Quantifying Structural Uncertainty in Hydrologic Models

by

Rezgar Arabzadeh

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Civil Engineering - Water

Waterloo, Ontario, Canada, 2026

© Rezgar Arabzadeh 2026

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	M. Reza Najafi Associate Professor, Dept. of Civil and Environmental Engineering, Western University
Supervisor(s):	James R. Craig Professor, Dept. of Civil and Environmental Engineering, University of Waterloo Bryan A. Tolson Professor, Dept. of Civil and Environmental Engineering, University of Waterloo
Internal Members:	Alana Lund Assistant Professor, Dept. of Civil and Environmental Engineering, University of Waterloo Bruce MacVicar Associate Professor, Dept. of Civil and Environmental Engineering, University of Waterloo
Internal-External Member:	Chris Fletcher Associate Professor, Dept. of Geography and Environmental Management, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis comprises three manuscripts, each presented as a separate chapter and each co-authored as listed below.

Chapter 2 has been published as:

> Arabzadeh, R., Romero-Cuellar, J., Craig, J.R., Tolson, B.A., Chlumsky, R. (2026). A surrogate-aided approach for accelerated Bayesian calibration of hydrologic models. *Environmental Modelling and Software*, 198, 106894. https://doi.org/10.1016/j.envsoft.2026.106894.

Chapter 3 is submitted to *Journal of Hydrology* as:

> Arabzadeh, R., Craig, J.R., Tolson, B.A., Romero-Cuellar, J., Chlumsky, R. Simultaneous Bayesian quantification and decomposition of structural, parameter, and forcing uncertainty in hydrologic models: A blended model approach. *Journal of Hydrology*.

Chapter 4 is submitted to *Nature Communications* as:

> Arabzadeh, R., Craig, J.R., Tolson, B.A. Continental-scale decomposition of hydrologic model uncertainty across 2,869 North American watersheds. *Nature Communications*.

For all three manuscripts, as lead author I was responsible for conceptualization, methodology, software development, formal analysis, data curation, validation, visualization, and writing the original draft. Co-authors contributed to conceptualization, methodology, validation, supervision, funding acquisition, and review and editing.

Abstract

Hydrologic models are essential tools for understanding watershed processes and supporting water resource management. However, their predictions are inherently uncertain due to imperfect model structures (structural uncertainty), parameter estimation challenges (parameter uncertainty), and limitations in observational data and model forcings (input uncertainty). Bayesian inference has become a widely used framework for quantifying these uncertainties because it enables probabilistic parameter estimation and prediction while formally incorporating prior information and observational evidence. Despite these advantages, the application of Bayesian methods to complex hydrologic models remains computationally demanding, and the resulting predictive uncertainty often represents a combination of multiple uncertainty sources (including input, parameter, and structural uncertainties) that are difficult to interpret individually. These limitations reduce the effectiveness of Bayesian uncertainty analysis as a diagnostic tool for improving hydrologic models.

This thesis develops methodological advances to improve the efficiency and interpretability of Bayesian uncertainty quantification in hydrologic modeling. The research focuses on two challenges: improving the computational feasibility of Bayesian inference for complex models and separating the sources of uncertainty represented within Bayesian predictive distributions. To address these challenges, new methods are developed and evaluated using both regional and continental-scale hydrologic datasets.

1. A machine learning–assisted framework is developed to improve the efficiency of Bayesian joint inference for hydrologic models. The proposed approach integrates machine learning techniques with Bayesian calibration to facilitate exploration of complex posterior parameter distributions and reduce the computational burden associated with traditional sampling methods. The framework is evaluated using twelve watersheds from the MOPEX dataset and demonstrates improved inference performance while maintaining reliable uncertainty quantification.

2. A variance decomposition methodology is introduced to identify and quantify the sources of uncertainty embedded within Bayesian predictions. While Bayesian calibration provides probabilistic estimates of model outputs, it does not directly attribute predictive uncertainty to individual components of the modeling framework. The proposed method decomposes posterior predictive uncertainty into interpretable components, enabling a clearer understanding of how different aspects of the modeling process contribute to overall uncertainty.

3. The proposed uncertainty decomposition framework is applied to a large-scale hydrologic analysis across approximately 3,000 watersheds in North America. This continental-scale application enables the systematic evaluation of spatial patterns in hydrologic model uncertainty and reveals how dominant uncertainty sources vary across hydroclimatic and physiographic regions.

Together, the contributions of this thesis improve both the computational efficiency and the interpretability of Bayesian uncertainty estimates in hydrologic modeling. The proposed approaches provide tools for diagnosing uncertainty sources and evaluating model reliability, which can support more transparent hydrological predictions across a range of environmental and water resource applications.

Acknowledgements

I owe my deepest gratitude to my supervisors, Professor James R. Craig and Professor Bryan A. Tolson, whose guidance shaped every stage of this thesis. Their willingness to debate ideas, push back, and give me the freedom to pursue the questions I cared about made this work possible. Whatever is good here owes much to them; the shortcomings are mine alone.

I thank the members of my examining committee, Professor Mohammad Reza Najafi (external examiner, Western University), Professor Alana Lund and Professor Bruce MacVicar (internal members), and Professor Chris Fletcher (internal-external member), for the time and care they invested in reading this thesis and for the thoughtful questions they raised at the defence.

To my friends, Aram, Sina, Mahkameh, Shae, Robert, and Jonathan: your companionship made the long stretches lighter and the difficult weeks bearable. Jonathan, in particular, has been both a brilliant collaborator on much of the work surrounding this thesis and a friend.

To my parents, Nabi and Marzieh, I owe more than words can carry. Their belief in education, their quiet sacrifices, and their unwavering encouragement across distances and years are the foundation of everything I have managed to do.

To my siblings, Foozieh, Fatheme, Fahmideh, and Azad: thank you for being my first companions and lifelong friends. Your steady support and love, even from across continents, have been a quiet strength behind every page of this work.

Most of all, to my lovely wife Parisa: none of this would have been possible without you. You have lived every late night, every dead end, and every small victory of this PhD alongside me, with patience and grace I will spend a lifetime trying to repay. And to my darling Connie: you arrived in the middle of all of this and made everything matter more. This thesis is for the two of you.

Dedication

To *Parisa* and *Connie*

and,

To the brave women and men of Iran and Rojhelat Kurdistan,
who continue to resist, and to light the way,

for:

Jin, Jiyan, Azadî

Table of Contents

Examining Committee Membership ii

Author's Declaration iii

Statement of Contributions iv

Abstract v

Acknowledgements vii

Dedication viii

List of Figures xiii

List of Tables xvi

List of Abbreviations xviii

List of Symbols xxi

1 Introduction 1

1.1 Background and Motivation 1

1.2 Sources of Hydrologic Predictive Uncertainty 2

1.3 The Methodological Gap 3

1.4 Thesis Objectives and Contributions 5
1.5 Thesis Structure 7

2 A Surrogate-Aided Approach for Accelerated Bayesian Calibration of Hydrologic Models 8

2.1 Introduction 8
2.2 Material and Methods 10
2.2.1 Experimental Design 10
2.2.2 Proposed Workflow 11
2.2.3 MOPEX12 Case Study Watersheds 20
2.3 Results 22
2.3.1 Convergence Assessment at Guadalupe Watershed 22
2.3.2 Gelman–Rubin Metric Across MOPEX12 24
2.3.3 Model Quantile Posteriors Analysis 27
2.4 Discussion 30
2.5 Conclusion 31

3 Simultaneous Bayesian Quantification and Decomposition of Structural, Parameter, and Forcing Uncertainty in Hydrologic Models: A Blended Model Approach 34

3.1 Introduction 35
3.2 Methodology 38
3.2.1 Case studies 38
3.2.2 Hydrologic Model 39
3.2.3 Bayesian Calibration Framework 40
3.2.4 Uncertainty decomposition 45
3.2.5 Probabilistic Performance Metrics 49
3.3 Results 50
3.3.1 Probabilistic Model Performance 50

3.3.2 Temporal Dynamics of Uncertainty Sources . . . 53
3.3.3 Cross-Catchment Summary of Uncertainty Partitioning . . . 56
3.3.4 Relationships Between Uncertainty Sources and Basin Attributes . 56
3.4 Discussion . . . 58
3.4.1 Decomposition Using the Blended Model Structure . . . 58
3.4.2 Forcing Uncertainty, Hydroclimatic Regime, and the Texas Catchments 59
3.4.3 Methodological Considerations and Limitations . . . 59
3.4.4 Implications for Practice and Future Directions . . . 60
3.5 Conclusions . . . 61

4 Continental-Scale Decomposition of Hydrologic Model Uncertainty Across 2,869 North American Watersheds 63
4.1 Introduction . . . 64
4.2 Results . . . 65
4.2.1 Probabilistic skill across hydroclimatic regimes . . . 66
4.2.2 Structural uncertainty dominates the continental budget . . . 66
4.2.3 Climate, not topography, controls the budget . . . 68
4.2.4 Posterior weight distributions form a continental atlas of process choice and confidence . . . 72
4.2.5 Seasonal expansion during summer–fall recession . . . 78
4.3 Discussion . . . 79
4.4 Methods . . . 83
4.4.1 Study domain and data . . . 84
4.4.2 Blended hydrologic model and forcing perturbation . . . 85
4.4.3 Uncertainty decomposition and evaluation . . . 86

5 Conclusions 89
5.1 Summary . . . 89
5.2 Contributions . . . 90
5.3 Limitations and Future Research . . . 91

References **93**

Appendices **104**

Appendix 1.A: SVR Training Sample Size Selection 104

Appendix 1.B: Residual Diagnostics for Gaussian Assumption 104

Appendix 2.A: Blended-model parameter list 106

Appendix 2.B: Aitchison effect size for compositional comparison 106

List of Figures

2.1 Surrogate error model methodology workflow 16
2.2 Geographic locations of MOPEX12 sites 21
2.3 Maximum Gelman–Rubin diagnostic values, Guadalupe 23
2.4 Simulated streamflow predictive uncertainty, Guadalupe 25
2.5 SHA convergence advantage across MOPEX12 26
2.6 CRPS improvement at Guadalupe 28
2.7 CRPS values across MOPEX12 29
2.8 Reliability and relative sharpness across MOPEX12 29

3.1 Locations of the 12 MOPEX watersheds. 39
3.2 Smooth transformation $\Theta'(\theta)$ versus θ (unbounded parameter). The factor $\alpha = 0.99$ is the fixed value used throughout this study; the green line is the regular clamping step function for comparison. 46
3.3 Probabilistic model performance across the 12 MOPEX catchments, evaluated using four complementary metrics: CRPS skill score, Kling–Gupta Efficiency (KGE), reliability (empirical coverage of the 95% prediction interval), and sharpness (concentration of predictive distributions). All metrics are oriented so that higher values indicate better performance. 51
3.4 Time-varying uncertainty decomposition across the 12 MOPEX catchments over the climatological hydrological year. Each panel shows the stacked proportional contributions of parameter (orange), forcing (yellow), structural (dark red), and unexplained (grey) variance, computed independently for each Julian calendar day using MANOVA applied to 5,000 posterior parameter samples. The climatological 50th-percentile daily observed streamflow is overlaid (black line, right axis). 54

3.5 Distribution of annual mean uncertainty proportions across the 12 MOPEX catchments for the four variance components. Violin plots show the full cross-catchment distribution, with individual catchment values overlaid as dots. The horizontal line within each violin denotes the median. 55

3.6 Pearson correlation coefficients between the annual mean uncertainty proportions for each source and five basin physiographic descriptors, computed across the 12 MOPEX catchments. Cell colour intensity indicates correlation magnitude; cell values show the coefficient. 57

4.1 Scaled continuous ranked probability score (CRPS) across 2,869 North American watersheds. Lower values (blue) indicate better probabilistic prediction skill. Scaled CRPS: mean daily CRPS normalized by mean observed streamflow. 67

4.2 Annual-mean uncertainty attribution across 2,869 North American watersheds. Color: proportion of total explained variance attributed to parameter (top), structural (middle), and forcing (bottom) uncertainty. 69

4.3 Ternary diagram of uncertainty sources (Par: parameter uncertainty; Str: structural uncertainty; Frc: forcing uncertainty), partitioning coloured by (a) mean catchment elevation (m), (b) longitude of the watershed centroid (°), (c) monthly precipitation coefficient of variation (CV), and (d) mean annual air temperature (°C). 71

4.4 Probabilistic atlas of process algorithm choice across 2,869 North American watersheds. (a) Infiltration, (b) quickflow (interflow), (c) evapotranspiration, (d) deep baseflow, (e) snow balance. Each panel shows only the basins where the labelled algorithm carries the maximum posterior weight (n = number of basins for which that algorithm is the dominant choice); colour indicates the IQR of the posterior weight distribution at each basin (pink: low IQR, high confidence in the dominant choice; dark green/blue: high IQR, lower confidence). 73

4.5 Algorithm-level identifiability and its landscape associations across 2,869 watersheds. Left: violin plots of the posterior weight IQR for each algorithm in (a) infiltration, (b) quickflow, (c) evapotranspiration, (d) deep baseflow, (e) snow balance. Lower IQR indicates more decisive structural preference. Right: rescaled distributions of the basin attributes most strongly associated with the IQR for each process group. 75

4.6 Correlogram of R^2 between the IQR of each posterior algorithm weight and 16 landscape attributes (n = 2,869). Colour: R^2 (pink = 0, dark blue = 1). Stars indicate significance level (* $q < 0.05$, ** $q < 0.01$, *** $q < 0.001$). Attributes with > 2 orders-of-magnitude range are log-transformed. 77

4.7 Time-varying uncertainty decomposition at three representative gauges and continental synthesis. (a) Gauge 01010000, St. John River at Ninemile Bridge, Maine (snow-influenced eastern). (b) Gauge 07186000, Spring River near Waco, Missouri (semi-arid central). (c) Gauge 01668000, Rappahannock River near Fredericksburg, Virginia (humid eastern). Stacked area: forcing (yellow), parameter (orange), structural (dark red), unexplained (grey). Black line: mean daily streamflow (right axis). (d) Continental map of the ratio of summer–fall (June–November) to winter–spring (December–May) structural variance share across all 2,869 basins; values 1–2 (pink) indicate near-stationary structural uncertainty, values 4–6 (dark blue) indicate four- to sixfold seasonal expansion. 80

5.1 SVR mean absolute error vs. sample size 105

5.2 Residual diagnostics, Guadalupe 107

5.3 Compositional comparison of the model used in this study against (a) blended model v2 and (b) blended model v2 + BASE_LINEAR, across 12 USGS gauges. Bar height: Aitchison effect size (centroid shift in CLR). Colour: FDR-adjusted significance level. Black line: rescaled $-\log_{10} q$. Symbols: hierarchical clustering. 110

List of Tables

2.1 Hydrologic and error model parameters to be inferred. Filled circles (•) indicate parameters directly inferred by MCMC samplers; open circles (∘) indicate parameters generated by the surrogate model conditional on MCMC samples. 18

2.2 The MOPEX12 watershed specifications. 21

2.3 Computational cost comparison. Values represent median ± 2.5 SD across 12 MOPEX watersheds. 31

3.1 The MOPEX12 watershed specifications. 38

4.1 Process algorithms in the blended hydrologic model. Each process group contains multiple algorithm options whose weights are jointly inferred. Code identifiers are those used in the Raven framework configuration files. 86

5.1 List of Raven parameters inferred via the MCMC sampler in Chapter 3. . . 108

List of Abbreviations

ANN Artificial Neural Network xvii

AR(1) First-order autoregressive process xvii

ARNO ARNO infiltration / saturation-excess scheme xvii

BATEA Bayesian Total Error Analysis xvii

CAMELS Catchment Attributes and Meteorology for Large-sample Studies xvii

CDF Cumulative Distribution Function xvii

CDF$_U$ Empirical CDF of the Probability Integral Transform xvii

CLR Centred Log-Ratio xvii

CRPS Continuous Ranked Probability Score xvii

CV Coefficient of Variation xvii

DDS Dynamically Dimensioned Search xvii

DEM Digital Elevation Model xvii

DREAM Differential Evolution Adaptive Metropolis xvii

ESS Effective Sample Size xvii

ET Evapotranspiration xvii

FDR False Discovery Rate xvii

GCM General Circulation Model xvii

GHA Gaussian Heteroscedastic Autocorrelated error model xvii

GLUE Generalized Likelihood Uncertainty Estimation xvii

GR Gelman–Rubin diagnostic xvii

GR4J Génie Rural à 4 paramètres Journalier (4-parameter daily lumped rainfall–runoff model) xvii

HBV Hydrologiska Byråns Vattenbalansavdelning (conceptual hydrologic model) xvii

HMC Hamiltonian Monte Carlo xvii

HMETS Hydrology Model – École de Technologie Supérieure xvii

HRU Hydrologic Response Unit xvii

HYSETS Hydrometeorological Sandbox – ETS xvii

IBUNE Integrated Bayesian Uncertainty Estimator xvii

ILR Isometric Log-Ratio xvii

IQR Interquartile Range xvii

KGE Kling–Gupta Efficiency xvii

LSTM Long Short-Term Memory neural network xvii

MAE Mean Absolute Error xvii

MANOVA Multivariate Analysis of Variance xvii

MaxIt Maximum number of iterations xvii

MCMC Markov Chain Monte Carlo xvii

ML Machine Learning xvii

MOPEX Model Parameter Estimation Experiment xvii

NSE Nash–Sutcliffe Efficiency xvii

NUTS No-U-Turn Sampler xvii

PDF Probability Density Function xvii

PET Potential Evapotranspiration xvii

REM Residual Error Model xvii

RMSE Root Mean Squared Error xvii

SHA Surrogate-aided Heteroscedastic Autocorrelated error model xvii

SVM Support Vector Machine xvii

SVR Support Vector Regression xvii

THA t-distributed Heteroscedastic Autocorrelated error model xvii

TOPMODEL TOPography-based hydrological MODEL xvii

USGS United States Geological Survey xvii

VIC Variable Infiltration Capacity model xvii

List of Symbols

$a,\ b$ intercept and slope of the heteroscedastic variance model $\sigma_\epsilon(t) = a + bQ_t$ xvii

$\boldsymbol{\alpha}$ forcing perturbation parameter vector xvii

$\boldsymbol{B}$ MANOVA design (group-membership) matrix xvii

$c,\ k$ midpoint and steepness parameters of the tanh transformation that maps unbounded NUTS proposals to admissible parameter ranges xvii

$\mathcal{D}$ observational data (streamflow time series) xvii

δ surrogate-model hyperparameter vector xvii

ϵ_t raw model residual at time t xvii

η_t standardized residual at time t xvii

EV_g^t percentage of explained variance attributable to group g at Julian day t xvii

$F_t(x)$ posterior predictive cumulative distribution function at time t xvii

$\mathcal{G}_g$ index set of parameters belonging to functional group $g \in \{\text{parameter, structure, forcing}\}$ xvii

M mass matrix of the HMC kinetic energy term xvii

N number of observations / training samples xvii

N_{jday} number of Julian calendar days (typically 365) xvii

N_{param} number of inferred parameters per watershed xvii

$N_{\text{realization}}$ number of posterior realizations used in variance decomposition xvii

ν degrees-of-freedom parameter of the Student-t residual distribution xvii

N_{year} number of simulation years used for variance decomposition xvii

$\boldsymbol{\omega}$ residual error model parameter vector xvii

ϕ_η lag-1 autocorrelation coefficient of the standardized residuals xvii

Ψ joint hydrologic and error-model parameter vector $[\theta, \omega]$ xvii

q Benjamini–Hochberg-adjusted p-value (false discovery rate) xvii

$\hat{Q}_t$ simulated streamflow at time t xvii

Q_t^o observed streamflow at time t xvii

$\boldsymbol{r}$ auxiliary momentum vector in Hamiltonian Monte Carlo xvii

R^2 coefficient of determination xvii

$\hat{R}$ Gelman–Rubin convergence statistic xvii

s_0, S_0 initial hydrologic model state xvii

σ_t, $\sigma_\epsilon(t)$ time-varying residual standard deviation (heteroscedastic component) xvii

SS_g^t sum of squares attributable to parameter group g at Julian day t xvii

$\boldsymbol{\Theta}$ full inferred parameter vector $\{\boldsymbol{w}', \boldsymbol{\theta}, \boldsymbol{\alpha}\}$ xvii

$\boldsymbol{\theta}$ hydrologic process parameter vector xvii

$\boldsymbol{w}$, $\boldsymbol{w}'$ blended-model process weights and weight-generating parameters xvii

$\boldsymbol{x}$, $\boldsymbol{x}'$ raw and perturbed forcing time series xvii

$\boldsymbol{Y}$ mean paired difference vector in ILR coordinates (Hotelling test) xvii

y_t, $\hat{y}_t$ observed and simulated response (streamflow) at time t xvii

Chapter 1

Introduction

1.1 Background and Motivation

When a calibrated hydrologic model produces a 95% predictive interval for tomorrow's streamflow, the width of that interval reflects what the model does not know about the catchment's underlying physics, the values of its parameters, and the rainfall and temperature data driving it. From the prediction itself, however, a practitioner cannot determine which of these sources of uncertainty is responsible for the width of the interval. The total uncertainty is quantified, but its origin is not.

This distinction matters because the corrective actions that reduce each kind of uncertainty are entirely different. Parameter-dominated uncertainty is best addressed through additional streamflow observations that tighten the calibration. Structural uncertainty calls for model development, the introduction of better process algorithms, alternative formulations, or richer model topologies. Forcing-dominated uncertainty motivates investment in precipitation gauges, radar products, better weather models, or better satellite-derived inputs. Without an attribution, the practitioner has no principled basis on which to choose between these very differently scaled interventions, and resources allocated to uncertainty reduction cannot be directed where they are most effective.

Bayesian inference has given hydrology a principled framework for quantifying predictive uncertainty, and a generation of careful work has applied it to thousands of catchments worldwide (Kuczera and Parent, 1998; Kavetski et al., 2006; Vrugt et al., 2009). But two limitations have prevented Bayesian uncertainty analysis from telling the practitioner what to do next. The first is computational. Joint inference over hydrologic and statistical error model parameters routinely requires hundreds of thousands of model evaluations per

site, with Markov Chain Monte Carlo (MCMC) chains converging slowly, if at all, when the posterior geometry is complicated by parameter interactions or heavy-tailed residuals (Renard et al., 2010; Schoups and Vrugt, 2010). For the continental-scale and operational applications that the field increasingly demands, this cost is prohibitive. The second is interpretive. The Bayesian posterior typically conflates structural, parameter, and forcing uncertainty into a single, lumped predictive distribution. The total spread is faithfully quantified, but the modeller cannot decompose it from the posterior alone. The two limitations are linked, since efficient inference is the prerequisite for the large-sample studies that make a decomposition statistically meaningful.

This thesis addresses both limitations within a single methodological pipeline. It introduces a machine-learning-assisted Bayesian framework that halves the sampling cost of joint inference; a variance decomposition methodology that converts the lumped posterior into an interpretable, time-varying budget of structural, parameter, and forcing contributions; and a continental-scale application of that framework to 2,869 North American watersheds that reveals how the dominant uncertainty source varies systematically with climate and season. Together, these contributions allow the lumped predictive spread of a Bayesian analysis to be attributed to its underlying sources and used to guide hydrologic model improvement.

1.2 Sources of Hydrologic Predictive Uncertainty

Predictive uncertainty in hydrologic modelling arises from four broad sources, three of which can be explicitly parameterized within a process-based modelling framework, and a fourth that is typically treated implicitly through the likelihood formulation.

Structural uncertainty stems from the simplifications and approximations embedded in any mathematical representation of catchment hydrology. A lumped conceptual model treats a heterogeneous watershed as a single homogeneous unit; a particular subsurface flow algorithm assumes a linear or power-law storage relationship; an evapotranspiration scheme adopts one of several competing formulations. Each of these choices introduces structural error whenever the chosen representation diverges from the underlying physical process (Clark et al., 2011; Fenicia et al., 2008). Because no single model can represent every catchment in every regime without compromise, structural uncertainty is irreducible in any fixed-structure framework and is most consequential when the dominant runoff-generating processes shift seasonally or across catchment types.

Parameter uncertainty arises because the numerical coefficients of a hydrologic

model, transmissivities, storage capacities, routing coefficients, melt factors, are not directly observable and must be inferred from streamflow records through calibration. The inverse problem is typically ill-posed, since many different parameter combinations produce nearly equivalent fits to the observed hydrograph, a phenomenon known as equifinality (Beven and Binley, 1992; Beven, 2006). Even with a long observational record, the posterior of a calibrated process-based model retains substantial spread, and that spread propagates directly into predictive uncertainty.

Forcing (input) uncertainty reflects errors in the meteorological inputs, principally precipitation and temperature, that drive the model. Gauge undercatch, sparse spatial coverage, satellite retrieval bias, and orographic effects all contribute to systematic and random errors in catchment-scale forcing estimates (McMillan et al., 2012; Renard et al., 2011), and because hydrologic models are nonlinear, even modest forcing errors can produce disproportionately large streamflow deviations whose magnitude depends on flow regime, basin characteristics, and time of year.

Observational uncertainty in streamflow records, arising from rating-curve extrapolation, ice effects, and gauge operation, is the fourth principal source. It is acknowledged throughout the thesis but treated implicitly through the likelihood function rather than parameterized as a separate source, consistent with common practice in Bayesian hydrologic inference (Westerberg et al., 2011; McMillan et al., 2010). Under this treatment, observational error is absorbed into the residual term of the likelihood. It is not assigned a separate share in the decompositions that follow, nor is it assumed to be distributed evenly among the other three sources.

These three explicitly parameterized sources, structure, parameters, and forcings, are not independent. Structural error can be partially compensated by biased parameter estimates; forcing errors can propagate into apparent parameter spread; and parameter non-uniqueness can mask structural inadequacy. Disentangling them is a long-standing methodological challenge that has been recognized in the community but rarely solved in a unified, well-posed manner (Renard et al., 2010; Blöschl et al., 2019). This disentanglement, and what it reveals about where to invest in uncertainty reduction, motivates the methodological contributions of this thesis.

1.3 The Methodological Gap

The two limitations introduced in Section 1.1, computational cost and the conflation of uncertainty sources, are not independent. They share a common origin in the way Bayesian

inference is typically formulated for process-based hydrologic models, and they have to be addressed together.

On the computational side, the standard practice is to perform joint inference over the full vector of hydrologic and error model parameters. The error model, the statistical description of the discrepancy between simulated and observed streamflow, contains its own parameters governing residual heteroscedasticity, autocorrelation, and distributional shape (Evin et al., 2013; Schoups and Vrugt, 2010). When these are sampled jointly with the hydrologic parameters, the high-dimensional posterior often exhibits strong cross-correlations and elongated geometries that slow MCMC mixing (Gelman et al., 2013). A series of decoupling strategies, post-processing residual error model fitting (Woldemeskel et al., 2018; Romero-Cuellar et al., 2024), adaptive Metropolis (Haario et al., 2001), differential-evolution samplers (Vrugt et al., 2009; Laloy and Vrugt, 2012), and likelihood-free methods (Ong et al., 2018), have been developed to mitigate this cost. Each involves trade-offs between probabilistic coherence and computational tractability, and none yields the kind of substantial, scale-changing reduction in sampling cost that would make continental-scale Bayesian decomposition routinely feasible, because they reduce the per-step cost or improve mixing without removing the error model parameters from the sampled space.

