherently uncertain due to a range of contributing factors, including model structural errors, parameter uncertainty, data limitations, and forcing variability. Failure to account for these uncertainties can result in either overly conservative designs or underestimated flood risks. Accurately characterizing and transparently communicating this uncertainty is essential to ensure the reliability and usability of flood frequency estimates in engineering design and flood risk management.~~

~~In this study, we proposed and tested an uncertainty quantification framework for simulation-based flood frequency analysis (UQ-flood), which explicitly accounts for hydrologic model errors through the integration of a statistical residual error model (REM). The framework corrects systematic biases in simulated flows and propagates predictive uncertainty into flood quantile estimates. It combines a process-based hydrologic model, a stochastic weather generator, and a residual error model to characterize uncertainty of extreme events. Central to this framework is the error model, which captures both systematic biases and unexplained variability in hydrologic simulations.~~

~~As a proof of concept, we applied the proposed UQ-flood framework to three Canadian watersheds with contrasting climatic and physiographic characteristics. To evaluate its performance, we compared UQ-flood against traditional statistical flood frequency analysis (FFA) methods based on the Generalized Extreme Value (GEV) distribution and Bayesian inference. The key findings are as follows:~~

1. ~~The proposed UQ-flood framework produces flood magnitude estimates that are consistent with those obtained from statistical methods applied to long observational records (e.g., 100 years), but with considerably narrower and more informative uncertainty bounds, even when using shorter 30-year records.~~
2. ~~When applied to short observational records (e.g., 30 years), statistical FFA yields wide and often impractical uncertainty ranges. In contrast, the UQ-flood framework provides robust and actionable estimates.~~
3. ~~Continuous simulation without incorporating an explicit error model typically leads to systematic bias flood quantiles, underscoring the critical importance of accounting for model error in simulation-based frequency analysis.~~

~~These results suggest that explicitly modelling hydrologic simulation error at the annual maxima scale can make continuous simulation more useful for engineering flood-frequency applications. UQ-flood offers a process-consistent and practical~~

~~alternative for at-site flood risk estimation, particularly where observed records are short and conventional uncertainty intervals are too broad to guide decisions. This capability is especially valuable in regions where Regional Flood Frequency Analysis (RFFA) is limited by sparse gauging networks, heterogeneous catchment characteristics, or a lack of suitable donor catchments. Future work should evaluate the framework across a wider range of hydroclimatic regimes, incorporate additional uncertainty sources such as rating-curve and weather-generator parameter uncertainty, and explore mechanism-aware residual error models conditioned on flood-generating processes.~~

Code and data availability

The stochastic weather generator CoSMoS is available for download from its Github repository: https://github.com/COSMOS. The Raven hydrological modelling framework can be accessed at https://raven.uwaterloo.ca/. Historical weather data were obtained from Environment and Climate Change Canada (ECCC) and are publicly available at https://climate.weather.gc.ca/. Streamflow data were sourced from the Water Survey of Canada and can be downloaded via https://dd.weather.gc.ca/hydrometric/csv/. The UQ-flood scripts and datasets used in the analyses presented in this manuscript are openly available at: https://github.com/rarabzad/REM_UQ_EXTREME/tree/main.

Author contributions

JRC: Writing original draft, investigation, visualization, software, and methodology
JRCr: Conceptualization, funding acquisition, project administration, supervision, writing, investigation and methodology
BAT: Supervision, writing, investigation and methodology
RA: Software and data curation

Competing interest

The authors declare that they have no conflict of interest.

Financial support

This work was funded under the Natural Resources Canada Flood Hazard Identification and Mapping Program (FHIMP) project FHIMP23-27P2.

References