Overall assessment

This technical note addresses a focused and practically relevant question: since large hydrological model ensembles are typically calibrated deterministically, does adopting Bayesian parameter estimation materially change (1) inferences about the role of model complexity and (2) ensemble predictive skill? The experimental design is appropriate, and the main empirical message is clear: Bayesian parameter estimation improves single-model performance most noticeably for higher-parameter/more complex models but yields only marginal gains once multi-model averaging is applied.

That said, several methodological choices could influence these conclusions and would benefit from clearer documentation and/or targeted sensitivity tests.

Comments:

1. Justify and test the posterior sample size (200 draws). You draw 200 parameter sets from the MCMC output to represent the posterior. Please justify this choice (e.g., computational budget vs diminishing returns) and add a brief sensitivity check (e.g., 100 vs 200 vs 500 draws for a subset of basins/models) to demonstrate stability.
2. Clarify/justify the KGE-based “likelihood” used for DREAM. Since KGE is not a conventional likelihood, please briefly explain why this informal likelihood is appropriate for your goals and discuss known modes of failure.
3. Writers can report computational cost and MCMC configuration. Given the emphasis on computational expense, please report the key DREAM settings and approximate compute: number of chains, iterations, burn-in/thinning (if any), and typical runtime per basin/model (or total core-hours).
4. Define calibration/validation windows and the “shifted” calibration rule. Please state explicitly how calibration/validation periods are chosen in the two data regimes, and how the calibration window is shifted when discharge records are incomplete (rule used, and how many basins are affected).
5. Missing data handling. Please clarify how missing discharge values are treated in calibration and validation (e.g., removed days, gap-filling, objective computed only on overlapping periods), and whether forcing inputs ever contain gaps and how those are handled.