On the interpretive side, the conventional Bayesian formulation typically fixes a single model structure and infers only its parameters. Structural uncertainty is then either ignored, approximated by a residual error term, or addressed externally through multi-model ensembles in which a discrete suite of competing structures is calibrated separately and their predictions averaged or weighted (Ajami et al., 2007; Thiboult et al., 2016; Troin et al., 2022). These approaches have produced important insights, but they treat structure as a discrete categorical choice rather than as a continuous source of inferential uncertainty. Continuous treatments of model error have also been proposed, including Gaussian-process discrepancy terms and related Bayesian model-discovery formulations, although these characterize structural error in the model output space rather than within the process representation itself. Renard et al. (2010) showed that simultaneously inferring separate input and structural error models from streamflow alone is, in general, ill-posed, because different combinations of input error, structural error, and parameter compensation can produce identical residual signatures, making the decomposition non-unique without independent constraints on at least one source.

Existing approaches therefore involve a trade-off. Tractable inferences address only one or two uncertainty sources at a time and accept that the decomposition will be incomplete. Ambitious decompositions attempt full multi-source inference but become statistically ill-conditioned, producing unstable estimates that the analyst cannot trust. A single methodological pipeline that delivers tractable inference and a well-posed decomposition,

efficient enough to apply at continental scale and structured so that the three uncertainty sources separate by construction rather than by post hoc statistical adjustment, has not previously been available. Developing such a pipeline is the aim of this thesis.

1.4 Thesis Objectives and Contributions

Motivated by the gap above, this thesis pursues four connected objectives:

- To exploit the conditional structure between hydrologic and error model parameters and develop a machine-learning-assisted Bayesian framework that materially reduces the sampling cost of joint inference without sacrificing probabilistic coherence;
- To formulate structural, parameter, and forcing uncertainty as components of a single well-posed Bayesian inference and to develop a variance decomposition that resolves their contributions to total predictive spread for each Julian calendar day;
- To deploy this combined framework at continental scale across thousands of watersheds and characterize how the dominant uncertainty source varies with hydroclimatic setting, physiography, and season;
- To translate the resulting uncertainty budgets into practical guidance, identifying for each landscape regime which intervention (additional observations, structural refinement, or improved forcing data) will yield the largest reduction in predictive uncertainty. Note that the thesis does not empirically evaluate how much any such intervention reduces predictive uncertainty.

The thesis answers these objectives through three methodological contributions, each developed into a peer-reviewed manuscript and presented as a separate chapter.

Contribution 1 (Chapter 2): A surrogate-aided framework for accelerated Bayesian calibration of hydrologic models. Published in *Environmental Modelling & Software* (2026).

Chapter 2 establishes that the computational bottleneck of Bayesian hydrologic calibration can be cut roughly in half by exploiting the fact that error model parameters (the coefficients governing residual heteroscedasticity and autocorrelation in the likelihood function) are deterministically conditional on hydrologic parameters once a model structure

and observation record are fixed. A support vector regression surrogate is trained offline on approximately one thousand pre-computed hydrologic parameter realizations, learns this conditional mapping, and is then deployed inside the No-U-Turn Sampler (NUTS) to predict error model parameters at every MCMC step rather than sampling them. Applied to a GR4J–CEMANeige configuration across the 12 MOPEX benchmark catchments, the framework achieves convergence with 49–52% fewer MCMC samples than conventional joint-inference benchmarks, halving wall-clock runtime while preserving or improving predictive skill. This efficiency gain is what makes the continental-scale work in Chapter 4 practical.

Contribution 2 (Chapter 3): Simultaneous Bayesian quantification and decomposition of structural, parameter, and forcing uncertainty in hydrologic models. Under preparation for submission to *Journal of Hydrology*.

Chapter 3 introduces the inferential framework that converts the joint Bayesian posterior into an interpretable, time-varying budget. Building on the blended modelling concept of Mai et al. (2020), in which the discrete choice between competing process algorithms is replaced by a continuously weighted average, structural uncertainty is reformulated as uncertainty in the weight-generating parameters. All three sources, structural (weights), parameter (process coefficients), and forcing (input perturbation), thus become components of a single Bayesian inference, and a multivariate analysis of variance (MANOVA) applied to the posterior ensemble for every Julian calendar day produces a continuous decomposition across the hydrological year. On the MOPEX12 sample, parameter uncertainty dominates the annual mean budget (median $\approx$ 53%), but structural uncertainty triples during summer–fall recession when competing baseflow algorithms diverge most strongly. The chapter establishes the methodology, validates it on a diverse but limited sample, and motivates the larger-scale deployment that follows.

Contribution 3 (Chapter 4): Continental-scale decomposition of hydrologic model uncertainty across 2,869 North American watersheds. Under preparation for submission to *Nature Communications*.

Chapter 4 deploys the framework across 2,869 watersheds drawn from the HYSETS database, a sample two orders of magnitude larger than any previously published multi-source decomposition study. The 430 million model simulations required are made possible by distributing computational jobs over the Digital Research Alliance of Canada high-performance computing clusters. Three findings emerge that small-sample studies cannot

resolve. First, structural uncertainty (continental median 42%) exceeds parameter uncertainty (36%) and forcing uncertainty (19%) at the continental scale, in contrast to the long-standing assumption that parameter uncertainty dominates hydrologic predictions. Second, climate, principally mean annual precipitation, precipitation variability, and temperature seasonality, explains roughly two-thirds of the spatial variance in structural share, while topography explains comparatively little; the uncertainty budget is therefore predictable from hydroclimatic setting alone. Third, structural uncertainty is non-stationary in time. At basins with prolonged summer recessions, whether snow-influenced or humid-eastern, it expands four- to six-fold during the recession season, while at semi-arid basins with flashy hydrographs the expansion is muted, indicating that the adequacy of any fixed process representation depends on the flow regime. Posterior weight distributions further yield a continental atlas of which process algorithm the data prefer at each location and how confidently.

1.5 Thesis Structure

The remainder of the thesis is organized as follows. Chapter 2 presents the surrogate-aided Bayesian calibration framework and its application to the MOPEX12 catchments. Chapter 3 develops the simultaneous structural–parameter–forcing decomposition methodology and validates it on the same MOPEX12 sample. Chapter 4 deploys the methodology at continental scale across 2,869 watersheds. Chapter 5 synthesizes the contributions and identifies directions for further research. The chapters are written to stand alone as journal manuscripts.

Chapter 2

A Surrogate-Aided Approach for Accelerated Bayesian Calibration of Hydrologic Models

Preface

A version of this chapter has been published in *Environmental Modelling and Software* (Arabzadeh et al., 2026). Minor formatting changes have been made for thesis presentation.

2.1 Introduction

Hydrologic modeling is widely used for understanding water systems and supporting decisions related to water management, flood forecasting, and climate adaptation. Whether using lumped conceptual models or spatially distributed process-based systems, hydrologists rely on models to interpret how rainfall becomes runoff and to predict how systems might behave under future scenarios. Recent efforts toward unified process-based frameworks seek to embed multiple component models under a single coding environment (Clark et al., 2015; Craig et al., 2020), but they still face challenges in calibration speed and uncertainty quantification.

A large body of work has emerged around understanding these uncertainties, particularly in the context of parameter estimation. A useful way to frame this uncertainty

landscape is through the method of multiple working hypotheses, which encourages testing multiple competing model structures in parallel (Clark et al., 2011). Models typically require calibration, which is the adjustment of parameters so that the simulated outputs align with observations, often streamflow at a watershed outlet. But multiple parameter sets can yield similar model performance—a challenge known as equifinality (Beven and Binley, 1992). In such cases, it becomes difficult to determine which parameters are "correct," and this ambiguity can limit confidence in both the model and its predictions. This issue is compounded when limited data are available or when models are applied in data-scarce environments. Large-sample hydrology studies have highlighted the tension between in-depth model testing and broad spatial coverage (Gupta et al., 2014b), motivating methods that can scale computationally without sacrificing rigor.

Bayesian inference has become a widely used tool to address some of these challenges. By framing model calibration as a probabilistic process, it allows uncertainty in model parameters—and, by extension, model predictions—to be formally quantified (Kuczera and Parent, 1998; Kavetski et al., 2006; Vrugt et al., 2009). Instead of seeking a single best-fit parameter set, Bayesian calibration yields a posterior distribution, highlighting a range of plausible solutions. This approach improves transparency and allows researchers to explore questions about model identifiability, predictive skill, and the structure of the residual errors (Renard et al., 2010; Thyer et al., 2009).

However, applying Bayesian methods to hydrologic models is not easy. One of the key ingredients in any Bayesian calibration is the *error model*—the statistical description of the difference between observed and simulated values. These models include parameters of their own, which must also be calibrated. Error models matter because they define the likelihood function used during inference, and their configuration has a strong influence on the results (Smith et al., 2015). Assessing the structural adequacy of rainfall–runoff models has been highlighted as critical for credible predictions (Gupta et al., 2014a), particularly when error formulations fail to separate input, structural, and parameter sources of uncertainty. But in practice, calibrating both hydrologic model parameters and error model parameters can lead to complex interactions, making the inference process slow or unstable (Marshall et al., 2005; Moges et al., 2021). Additionally, overly complex error models may degrade performance if their assumptions do not align well with the data (Schoups and Vrugt, 2010; Vrugt, 2024).

Researchers have made important strides in improving error models and the efficiency of inference techniques, such as adaptive MCMC (Haario et al., 2001), likelihood-free methods (Ong et al., 2018), and surrogate modeling (Razavi et al., 2012). However, many of these solutions still require simultaneous estimation of all model parameters, which can be computationally intensive and difficult to manage, especially for larger or more detailed

models. Inference is also sensitive to how well the error model is specified, since a mischaracterized error model can bias the hydrologic parameters and affect the realism of the results (Renard et al., 2010; Samadi et al., 2018).

While joint inference techniques have been successfully applied to the inference of hydrologic model parameters, the associated computational challenges have inspired various decoupling strategies. Traditional post-processing approaches, such as deterministic calibration followed by residual error model (REM) fitting, exploit the conditional dependence of error parameters on hydrologic parameters to simplify workflows (Woldemeskel et al., 2018; Romero-Cuellar et al., 2024). These methods, including the residual error model post-processor (REM-PP), avoid the computational burden of joint inference by treating error estimation as a downstream correction (Romero-Cuellar et al., 2024). While efficient, such sequential approaches sacrifice Bayesian coherence by not truly propagating hydrologic parameter uncertainty to the error model, potentially biasing predictive uncertainty (Renard et al., 2010).

Building on these conditional approaches, this work aims to combine the efficiency of post-processing with the rigor of joint inference. Like REM-PP, the proposed methodology preserves computational tractability by exploiting the deterministic relationship between hydrologic parameters and error model parameters. However, instead of treating error estimation as a separate step, a machine learning surrogate dynamically predicts REM parameters during MCMC sampling. This technique mirrors the efficiency gains of decoupled methods (e.g., avoiding unstable parameter interactions) while maintaining the probabilistic consistency of full Bayesian inference (Evin et al., 2013).

Therefore, this chapter introduces a methodology that facilitates the efficient inference of hydrologic model parameters and error model parameters. A formal probabilistic framework is adopted and the inference of error model-specific parameters and hydrologic model parameters is decoupled. The MCMC inference learns about the degree to which calibrated hydrologic model parameters are correlated. However, the decoupling of the error model recognizes that error model parameters are completely dependent on the set of hydrologic model parameters. Motivated by this conditional relationship, the key contribution of this work is the demonstration of how a machine learning model within an MCMC sampler can directly estimate error model parameters as a function of the hydrologic model parameter values.

2.2 Material and Methods

2.2.1 Experimental Design

In this study, the Surrogate-aided Heteroscedastic Autocorrelated error model (SHA) is implemented using a Gaussian distribution for the model residuals representation. This is benchmarked against three commonly used error models: a Heteroscedastic Autocorrelated error model with a Student's t-distribution for residuals (THA), a homoscedastic, uncorrelated error model with a Gaussian distribution (termed G), and a Gaussian Heteroscedastic Autocorrelated model (termed GHA). These approaches allow benchmarking of SHA against conventional error models used in the Bayesian inference of parametric hydrologic models. See Section 2.2.2 for details of the SHA, THA, G, and GHA error models. While multi-objective calibration strategies have been proposed to balance competing performance criteria (Fenicia et al., 2007), they still generally require joint estimation of model and error parameters, leading to increased computational loads.

2.2.2 Proposed Workflow

Generally, the inference of hydrologic model parameters is more challenging than the parameter inference for simple statistical models with a similar number of parameters due to several reasons, including nonlinear processes, data scarcity, and structural uncertainties. Here, interactions refer to the statistically significant relationships between error model parameters and water storage compartments. This increased challenge makes it extremely hard—or usually impossible—to derive analytical solutions for calculating the joint posterior distribution. As such, this joint inference problem is typically solved using a numerical sampling approach.

Markov Chain Monte Carlo (MCMC) samplers are the standard numerical sampling approach for estimating the joint posterior distribution (Evin et al., 2013). However, they can be problematic for the same reasons that prohibit an analytical solution. Additionally, dynamic error model parameters are often more difficult to infer than static ones, leading to undesired interactions with water balance compartments. To avoid this interaction issue, one may consider the error model parameters to be constant during the joint inference process, simplifying the complexity of estimation (McInerney et al., 2017).

Bayesian inference estimates the posterior distribution of parameters by combining prior distributions (initial beliefs) with the likelihood function (which measures how well the parameters explain the observed data), and MCMC provides the numerical sampling

used to approximate it. MCMC explores the parameter space by sampling more frequently from regions with higher posterior probability, guided by a proposal function. In hydrologic modeling, focusing the proposal function solely on the hydrologic model parameters can improve convergence and is often a practical choice. However, this approach may lead to distorted posterior distributions since it neglects the joint inference of the error model parameters.

Likelihood Function Configurations

The likelihood functions used in this study are based upon the formulations proposed by Evin et al. (2013), which incorporate model residuals autocorrelation and heteroscedasticity in the error model.

Let the model residuals be defined as below:

$$\epsilon_t = Q_t^o - \hat{Q}_t(S_0, \theta, X_t^o) \tag{2.1}$$

where Q_t^o is the streamflow observation, and $\hat{Q}_t$ is the streamflow from the model given the model forcings, X_t^o, initial conditions, S_0, and some parameters θ.

To address the heteroscedasticity in the residuals, standardized residuals (η_t) are used as below:

$$\eta_t = \frac{\epsilon_t}{\sigma_\epsilon(t)} \quad \text{s.t.} \quad 1 \leq t \leq N \tag{2.2}$$

where $\sigma_\epsilon(t)$ is the unknown variance of the error in time step t. Using this treatment, the variance stabilizes with respect to the flow magnitude; thus, it is desirable to correlate the residuals with flow magnitude using a transformation of the form $\sigma_\epsilon(t) = f(Q_t, a, b)$. Here, the following form is used (Evin et al., 2013):

$$\sigma_\epsilon(t) = a + b\, Q_t \tag{2.3}$$

where a and b are the parameters of the transformation. This linear flow–variance relationship follows the standard formulation of Evin et al. (2013). Although nonlinear (e.g., power-law) alternatives are possible, they were not tested here.

Similarly, lag-1 autocorrelated residuals can be addressed by assuming the following relation between each pair of consecutive residuals linked by a ϕ_η multiplier:

$$\eta_t = \phi_\eta \eta_{t-1} + \nu_t \quad \text{s.t.} \quad \nu_t \sim \mathcal{N}(0, \sigma_\nu^2) \tag{2.4}$$

A lag-1 (first-order) autoregressive structure is adopted following Evin et al. (2013), as it captures the dominant short-range residual dependence at the daily time step while keeping the error model parsimonious.

Since the likelihood function is a joint PDF of the residuals, it must account for the transformation described in Equation 2.3 and Equation 2.4 (Evin et al., 2013):

$$\epsilon_t = f(\eta_t) \tag{2.5}$$

First, the absolute value of the determinant of the Jacobian matrix of the transformation used in Equation 2.3 is calculated (Evin et al., 2013):

$$p(\epsilon_t, \eta_t) = \det \left| \frac{\partial \eta_t}{\partial \epsilon_t} \right| p(\eta_t) = \prod_{t=1}^{N} \frac{1}{\sigma_{\epsilon_t}} \tag{2.6}$$

Then the likelihood function is derived as the joint distribution of all standardized residuals as below:

$$p(\eta_1, \ldots, \eta_N) = \frac{p(\eta_1)}{\sigma_\epsilon(1)} \prod_{t=2}^{N} \frac{p(\eta_t \mid \eta_{t-1})}{\sigma_\epsilon(t)} \tag{2.7}$$

Assuming a Gaussian distribution for the random component of residuals, the conditional probability of the transformed residuals, $p(\eta_t \mid \eta_{t-1})$, can be expressed as $p(\eta_t \mid \eta_{t-1}) = \mathcal{N}(\phi_\eta \eta_{t-1}, 1 - \phi_\eta^2)$ (Evin et al., 2013). So, the likelihood function becomes:

$$p(\eta_t, a, b, \phi_\eta) = \frac{p(\eta_1)}{a + b\hat{Q}_1} \prod_{t=2}^{N} \frac{\mathcal{N}(\phi_\eta \eta_{t-1}, 1 - \phi_\eta^2)}{a + b\hat{Q}_t} \tag{2.8}$$

Here the likelihood function can be configured by a set of parameters, $\omega = [\phi_\eta, a, b]$, and a Gaussian distribution to measure the likelihood of simulation and observations given a set of ω. Hereafter, the error model that configures the above-mentioned likelihood function using a Gaussian distribution with autocorrelated and heteroscedastic residuals is referred to as GHA.

The homoscedastic and independent residual Gaussian distribution likelihood function (G) is also accessible when $\phi_\eta = 0$ and $b = 0$.

Likewise, a likelihood function is derived that utilizes heteroscedastic and autocorrelated residuals with a Student-t distribution model to characterize the residuals. Based on

the logic described above, a likelihood function can be derived with a heteroscedastic and autocorrelated error model with Student-t distribution (THA) for residuals as below:

$$p(\epsilon_t, a, b, \nu, \phi_\eta) = \left[\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{1/2}}\right]^{N-1} \prod_{t=2}^{N} \left[\frac{1}{a+b\hat{Q}_t}\left(1+\frac{(\epsilon_t-\phi_\eta\epsilon_{t-1})^2}{\nu(a+b\hat{Q}_t)^2}\right)^{-\frac{\nu-1}{2}}\right] \tag{2.9}$$

where ν is the Student-t degree of freedom. Thus, let $\omega = [\phi_\eta, a, b, \nu]$.

Surrogate Model Development

From Equation 2.1, the model residuals are defined as the difference between the observations and simulation streamflow quantities. In deterministic applications, ϵ_t can be considered as a vector of residuals—the difference between simulations and observations. In contrast, when calibrating models probabilistically, a likelihood function is used to measure how well the model performs based on the observations given a set of parameters which require an error model to describe the characteristics of the raw errors, typically expressed as a distribution. Given residuals $Q_t^o - \hat{Q}_t$, conditioned on some initial state of the system, S_0, hydrologic model exclusive parameters, θ, and observation forcing functions, X_t^o, the log-likelihood function can be calculated as below:

$$L(\Psi = [\omega, \theta]) = \log(p((Q_t^o \mid S_0, \theta, X_t^o) \mid \omega)) \tag{2.10}$$

where ω is the log-likelihood-function parameters, including the error model parameter(s) and other likelihood function parameter(s). Within the Bayesian framework, it is common to have the errors systematically characterized through "formal likelihood functions" (Schoups and Vrugt, 2010). Formal likelihood functions typically encapsulate the probabilistic characteristics of model errors, alongside other components such as autocorrelation and heteroscedasticity, where relevant.

The discourse surrounding the choice and suitable application of formal likelihood functions has been a focal point in the Bayesian inference for hydrologic modeling literature (e.g., Mantovan and Todini, 2006; Beven et al., 2008; Stedinger et al., 2008; Vrugt et al., 2022). Given the ongoing debate regarding formal likelihood functions, there is a necessity to appropriately assess and choose between formal likelihood models.

Given that the GR4J model configuration includes four parameters, the log-likelihood function of the hydrologic model can be expressed as a joint distribution of the hydrologic model and error model parameters.

If $f_{\text{pri}}(\Psi)$ is the joint distribution of the parameter priors, it is possible to express the posterior distribution using Bayes' theorem as below:

$$f^{L}_{\text{post}}(\Psi = [\theta, \omega], Q^{o}_{t}, X^{o}_{t}) = \frac{p_{Q^{o}_{t}|\Psi}(Q^{o}_{t}, \Psi, X^{o}_{t}) f_{\text{pri}}(\Psi)}{p_{Q^{o}_{t}|\Psi'}(Q^{o\prime}_{t}, \Psi', X^{o\prime}_{t}) f_{\text{pri}}(\Psi')} \tag{2.11}$$

This study focuses on the vector of ω parameter(s), which configures the log-likelihood function. It is common that both error model parameters, ω, and hydrologic parameters, θ, are inferred concurrently. However, challenges related to inference algorithm performance—such as the high sensitivity of the log-likelihood function to the ω and ω interactions with hydrologic model components—have been reported in the past (Renard et al., 2010). Commonly, inference algorithm performance and convergence become problematic when using a joint proposal function for both log-likelihood function parameters and hydrologic model parameters. Due to the inherent stochasticity of the samplers' proposals and the curse of dimensionality, water budget components become more volatile, generally resulting in erratic and poor MCMC sampler acceptance rate (Renard et al., 2010).

The stochastic aspect of the proposal function allows the sampler to explore the space of error model parameters. However, given the proposal function's explorative capability, achieving convergence can be delayed, particularly in the initial sampling stages. This delay may occur because the proposal function generates poorly informed ω samples.

In this study, instead of the joint inference of $[\theta, \omega]$, the proposal sampling process between θ and ω is decoupled, and ω is deterministically inferred conditional on the values of the hydrologic parameters θ as below:

$$\omega^{t}_{i} = M_{i}(\theta^{t}, \delta) \tag{2.12}$$

where M_i represents a deterministic function that maps θ^t at the t-th stage of the sampling process, given some additional calibration parameters δ. Residual emulator error is not considered in this process as a source of uncertainty. Instead, residual errors are assumed to be 100% dependent on θ. This assumption could function to narrow the predictive intervals since the method assumes no error model structure uncertainty and no uncertainty in the error model parameters (e.g., due to computing parameters on a finite sample of residuals). This essentially serves as a surrogate proposal model that deterministically maps the sampled hydrologic parameters, θ, to the likelihood function parameter(s), ω. This could be done through a statistically trained machine learning model that infers a relatively optimal ω based on θ as an input to the model, M_i. Note that the surrogate emulates this parameter-to-parameter mapping only. It does not model the sampler's

trajectory through the parameter space. Thus, the parameter posteriors can be updated as below, incorporating the parameter conditioning described in Equation 2.12:

$$f^{L}_{\text{post}}(\theta, \omega, \epsilon_t, x_t) = \frac{p_{Q_t^o|\theta,\omega}(Q_t^o, X_t^o, (\theta \mid \Psi)) f_{\text{pri}}(\Psi)}{p_{Q_t^{o\prime}|\theta',\omega'}(Q_t^{o\prime}, X_t^{o\prime}, (\theta' \mid \Psi')) f_{\text{pri}}(\Psi')} \tag{2.13}$$

In order to define the surrogate model, M_i, it is essential to have a training dataset. This dataset allows adjustment of the vector of δ parameter(s), needed to configure the model M_i, which in turn enables encompassing both hydrological parameters and their associated error model parameters during model calibration. To this end, N sets of hydrologic parameters, θ, are sampled from the parameters' prior. These N-sets of parameters are evaluated by the GR4J hydrologic model. Subsequently, each parameter set, θ_s^i, yields a corresponding vector of residuals, ϵ_t^i, for which an error distribution with a vector of parameters of ω_s^i can be fitted. In summary, the following transformation is applied: $\theta_s^i \rightarrow \omega_s^i$ such that $1 \leq i \leq N$.

This methodology relies solely on the N initial training parameter sets, since additional information does not contribute significantly to improving the surrogate model and only increases computational cost. However, the approach could potentially be extended by adaptively incorporating future data points as the MCMC progresses, which may further refine the surrogate model parameters—particularly when the SVR tends to improve for larger sample sizes. Figure 2.1 summarizes the proposed methodology workflow and details these steps. This workflow consists of three main stages: (1) hydrologic parameters training set generation and evaluation, (2) surrogate model deployment, and (3) uncertainty analysis to derive posteriors of likelihood function and hydrologic model parameters.

In the training set generation, N-sets of hydrologic parameters are generated by Monte Carlo sampling, drawn uniformly within the parameter bounds, θ_s^n where $n \in [1 \ldots N]$; each set is then evaluated in a Raven GR4J model emulation. No prior calibration is required to generate the training set since the hydrologic parameters are sampled directly from their feasible ranges. Then, using each model output instance, the error model parameters, ω_s^n, are derived and processed to the next stage. In the next step, in the surrogate model deployment, the surrogate model is trained on the θ_s^n values where each member of ω_s^n is set as the surrogate model response and the vector θ_s^n as the regressors for all $n \in [1 \ldots N]$.

In this work, a support vector regression (SVR) model is used as the surrogate model. SVR was chosen for its proven efficiency in high-dimensional regression tasks with limited training data (Liong and Sivapragasam, 2002), aligning with the goal of mapping hydrologic parameters (θ) to error parameters (ω) using $N = 1000$ samples. SVR employs an ϵ-insensitive loss function that treats deviations between predicted and true ω within $\pm\epsilon$ as

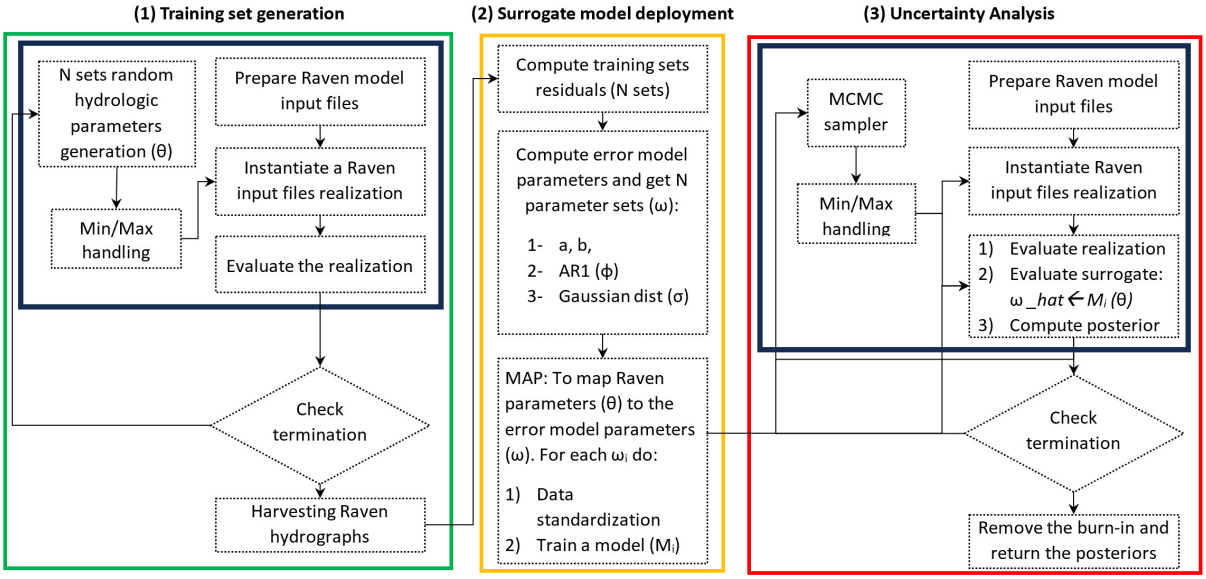


Figure 2.1: Surrogate error model methodology workflow. In the left panel, training sets are generated using a Monte Carlo simulation considering the parameter bounds. In the middle panel, a Support Vector Machine (SVM) model is trained over the training set from the left panel. In the right panel the SVM is deployed in the MCMC sampling procedure to infer the error model parameters conditional on the sampled hydrologic model parameters.

zero loss; by ignoring small errors, the regressor focuses on larger, systematic discrepancies and reduces sensitivity to random noise in (θ, ω) pairs produced by MCMC sampling (Smola and Schölkopf, 2004). While deep learning alternatives offer greater flexibility, they typically require larger training sets and extensive hyperparameter tuning (Razavi et al., 2012), increasing computational overhead without guaranteed accuracy gains for this specific application.

