Response to reviewer

An Uncertainty Quantification Framework for Simulation-based Flood Frequency Analysis

We are grateful to the reviewer for carefully reading our manuscript and for the constructive critique. Please see below, the notation, for a point-by-point response to the reviewers' comments and concerns.

Notation:

Bold font: referee's comment.

Black font: authors' comments.

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

Reviewer request:

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

Response:

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

Manuscript changes:

3.1 Hydrologic Model Development: Add: "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."

Comment 2 — Sufficiency of the AR residual-error component

Reviewer request:

Was the AR component added based on diagnostics or subjectively?

Response:

We based the residual-error model (REM) structure on diagnostics of the raw residuals (observation minus simulation), not on a subjective choice. Specifically, we:

- Examined variance behavior (time-varying plots and segment-wise summaries) to check for heteroscedasticity.
- Computed sample autocorrelation and partial autocorrelation functions (ACF/PACF) of the AMAX residuals to assess serial dependence.
- Assessed residual distributions using histograms and Q-Q plots to check symmetry and tail behavior.

For the annual-maxima (AMAX) residuals at our sites, these diagnostics indicated negligible serial dependence; consequently, an AR term is not required to capture dependence in our AMAX application. We nevertheless kept a permissive AR(1) option in the REM so that the framework can accommodate sites or use-cases where residual autocorrelation is non-negligible; in our case, the estimated AR coefficient was near zero and did not materially affect results. We will add a summary table/figure of the ACF/PACF and a one-line note in the Methods clarifying that the REM structure was diagnostics-driven.

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

Reviewer request:

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

Response:

We appreciate the point. Our design uses the residual-error model (REM) to represent predictive (residual) uncertainty, which aggregates the effects of forcing, parameter, and structural model errors—consistent with established post-processing paradigms (e.g., hydrologic uncertainty processors and related Bayesian/post-processing frameworks). In this view, explicit parameter-ensemble bootstrapping is not required for the present paper's scope because parametric effects are indirectly captured by the REM's predictive distribution fitted to simulation-observation discrepancies. We will make this assumption explicit in 3.2 Residual Error Modelling and flag full

parameter-uncertainty propagation as a future extension rather than adding a new sensitivity analysis here.

Manuscript changes:

3.2 Residual Error Modelling: Add 2–3 sentences clarifying that the REM absorbs aggregate discrepancy (including parametric effects), and that explicit parameter ensembles are beyond scope but are a natural next step in future work.

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

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

Response:

Agreed. We will cite Velásquez & Krajewski (2024), who demonstrate—via B17C and Monte-Carlo scenarios—that measurement/rating-curve errors can materially alter flood quantiles, at times exceeding 50% and up to 100% for long return periods, and we will acknowledge that our current bands do not propagate rating-curve error. We will frame this as an important limitation and a priority for future development (e.g., coupling a stage–discharge error model with the REM).

Manuscript changes:

Discussion: Add 2–3 sentences citing Velásquez & Krajewski (2024) and noting that rating-curve uncertainty is not yet propagated but is slated for future work.

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

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

Response:

We agree the idea is interesting. Detecting distinct flood-generation mechanisms would require event classification and mechanism-conditional REM parameters (e.g., conditioning on seasonal/weather-pattern states). That extension is beyond the scope of our present comparison between UQ-flood and GEV uncertainty; we will flag it as future work.

Manuscript changes:

Discussion (last paragraph): Add a sentence on mechanism-aware/conditional REM as a promising extension.

Minor comments — Authors' response

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