Calibration Parameters and Inference Algorithm

The sampling performed in this study includes both the hydrologic model parameters and additional parameters ($\theta = \{X_{1-4}, snow_{coef}, temp_{lapse}, precip_{lapse}\}$) associated with the snow model and forcing adjustment parameters (air–snow coefficient, temperature lapse rate, precipitation lapse rate). Thus, the experiments presented here reflect an emulator of GR4J with the CEMANeige snow model (Valéry et al., 2014) with lapse-rate adjustments rather than a "vanilla" GR4J-only calibration. Table 2.1 lists the parameters calibrated across the error model experiments.

Table 2.1: Hydrologic and error model parameters to be inferred. Filled circles (•) indicate parameters directly inferred by MCMC samplers; open circles (○) indicate parameters generated by the surrogate model conditional on MCMC samples.

Parameter	Prior	Range	Unit	Error models			
				G	GHA	THA	SHA
X_1 (topsoil thickness)	vague	$[0, 1]$	m	•	•	•	•
X_2 (groundwater exchange rate)	vague	$[-5, 5]$	mm/d	•	•	•	•
X_3 (reference storage)	vague	$[5, 200]$	mm	•	•	•	•
X_4 (time routing)	vague	$[0.1, 5.0]$	d	•	•	•	•
Air–snow coefficient ($snow_{coef}$)	vague	$[0.2, 1.5]$	1/d	•	•	•	•
Precipitation lapse ($precip_{lapse}$)	vague	$[0, 10]$	mm/d/km	•	•	•	•
Temperature lapse ($temp_{lapse}$)	vague	$[0, 7]$	°C/km	•	•	•	•
a (transformation parameter)	vague	$[-1000, 1000]$	–		•	•	○
b (transformation parameter)	vague	$(0, 0.2]$	–		•	•	○
σ (Gaussian distribution parameter)	vague	$(0, 100]$	m^3/d	•	•		○
ϕ (AR1 parameter)	vague	$[0.2, 0.9]$	–		•	•	○
ν (t-Student parameter)	vague	$[2, 30]$	–			•	

The analyses were conducted within the R statistical environment (R Core Team, 2022) on a machine equipped with a 24-core processor operating at 2.4 GHz and 20 GB of RAM.

This configuration provided the necessary computational power to handle the resource-intensive tasks involved in the study. The LaplacesDemon R package (Statisticat LLC, 2021) is utilized for Bayesian inference, which offers a comprehensive environment with a wide range of algorithms suitable for various problems, from simple to advanced, all implemented in R.

The No-U-Turn Sampler (NUTS) from LaplacesDemon for parameter inference is employed. NUTS is an extension of Hamiltonian Monte Carlo (HMC) that automatically determines the appropriate number of steps to take during sampling. This feature eliminates the need for manual tuning of the trajectory length, which can be challenging and time-consuming in traditional HMC implementations. NUTS dynamically adapts the trajectory length based on the geometry of the posterior distribution, ensuring that the sampler does not retrace its steps—a situation referred to as "U-turns." This capability enhances the efficiency and accuracy of the sampling process, allowing NUTS to explore the parameter space effectively and reduce autocorrelation among samples.

Diagnostic Metrics

Two key metrics are utilized to assess the performance of the inference methods: the Continuous Ranked Probability Score (CRPS) (Hersbach, 2000) and the Gelman–Rubin diagnostic (Gelman and Rubin, 1992).

The Continuous Ranked Probability Score (CRPS) is employed to evaluate the accuracy of the probabilistic predictions. CRPS measures the quadratic difference between the cumulative distribution of the prediction and the cumulative distribution of the observed outcomes. A lower CRPS value indicates a closer match between the predictions and the observations, with a perfect score of zero representing an ideal forecast. Mathematically, the CRPS for a given observation–forecast pair is expressed as:

$$\mathrm{CRPS}(k) = \int_{-\infty}^{\infty} \left(F_k'(x) - F_k^o(x)\right)^2 dx \tag{2.14}$$

where $F_k'(x)$ represents the cumulative distribution function (CDF) of the forecast for the kth realization, and $F_k^o(x)$ is the CDF corresponding to the observed outcome. In this study, the probabilistic predictions of streamflow are generated using 10% of the budget samples drawn randomly from the posterior distribution after burn-in.

The range of the CRPS extends from 0 to infinity. For instance, a CRPS of zero indicates a perfect prediction, meaning the predicted CDF matches the observed distribution exactly. The higher the CRPS, the larger the discrepancy between the prediction and the

observations. There is no upper bound, as the score grows larger as the predicted and observed deviations increase. Since CRPS is computed for each streamflow prediction on each day, the average CRPS is computed and reported across the entire calibration period.

To further isolate the two components that contribute to CRPS—sharpness and reliability—two complementary metrics frequently used in probabilistic hydrologic forecasting are adopted (Gneiting et al., 2007). Sharpness measures the concentration of the predictive distribution relative to the mean observed discharge, while reliability measures the proportion of observations captured within the predictive credible interval. They are defined as follows:

$$\text{Relative Sharpness} = 1 - \frac{\sum_i (U_i - L_i)}{\sum_i Q_i} \tag{2.15}$$

$$\text{Reliability} = \frac{1}{N} \sum_i \mathbf{1}\{L_i \leq Q_i \leq U_i\} \tag{2.16}$$

where U_i and L_i denote the 97.5th and 2.5th percentiles of the simulated discharge at time step i, Q_i is the observed discharge, and N is the number of time steps. For both metrics, values closer to one indicate better performance—perfect coverage for reliability and narrower, more confident predictions for sharpness.

Additionally, the Gelman–Rubin (GR) statistic is employed to assess the convergence of the MCMC simulations. This metric compares the variance within each MCMC chain to the variance between chains. A value of GR close to 1 indicates that the chains have converged, meaning that the parameter estimates are reliable, and the simulation has sampled from the target posterior distribution. The Gelman–Rubin metric is calculated as:

$$\widehat{GR} = \sqrt{\frac{\widehat{V}}{W}} \tag{2.17}$$

where $\widehat{V}$ is the estimated variance of the target distribution, and W is the average within-chain variance.

In this study, each of the four experiments (SHA, THA, GHA, and G) was run for three MCMC chains with 50 different sample sizes ranging from 1,000 to 50,000, incremented by 1,000. This yielded a total of 600 experiments, and 15.3 million hydrologic model runs for each of the 12 MOPEX basins.

Each model run took approximately 0.3 s. However, by utilizing 24 cores for parallel computation, all experiments were completed for all 12 basins in roughly four weeks. This computational effort was necessary to ensure the robustness of the results across different sample sizes and chains. Note that MCMC sampler burn-in is defined as the sampling

results prior to the sampler reaching the high density regions of the posterior, and this can be specified by a threshold value of the Gelman–Rubin statistic.

2.2.3 MOPEX12 Case Study Watersheds

In this study, MOPEX12 watersheds were used to test the proposed methodologies. The Model Parameter Estimation Experiment (MOPEX) (Duan et al., 2006) is an international initiative focused on advancing parameter estimations in hydrologic models. The MOPEX database encompasses 438 catchments across the continental United States. Specifically, during the second and third MOPEX workshops, 12 of these catchments were carefully chosen to form a benchmark set, referred to in this study as "MOPEX12". For comprehensive information on the second and third MOPEX workshops, readers can refer to Schaake et al. (2006).

The MOPEX12 catchments, covering the southeastern portion of the United States as shown in Figure 2.2, were selected due to their diverse hydrologic characteristics, making them a suitable set for evaluating various model structures under a range of hydrologic conditions.

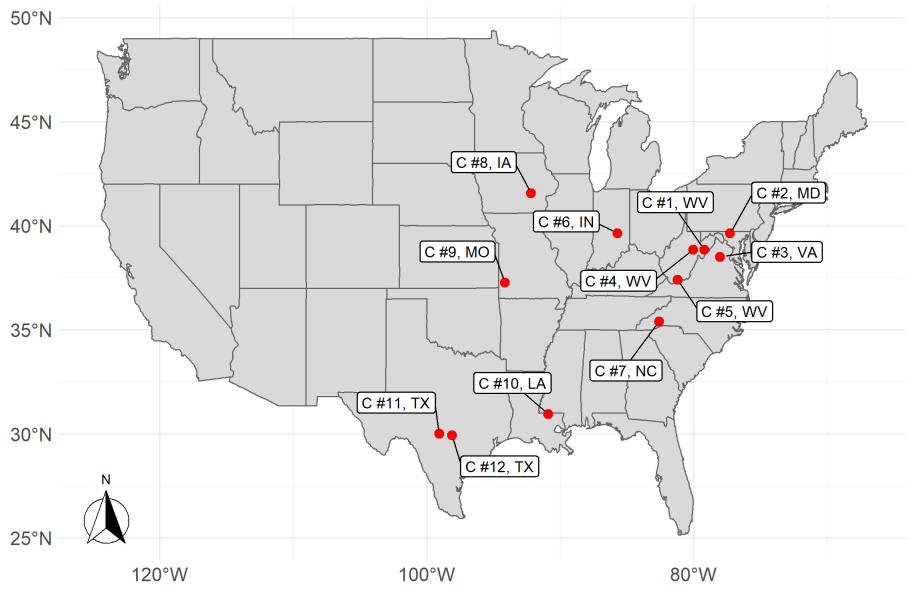


Figure 2.2: Geographic locations of MOPEX12 sites across the United States. The numbers on the map correspond to the numbering in Table 2.2.

Details regarding the catchment characteristics of MOPEX12 can be found in Table 2.2, derived from 56 years of available forcing data spanning 1948–2003. This table includes major basin characteristics such as mean annual temperature, precipitation, and evapotranspiration, and specifies the geophysical properties necessary for setting up the hydrologic model.

Table 2.2: The MOPEX12 watershed specifications.

ID	USGS gauge ID	Basin	State	Lat. (°)	Lon. (°)	Area (km^2)	Elev. (m)	Slope (°)	P (mm)	PET (mm)	Q (m^3/s)
1	1608500	Potomac	WV	38.85	−79.21	3782	661.4	11.0	935	762	1414
2	1643000	Monocacy	MD	39.63	−77.26	2115	194.6	3.1	1059	898	847
3	1668000	Rappahannock	VA	38.50	−78.04	4129	207.3	4.2	1078	921	1481
4	3054500	Tygart Valley	WV	38.84	−80.05	2366	718.7	8.6	1318	711	1755
5	3179000	Bluestone	WV	37.39	−81.20	1023	793.2	8.3	984	741	418
6	3364000	East Fork	IN	39.64	−85.70	4419	268.8	0.6	1032	856	1681
7	3451500	French Broad	NC	35.39	−82.60	2446	829.8	8.8	1510	820	1831
8	5455500	English	IA	41.56	−92.18	1486	253.5	1.6	867	995	357
9	7186000	Spring	MO	37.26	−94.15	3013	331.5	0.8	1079	1095	856
10	7378500	Amite	LA	30.95	−90.91	3314	82.4	1.0	1575	1074	1831
11	8167500	Guadalupe	TX	30.00	−99.06	3404	544.4	2.7	772	1529	359
12	8172000	San Marcos	TX	29.93	−98.10	2169	298.8	1.8	844	1449	361

All 12 MOPEX catchments were modeled as single basins, each represented by a single Hydrologic Response Unit (HRU) within the Raven framework (Craig et al., 2020). By treating the entire catchment as a homogeneous unit, the lumped model focuses on aggregated hydrologic behavior, which can reduce computational complexity but may overlook finer-scale processes that are spatially distributed across the catchment.

2.3 Results

Results below are first evaluated in Section 2.3.1, which examines the MCMC sampling convergence (GR metric). Then, in Section 2.3.3, regardless of successfully achieving numerical convergence of the GR metric, the randomly selected 10% of all sampling experiment parameter sets are utilized to make probabilistic predictions as if they had converged (see Shafii et al., 2014). This comparison of probabilistic performance deliberately evaluates predictive quality at matched computational budgets rather than only at full convergence, so that the error models are compared on an equal computational cost basis (see Shafii et al., 2014). In both sections, a detailed look at an example of the Guadalupe watershed is presented and then followed by a summary analysis across all 12 case study watersheds.

2.3.1 Convergence Assessment at Guadalupe Watershed

Figure 2.3 shows the Gelman–Rubin metric calculated for different sample sizes across the four methods (SHA, THA, GHA, and G). The results shown here factor in the N-set training data simulation, where all SHA data points are adjusted by $N = 1000$ to account for the simulations needed to prepare the training data for the SVR setup.

The training set size ($N = 1000$) was empirically determined as the minimum required to achieve robust SVR performance, yielding mean $R^2 = 0.89 \pm 0.06$ for ω prediction across all MOPEX12 watersheds. SVR was adopted over other alternatives because the $\theta \rightarrow \omega$ mapping is nonlinear, which a linear method such as multiple linear regression cannot capture, and because SVR performs reliably in high-dimensional regression with limited training data (Liong and Sivapragasam, 2002). However, more complex applications—such as models with higher-dimensional parameter spaces (e.g., HBV with 15+ parameters) or multi-component error models—may require many more samples to train the model.

Figure 2.3 evaluates the convergence behavior of the four error model configurations during hydrological model parameter inference for the Guadalupe watershed, using the GR metric. All error models show decreasing GR trends, but their convergence rates vary mainly due to differences in error-model complexity. The surrogate model with heteroscedastic and autocorrelated residuals (SHA) achieves the fastest convergence, dropping below 1.1 at a sample size of roughly 40,000. This supports the idea that informed use of the parameter space can improve convergence in complex Bayesian inference problems. In contrast, the THA (t-distributed heteroscedastic/autocorrelated) and GHA (Gaussian heteroscedastic) models exhibit slower convergence, with GR values persisting above 1.1 for all sample sizes evaluated. This aligns with Gelman et al. (2013), who emphasized that complex error structures (e.g., heavy-tailed distributions in THA) introduce parameter correlations and elongated posterior geometries, challenging MCMC mixing. The t-distribution in THA, while robust to outliers, increases sampling inefficiency due to its non-Gaussian tails. Similarly, GHA's slower convergence mirrors the findings of Renard et al. (2010), who noted that overly simplistic error assumptions conflate input, structural, and parameter uncertainties, complicating predictive reliability.

Methods such as Laloy and Vrugt (2012), which combine parameter-space exploration with adaptation strategies, improve MCMC efficiency in high-dimensional hydrologic models. SHA provides efficiency gains through a different route. SHA also achieves efficiency gains within a spatially coherent hierarchical framework, without incurring the significant computational overhead highlighted by Schoups and Vrugt (2010) when formalizing likelihoods for complex errors like heteroscedasticity and autocorrelation.

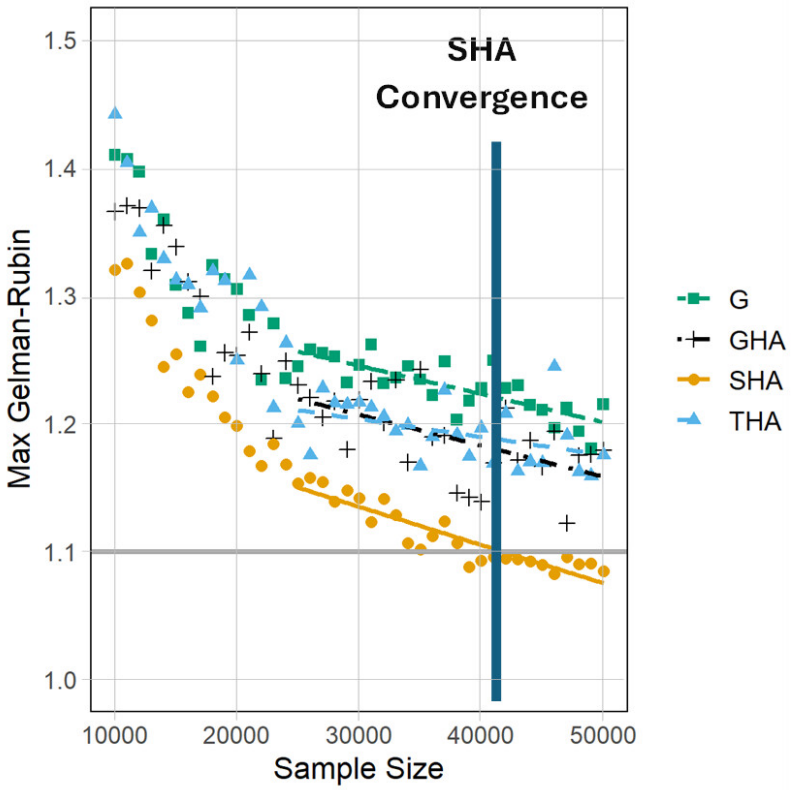


Figure 2.3: Maximum Gelman–Rubin diagnostic values across seven hydrologic model parameters inferred directly by the MCMC sampler as a function of sample size (model evaluations) for the Guadalupe watershed. The horizontal line at 1.1 indicates the convergence threshold (Gelman et al., 2004). SHA: surrogate error model with heteroscedastic and autocorrelated residuals; THA: t-distribution error model with heteroscedasticity and autocorrelation; G: Gaussian error model with homoscedastic and uncorrelated errors; GHA: Gaussian error model with heteroscedastic and autocorrelated errors.

The efficiency of the SHA model is relevant to operational hydrology, where rapid updates to operational models are needed or model calibration is computationally expensive. This method is suited to real-time streamflow forecasting and large-scale hydrologic modeling that require repeated recalibration. In these settings, reducing sampling requirements directly shortens update times and lowers computing costs, enabling more frequent recalibration and faster delivery of probabilistic forecasts. However, as cautioned by Renard et al. (2010), oversimplified error assumptions can lead to biased uncertainty estimates, particularly under nonstationary climatic conditions, which emphasizes the need to balance model complexity with spatial coherence to improve projection robustness (Pan et al., 2019).

To visualize the practical implications of these probabilistic diagnostics, Figure 2.4 displays a representative hydrograph for Guadalupe, Texas, simulated using posterior samples from the SHA method. The shaded region represents the 95% credible interval, while the solid line denotes observed discharge. The predictive band successfully tracks both baseflow and peak events, with varying width that reflects flow-dependent uncertainty. This example illustrates how the surrogate-aided inference translates into realistic uncertainty representation in simulated runoff, complementing the aggregated performance metrics shown in Figure 2.6.

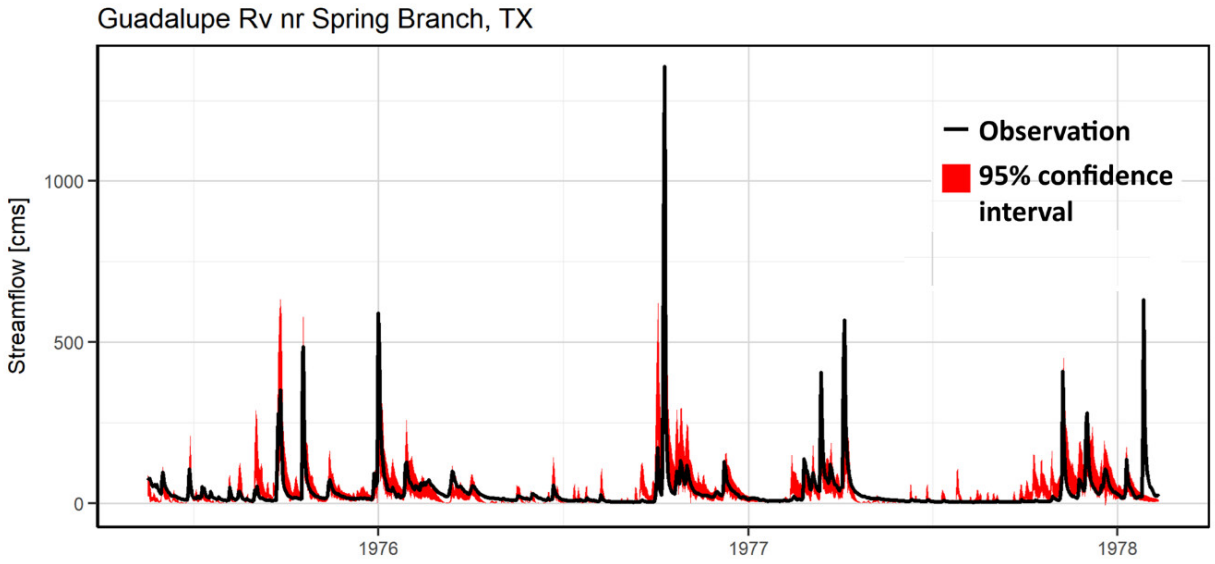


Figure 2.4: Simulated streamflow predictive uncertainty from the SHA method and its correspondence with the observed time series from 1975 to 1978.

2.3.2 Gelman–Rubin Metric Across MOPEX12

All MOPEX12 watersheds were analyzed and, for each parameter, the difference in sample size between the point where the SHA trendline crossed the Gelman–Rubin convergence threshold ($GR = 1.1$) and the point where the fastest-converging method among GHA, THA, and G crossed the same threshold was computed.

Figure 2.5 presents a boxplot summarizing these differences and showing the variability across parameters. The boxplot conventions used here are standard: the central line is the median, the box edges mark the 25th and 75th percentiles, whiskers extend to $1.5 \times$ IQR, and open circles indicate outliers. Across parameters, the central tendency is on the order of $\sim$ 10,000 samples, though per-parameter medians vary substantially. The reported "overall" median is a compact summary; individual parameter medians (e.g., $X_1 \approx 31{,}000$; $X_4 \approx 7{,}000$) should be read from the per-parameter descriptions. However, there is some variability, as indicated by the spread of the boxes and whiskers, with parameters such as X_2 and X_3 showing slightly larger differences. Figure 2.5 summarizes the sample-size requirements of SHA relative to the other methods. The method consistently requires fewer samples to achieve convergence compared to other methods across various watersheds, confirming computational advantage in large-scale applications (see Figure 2.7).

For the topsoil thickness parameter (X_1), SHA demonstrates the largest convergence advantage, with sample size differences ranging from approximately 14,000 to nearly 49,000 and a median near 31,000. This indicates a substantial improvement in convergence, though variability across watersheds remains considerable. By contrast, the reference storage (X_3) and time routing (X_4) parameters exhibit much tighter distributions. For instance, X_4 spans about 4,000 to 9,000 samples with a median near 7,000, suggesting more consistent performance and fewer samples required across watersheds.

The X_2 parameter shows intermediate variability, with values ranging from roughly 15,000 to 24,000 and a median around 19,000.

The air–snow coefficient also displays moderate variability, ranging from about 20,000 to 28,500 with a median near 22,000. In contrast, the lapse-rate parameters are notably more stable. The precipitation lapse-rate spans approximately 11,000 to 18,000 with a median near 12,000, while the temperature lapse-rate shows one of the most consistent patterns, ranging from about 5,500 to 15,500 with a median close to 12,000.

Taken together, these results indicate that while parameters such as X_1 and the air–snow coefficient reveal wider variability, most parameters exhibit more stable convergence behavior. Overall, SHA achieves substantially faster convergence than alternative methods across the parameter set, with typical gains equivalent to many thousands of samples.

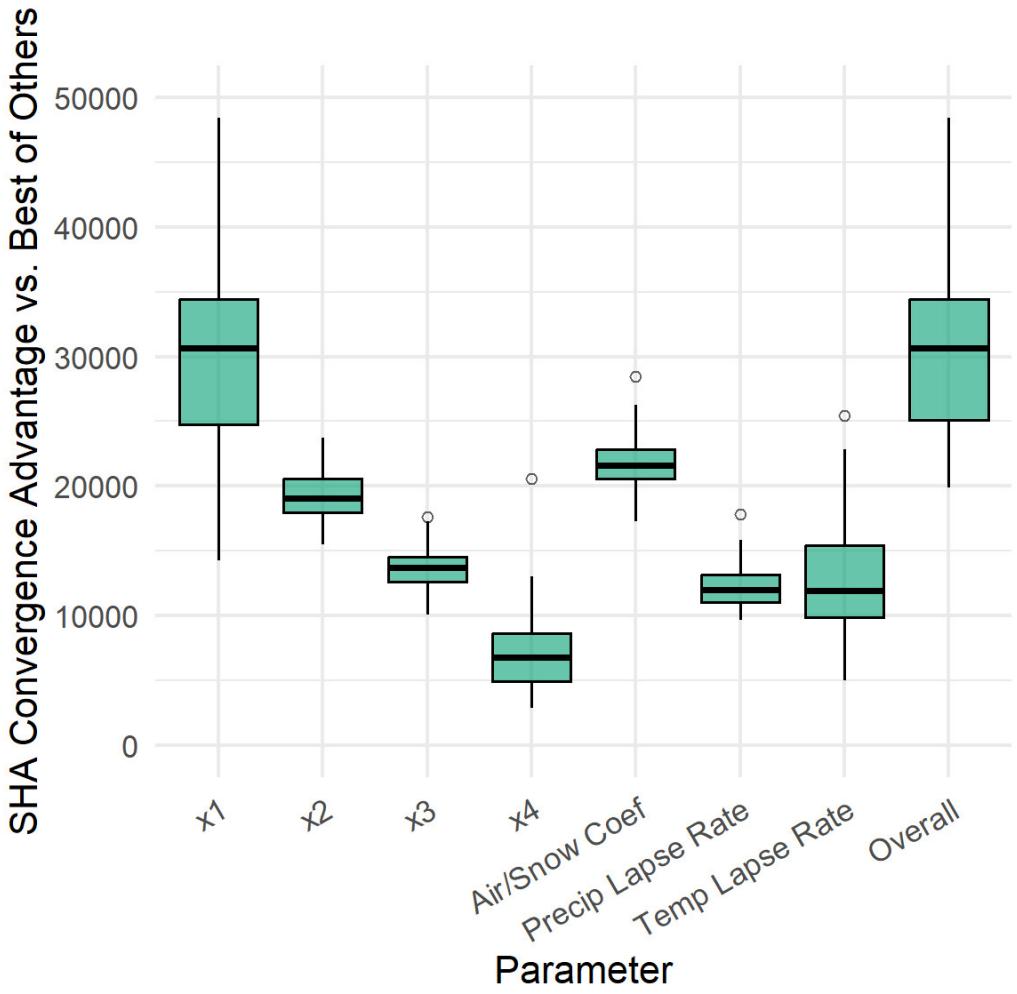


Figure 2.5: The distance between the SHA and average THA, GHA, and G trendlines at the Gelman–Rubin cutoff point for all 12 MOPEX watersheds.

2.3.3 Model Quantile Posteriors Analysis

In this section, the model diagnostics metric, CRPS, is presented across the conducted experiments. First, the CRPS metric is examined at a local scale, and then it is summarized across the case study watersheds.

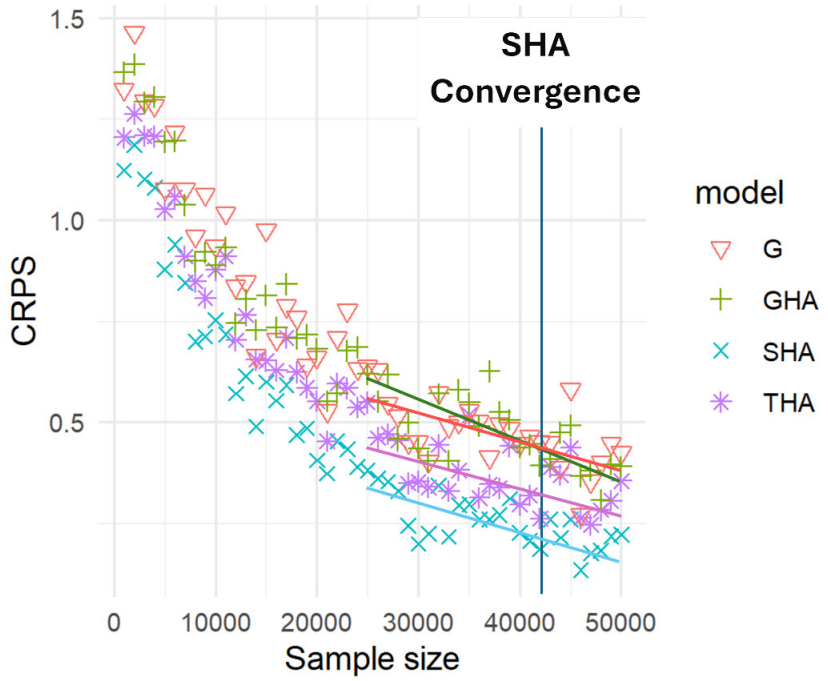


Figure 2.6: CRPS improvement during the course of sample size incrementation at Guadalupe. Note that CRPS values greater than one have not been shown in this figure.

All corresponding hydrographs for the model posterior, randomly sampled from 10% of the chains after the burn-in period, were collected and the resulting CRPS at the Guadalupe watershed was computed and is presented in Figure 2.6. The figure suggests that SHA consistently outperforms the other methods (THA, GHA, and G), especially as sample size increases. This performance is attributed to how SHA generates more informative error-model parameters during the inference process, which is achieved using the surrogate model.

To complement the CRPS results and provide a more detailed assessment of probabilistic performance, Figure 2.8 presents the reliability and relative sharpness scores computed

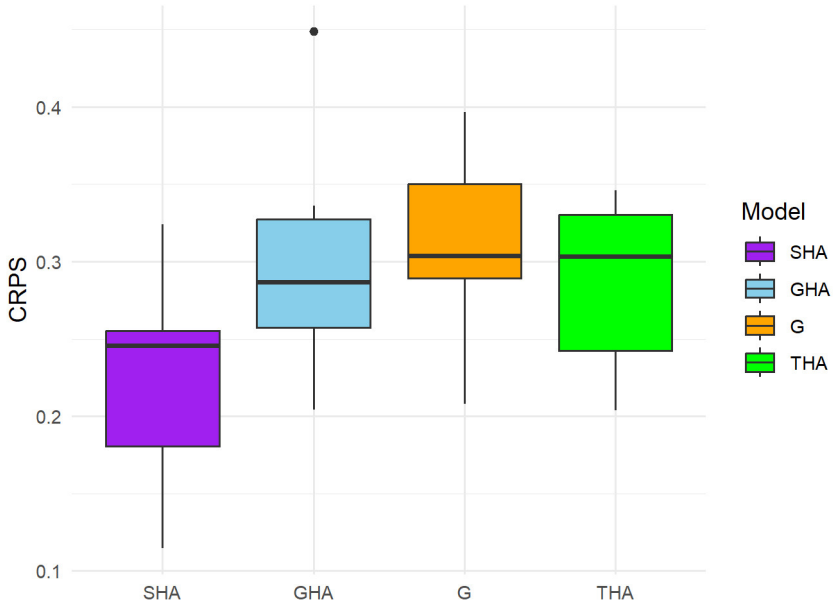


Figure 2.7: CRPS values for all 12 MOPEX watersheds at the intersection of the SHA trendline and GR value of 1.1.

for all MOPEX12 watersheds. While CRPS summarizes both properties in a single metric, the explicit evaluation of reliability and sharpness clarifies the trade-off between predictive interval coverage and concentration. High reliability indicates that the predictive intervals successfully enclose most of the observed flows, while high sharpness corresponds to narrower credible intervals and thus greater confidence in the forecasts.

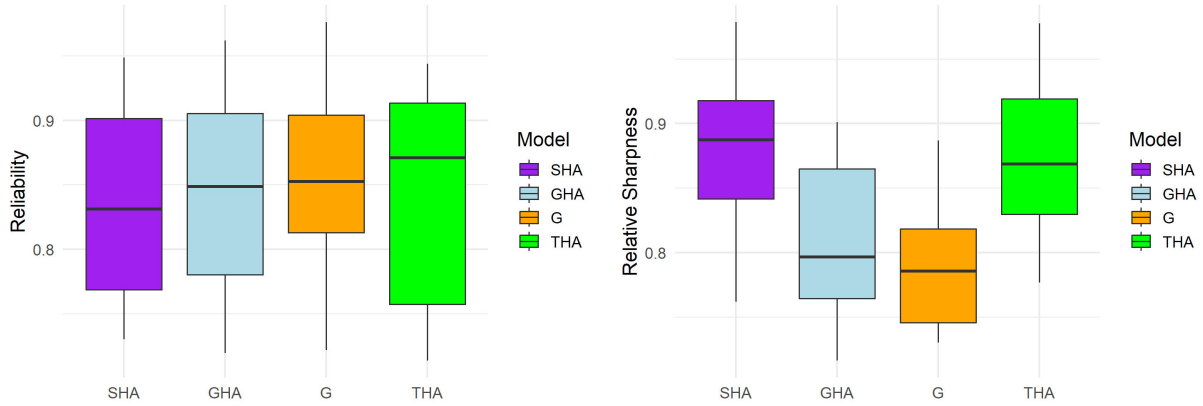


Figure 2.8: Reliability and relative sharpness at MOPEX12 locations when SHA is considered converged.

The boxplots in Figure 2.8 show that the surrogate-aided SHA method achieves a favorable balance between these two competing objectives. Models such as G tend to yield high reliability but low sharpness, implying broader yet well-calibrated predictive intervals. In contrast, SHA and THA provide sharper, more concentrated distributions while maintaining comparable reliability. This indicates that the SHA approach constrains uncertainty without compromising predictive coverage.

2.4 Discussion

The surrogate model in the SHA approach, as described in the methodology section, takes hydrologic model parameters as input and determines the corresponding error model parameters in a way that is more tailored to the specific dynamics of the watershed. This process leads to less erratic behavior in the error model, particularly in the early stages of the inference process (iterations $< 5{,}000$). By representing more accurate errors from the

outset, SHA is able to produce better representations of the residuals, which ultimately leads to a reduction in predictive error, as evidenced by the lower CRPS values across the entire range of sample sizes.

In contrast, GHA, THA, and G are more susceptible to higher variability and noise in their likelihood functions, which is attributed to the error-model structure and sampler configuration. This results in higher CRPS values, reflecting poorer predictive accuracy and greater variability compared to SHA. Particularly in the initial iterations, when chains have not fully converged, SHA's ability to stabilize the inference process using more informed log-likelihood function parameters appears to provide an advantage.

As sample sizes increase, all methods show improvement, with CRPS gradually decreasing, but SHA's more consistent and smoother decline in CRPS demonstrates that it is able to capture the underlying dynamics of the watershed with fewer samples. This indicates a more efficient exploration of the parameter space and quicker convergence toward accurate predictions, especially in the early stages of the simulation.

In summary, SHA's enhanced performance can be attributed to its effective use of the surrogate model, which translates hydrologic model parameters into error model parameters. This benefit extends beyond improved starting values since the surrogate supplies a consistent error-model parameterization at every MCMC step rather than only at initialization. This results in less stochastic behavior in the likelihood estimates and better representation of residuals, leading to lower CRPS values and improved predictive accuracy, particularly in the early stages of inference.

Looking at the CRPS metric across MOPEX12 catchments, the boxplot in Figure 2.7 compares this metric across the four error models—SHA, THA, GHA, and G—based on model posterior quantiles, calculated for all catchments.

In this analysis, the results are drawn from experiments where SHA crosses the 1.1 Gelman–Rubin convergence threshold, and all other methods (THA, GHA, and G) were evaluated using the same sample size for fair comparison. SHA shows the strongest performance, with the lowest median CRPS and a narrower interquartile range, suggesting it provides more accurate and consistent predictions. In contrast, THA, GHA, and G display higher median CRPS values, indicating lower predictive accuracy and greater variability across experiments. This suggests that SHA achieves convergence faster and delivers more reliable forecast outcomes than the other approaches.

The SHA framework demonstrates significant computational efficiency despite its 1,000-simulation offline training phase. As quantified in Table 2.3, SHA achieves convergence with only $38{,}350 \pm 3{,}795$ MCMC samples—approximately 49.2% fewer samples than THA ($75{,}430 \pm 7{,}230$) and 52.1% fewer than GHA ($80{,}130 \pm 6{,}358$). This sample reduction

translates to substantial runtime savings: SHA completes calibration in just 6.5 ± 0.6 hours compared to 12.6 ± 1.2 hours for THA (48.4% faster) and 13.4 ± 1.1 hours for GHA (51.5% faster). These gains align with surrogate-assisted inference principles, where pre-computation amortizes costs during sampling (Razavi et al., 2012), particularly benefiting multi-chain applications.

Table 2.3: Computational cost comparison. Values represent median ± 2.5 SD across 12 MOPEX watersheds.

Model	Offline Simulations	MCMC Samples to Converge	Total Runtime
SHA	1,000	38,350 (±3,795)	6.5 ± 0.6 h
THA	0	75,430 (±7,230)	12.6 ± 1.2 h
GHA	0	80,130 (±6,358)	13.4 ± 1.1 h
G	0	79,341 (±5,770)	13.2 ± 1.0 h

2.5 Conclusion

This study presents a surrogate-aided error model that improves the efficiency of Bayesian calibration in hydrologic models by decoupling the inference of hydrologic and error-model parameters. The methodology uses support vector regression (SVR) to estimate log-likelihood function parameters, reducing computational cost while maintaining or improving predictive accuracy. Applied to the GR4J model across MOPEX12 watersheds, the surrogate-aided approach demonstrated faster convergence, requiring fewer samples than traditional calibration methods. Additionally, it consistently outperformed conventional approaches in predictive accuracy, as evidenced by CRPS values.

The key findings of this study include:

- The surrogate-aided approach enhances the convergence of Bayesian calibration by simplifying the joint inference process, reducing the need for extensive computational resources.
- Support vector regression (SVR) proved effective as a surrogate model, with scope to scale to other hydrologic frameworks.
- The approach reduces sample-size requirements, making it more efficient for large-scale hydrologic applications.

- Consistently improved predictive accuracy was observed, suggesting this methodology can deliver more reliable hydrologic predictions.

The surrogate error model proposed in this study provides a framework for efficient Bayesian calibration by decoupling the inference of hydrologic parameters from error-model parameters. The efficiency of the SHA model demonstrates the potential of surrogate-based frameworks to improve the efficiency of operational hydrologic applications, particularly those requiring frequent updates or involving high computational demand. By reducing sampling effort, the approach accelerates calibration and facilitates the timely production of probabilistic forecasts. However, several directions remain for future work. For instance, incorporating nonstationarity conditions—such as climate change impacts or evolving catchment characteristics—into the surrogate model can help maintain its relevance under varying environmental conditions (Blöschl et al., 2019).

While SHA's decoupling principle remains generalizable, SVR performance may degrade for problems with a higher-dimensional parameter space without dimensionality reduction. Future work should test SHA for hydrologic models with more parameters, such as HBV-EC (Moore, 1993), which offer a better representation of underlying physical processes. Additionally, exploring alternative distributions for representing residuals in the likelihood function could further refine its accuracy and broaden its use in more complex hydrological systems. Future work could also integrate limited-memory acceleration strategies for MCMC (Kuczera et al., 2010) into the surrogate framework to further reduce runtimes, especially for high-parameter models. These enhancements will help determine whether the surrogate model can maintain its computational efficiency and performance in more parameter-intensive settings.

Another direction for future work is to extend the method to ungauged basins, where data scarcity poses challenges, by using regional information or transfer-learning techniques (Hrachowitz et al., 2013). Additionally, diagnosing and analyzing raw errors more comprehensively to inform the surrogate error model could be paired with deep learning approaches, offering a more general and flexible framework for error modeling. These directions would extend the method to a wider range of hydrological modeling settings (Jiang et al., 2019).

The proposed surrogate error model has several practical implications for hydrologic modeling and broader applications. One benefit is the time saved in the daily workflow of practitioners by simplifying and accelerating the Bayesian calibration process. This method is particularly suited to large-scale and global models, where traditional calibration can be computationally prohibitive (Beven and Binley, 1992; Clark et al., 2017; Tsai et al., 2021). By integrating this approach into a workflow that allows model sharing, it aligns well with

the initiative to generate models across large regions (Kratzert et al., 2023) to facilitate broader adoption and contribute to improved hydrologic predictions.

Chapter 3

Simultaneous Bayesian Quantification and Decomposition of Structural, Parameter, and Forcing Uncertainty in Hydrologic Models: A Blended Model Approach

Preface

This chapter presents a unified Bayesian framework that jointly quantifies and temporally decomposes structural, parameter, and forcing uncertainty in streamflow prediction. By reformulating structural uncertainty as parameter uncertainty through the blended-model approach of Mai et al. (2020), the framework circumvents the identifiability obstacles that have constrained prior decomposition attempts and yields a continuous, time-varying uncertainty budget across the hydrological year. A version of this work, with minor changes, has been submitted to the *Journal of Hydrology*, with co-authors J. R. Craig, B. A. Tolson, J. Romero-Cuellar, and R. Chlumsky.

3.1 Introduction

Hydrologic modeling is widely used for understanding and predicting the behavior of water systems, including flood forecasting, water resource management, and assessing the impacts of climate change (Beven, 2012; Gupta et al., 2014b). These watershed models, which range widely in complexity, simulate the processes of the hydrological cycle, including precipitation, evapotranspiration, infiltration, and runoff generation (Beven and Binley, 1992; Gupta et al., 1998). Despite significant advancements in computational power and model sophistication, a persistent challenge remains, namely the inherent uncertainty associated with hydrologic predictions, which can substantially influence their reliability and practical utility (Beven, 2006; Montanari, 2007). Addressing these uncertainties is a necessity for informed decision-making in water-related policy and infrastructure planning.

Prediction uncertainty in hydrologic modeling arises from multiple sources. Structural uncertainty stems from simplifications and assumptions in model formulations, which may fail to fully capture the complexity of natural hydrological processes (Clark et al., 2011; Fenicia et al., 2008). For instance, lumped models, while computationally efficient, often disregard spatial variability in watershed characteristics. Parameter uncertainty, another major source, emerges from the difficulty in accurately estimating model parameters, particularly when observational data are sparse or noisy (Vrugt et al., 2008; Beven, 2009). Many hydrologic models rely on calibration techniques to optimally estimate parameters, but these methods typically yield non-unique solutions, further complicating the interpretation of results. Input data uncertainty, including errors in precipitation, temperature, and other meteorological forcings, also contributes, as even minor inaccuracies can propagate through the model and lead to substantial deviations in predictions (McMillan et al., 2012; Renard et al., 2011). Finally, observational uncertainty, which encompasses measurement errors in streamflow and other hydrological variables used in model calibration and validation, can potentially bias model performance (Westerberg et al., 2011; McMillan et al., 2010). Of these four sources, the present study focuses on the three that can be explicitly parameterized within a process-based modeling framework, structural, parameter, and forcing uncertainty, while acknowledging that observational uncertainty in streamflow records is treated implicitly through the likelihood formulation.

Efforts to quantify these uncertainties have led to the development of various methodological approaches. Monte Carlo simulations, for example, have been widely employed to assess parameter uncertainty by generating ensembles of model outputs based on randomized parameter sets (Kuczera and Parent, 1998; Beven and Freer, 2001). These simulations provide a probabilistic distribution of possible outcomes, offering insights into the sensitivity of model predictions to parameter variation. Bayesian methods have also gained

prominence, providing a framework for incorporating prior knowledge and observational data to estimate posterior distributions of model parameters and outputs (Kavetski et al., 2006; Vrugt et al., 2009). These approaches are particularly suited to cases where data are limited, as they allow for the integration of soft data such as expert judgment and historical precedent.

Despite these advancements, a significant gap remains. The hydrologic modeling community lacks the ability to simultaneously disentangle and quantify the relative influence of the three key uncertainty sources (Blöschl et al., 2019). Past investigations often address individual uncertainty components in isolation, potentially overlooking the interactions and compounded effects among them (Renard et al., 2011; Liu and Gupta, 2007). For example, structural and parameter uncertainties are frequently interdependent, as the choice of model structure influences parameter identifiability and sensitivity (Gupta et al., 2012; Clark et al., 2011). Similarly, input data errors can propagate through the model, affecting both parameter estimation and output predictions (McMillan et al., 2012; Del Giudice et al., 2016). Neglecting interaction effects can lead to both under- and over-estimations of total uncertainty, and therefore over- or under-confidence in model predictions, with potentially severe consequences for water management and flood risk assessment.

Recent studies have highlighted the need for integrated frameworks capable of concurrently addressing multiple uncertainty sources. An influential contribution in this direction was Montanari and Koutsoyiannis (2007), who proposed a Bayesian total error analysis framework that characterizes model structural error through storm-dependent random variation of model parameters, an explicit acknowledgement that conceptual rainfall–runoff models are inherently stochastic rather than deterministic representations of catchment dynamics. In that framework, structural inadequacy is expressed through the parameters themselves, whose values are allowed to vary randomly between storm events so that the magnitude of this variation quantifies the error of a fixed model structure. Building on the related BATEA framework of Kavetski et al. (2006), this approach demonstrated that realistic uncertainty bounds require explicit, separate characterization of input, parameter, and structural errors, not a lumped residual error model. Renard et al. (2010) subsequently formalized the fundamental challenge of this decomposition, showing that when input and structural error models are inferred simultaneously from streamflow data alone, the inverse problem is ill-posed unless sufficiently informative priors on data quality are available. This identifiability limitation means that, in practice, meaningful partitioning of uncertainty is extremely difficult without independent constraints on at least one error source.

Renard et al. (2011) attempted to resolve this by using conditional rainfall simulation to constrain the input error model, demonstrating that reliable decomposition becomes tractable when rainfall uncertainty is characterized independently. Ajami et al. (2007) pro-

posed the Integrated Bayesian Uncertainty Estimator (IBUNE), which simultaneously confronts input, parameter, and model structural uncertainty within a multi-model Bayesian framework, representing an important operational step toward joint quantification. In the forecasting domain, Thiboult et al. (2016) demonstrated that combining ensemble Kalman filtering, multi-model approaches, and meteorological ensemble forcing can account for three distinct uncertainty sources, initial conditions, model structure, and meteorological forcing, with their combination outperforming any single approach. For instance, Salamon and Feyen (2010) proposed a methodology combining multiplicative error models with sequential data assimilation to separate input and structural uncertainties. Similarly, Vrugt et al. (2008) introduced the Differential Evolution Adaptive Metropolis (DREAM) algorithm, a Markov chain Monte Carlo scheme that efficiently estimates posterior distributions of model parameters while explicitly treating precipitation input uncertainty, demonstrating that ignoring forcing error substantially alters inferred parameter distributions. More recently, Peng et al. (2025) proposed a hierarchical Bayesian inference framework that keeps satellite-based precipitation and model-parameter uncertainty sources separable in ensemble streamflow prediction, demonstrating that explicit decoupling of forcing and parameter uncertainty becomes tractable when an independently characterized precipitation ensemble is supplied as a prior. This contribution illustrates the value of unified probabilistic inference, but, like the satellite-precipitation context that motivates it, depends on access to an external uncertainty constraint on at least one error source. While these methods represent important steps forward, they often address only a subset of uncertainty sources, treat sources as stationary over time, and do not provide a continuous, temporally resolved picture of how the relative contributions of different sources evolve across the hydrological year. The absence of a unified inference framework that jointly estimates all three uncertainty sources without encountering the identifiability trade-offs that arise when input and structural error models are fitted separately to the same streamflow record (Renard et al., 2010) remains an unresolved challenge in operational hydrology (Beven, 2009; Montanari et al., 2013).

In response to this gap, this study proposes a novel methodology to quantify and separate predictive uncertainty in streamflow modeling by addressing the three dominant sources of error in traditional process-based hydrologic models: parameters, structure, and forcings. The central enabling concept is the blended modelling approach of Mai et al. (2020), which replaces the discrete selection of a single model structure which represents a given hydrologic process with a continuously weighted average of multiple process algorithms. By encoding structural uncertainty as uncertainty in the weight-generating parameters, the framework circumvents the ill-posed simultaneous estimation of separate input and structural error models that has limited prior decomposition approaches (Renard

et al., 2010): all three sources become parameters in the same Bayesian inference, yielding a well-posed decomposition by construction. The identifiability trade-off is avoided because no separate structural error model needs to be inferred from the streamflow residuals. Structural uncertainty instead enters through the process weights, whose influence on simulated streamflow is mechanistically distinct from that of the forcing perturbation parameters. Building on this, we make three specific contributions. First, we develop a single Bayesian calibration framework that simultaneously infers weight-generating parameters (structural uncertainty), hydrologic process parameters (parameter uncertainty), and meteorological perturbation parameters (forcing uncertainty), yielding a well-posed inference that circumvents the identifiability problem inherent in separately fitted error models. Second, we apply a multivariate variance decomposition (MANOVA) to the resulting posterior ensemble to partition total predictive uncertainty among these three sources for every Julian calendar day, yielding a time-varying uncertainty budget across the hydrological year. Third, we demonstrate this framework on 12 benchmark MOPEX watersheds spanning a range of hydroclimatic regimes, providing insight into how the dominant source of uncertainty shifts seasonally and across catchment types.

3.2 Methodology

3.2.1 Case studies

In this study, we used a subset of 12 watersheds from the MOPEX dataset to evaluate the proposed methodologies. The Model Parameter Estimation Experiment (MOPEX) (Duan et al., 2006) is a global initiative aimed at improving parameter estimation techniques in hydrologic modeling. The full MOPEX database includes 438 catchments across the continental United States.

Among these, 12 catchments were selected during the second and third MOPEX workshops (Schaake et al., 2006) to serve as a benchmark dataset, referred to in this study as MOPEX12 (see Table 3.1 for MOPEX12 watersheds specifications). This selection was based on their diverse hydrologic behaviors, making them well-suited for testing different modeling frameworks under varying hydrologic conditions. These catchments are primarily located in the eastern United States, a region characterized by a wide range of climate and watershed responses.

In this study, we use daily data for both meteorological forcing functions and streamflow time series to maintain consistency across all stages of modeling and inference. The forcing

Table 3.1: The MOPEX12 watershed specifications.

ID	USGS gauge ID	Basin	State	Lat. (°)	Lon. (°)	Area (km^2)	Elev. (m)	Slope (°)	P (mm)	PET (mm)	T (°C)	Q (m^3/s)
1	1608500	Potomac	WV	38.85	−79.21	3,782	661.4	11	935	762	10.7	1,414
2	1643000	Monocacy	MD	39.63	−77.26	2,115	194.6	3.1	1,059	898	11.5	847
3	1668000	Rappahannock	VA	38.50	−78.04	4,129	207.3	4.2	1,078	921	12.5	1,481
4	3054500	Tygart Valley	WV	38.84	−80.05	2,366	718.7	8.6	1,318	711	9.8	1,755
5	3179000	Bluestone	WV	37.39	−81.20	1,023	793.2	8.3	984	741	11.0	418
6	3364000	East Fork	IN	39.64	−85.70	4,419	268.8	0.6	1,032	856	11.0	1,681
7	3451500	French Broad	NC	35.39	−82.60	2,446	829.8	8.8	1,510	820	12.5	1,831
8	5455500	English	IA	41.56	−92.18	1,486	253.5	1.6	867	995	9.8	357
9	7186000	Spring	MO	37.26	−94.15	3,013	331.5	0.8	1,079	1,095	13.7	856
10	7378500	Amite	LA	30.95	−90.91	3,314	82.4	1.0	1,575	1,074	18.9	1,831
11	8167500	Guadalupe	TX	30.00	−99.06	3,404	544.4	2.7	772	1,529	18.2	359
12	8172000	San Marcos	TX	29.93	−98.10	2,169	298.8	1.8	844	1,449	19.5	361

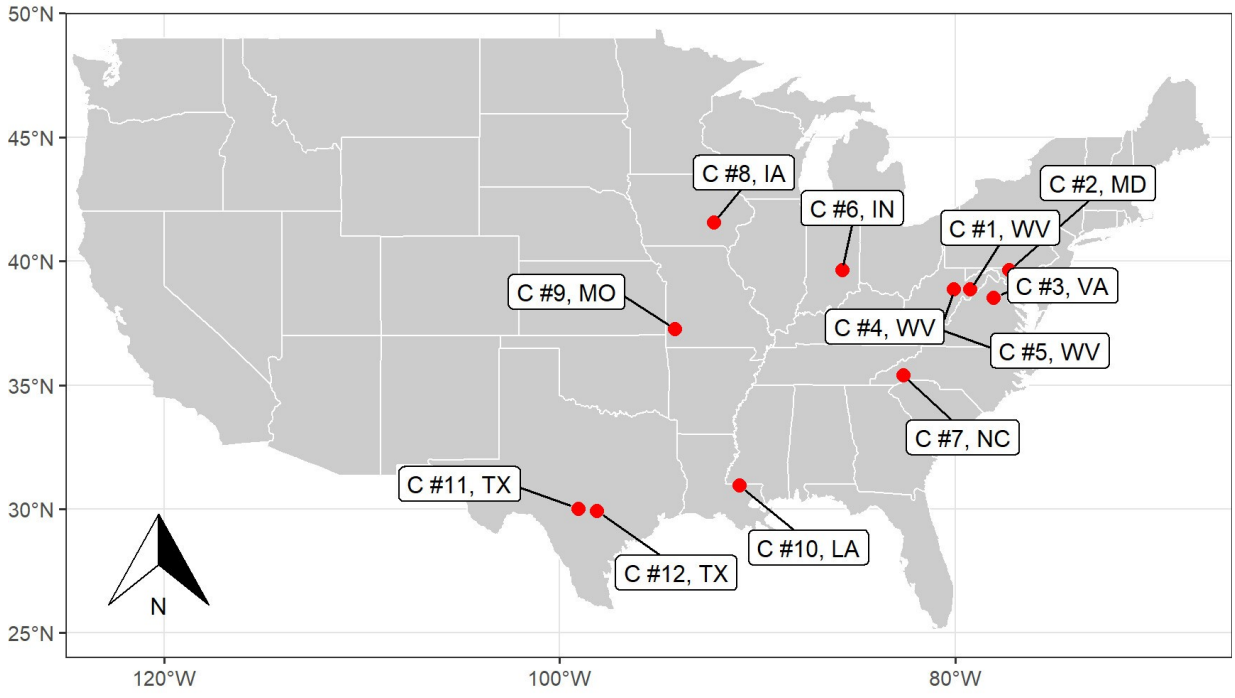


Figure 3.1: Locations of the 12 MOPEX watersheds.

data includes daily minimum temperature, daily maximum temperature, and precipitation, and includes the full period from January 1, 1948 to January 1, 2004.

3.2.2 Hydrologic Model

The model used in this study is the original "blended model" used in Mai et al. (2020) and Chlumsky et al. (2021). In this approach, model structure uncertainty is addressed by introducing additional parameters that serve as weights in a weighted averaging scheme applied to internal model fluxes. These fluxes are estimated using a weighted average of multiple process options within the model, allowing for continuous exploration of the model structure space, rather than relying on discrete structural options. The inclusion of these extra parameters can provide insight into both model inadequacy and process identifiability (Chlumsky et al., 2021, 2024). The averaging within the blended model differs from Bayesian Model Averaging, which combines a discrete set of separately calibrated models through posterior model weights. For example, consider a modular hydrologic model where each hydrologic process can be represented by multiple algorithmic options. Instead of selecting a single formulation for each process, a blended model can be constructed by assigning weights to different process algorithms, effectively enabling a continuous blending between process options. Hydrological simulations were performed continuously over a 56-year period (January 1, 1948–January 1, 2004) using lumped model configurations for all 12 MOPEX catchments, ensuring well-initialized states and temporal coherence across watersheds.

Here, it is used to convert the model structural uncertainty assessment approach into an equivalent parameter uncertainty assessment approach. The basics of the internal calculations of hydrologic fluxes in the blended model, discussed in detail within Mai et al. (2020) or Chlumsky et al. (2021), are briefly described here. Let $M_{i,j}$ represent the j-th algorithm for calculating water fluxes for the i-th process (e.g., evaporation or infiltration) within a hydrologic model. The water flux for that process, f_i, is calculated as a weighted combination of the flux from each individual process algorithm:

$$f_i = \sum_{j=1}^{l} w_{i,j}\, M_{i,j}, \qquad \text{s.t.} \quad \sum_{j=1}^{l} w_{i,j} = 1. \tag{3.1}$$

The constraint noted here ensures that the weights sum to unity, maintaining the physical consistency of the model, and ensuring that the weighted flux is bounded by the individual flux estimates. When generating weights, this constraint must be explicitly enforced during any inference process. This weighted flux estimation is supported by the Raven Hydrological Modelling Framework (Craig et al., 2020) used here.

3.2.3 Bayesian Calibration Framework

With the blended hydrologic model, we aim to simultaneously infer process weights, $\boldsymbol{w}$, alongside other hydrologic model parameters and error model parameters. Throughout this section $\boldsymbol{w}$ denotes the full set of process weights while $\boldsymbol{w}'$ describes the weight-generating parameters ($\boldsymbol{w}' \to \boldsymbol{w}$, and this conversion is described in Section 3.2.3). The resulting inference vector is $\boldsymbol{\Theta} = \{\boldsymbol{w}', \boldsymbol{\theta}, \boldsymbol{\alpha}\}$. The symbol $\boldsymbol{w}$ without a prime refers to the final computed weights $w_{i,j}$ satisfying $\sum_{j=1}^{l} w_{i,j} = 1$, which account for structural uncertainty. The vector $\boldsymbol{\theta}$ holds the model parameters (parameter uncertainty) and $\boldsymbol{\alpha}$ the forcing function perturbation parameters (forcing uncertainty), all inferred using a Bayesian calibration framework. This approach updates prior beliefs about the entire set of parameters $\{\boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\alpha}\}$ using observed data, yielding a posterior distribution that combines prior information with the likelihood, and reflects updated beliefs about the parameters given the data.

For a given time series of streamflow observations $\boldsymbol{y} = \{y_t\}_{t=1}^{T}$ and corresponding model forcings $\boldsymbol{x} = \{x_t\}_{t=1}^{T}$ (e.g., temperature and precipitation), the posterior distribution of the model parameters $\boldsymbol{\Theta} = \{\boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\alpha}\}$ is expressed as:

$$p(\boldsymbol{\Theta} \mid \boldsymbol{y}, \boldsymbol{x}) \propto p(\boldsymbol{y} \mid \boldsymbol{\Theta}, \boldsymbol{x}, s_0)\, p(\boldsymbol{\Theta}), \tag{3.2}$$

where $p(\boldsymbol{\Theta})$ denotes the prior distributions of the parameters, and $p(\boldsymbol{y} \mid \boldsymbol{\Theta}, \boldsymbol{x}, s_0)$ is the likelihood function, quantifying the likelihood of the observed data given the parameters and some initial conditions s_0. Here $\boldsymbol{y}$ denotes the observed streamflow series, equivalent to the symbol Q used in the preceding chapter. To quantify the discrepancy between observed and simulated streamflow values, we adopt a homoscedastic Gaussian likelihood, assuming that residuals between observed and simulated streamflow are independent and identically distributed with constant variance σ^2. While this is a simplifying assumption, streamflow residuals are often heteroscedastic and temporally correlated, it provides a tractable starting point consistent with prior studies (Kavetski et al., 2006) and the primary focus here is on the relative partitioning of uncertainty sources rather than on precise absolute uncertainty bounds. The Gaussian and Student-t forms were chosen for their analytical tractability and, in the Student-t case, robustness to heavy-tailed residuals. More flexible forms such as skew-normal or censored Gaussian distributions are better suited to intermittent semi-arid flows but were not adopted here to keep the likelihood consistent across the continental sample. Thus, the likelihood function is formulated as:

$$p(\boldsymbol{y} \mid \boldsymbol{\Theta}, \boldsymbol{x}) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\big(y_t - \hat{y}_t(\boldsymbol{\Theta}, x_t, s_0)\big)^2}{2\sigma^2} \right]. \tag{3.3}$$

Here, $\hat{y}_t(\boldsymbol{\Theta}, x_t, s_0)$ denotes the simulated streamflow at time t given the parameter set $\boldsymbol{\Theta}$, forcings x_t, and initial conditions s_0, while σ^2 represents the constant error variance across all time steps.

While specifying priors for hydrologic process parameters is relatively straightforward, typically bounded uniform distributions reflecting known feasible ranges, the process weights require special treatment due to their normalization constraint. The process weights $w_{i,j}$ are subject to the normalization constraint. In the blended model structure mentioned earlier, we must ensure the weights sum to one. Although the weights are inferred as static over the calibration period, their effect on the uncertainty budget varies seasonally because the contributing process algorithms activate at different times of year, and interactions among process algorithms are absorbed into the structural component rather than resolved individually.

To enforce this constraint, we employ the pie-sharing distribution method (Mai et al., 2020), which provides an unbiased method to sample $N+1$ weights that sum to one via the use of weight-generating parameters sampled from a uniform distribution. This approach is analogous to randomly subdividing a pie into $N+1$ slices with identically distributed volumes. The pie sharing method uses N random weight-generating parameters $\boldsymbol{w'}$, each between 0 and 1, to compute $N+1$ unbiased weights summing to one. Thus, the adjusted vector of inference variables becomes $\boldsymbol{\Theta} = \{\boldsymbol{w'}, \boldsymbol{\theta}, \boldsymbol{\alpha}\}$.

In this study we used a bounded uniform distribution for the prior of each parameter. Such priors are weakly informative rather than strictly uninformative, since the choice of feasible range itself encodes prior information. The bounds adopted here (Appendix 2.A) follow established usage of the blended model. For the process parameters $\boldsymbol{\theta}$, we therefore assign uniform priors over their feasible ranges:

$$\theta_k \sim \text{Uniform}(a_k, b_k), \tag{3.4}$$

where a_k and b_k define the bounds of the uniform distribution for parameter θ_k (see Appendix 2.A for further details).

Inference algorithm

To perform Bayesian inference, we utilize the No-U-Turn Sampler (NUTS) (Hoffman and Gelman, 2014), an adaptive variant of Hamiltonian Monte Carlo (HMC), which is a Markov Chain Monte Carlo (MCMC) algorithm well-suited for high-dimensional problems. The NUTS algorithm employed in this study is implemented in the LaplacesDemon R package

Algorithm: Hamiltonian Monte Carlo Sampling Procedure.

Input: Data $(\boldsymbol{y}, \boldsymbol{x})$, initial parameter $\boldsymbol{\Theta}_0$, mass matrix M.
Output: Posterior samples of $\boldsymbol{\Theta}$.

1: Initialize the parameter set $\boldsymbol{\Theta}_0$.
2: Propose a new candidate $\boldsymbol{\Theta}^*$ using Hamiltonian dynamics: $p(\boldsymbol{\Theta}, \boldsymbol{r}) \propto \exp\big(-U(\boldsymbol{\Theta})\big) \cdot \exp\big(-K(\boldsymbol{r})\big)$, where $U(\boldsymbol{\Theta}) = -\log p(\boldsymbol{\Theta} \mid \boldsymbol{y}, \boldsymbol{x})$ is the potential energy and $K(\boldsymbol{r})$ is the kinetic energy with $\boldsymbol{r} \sim \mathcal{N}(0, M)$.
3: Numerically simulate a trajectory using leapfrog integration.
4: Adaptively determine the trajectory length to avoid U-turns.
5: Accept or reject the proposal based on Metropolis–Hastings.
6: **repeat** until $\hat{R} < 1.1$ (Gelman–Rubin) and minimum ESS threshold (set to 1000 in this study) is reached.

(LaplacesDemon Core Team, 2023), which offers a comprehensive environment for Bayesian inference. The NUTS algorithm operates as follows.

The NUTS algorithm is particularly suitable for estimating $\boldsymbol{\Theta}$ due to its efficiency in exploring complex posterior distributions while maintaining computational feasibility. For each watershed, three independent chains were run, each drawing 100,000 posterior samples, leading to a total computational burden of $12 \times 3 \times 100{,}000 = 3{,}600{,}000$ model simulations to infer parameters across the case studies. Each individual 56-year model simulation required approximately 1.3 seconds on the Digital Research Alliance of Canada (formerly Compute Canada) high-performance computing clusters.

Forcing uncertainty considerations

Representing uncertainties in meteorological inputs, such as temperature and precipitation, is important for reliable hydrologic modeling. Here, we have implemented a perturbation approach that introduces stochastic variability into these inputs, governed by six parameters in total (two per forcing variable, applied to precipitation, minimum temperature, and maximum temperature): E and β. The parameter E controls the magnitude of the perturbation, while β adjusts its skewness. This method replicates approaches used in previous studies to address input uncertainties in hydrologic simulations (Nijssen and Lettenmaier, 2004; Wongchuig et al., 2019).

This perturbation technique modifies each time series of meteorological inputs by applying either a multiplicative or an additive factor, introducing controlled variability while preserving inter-variable collinearity. The perturbation scheme applies variable-specific transformations that reflect the statistical characteristics of each meteorological input. Temperatures are perturbed using a multiplicative formulation, while precipitation is also perturbed multiplicatively to preserve its strictly non-negative and skewed distribution. For each day t, a multivariate standard normal random vector

$$\boldsymbol{\epsilon}_t = \left(\epsilon_{t,\mathrm{pr}},\, \epsilon_{t,\mathrm{tmin}},\, \epsilon_{t,\mathrm{tmax}}\right)^\top \sim \mathcal{N}(0,\, R) \tag{3.5}$$

is drawn, where R is the empirical Pearson correlation matrix estimated from the historical standardized anomalies of the input series. This joint sampling preserves inter-variable correlations across the perturbed series. Dependence between precipitation and temperature is therefore represented to the extent that it is captured by this empirical correlation structure. Uncertainty in the dependence itself (for example, departures from the historical correlation) is not perturbed separately and would, if present, be absorbed into the forcing component of the decomposition. For a given precipitation time series $x_{t,\mathrm{pr}}$ at time step t, the perturbed value $x'_{t,\mathrm{pr}}$ is computed as

$$x'_{t,\mathrm{pr}} = (1+\beta_{\mathrm{pr}})\sqrt{E_{\mathrm{pr}}^2+1}\,\exp\!\left(\sqrt{\log\left(E_{\mathrm{pr}}^2+1\right)}\,\epsilon_{t,\mathrm{pr}}\right) x_{t,\mathrm{pr}}. \tag{3.6}$$

This formulation ensures that the perturbations have a log-normal distribution, preserving the non-negative nature of variables like precipitation and capturing the asymmetric characteristics of meteorological uncertainties. To maintain reproducibility in the perturbations, the random seed for generating $\boldsymbol{\epsilon}_t$ is deterministically set based on the values of E and β:

$$\mathrm{seed} = \lfloor E \cdot M_1 + \beta \cdot M_2 \rfloor, \tag{3.7}$$

where M_1 and M_2 are large constants used to ensure sufficient variability in seed number generation, and $\lfloor \cdot \rfloor$ represents conversion to an integer. This deterministic seeding ensures that the same perturbations can be reproduced given the same parameter values, thereby facilitating consistent calibration and validation processes. This feature is particularly important in the Bayesian inference of forcing perturbation parameters, especially when the rejection step causes the algorithm to re-evaluate the same parameter sets where the same vector of noise time series is needed.

Likewise, we can formulate multiplicative noise for minimum and maximum temperature (Wongchuig et al., 2019) as:

$$\begin{aligned} x'_{t,\mathrm{tmax}} &= \left(\beta_{\mathrm{tmax}} + E_{\mathrm{tmax}}\,\epsilon_{t,\mathrm{tmax}}\right) x_{t,\mathrm{tmax}}, \\ x'_{t,\mathrm{tmin}} &= \left(\beta_{\mathrm{tmin}} + E_{\mathrm{tmin}}\,\epsilon_{t,\mathrm{tmin}}\right) x_{t,\mathrm{tmin}}, \end{aligned} \tag{3.8}$$

where E controls the scale (variance) of the noise, while β adjusts its mean (bias).

Handling bounded parameters

Most hydrologic parameters are bounded between lower and upper limits due to physical constraints and/or expert judgement. To ensure that the gradient-based No-U-Turn Sampler respects these bounds during Bayesian inference, a smooth transformation based on the hyperbolic tangent function (tanh) is employed. This transformation softly and deterministically constrains parameter values within a specified interval.

Unlike rejection-based approaches, e.g., DREAM (Laloy and Vrugt, 2012), which may introduce inefficiencies due to repeated proposal rejection, the proposed transformation is infinitely differentiable (C^∞). This property is critical for gradient-based samplers, enabling stable and efficient posterior exploration. The use of smooth transformations to map unbounded variables into bounded parameter spaces is well established in probabilistic programming.

Let $\Theta_i \in \mathbb{R}$ denote the i-th hydrologic model parameter, with admissible bounds $a_i < b_i$. The transformation is defined as

$$\Theta'(\Theta_i, a_i, b_i) = a_i + (b_i - a_i) \cdot \tfrac{1}{2}\big(1 + \tanh\big(k(\Theta_i - c)\big)\big), \tag{3.9}$$

where

$$c = \tfrac{\Theta_{\max}+\Theta_{\min}}{2}, \qquad k = \tfrac{\operatorname{atanh}(\alpha)}{(\Theta_{\max}-\Theta_{\min})/2}. \tag{3.10}$$

Here, c represents the midpoint of the expected input domain, while k controls the steepness of the transition. The parameter k is selected such that approximately $\alpha\%$ of the input domain behaves nearly linearly, with curvature confined to the outermost $(1-\alpha)/2\%$ near each bound.

The transformation is strictly monotone and asymptotically approaches the bounds without ever reaching them, i.e., $\Theta'(\Theta_i) \to a_i$ as $\Theta_i \to -\infty$, and $\Theta'(\Theta_i) \to b_i$ as $\Theta_i \to +\infty$. As a result, the ordering of parameter proposals is preserved, and hard truncation at the boundaries is avoided.

Figure 3.2 illustrates the behavior of the transformation over an example input domain where $\Theta_i \in [-\infty, +\infty]$ and the output is bounded within $[0, 1]$. The function remains nearly linear across the central portion of the domain, while smooth transitions occur near the bounds. The function never exactly reaches the limits, which ensures non-zero gradients throughout the domain. This property provides informative gradient signals even near the feasible boundaries, allowing NUTS to efficiently and stably explore the posterior distribution, particularly in cases where boundary behavior is influential.

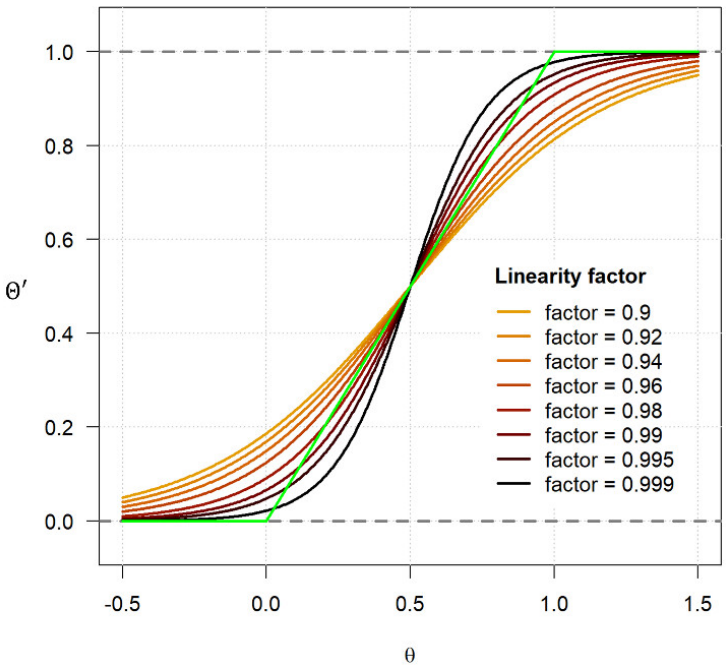


Figure 3.2: Smooth transformation $\Theta'(\theta)$ versus θ (unbounded parameter). The factor $\alpha = 0.99$ is the fixed value used throughout this study; the green line is the regular clamping step function for comparison.

3.2.4 Uncertainty decomposition

The Bayesian calibration described above yields a posterior distribution over the full parameter vector $\boldsymbol{\Theta} = \{\boldsymbol{w}', \boldsymbol{\theta}, \boldsymbol{\alpha}\}$. Because each subset of parameters corresponds to a distinct source of uncertainty, process weights $\boldsymbol{w}'$ to structural uncertainty, hydrologic parameters $\boldsymbol{\theta}$ to parameter uncertainty, and perturbation parameters $\boldsymbol{\alpha}$ to forcing uncertainty, the spread in posterior predictions can be attributed back to these sources. The key idea is to draw a large ensemble of posterior parameter samples, propagate each sample through the hydrologic model to obtain a corresponding streamflow simulation, and then use a variance decomposition to partition the total spread in simulated streamflow among the three parameter groups. This construction is conceptually similar to variance-based global sensitivity analysis such as the Sobol' decomposition (Saltelli et al., 2008), with the important difference that parameter realizations are drawn from the joint posterior rather than from independent prior distributions, so the resulting partition reflects calibrated rather than prior uncertainty. By repeating this decomposition independently for every Julian calendar day, we obtain a time-varying picture of how each uncertainty source contributes to prediction uncertainty across the hydrological year.

Let N_{param} denote the total number of transformed model parameters and $N_{\text{realization}}$ the number of posterior samples $\boldsymbol{\Theta}^s$. Each sampled parameter set $\boldsymbol{\Theta}^s$ is transformed using equations (3.9)–(3.10) and used to simulate daily streamflow for the period 1948–2003 using the Raven hydrological modelling framework (Craig et al., 2020).

To facilitate subsequent analyses, the ensemble of simulated hydrographs is reshaped into a three-dimensional tensor with dimensions $\left(N_{\text{realization}}, N_{\text{year}}, N_{\text{jday}}\right)$, where N_{year} is the number of years in the simulation period, and N_{jday} is the number of days in a year (in this study 365). This tensor structure allows for efficient computation of statistical metrics, such as applying the variance analysis for every Julian calendar day independently. Schematically:

$$\boldsymbol{\Theta}^s = \begin{bmatrix} \Theta^s_{1,1} & \cdots & \Theta^s_{1,N_{\text{param}}} \\ \vdots & \ddots & \vdots \\ \Theta^s_{N_{\text{realization}},1} & \cdots & \Theta^s_{N_{\text{realization}},N_{\text{param}}} \end{bmatrix} \Rightarrow \boldsymbol{y}^t = \begin{bmatrix} y^t_{1,1} & \cdots & y^t_{1,N_{\text{param}}} \\ \vdots & \ddots & \vdots \\ y^t_{N_{\text{realization}},1} & \cdots & y^t_{N_{\text{realization}},N_{\text{param}}} \end{bmatrix}. \tag{3.11}$$

To attribute uncertainty in simulated streamflow to specific sources, we assume an approximately linear relationship exists between the parameter realizations $\boldsymbol{\Theta}^s$ and the corresponding simulated streamflow $\boldsymbol{y}^t$ for each calendar day, under the additional assumption that the simulated flows across different years for a given calendar day are statistically

independent. Based on this assumption, a multivariate analysis of variance (MANOVA; Krzanowski, 2000) is performed for each Julian day assuming the following linear model:

$$\boldsymbol{y}^t = \boldsymbol{B}\,\boldsymbol{\Theta}^s + \boldsymbol{\varepsilon} \qquad \forall\, t \in \{1, \ldots, N_{\text{jday}}\}, \tag{3.12}$$

where $\boldsymbol{y}^t$ is the $M \times 1$ vector of responses (simulated hydrographs in a given calendar day for all ensemble members), $\boldsymbol{\Theta}^s$ is the matrix of predictors (all transformed parameter set realizations), $\boldsymbol{B}$ is a matrix of coefficients, and $\boldsymbol{\varepsilon}$ is the matrix of residuals. The parameter–streamflow relationship is in general nonlinear. The linear MANOVA is applied per Julian day, over which we assume the relationship is approximately linear within the posterior ensemble. This formulation allows the decomposition of total variance in daily streamflow into components attributable to each parameter. Because each parameter is associated with structural, forcing, or parameter uncertainty, this enables temporal characterization of uncertainty type proportions across the hydrological year.

Let y_k^t be the simulated streamflow on calendar day t in year k, with $k \in \{1, 2, \ldots, N_{\text{year}}\}$; $\hat{y}_k^{t,i}$ is the MANOVA-fitted streamflow on day t in year k explained by parameter Θ_i for $1 \leq i \leq N_{\text{param}}$, obtained from the MANOVA model. N_{param} is the total number of parameters (in the model setup used in this study $N_{\text{param}} = 51$); each parameter is pre-assigned to one of three groups $g \in \{\text{parameter}, \text{structure}, \text{forcing}\}$ based on its role in the model ($\mathcal{G}_g \subset \{1, \ldots, N_{\text{param}}\}$ is the set of parameter indices associated with group g). EV_g^t is the explained variance percentage attributed to group g on calendar day t; SS_g^t denotes the sum of squared deviations between each parameter's MANOVA-fitted streamflow and the ensemble mean, summed over years k attributable to group g for day t; $\text{SS}_{\text{residual}}^t$ is the residual (unexplained) sum of squares for day t; and $\text{EV}_{\text{unexplained}}^t$ represents the proportion of variance not explained by any group on day t. The adequacy of the linearity assumption is assessed by inspecting the per-Julian-day variance contribution of the unexplained portion to the total variance. The unexplained portion is the residual of the linear MANOVA model. It should not be interpreted as observational (data) uncertainty, which is treated implicitly through the likelihood function and does not appear as a separate component in the uncertainty decomposition.

The sum of squared deviations can be computed for each parameter Θ_i, measuring how much the individual parameter contributes to the variability in streamflow predictions. The total sum of squares of each parameter group is calculated as

$$\text{SS}_g^t = \sum_{i \in \mathcal{G}_g} \sum_{k=1}^{N_{\text{year}}} \left(\hat{y}_k^{t,i} - \bar{y}_k^t\right)^2 \qquad \forall\, g \in \{\text{parameter}, \text{structure}, \text{forcing}\}. \tag{3.13}$$

The total sum of squares for day t equals the sum of all group contributions plus the residual:

$$\mathrm{SS}^t_{\mathrm{total}} = \mathrm{SS}^t_{\mathrm{residual}} + \sum_g \mathrm{SS}^t_g. \tag{3.14}$$

Then the explained variance percentage for each group g is computed as

$$\mathrm{EV}^t_g = \frac{\mathrm{SS}^t_g}{\mathrm{SS}^t_{\mathrm{total}}} \times 100. \tag{3.15}$$

Finally, the unexplained variance percentage is calculated as

$$\mathrm{EV}^t_{\mathrm{unexplained}} = \frac{\mathrm{SS}^t_{\mathrm{residual}}}{\mathrm{SS}^t_{\mathrm{total}}} \times 100 = 100 - \sum_g \mathrm{EV}^t_g. \tag{3.16}$$

This decomposition is performed independently for each calendar day $t \in \{1, 2, \ldots, N_{\mathrm{jday}}\}$, allowing us to analyze how the relative influence of each uncertainty source varies over the course of the year.

3.2.5 Probabilistic Performance Metrics

Probabilistic model performance was evaluated using four complementary metrics: the Continuous Ranked Probability Score (CRPS) skill score, the Kling–Gupta Efficiency (KGE), reliability, and sharpness. These metrics assess probabilistic accuracy, deterministic accuracy of the ensemble mean, calibration, and predictive concentration, respectively, and are defined such that higher values indicate better performance. The CRPS is a strictly proper scoring rule that evaluates the full predictive distribution against observations and is defined as (Hersbach, 2000):

$$\mathrm{CRPS}\big(F_t, y_t\big) = \int_{-\infty}^{\infty} \big(F_t(x) - \mathbf{1}(x \geq y_t)\big)^2 \, \mathrm{d}x, \tag{3.17}$$

where $F_t(x)$ is the cumulative distribution function (CDF) of the predictive distribution at time t, y_t is the observed value, and $\mathbf{1}(\cdot)$ is the indicator function.

The CRPS skill score quantifies forecast accuracy relative to a climatological benchmark and is defined as

$$\mathrm{CRPS}_{\mathrm{skill}} = 1 - \frac{\mathrm{CRPS}_{\mathrm{model}}}{\mathrm{CRPS}_{\mathrm{ref}}}, \tag{3.18}$$

where $\mathrm{CRPS}_{\mathrm{model}}$ is the temporal mean CRPS of the model ensemble and $\mathrm{CRPS}_{\mathrm{ref}}$ is the CRPS of a reference forecast based on mean observed discharge. A value of 1 indicates perfect performance, while values near 0 indicate no improvement over climatology.

Reliability was defined as the empirical coverage of the 95% prediction interval (McInerney et al., 2017):

$$\text{Reliability} = \frac{1}{N}\sum_{t=1}^{N} \mathbf{1}\big(Q_{0.025,t} \le y_t \le Q_{0.975,t}\big), \tag{3.19}$$

where $Q_{0.025,t}$ and $Q_{0.975,t}$ are the 2.5th and 97.5th percentiles of the posterior predictive distribution at time t, y_t is the observed value, $\mathbf{1}(\cdot)$ is the binary indicator function, and N is the number of observations. With this specific definition of reliability, a perfectly calibrated model considering reliability only would have a reliability of 0.95.

The mean-flow relative sharpness measures the concentration of predictive distributions and is defined as (McInerney et al., 2017):

$$\text{Sharpness} = 1 - \frac{1}{N\,\bar{y}}\sum_{t=1}^{N}\big(Q_{0.975,t} - Q_{0.025,t}\big), \tag{3.20}$$

where $\bar{y}$ is the mean observed discharge. This formulation assigns high values of sharpness to narrow prediction intervals. A perfect model considering only sharpness as defined above has no prediction uncertainty and thus sharpness equal to 1.

To complement the probabilistic metrics above with a deterministic measure of ensemble-mean accuracy, we additionally report the Kling–Gupta Efficiency (KGE) (Gupta et al., 2009), which decomposes model performance into correlation, variability, and bias components and is widely used as a benchmark in hydrologic modelling. KGE is computed from the posterior ensemble mean streamflow $\bar{y}$ and the observed series y as

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha_{\mathrm{KGE}} - 1)^2 + (\beta_{\mathrm{KGE}} - 1)^2}, \tag{3.21}$$

where r is the Pearson correlation coefficient between the simulated ensemble mean and the observations, $\alpha_{\mathrm{KGE}} = \sigma_{\mathrm{sim}}/\sigma_{\mathrm{obs}}$ is the ratio of simulated to observed standard deviations, and $\beta_{\mathrm{KGE}} = \mu_{\mathrm{sim}}/\mu_{\mathrm{obs}}$ is the ratio of simulated to observed means. A value of $\mathrm{KGE} = 1$ indicates perfect agreement; values approaching 1 therefore indicate stronger deterministic skill.

3.3 Results

3.3.1 Probabilistic Model Performance

Before attributing predictive uncertainty to its constituent sources, we evaluate the overall probabilistic performance of the calibrated blended model across the 12 MOPEX watersheds using CRPS skill score, KGE, reliability, and sharpness (Figure 3.3). All four metrics are oriented so that higher values indicate better performance.

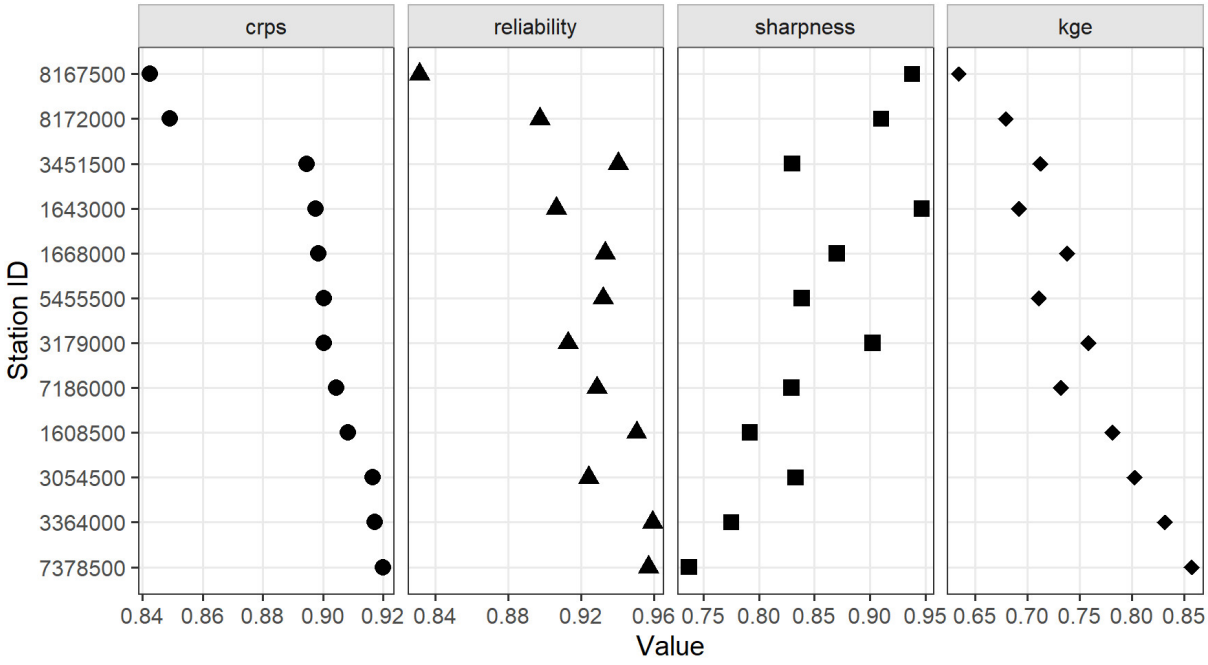


Figure 3.3: Probabilistic model performance across the 12 MOPEX catchments, evaluated using four complementary metrics: CRPS skill score, Kling–Gupta Efficiency (KGE), reliability (empirical coverage of the 95% prediction interval), and sharpness (concentration of predictive distributions). All metrics are oriented so that higher values indicate better performance.

CRPS skill scores range from approximately 0.84 to 0.92 across the 12 catchments, indicating that the blended model substantially outperforms a climatological reference forecast at every site. The two Texas catchments (stations 8167500 and 8172000, corresponding to C#11 and C#12 in Figure 3.1) achieve the lowest skill ($\sim$ 0.84–0.85), while the remaining 10 eastern catchments cluster more tightly between 0.89 and 0.92. This spatial

gradient in CRPS skill is consistent with the semi-arid, episodic precipitation regime of the Texas sites, where rare but intense convective events generate flashy hydrographs that are structurally challenging for process-based models calibrated primarily on long-term climatological signal.

The Kling–Gupta Efficiency (KGE) shows consistent deterministic performance across the 12 catchments, with values ranging from approximately 0.63 to 0.85 and the same hydroclimatic gradient observed for the CRPS skill score. The 10 humid eastern catchments achieve KGE values broadly between 0.70 and 0.85, indicating well-balanced performance in correlation, variability, and bias of the ensemble mean against observations. The two Texas catchments again sit at the lower end ($\sim$ 0.63–0.68), reflecting the same difficulty that the probabilistic metrics expose.

The agreement between KGE and CRPS skill score is reassuring, as the deterministic ranking of catchments aligns with their probabilistic ranking (Romero-Cuellar et al., 2024), indicating that the ensemble mean retains the central-tendency information of the posterior predictive distribution. KGE values in the 0.70–0.85 range are consistent with the better-performing tier of large-sample hydrology benchmarks reported by Knoben et al. (2019) and Troin et al. (2022), supporting the conclusion that the blended model used here provides competitive deterministic skill in addition to its probabilistic performance.

Reliability ranges from approximately 0.84 to 0.96, with most sites achieving coverage near the nominal 0.95 level. One Texas catchment (8167500) shows the lowest reliability ($\sim$ 0.84), while the other (8172000) sits at $\sim$ 0.90; both Texas sites lie at the high end of the sharpness range ($\sim$ 0.92–0.94), reflecting narrow prediction intervals that fail to encompass observed variability. Across the 10 eastern catchments, sharpness ranges from approximately 0.74 to 0.94 alongside near-nominal reliability. The reliability–sharpness trade-off at the Texas catchments is a hallmark of overconfident predictive distributions, and by attributing a persistent share of predictive variance to forcing year-round, the decomposition correctly identifies that these catchments are controlled by inherently unpredictable episodic meteorological forcing, pinpointing the Gaussian homoscedastic likelihood as the proximate source of the calibration failure (Schoups and Vrugt, 2010).

3.3.2 Temporal Dynamics of Uncertainty Sources

Figure 3.4 presents the daily temporal decomposition of predictive uncertainty across all 12 MOPEX watersheds over the climatological hydrological year. Each panel shows the stacked proportional contributions of forcing (yellow), parameter (orange), structural (dark

red), and unexplained (grey) uncertainty, alongside the climatological 50th-percentile observed hydrograph (black line).

Parameter uncertainty (orange) is the dominant source of predictive variance at the majority of catchments and throughout the year, typically accounting for 40–60% of total uncertainty in the more humid eastern sites; in the elevated-forcing cluster (Fig. 3.4a, d, i, j, k, l), parameter uncertainty remains substantial ($\sim$ 30–55%) but is no longer unambiguously dominant relative to forcing uncertainty. This persistent dominance reflects the high-dimensional parameter space of the blended model and is consistent with the interpretation that the long observational record constrains the seasonal mean behavior of the model, but fails to uniquely resolve individual parameter combinations, a classic manifestation of equifinality in hydrologic modelling (Beven, 2006).

Forcing uncertainty (yellow) is the second-largest contributor across most catchments, typically contributing 20–40% of total variance, and exhibits the most pronounced seasonal dynamics of the three attributed sources. In several humid eastern catchments (e.g., Fig. 3.4e, h, b), forcing uncertainty amplifies during the winter–spring high-flow season, when heavy frontal precipitation generates the largest streamflows of the year. This elevated forcing contribution attenuates during summer–fall months as precipitation inputs diminish. In contrast, a distinct cluster of catchments, encompassing the two Texas sites (Fig. 3.4k, l) together with the Spring (i), Amite (j), Tygart Valley (d), and Potomac (a) basins, exhibits elevated forcing uncertainty year-round, with forcing contributions consistently in the 30–50% range and only modest seasonal modulation. The shared signal across this cluster is better characterized as a persistent elevation of forcing uncertainty driven by a combination of semi-arid episodic precipitation (Texas), high-relief orographic variability (Tygart Valley, Potomac), and land–atmosphere coupling in the central and lower-latitude sites (Spring, Amite), rather than as a clean Texas-versus-eastern dichotomy.

Structural uncertainty (dark red) contributes approximately 10–25% of total variance and is distinguished from the other components by a consistent summer–fall amplification visible across multiple catchments (e.g., Fig. 3.4e, g, h), coinciding with the low-flow recession period. This pattern is physically interpretable. During summer–fall recession, baseflow generation and subsurface drainage dominate catchment response, and competing structural algorithms diverge most strongly in their flux estimates, amplifying structural spread. Conversely, during winter–spring high-flow periods, the large streamflow signal strongly constrains which structural configurations are consistent with observations.

The unexplained variance (grey) remains consistently the smallest component, staying below approximately 12% at most sites throughout the year. Its low magnitude confirms that the linear MANOVA model provides a broadly adequate approximation of the

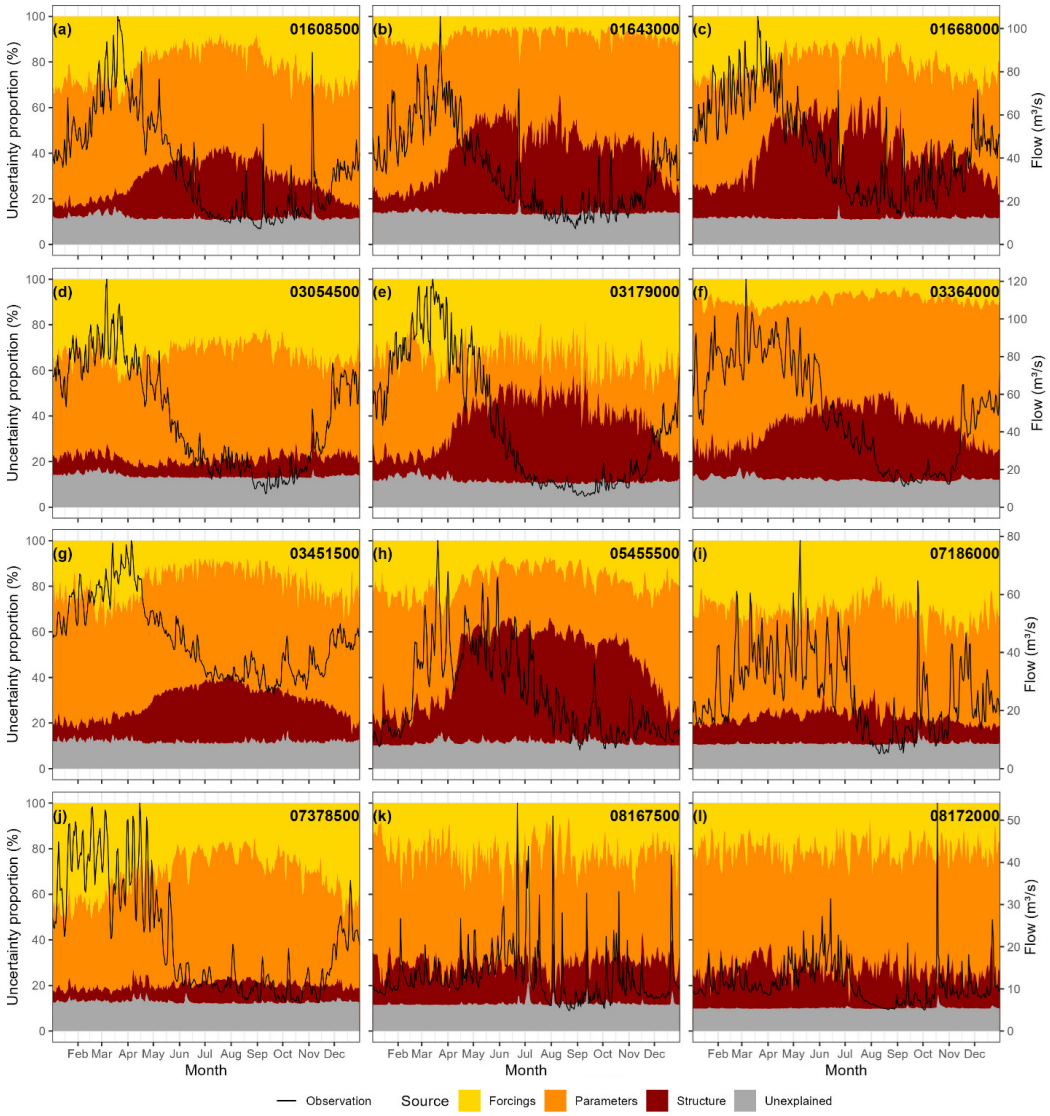


Figure 3.4: Time-varying uncertainty decomposition across the 12 MOPEX catchments over the climatological hydrological year. Each panel shows the stacked proportional contributions of parameter (orange), forcing (yellow), structural (dark red), and unexplained (grey) variance, computed independently for each Julian calendar day using MANOVA applied to 5,000 posterior parameter samples. The climatological 50th-percentile daily observed streamflow is overlaid (black line, right axis).

nonlinear parameter–streamflow relationship within the posterior ensemble. Elevated unexplained fractions at the Texas catchments suggest that the linearity assumption may be less well satisfied in flashy, semi-arid settings.

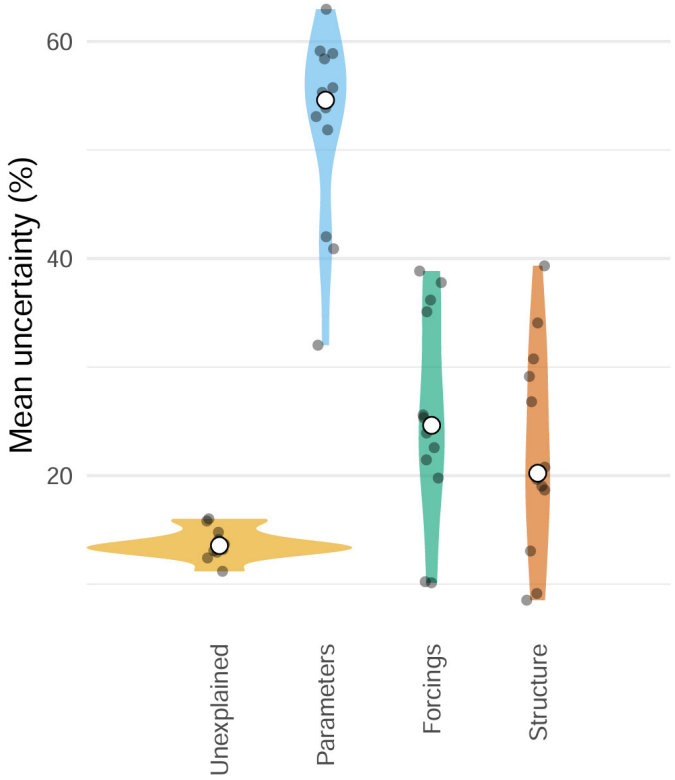


Figure 3.5: Distribution of annual mean uncertainty proportions across the 12 MOPEX catchments for the four variance components. Violin plots show the full cross-catchment distribution, with individual catchment values overlaid as dots. The horizontal line within each violin denotes the median.

3.3.3 Cross-Catchment Summary of Uncertainty Partitioning

Figure 3.5 summarizes the annual mean uncertainty proportions across all 12 catchments. Parameter uncertainty is the dominant source at every site, with a median annual contribution of approximately 53% and a range of 30–65%. Most catchments cluster tightly

between 50 and 60%, with three lower-end outliers ($\sim$ 30–42%) drawn from the elevated-forcing cluster identified above.

Forcing uncertainty exhibits the greatest inter-catchment variability of the three attributed components, with a median of approximately 27% and a range spanning from below 10% to nearly 38%. The distribution shows a distinct bimodal character, with a cluster of humid catchments near 20–30% and the elevated-forcing cluster (Texas, Spring, Amite, Tygart Valley, Potomac) sitting near 35–38%, consistent with the hydroclimatic gradient.

Structural uncertainty has a median of approximately 20% and a broader distribution (range $\sim$ 5–40%), indicating that the model's capacity to constrain its own structural configuration via Bayesian inference varies meaningfully across catchments. The unexplained component is consistently the smallest, with a median of approximately 11% and a very tight distribution (approximately 8–15%), confirming the adequacy of the MANOVA decomposition framework.

3.3.4 Relationships Between Uncertainty Sources and Basin Attributes

Figure 3.6 presents the Pearson correlation coefficients between the annual mean uncertainty proportions for each source and five basin descriptors: drainage area, mean elevation, mean channel slope, forest cover fraction, and urban fraction. These correlations are exploratory given the small sample size of 12 catchments and should be regarded as hypothesis-generating rather than conclusive.

Parameter uncertainty shows the strongest association with basin attributes. Its positive correlation with drainage area ($r = 0.47$) suggests that larger basins, which typically exhibit greater spatial heterogeneity, present a more complex and under-constrained parameter estimation problem. The negative correlations with forest fraction ($r = -0.27$), urban fraction ($r = -0.22$), elevation ($r = -0.17$), and slope ($r = -0.18$) may reflect the constraining effect of well-defined physiographic and land cover characteristics.

Forcing uncertainty exhibits its strongest positive correlation with forest fraction ($r = 0.50$), followed by elevation ($r = 0.35$) and slope ($r = 0.29$). The forest-fraction association likely reflects multiple sources of meteorological uncertainty in forested environments: canopy interception, transpiration sensitivity, and above-canopy gauge undercatch. The elevation and slope associations may reflect orographic amplification of precipitation variability. The negative correlation with drainage area ($r = -0.24$) likely arises from spatial averaging.

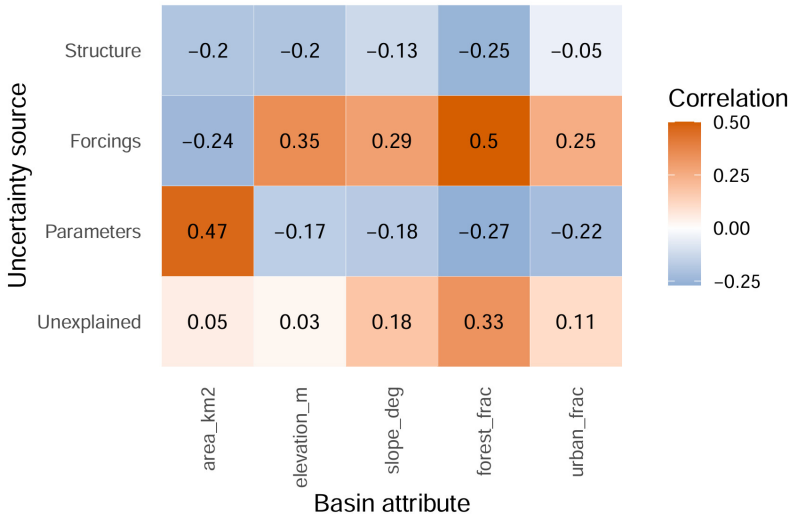


Figure 3.6: Pearson correlation coefficients between the annual mean uncertainty proportions for each source and five basin physiographic descriptors, computed across the 12 MOPEX catchments. Cell colour intensity indicates correlation magnitude; cell values show the coefficient.

Structural uncertainty is negatively correlated with all five basin attributes, most notably with forest fraction ($r = -0.25$) and drainage area ($r = -0.20$). This suggests that the blended model's weight-generating parameters are most effectively constrained by data in larger, well-vegetated catchments. The unexplained component shows consistently weak correlations with all attributes, confirming that the variance attribution is not strongly confounded by basin physiography.

3.4 Discussion

3.4.1 Decomposition Using the Blended Model Structure

The results of this study, parameter uncertainty dominant at $\sim 53\%$, followed by forcing at $\sim 27\%$ and structural uncertainty at $\sim 20\%$, are consistent with and add to existing understanding of predictive uncertainty in hydrological modelling. The dominance of parameter uncertainty is consistent with earlier multi-source decomposition studies (Renard et al., 2011; Moges et al., 2021). Results also reflect a well-documented reality, namely that even with long observational records, the high-dimensional posterior of a calibrated

process-based model retains substantial spread through equifinality (Beven, 2006). The mechanism by which this result is obtained differs from earlier studies. By encoding structural uncertainty within the weight-generating parameters, the blended model concentrates all three uncertainty sources into a single, well-posed Bayesian inference, circumventing the identifiability problem that has constrained prior decomposition attempts. Renard et al. (2010) demonstrated that simultaneous inference of separate input and structural error models from streamflow records alone is ill-posed without precise prior knowledge of data quality, and Salamon and Feyen (2010) showed that omitting structural error substantially underestimates total predictive uncertainty. The present framework avoids both problems by construction. A formal compositional comparison against two alternative blended-model variants (BM v2 of Chlumsky et al. (2024) and BM v2 + BASE_LINEAR), reported in Appendix 2.B, indicates that the three-way uncertainty partitioning is statistically indistinguishable from the BM v1 result at the majority of MOPEX12 catchments.

Comparing these results against quantitative benchmarks from the literature further clarifies their interpretation. Thiboult et al. (2016), working in a short-range forecasting context, found initial-condition uncertainty to be the dominant error source. That the present study finds parameter uncertainty dominant instead reflects the fundamentally different modelling regime of long-term simulation, in which initial conditions are well-constrained by a 55-year observational record. Troin et al. (2022), using an ensemble of 180 discrete model structures across 698 North American catchments, found that structural uncertainty varies markedly across climate zones and flow metrics. The summer–fall amplification of structural uncertainty identified in Figure 3.4 is consistent with findings from Clark et al. (2011) and Fenicia et al. (2008), who showed that structural errors are most consequential during low-flow events.

3.4.2 Forcing Uncertainty, Hydroclimatic Regime, and the Texas Catchments

Forcing uncertainty, though second in the overall hierarchy, exhibits the most pronounced seasonal dynamics and the strongest correlations with basin physiography of the three sources. Across the humid eastern catchments, forcing uncertainty amplifies markedly during the winter–spring high-flow season (Kavetski et al., 2006; McMillan et al., 2011); it attenuates during summer when reduced precipitation and lower baseflows diminish the absolute contribution of forcing-driven prediction spread. The positive correlation between forcing uncertainty and forest fraction ($r = 0.50$; Figure 3.6) is consistent with several physically plausible mechanisms: forested catchments are subject to above-canopy gauge

undercatch, canopy interception, and high transpiration sensitivity to temperature inputs (McMillan et al., 2012). The elevation and slope associations likely reflect orographic amplification of precipitation variability, consistent with Tang et al. (2023).

The two Texas catchments provide an informative illustration of the framework's diagnostic capability. These semi-arid sites exhibit elevated and persistent forcing uncertainty ($\sim$ 30–45% of total variance throughout the year), and they exhibit the characteristic reliability–sharpness trade-off of overconfident models. Where a conventional deterministic or lumped-error model would simply report poor performance, the uncertainty decomposition attributes the calibration failure specifically to meteorological forcing unpredictability and identifies the Gaussian homoscedastic likelihood as the proximate methodological mismatch (Schoups and Vrugt, 2010). Future extensions should explore skew-normal, log-normal, or censored Gaussian likelihoods for semi-arid regimes (Moges et al., 2021; Hassan et al., 2022).

3.4.3 Methodological Considerations and Limitations

The MANOVA-based variance decomposition rests on an approximate linearity assumption between posterior parameter samples and simulated streamflow for each Julian calendar day. The low unexplained variance (median $\sim$ 11%) confirms this assumption is broadly satisfied across the MOPEX12 sites. The elevated residuals at the Texas catchments ($\sim$ 15–20%), however, signal that the linear model is less adequate in highly nonlinear, threshold-driven flow regimes, where variance-based methods employing nonlinear surrogates (Saltelli et al., 2008) or random forest variable importance may provide sharper attribution. The Gaussian homoscedastic likelihood introduces a complementary limitation, in that systematic, temporally correlated structural errors may be partially absorbed into the parameter or residual components rather than attributed to structure. Autoregressive or Gaussian process error corrections (Del Giudice et al., 2016) could improve attribution in such cases, and the choice of likelihood function itself constitutes a structural modelling decision with material consequences for the uncertainty decomposition (Hassan et al., 2022).

A natural concern with any decomposition that reformulates structural uncertainty as parameter uncertainty is whether the reported partitioning is sensitive to the configuration of the blended model. To assess this, we repeated the inference using two alternative blended-model variants and compared the resulting three-part output compositions in compositional log-ratio space. The full procedure and station-level results are reported in Appendix 2.B, where Aitchison effect sizes and false discovery rate (FDR) (Benjamini and Hochberg, 1995) are used to control the expected proportion of false positives. Across

the 12 MOPEX catchments, only a small subset of stations exhibits statistically significant compositional shifts, concentrated among the weaker-performing sites identified in Section 3.3.2. We therefore interpret the BM v1 decomposition as a robust characterization of the uncertainty hierarchy at the majority of catchments.

The 12-catchment sample, though deliberately selected for hydroclimatic diversity (Schaake et al., 2006), is geographically confined to the eastern and south-central United States. Replication across the CAMELS (Addor et al., 2017) and/or the full MOPEX438 dataset would be required to establish whether the uncertainty hierarchy is robust across snow-dominated, monsoonal, and temperate oceanic regimes.

3.4.4 Implications for Practice and Future Directions

A central practical implication of this study is that the dominant uncertainty source directly identifies where investment in model improvement or data collection is most cost-effective. At most MOPEX12 catchments, parameter uncertainty accounts for over half of total predictive spread, implying that additional observations enabling tighter posterior constraint would reduce uncertainty more effectively than improved precipitation measurement networks or structural refinement. This aligns with the data-worth analysis perspective advocated by Renard et al. (2011) and Gupta et al. (2012).

The temporal dynamics of structural uncertainty also have direct operational implications. Structural uncertainty varies seasonally, peaking during summer–fall low-flow recession, while forcing uncertainty is most consequential during the winter–spring high-flow season when accurate flood predictions are critical. Applying a single static uncertainty envelope throughout the year therefore misrepresents risk systematically. Ensemble-based forecasting systems that propagate the full posterior distribution through the blended model in real time are directly suited to delivering time-varying uncertainty estimates to decision-makers.

Looking ahead, several directions merit investigation. Applying the framework under climate change scenarios would test whether the relative contributions of uncertainty sources shift as hydroclimatic regimes evolve (Troin et al., 2022). Extending the blended model to semi-distributed or distributed configurations would add a spatial dimension to the temporal decomposition presented here.

3.5 Conclusions

This study has presented a unified Bayesian framework for simultaneously quantifying and temporally decomposing the three dominant sources of predictive uncertainty in hydrological modelling: parameter, structural, and forcing uncertainty. By using the blended model structure of Mai et al. (2020), which reformulates structural uncertainty as an equivalent parameter uncertainty through continuously weighted process algorithms, all three uncertainty sources become amenable to joint Bayesian inference. The No-U-Turn Sampler (NUTS) was used to obtain posterior distributions over weight-generating parameters, hydrologic process parameters, and meteorological perturbation parameters simultaneously. A MANOVA-based variance decomposition was then applied to the resulting posterior ensemble for each Julian calendar day, yielding a time-varying uncertainty budget across the full hydrological year.

Applied to 12 benchmark MOPEX catchments spanning diverse hydroclimatic regimes across the conterminous United States over 1948–2004, the framework achieved CRPS skill scores of 0.84–0.92, KGE values of approximately 0.63–0.85, and near-nominal 95% prediction interval coverage, confirming that the blended model provides reliable probabilistic streamflow predictions across contrasting climate settings.

The uncertainty decomposition revealed three key findings. First, parameter uncertainty is the dominant source of predictive variance at every catchment, with median contribution of approximately 53% annually. Second, forcing uncertainty (median $\sim 27\%$) exhibits pronounced seasonal amplification during convective precipitation seasons and scales positively with forest cover fraction ($r = 0.50$) and elevation. Third, structural uncertainty (median $\sim 20\%$) peaks during summer–fall low-flow recession when competing baseflow and subsurface algorithms diverge most strongly.

These results have practical implications. The dominance of parameter uncertainty across all catchments suggests that additional streamflow observations providing stronger posterior constraint would yield greater uncertainty reductions than equivalent investments in forcing data quality or model structural refinement, except in forested and topographically complex catchments where forcing uncertainty is disproportionately large. The seasonal dynamics of structural uncertainty imply that static uncertainty envelopes are inadequate for operational flood forecasting. The blended model framework presented here provides a computationally tractable path toward this goal, and its extension to larger catchment samples (e.g., CAMELS, Addor et al., 2017), distributed model configurations, and multi-variable observational constraints is a natural next step for uncertainty decomposition in operational hydrology.

Funding

This work was funded by Natural Resources Canada's Floodplain Hazard and Mapping Program grant (FHIMP23-27P2), NSERC Discovery grants held by co-authors Craig (RGPIN-2017-03920; RGPIN-2023-03459) and Tolson (RGPIN-2022-03890), and by funding from co-author Craig's NSERC Tier II Canada Research Chair in Hydrologic Modeling and Analysis (CRC-2020-00176).

Chapter 4

Continental-Scale Decomposition of Hydrologic Model Uncertainty Across 2,869 North American Watersheds

Preface

This chapter deploys the methodological framework developed in Chapter 3 at continental scale across 2,869 watersheds drawn from the HYSETS database (Arsenault et al., 2020). A version of this work, with minor changes, has been submitted to *Nature Communications*, with co-authors J. R. Craig, B. A. Tolson, and J. Romero-Cuellar.

Overview

Hydrologic model predictions are uncertain, yet the relative contributions of structural, parameter, and forcing uncertainty have not been estimated at continental scale. Across 2,869 North American watersheds, we show that the long-standing assumption that parameter uncertainty dominates does not hold everywhere and at all times. Overall, structural uncertainty is the largest contributor to predictive spread (42% of variance), exceeding parameter uncertainty (36%) and forcing uncertainty (19%) at the continental median. The dominant source varies systematically with landscape, and hydroclimatic attributes

explain 66% of the spatial variance in structural share. Structural uncertainty is also non-stationary in time. At basins with prolonged summer recessions, whether snow-influenced or humid-eastern, it expands several-fold as competing subsurface algorithms diverge, while at semi-arid basins with flashy recessions this expansion is muted. Posterior analysis yields a continental atlas of which algorithm the data prefer at each location, organized predictably by climate and terrain. These results indicate that uncertainty reduction is a landscape- and season-specific problem, in which the most effective intervention, and the season in which it matters most, depend on where a basin sits in the hydroclimatic landscape.

4.1 Introduction

Flood forecasting, water resource allocation, and climate-change impact assessments often rest on uncertainty estimates that decision-makers expect to be reliable. Yet the relative contributions of the three predominantly epistemic sources of hydrologic predictive uncertainty—model structure, model parameters, and meteorological forcing, each reducible in principle through better models, data, or inference (Beven, 2006; Liu and Gupta, 2007; Montanari, 2007), are difficult to estimate jointly, a disentanglement problem identified as one of the 23 challenges in hydrology (Blöschl et al., 2019). Such relative contributions have never been estimated at the continental scale. Existing frameworks typically quantify only one or two sources at a time: parameter uncertainty through Bayesian inference (Kavetski et al., 2006; Vrugt et al., 2008), structural uncertainty through multi-model ensembles (Chlumsky et al., 2021, 2024; Troin et al., 2022), or forcing uncertainty through stochastic input perturbation (McMillan et al., 2011; Nijssen and Lettenmaier, 2004). Others lump all three into a single residual term that cannot be decomposed into its contributing sources (Evin et al., 2014). The practical cost of this gap is that resources for uncertainty reduction, such as additional gauges, better models, or improved forcing estimates, are allocated without knowing which would help most.

Which source of uncertainty dominates at a given location matters operationally, because it determines which intervention yields the greatest reduction in predictive uncertainty: additional streamflow observations when parameter uncertainty is limiting, model development when structural uncertainty dominates, or improved precipitation estimates when forcing uncertainty controls predictions (Markstrom et al., 2016). Without knowledge of the dominant source, these investments cannot be prioritized.

Previous decomposition attempts have faced two constraints. First, when input and structural error models are inferred simultaneously from streamflow alone, the inverse

problem is ill-posed (Renard et al., 2010), because different combinations of input error, structural error, and parameter compensation produce the same misfit, making the decomposition non-unique (Renard et al., 2011). Ajami et al. (2007) addressed this through the Integrated Bayesian Uncertainty Estimator (IBUNE); Thiboult et al. (2016) combined ensemble Kalman filtering with multi-model approaches; Salamon and Feyen (2010) paired multiplicative error models with sequential data assimilation. However, these methods either address a subset of sources, treat the decomposition as stationary, or require separately calibrated model suites that introduce additional complexities in the decomposition problems (Clark et al., 2017; Mendoza et al., 2015). Second, existing studies have been restricted to small samples, typically fewer than 20 catchments (Chlumsky et al., 2021, 2024; Salamon and Feyen, 2010), preventing generalizable conclusions about how landscape controls the uncertainty budget.

Here we address both limitations. A blended hydrologic model (Chlumsky et al., 2021, 2024; Mai et al., 2020) replaces discrete structural selection with continuously weighted process algorithms, encoding structural uncertainty as uncertainty in weight-generating parameters. This converts the three-source decomposition into a well-posed single Bayesian inference over $\boldsymbol{\Theta} = \{\boldsymbol{w}', \boldsymbol{\theta}, \boldsymbol{\alpha}\}$, corresponding to structural parameters, model input parameters, and forcing parameters, though residual trade-offs among $\boldsymbol{w}'$, $\boldsymbol{\theta}$, and $\boldsymbol{\alpha}$ within the joint posterior remain possible and motivate sensitivity testing (Renard et al., 2010). Mai et al. (2022) previously applied this model to map process sensitivities across North American basins. Here we apply a similar strategy, applying full Bayesian uncertainty decomposition across 2,869 HYSETS watersheds (Arsenault et al., 2020), applying MANOVA decomposition to the posterior ensemble for every Julian calendar day.

The results reveal three findings. First, structural uncertainty, not parameter uncertainty, is the largest contributor to predictive spread at the continental scale, though the dominant source varies regionally in a pattern predictable from basin attributes. Second, the posterior distributions of process algorithm weights provide a continental atlas of which algorithm the data prefer at each location and how confidently they prefer it; the interquartile range (IQR) of these distributions, rather than the weight values alone, becomes the central diagnostic of structural identifiability. Third, structural uncertainty is nonstationary in time. At basins with prolonged summer recessions, whether snow-influenced or humid-eastern, it expands several-fold during summer–fall recession when competing subsurface algorithms diverge, while at semi-arid basins with flashy recessions this expansion is muted, with the continental pattern visualized through the ratio of summer–fall to winter–spring structural variance share.

4.2 Results

The results are presented in five parts. We first establish that the blended model produces adequate probabilistic predictions across the 2,869-watershed sample, justifying the decomposition that follows. We then quantify the continental uncertainty budget, finding that structural uncertainty exceeds parameter and forcing uncertainty at the median. Next, we examine which landscape attributes predict the dominant uncertainty source, showing that climate variables outweigh topography. We then map the posterior process weights and their IQRs to produce a continental atlas of which algorithm the data prefer at each location and how confidently. Finally, we resolve the decomposition seasonally, showing that structural uncertainty expands sharply during summer–fall recession at basins with prolonged recession regimes (snow-influenced or humid-eastern) while remaining stable at semi-arid sites.

4.2.1 Probabilistic skill across hydroclimatic regimes

Before decomposing uncertainty, we assessed the blended model's predictive adequacy. Figure 4.1 maps the scaled continuous ranked probability score (CRPS; Hersbach, 2000). Across all 2,869 watersheds, the median scaled CRPS is 0.25 (IQR: 0.19–0.31; range: 0.12–0.79), indicating that the typical probabilistic prediction error is approximately 25% of mean observed streamflow. Three-quarters of watersheds achieve scaled CRPS below 0.31, and the best-performing quarter fall below 0.19. The model performs best across the southeast and Pacific Northwest coast (scaled CRPS predominantly below 0.20), shows moderate performance in the northeast and Great Lakes (0.20–0.30), and exhibits the weakest scores across the Great Plains and semi-arid interior (0.30–0.70, with dense clusters exceeding 0.60 in the semi-arid interior). This spatial gradient is consistent with other continental-scale models (Newman et al., 2015; Rakovec et al., 2019), and the weaker performance in semi-arid regions reflects the Gaussian likelihood's poor fit to intermittent, heavy-tailed streamflow distributions (Markstrom et al., 2016). With 75% of watersheds achieving scaled CRPS below 0.31, the blended model provides adequate probabilistic predictions across a wide range of hydroclimatic settings, supporting the uncertainty decomposition that follows. In analysis not shown here, performance was evaluated using additional quality metrics; these results also supported the conclusion that the predictive model was adequately skilled.

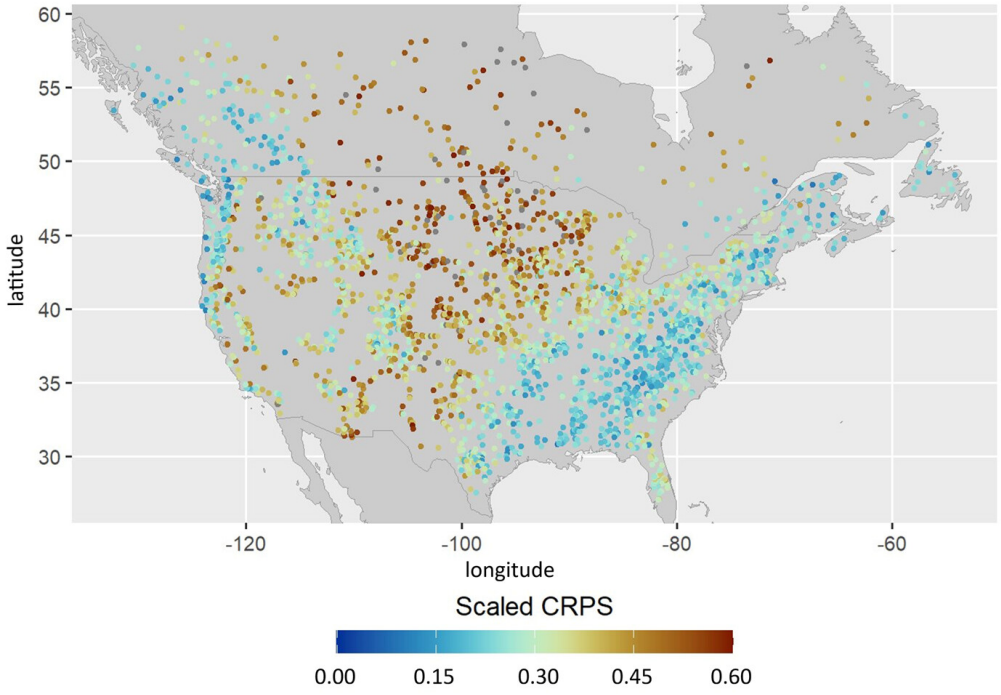


Figure 4.1: Scaled continuous ranked probability score (CRPS) across 2,869 North American watersheds. Lower values (blue) indicate better probabilistic prediction skill. Scaled CRPS: mean daily CRPS normalized by mean observed streamflow.

4.2.2 Structural uncertainty dominates the continental budget

Across all 2,869 watersheds, the median annual contribution of structural uncertainty to total predictive variance is 42% (IQR: 33–52%), followed by parameter uncertainty at 36% (IQR: 28–44%) and forcing uncertainty at 19% (IQR: 12–27%). With 2,869 watersheds, the 6-percentage-point separation between structural and parameter medians is highly unlikely to arise from sampling variability. This ranking runs against the conventional assumption that parameter uncertainty always dominates hydrologic predictions (Beven, 2009). At the continental scale, structural uncertainty is the single largest contributor. The remaining $\sim 4\%$ is unexplained variance (MANOVA residual), confirming the adequacy of the linear decomposition. Because the reported values are medians of per-watershed shares, the three medians need not sum to 100%. At each individual watershed, the three attributed shares and the residual sum to 100%. Observational uncertainty in the streamflow records is treated implicitly through the likelihood function and is not assigned a separate share in this decomposition.

Although structural uncertainty dominates the continental median, Figure 4.2 reveals pronounced regional variation. Parameter uncertainty (Fig. 4.2, top) is highest across the Great Lakes and northeast and northwest corridors, reaching 50–75%. These humid, energy-limited catchments generate consistent runoff from year to year, constraining the forcing parameters effectively but leaving the high-dimensional posterior of process parameters under-determined through equifinality (Beven, 2006).

Across the western and central continent, parameter uncertainty drops below 40%. Structural uncertainty (Fig. 4.2, middle) shows the complementary pattern, being most prominent across the Great Plains and interior mountains (40–70%), where competing process representations, VIC–Arno versus HBV infiltration, BASE_VIC versus TOPMODEL versus linear-analytic quickflow, produce sharply divergent predictions because the landscape does not impose a single dominant runoff pathway (Khatami et al., 2019). This region, which includes the prairie pothole region of central North America, is well-known as a place where one-size-fits-all model structures tend to fail (Ahmed et al., 2023). Forcing uncertainty (Fig. 4.2, bottom) is generally secondary at the annual scale (median 19%) but exceeds 25% across some of the humid southeast United States and Gulf Coast, where heavy, frequent rainfall amplifies the effect of forcing irregularities on streamflow, although it shows meaningful seasonal modulation, as discussed in the Seasonal dynamics section below.

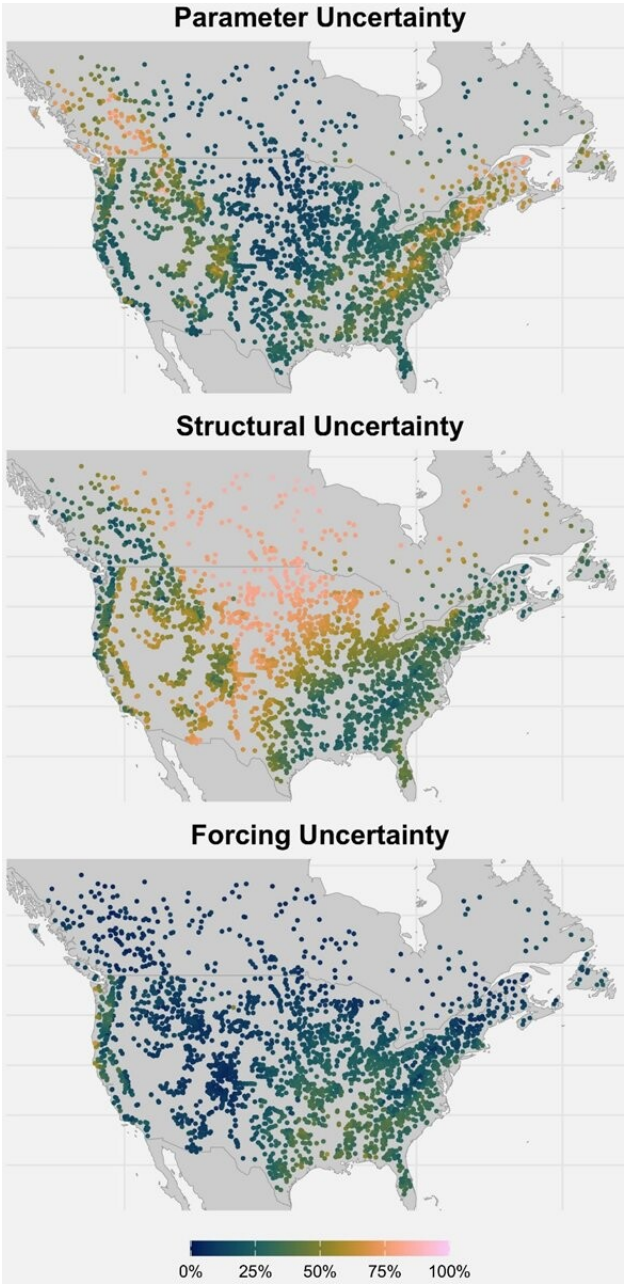


Figure 4.2: Annual-mean uncertainty attribution across 2,869 North American watersheds. Color: proportion of total explained variance attributed to parameter (top), structural (middle), and forcing (bottom) uncertainty.

4.2.3 Climate, not topography, controls the budget

To assess how basin physical characteristics shape the contributions of each uncertainty source, we examined ternary diagrams of the share of total variance attributable to each source. These reveal that basin attributes organize the uncertainty space (Fig. 4.3), though with important nuances that quantitative analysis clarifies. Elevation shows a visual gradient, but not toward the structural corner as might be expected. High-elevation watersheds cluster along the Parameter–Structure left edge of the triangle (high parameter share, moderate structural share, low forcing share), indicating that elevation primarily reduces forcing uncertainty rather than elevating structural uncertainty (Fig. 4.3a). This visual impression is confirmed quantitatively, as elevation alone has weak predictive power for structural share ($R^2 < 0.05$). This is consistent with expectations of freshet-dominated basins where the cumulative, rather than event-specific, snowfall magnitudes determine discharge magnitudes. This is echoed by the results for mean temperature, where cold catchments (< 0 °C) concentrate away from the forcing corner and somewhat toward the structural corner where snow process representation becomes the key structural decision, while warm catchments (> 15 °C) shift toward increased proportion of parameter and forcing contributions (Fig. 4.3d). Longitude reveals a more informative pattern. Central continental watersheds (roughly $-110°$ to $-90°$, the Great Plains) are the most structurally uncertain, shifting toward the Structural uncertainty corner, while far-eastern and maritime basins ($-70°$ to $-60°$) cluster toward the Forcing corner (Fig. 4.3b). Precipitation seasonality also organizes the space, with low-CV (uniform precipitation) catchments reaching higher into the structural region than high-CV (seasonal) catchments, although overlap is substantial (Fig. 4.3c).

If structural uncertainty dominates the continental budget but varies regionally, the natural question is what controls the variation. These are tendencies with overlap, not sharp divisions. Quantitatively, mean annual precipitation is the strongest single predictor of structural uncertainty share ($R^2 = 0.35$), followed by precipitation coefficient of variation ($R^2 = 0.28$), temperature seasonality ($R^2 = 0.22$), and precipitation event intensity ($R^2 = 0.20$). The precipitation regime explains far more variance than topographic factors (slope $R^2 = 0.09$; elevation $R^2 = 0.01$) or land cover attributes (forest fraction $R^2 = 0.10$), indicating that climate, not physiography, is the primary control on which uncertainty source dominates. A multiple regression combining precipitation and temperature attributes explains 66% of the variance in structural share across the 2,869-watershed in-sample fit, demonstrating that the uncertainty budget is effectively determinable from hydroclimatic setting alone. Because mean annual precipitation correlates with temperature seasonality, their combined R^2 reflects shared climatic information; out-of-sample

cross-validation, required to formally justify predictive extension to ungauged basins, is an important next step. This is consistent with Mai et al. (2022), who showed model sensitivities can be predicted from basin attributes with $R^2 > 0.80$. The predictability envelope is thus climate-driven. Drier catchments with more seasonal and intense precipitation have larger structural uncertainty, while parameter uncertainty dominates in wetter catchments with uniform precipitation.

4.2.4 Posterior weight distributions form a continental atlas of process choice and confidence

To address which structural choices are responsible for the spatial patterns above, we examined the posterior distributions of process algorithm weights at each watershed. Rather than mapping the posterior mode of every weight, which has a deterministic flavor, we adopted a probabilistic framing. For each process group (infiltration, quickflow, evapotranspiration, deep baseflow, snow balance), we identified at each location the algorithm carrying the maximum posterior weight (the dominant choice) and then coloured that location by the IQR of its posterior weight distribution. Low IQR indicates that the data confidently supports that algorithm at that location; high IQR indicates substantial residual ambiguity even in the dominant preference. The result is a continental atlas of process choice paired with a quantitative measure of how strongly the data support each choice (Fig. 4.4).

The atlas reveals four distinct patterns of process identifiability. First, the dominant infiltration algorithm shifts strongly with landscape (Fig. 4.4a). VIC–Arno is preferred at 2,010 of the 2,869 watersheds (70%), particularly across the agricultural Great Plains and central lowlands, with mostly moderate IQR (typically 0.10–0.20). HBV infiltration (Bergström and Forsman, 1973) is preferred at 599 watersheds (21%), distributed across elevated terrain and the prairie pothole region, at the lowest IQR of the three (typically below 0.15). HMETS infiltration is preferred only rarely ($n = 260$, 9%) and, as the violin plots of Figure 4.5a show distinctly, with the broadest IQR distribution of the three algorithms; even where the data prefer HMETS, that preference is comparatively weak. The infiltration choice is thus well supported by the data almost everywhere, with HMETS the least confidently identified of the three options.

Second, quickflow shows greater algorithmic ambiguity (Fig. 4.4b). BASE_VIC is preferred at 1,659 basins (58%), predominantly in the central and southern lowlands, at consistently low IQR (typically below 0.15), making it the most identifiable quickflow choice across the continent. By contrast, BASE_TOPMODEL ($n = 674$, 23%,

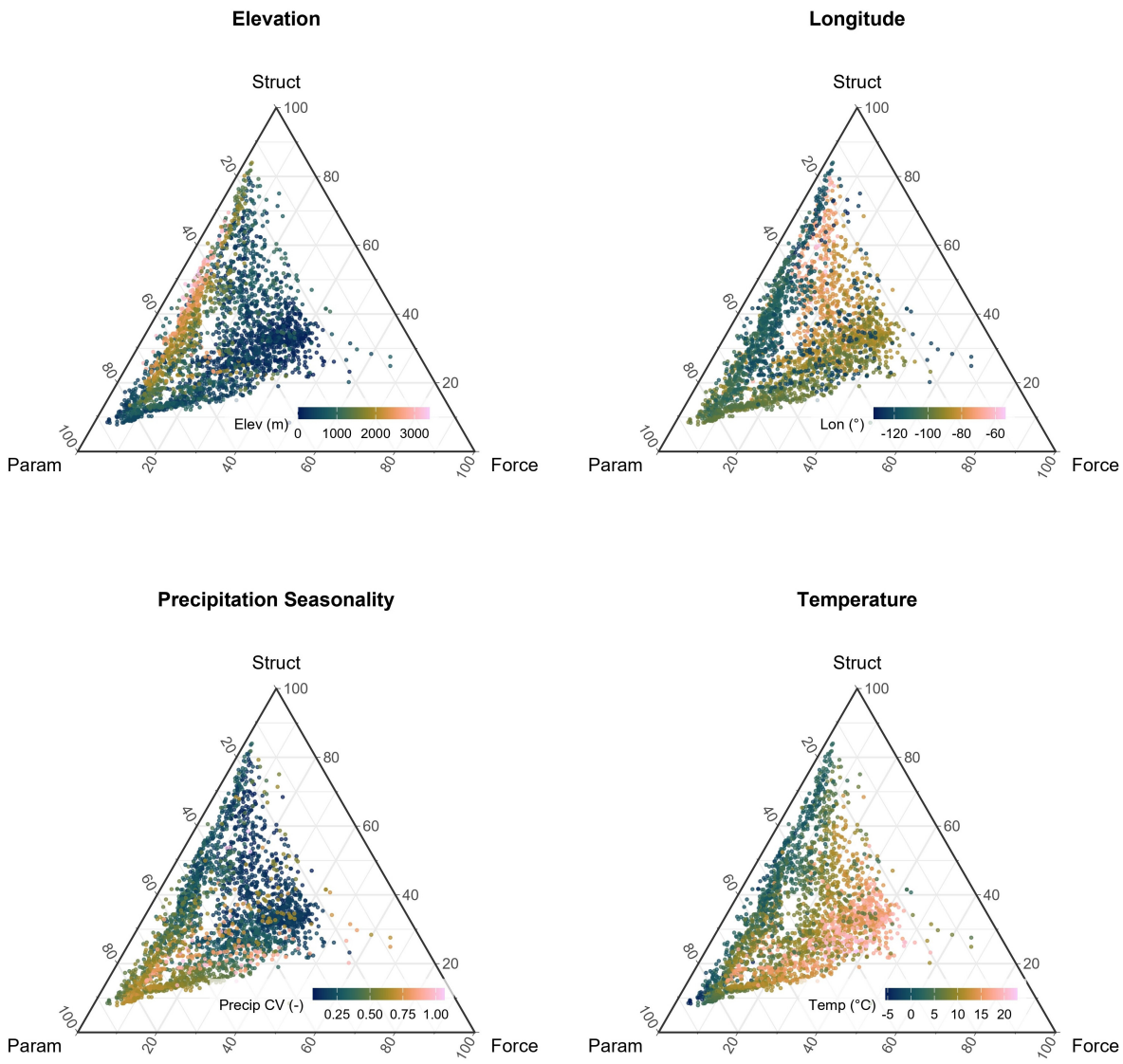


Figure 4.3: Ternary diagram of uncertainty sources (Par: parameter uncertainty; Str: structural uncertainty; Frc: forcing uncertainty), partitioning coloured by (a) mean catchment elevation (m), (b) longitude of the watershed centroid (°), (c) monthly precipitation coefficient of variation (CV), and (d) mean annual air temperature (°C).

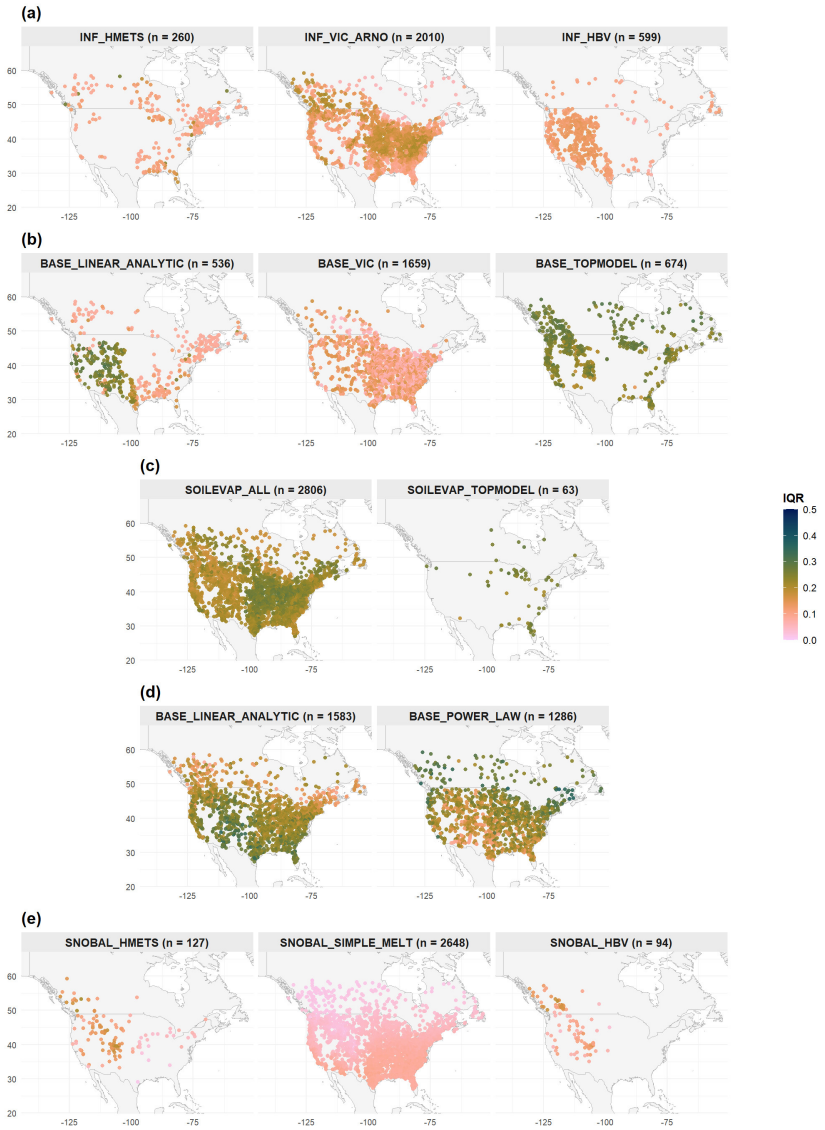


Figure 4.4: Probabilistic atlas of process algorithm choice across 2,869 North American watersheds. (a) Infiltration, (b) quickflow (interflow), (c) evapotranspiration, (d) deep baseflow, (e) snow balance. Each panel shows only the basins where the labelled algorithm carries the maximum posterior weight (n = number of basins for which that algorithm is the dominant choice); colour indicates the IQR of the posterior weight distribution at each basin (pink: low IQR, high confidence in the dominant choice; dark green/blue: high IQR, lower confidence).

concentrated in the Pacific Northwest, Appalachians, and parts of eastern Canada) and BASE_LINEAR_ANALYTIC (n = 536, 19%, mostly in the mountainous west) show notably higher IQR (typically 0.20–0.30; Fig. 4.5b). Even where TOPMODEL or LINEAR_ANALYTIC are most likely, residual posterior uncertainty in that preference remains substantial. Quickflow is therefore the process for which structural identifiability is most uneven across the continent, with confident identification in the cropland-dominated lowlands and substantial residual ambiguity elsewhere.

Third, evapotranspiration is dominated almost universally by SOILEVAP_ALL (preferred at 2,806 watersheds, 98%), with only 63 watersheds preferring SOILEVAP_TOPMODEL (Fig. 4.4c). The IQRs for both algorithms are moderately spread (medians $\sim$ 0.20–0.22; Fig. 4.5c), suggesting that while one algorithm is consistently preferred at the continental scale, the strength of that preference is moderate rather than sharp. This is consistent with Mai et al. (2022), who identified evapotranspiration as a less sensitive process than quickflow. Fourth, deep baseflow is split nearly evenly across the continent: BASE_LINEAR_ANALYTIC dominates at 1,583 watersheds and BASE_POWER_LAW at 1,286 (Fig. 4.4d), but in both cases IQRs are nearly identical and remain in the 0.20–0.35 range, among the highest of any process group together with evapotranspiration (Fig. 4.5d). The data express no decisive preference between linear and power-law representations of deep baseflow.

Snow balance shows the most clearly organized pattern (Fig. 4.4e). SNOBAL_SIMPLE_MELT dominates the continent (n = 2,648, 92%), preferred across most of the continent including basins with minimal snow influence, and at very low IQR (typically below 0.10). The more physically detailed SNOBAL_HMETS (n = 127, mountain west and Canadian prairie) and SNOBAL_HBV (n = 94, mainly mountainous and northern interior) cover the remaining basins; their IQRs in Figure 4.4e appear modestly higher (orange-tinted, mostly 0.10–0.15), but the violin plots in Figure 4.5e show all three snow algorithms have narrow, low IQR distributions overall. Snow balance is the process for which the data express the most decisive structural preferences continent-wide.

The right column of Figure 4.5 adds to this by showing the rescaled distributions of the basin attributes most strongly associated with each process group's IQR across the 2,869-watershed sample. Three patterns recur across panels and appear to link to the left-column IQR distributions. First, crop fraction is markedly bimodal in panels (a), (b), (c), and (e), with peaks near 0 (forested or rangeland basins) and near 1 (heavily cropped midwestern lowlands); this bimodality is the structural mechanism behind crop fraction's high explanatory power for INF_HBV, BASE_VIC quickflow, and SOILEVAP_ALL IQR, cropland status sharply differentiates which structural representations the data can support. Second, the topographic and morphological predictors (slope, elevation,

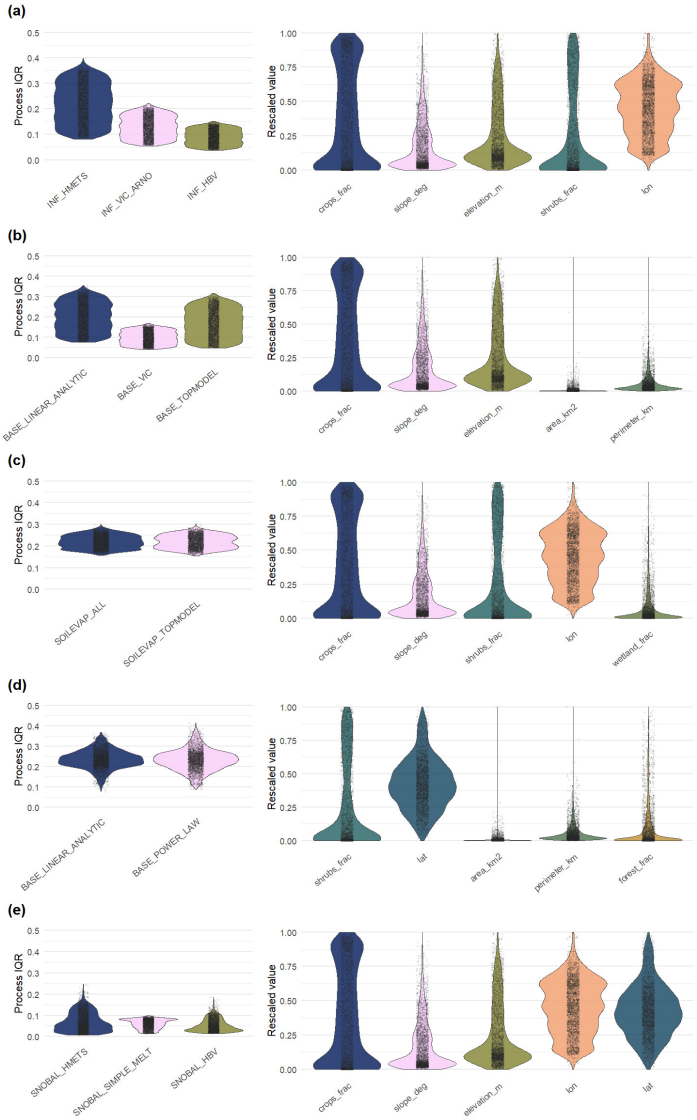


Figure 4.5: Algorithm-level identifiability and its landscape associations across 2,869 watersheds. Left: violin plots of the posterior weight IQR for each algorithm in (a) infiltration, (b) quickflow, (c) evapotranspiration, (d) deep baseflow, (e) snow balance. Lower IQR indicates more decisive structural preference. Right: rescaled distributions of the basin attributes most strongly associated with the IQR for each process group.

drainage area, perimeter, shrub fraction, forest fraction) are heavily right-skewed in every panel, indicating that the bulk of the 2,869 watersheds are flat, low-elevation, small, and lightly vegetated, with long upper tails of mountainous and extensive basins. The algorithms whose left-column IQR distributions extend higher, INF_HMETS in panel (a), BASE_LINEAR_ANALYTIC and BASE_TOPMODEL in panel (b), SNOBAL_HMETS and SNOBAL_HBV in panel (e), draw their dominant basins disproportionately from these upper tails, so the right-skew of the landscape projects directly onto an upward-shifted IQR. Third, the geographic predictors (longitude in panels (a), (c), and (e); latitude in panels (d) and (e)) are centred near mid-continent, indicating that variability in identifiability is organized along continental-scale climate gradients rather than concentrated in localized hot spots. Read together, the two columns of Figure 4.5 indicate that the shape of an algorithm's IQR distribution is a near-direct projection of where its dominant-basin set sits within these continent-wide attribute distributions. Tight, narrow IQR violins (BASE_VIC quickflow, SNOBAL_SIMPLE_MELT) correspond to algorithms whose dominant basins cluster densely in the bulk of the predictor distributions, while broad IQR violins (INF_HMETS) correspond to algorithms whose dominant basins occupy the sparse upper tails.

Taken together, these patterns demonstrate that model structure is not a free choice. The landscape strongly influences both which structures are appropriate (Clark et al., 2011) and how confidently the data can support each choice. One practical implication is that processes with persistently high IQR across the continent, most notably deep baseflow, are intrinsically poorly identifiable from daily streamflow alone, regardless of landscape, so improving identifiability for these processes would require data of a different kind, such as groundwater levels or storage anomalies. To probe which landscape attributes are most strongly associated with the IQRs themselves, i.e., with structural identifiability rather than algorithmic preference per se, we computed the R^2 between each algorithm's IQR and 16 landscape attributes (Fig. 4.6). The strongest associations are physically intuitive. Topographically driven processes show the tightest links. HBV infiltration's IQR is most strongly explained by shrub fraction ($R^2 = 0.73$), elevation (0.59), longitude (0.43), and crop fraction (0.46); SNOBAL_HMETS IQR by elevation (0.68); SNOBAL_HBV IQR by elevation (0.51) and longitude (0.44). Land cover and basin size dominate elsewhere. SOILEVAP_ALL IQR is most tied to crop fraction (0.63), and BASE_VIC quickflow IQR to drainage area (0.72). SNOBAL_SIMPLE_MELT IQR is overwhelmingly explained by latitude (0.71), reflecting the fundamentally latitude-driven nature of snowmelt regimes. In every case, the strength of structural preference, not just which algorithm is preferred, is systematically conditioned by the landscape. Reading Figure 4.6 by column complements this row-by-row picture. Elevation, crop fraction, slope, shrub fraction, longitude, latitude,

and drainage area form the warm columns, each containing multiple $R^2 > 0.4$ entries spanning several process groups, while aspect, Gravelius compactness, water fraction, snow/ice fraction, urban fraction, grass fraction, and wetland fraction are uniformly cool (R^2 mostly below 0.10), contributing little to identifiability despite many cells being statistically significant under Benjamini–Hochberg correction. Overall, a small set of climate and terrain attributes captures the bulk of structural identifiability continent-wide. This is encouraging for practical regionalization, since identifiability at ungauged basins can in principle be approximated from a small number of physically interpretable landscape variables.

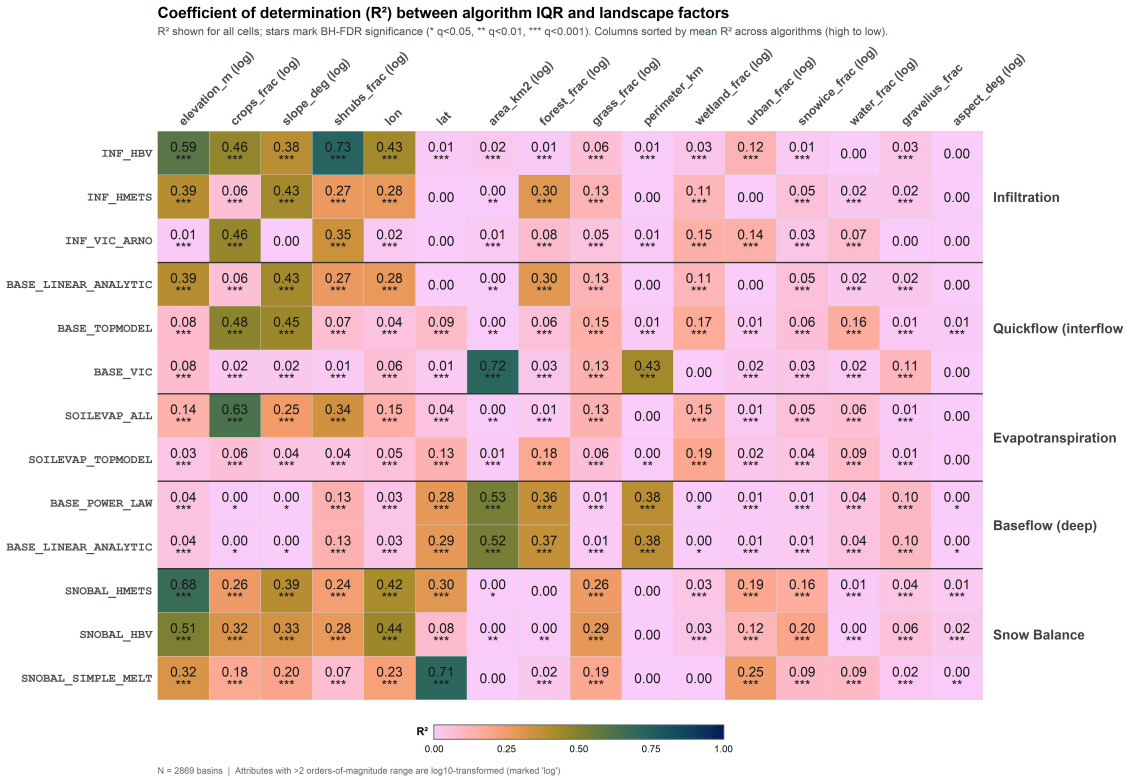


Figure 4.6: Correlogram of R^2 between the IQR of each posterior algorithm weight and 16 landscape attributes (n = 2,869). Colour: R^2 (pink = 0, dark blue = 1). Stars indicate significance level (* $q < 0.05$, ** $q < 0.01$, *** $q < 0.001$). Attributes with > 2 orders-of-magnitude range are log-transformed.

4.2.5 Seasonal expansion during summer–fall recession

The probabilistic atlas indicates where structural uncertainty originates in space and which algorithms drive it. These are, however, static annual-mean snapshots. Hydrologic systems cycle through wet and dry seasons, and the uncertainty budget cycles with them. We therefore ask whether structural uncertainty is stable through the year or whether it intensifies during particular regimes. To answer this, we present three representative gauges spanning distinct hydroclimatic regimes (Fig. 4.7a–c), together with a continental synthesis: a map of the ratio of summer–fall to winter–spring structural variance share across all 2,869 basins (Fig. 4.7d).

The three gauges show regime-dependent behavior. At gauge 01010000 (St. John River at Ninemile Bridge, Maine), a snow-influenced eastern catchment, structural uncertainty is about 8–12% in winter and expands four- to fivefold to 35–50% during summer–fall recession (June–November) (Fig. 4.7a). At gauge 01668000 (Rappahannock River near Fredericksburg, Virginia), a humid eastern catchment without snow influence, the expansion is even more pronounced, with structural uncertainty climbing from a winter baseline of 15–20% to 55–65% during summer recession (Fig. 4.7c). By contrast, at gauge 07186000 (Spring River near Waco, Missouri), a semi-arid central catchment, structural uncertainty remains stable at 10–20% throughout the year, with forcing uncertainty dominating (Fig. 4.7b). The shared feature of the two expanding gauges is not snow but a long, sustained recession period; the semi-arid gauge's flashy hydrograph offers no such recession window. The continental map (Fig. 4.7d) confirms this regime dependence systematically across all 2,869 basins. Ratios of 4–6× (dark blue) cluster across the Pacific Northwest, the western mountains, and snow-influenced parts of the northern continental interior, with scattered occurrences in the humid eastern uplands. Ratios near 1–2 (pink) occupy the southern continental interior, the lower Mississippi, Texas, the Gulf Coast, where structural uncertainty is comparatively stationary year-round. The picture parallels the spatial distribution of structural-share dominance in Figure 4.2: regions with seasonal flow regimes show the largest temporal expansion, while regions with uniform or flashy regimes show muted seasonal change.

The physical mechanism behind this regime dependence is the flow-regime sensitivity of algorithmic divergence. During high-flow periods, the streamflow signal is large enough to constrain which structural configurations are consistent with observations, since all competing algorithms must route substantial water volumes through the system, and their outputs converge. During recession, however, the fast-response signal disappears and slow-response processes take over. Linear-analytic baseflow drains storage exponentially toward zero; power-law baseflow sustains discharge through a longer-tailed nonlinear reces-

sion; VIC interflow generates threshold-driven pulses; TOPMODEL produces continuous saturated-area contributions. These differences are negligible when discharge is high but diverge sharply as flows recede, amplifying structural spread. At semi-arid basins where recessions are short and flashy, this mechanism has insufficient time to operate, explaining the muted seasonal expansion at gauge 07186000 and across the central plains in Figure 4.7d. Forcing uncertainty (yellow) typically rises in late winter as flow increases ($\sim$ 15–20%), briefly exceeds structural uncertainty as the secondary source, and then contracts during the spring peak despite high discharge, suggesting that posterior forcing parameters are well constrained by the strong peak-flow signal. Parameter uncertainty (orange) provides a seasonally stable backdrop ($\sim$ 35–65%), consistent with parameter non-uniqueness (Beven, 2006): the observational record constrains seasonal mean behavior but fails to uniquely resolve individual parameter combinations.

4.3 Discussion

Across 2,869 North American watersheds, structural uncertainty is the largest contributor to hydrologic predictive spread, with a continental median of 42% of total variance compared to 36% for parameter uncertainty and 19% for forcing uncertainty. This finding challenges the common assumption, built into most uncertainty analysis workflows, that parameter uncertainty will always dominate hydrologic predictions (Beven, 2009). The dominance of structural uncertainty means that, on average, how we represent hydrologic processes matters more for prediction spread than how we determine parameters or measure inputs.

This continental average, however, masks important regional variation with direct practical implications. In the humid, energy-limited northeast, where equifinality is most severe, parameter uncertainty dominates because persistent runoff generation constrains forcing but leaves the process parameter space under-determined (Beven, 2006). Across the central continental interior, particularly the Great Plains, and into the topographically complex west, the opposite holds. Climate variability and diverse terrain/geology activate competing structural representations, and the landscape's failure to impose a single dominant runoff pathway amplifies structural divergence. Previous decomposition studies, limited to smaller sets of watersheds (Chlumsky et al., 2021, 2024; Salamon and Feyen, 2010), could not resolve whether this spatial pattern was real or an artefact of site selection; our 2,869-basin sample establishes it as a robust, continental-scale phenomenon (Gupta et al., 2014b). The scaled CRPS map (Fig. 4.1) provides independent corroboration. While the median performance is satisfactory (scaled CRPS 0.25 across all 2,869 watersheds), the

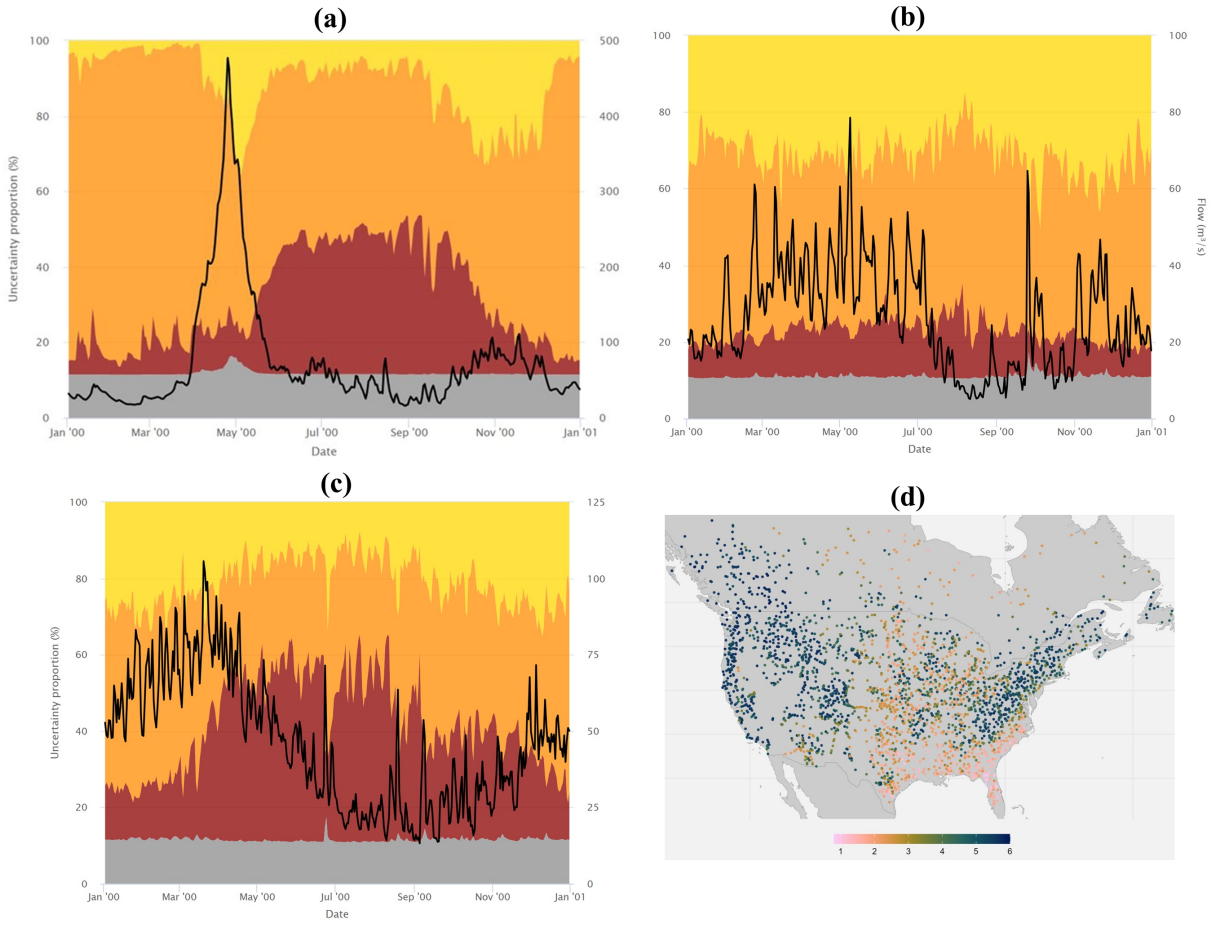


Figure 4.7: Time-varying uncertainty decomposition at three representative gauges and continental synthesis. (a) Gauge 01010000, St. John River at Ninemile Bridge, Maine (snow-influenced eastern). (b) Gauge 07186000, Spring River near Waco, Missouri (semi-arid central). (c) Gauge 01668000, Rappahannock River near Fredericksburg, Virginia (humid eastern). Stacked area: forcing (yellow), parameter (orange), structural (dark red), unexplained (grey). Black line: mean daily streamflow (right axis). (d) Continental map of the ratio of summer–fall (June–November) to winter–spring (December–May) structural variance share across all 2,869 basins; values 1–2 (pink) indicate near-stationary structural uncertainty, values 4–6 (dark blue) indicate four- to sixfold seasonal expansion.

weakest probabilistic performance (upper quartile, scaled CRPS > 0.31) concentrates in the regions where structural uncertainty dominates. The overlap makes physical sense, since where process representations diverge, the posterior predictive distribution widens and model skill degrades. It also implies that targeted structural improvement could both reduce uncertainty and improve prediction skill. The practical conclusion is that uncertainty reduction must be landscape-specific, with additional observations most valuable where parameter non-uniqueness dominates and improved model structures most valuable where the landscape is topographically complex. This aligns with the data-worth perspective of Renard et al. (2011) and Gupta et al. (2012).

These spatial patterns raise the question of which structural choices are responsible for the divergence, and how confidently those choices can be identified. The probabilistic framing in Figure 4.4, dominant algorithm at each location plus the IQR of its posterior weight, extends the sensitivity-based mapping of Mai et al. (2022), who identified quickflow as the most sensitive process and showed that process sensitivity can be predicted from basin attributes. Our results add two probabilistic dimensions. First, identifiability varies sharply by process. The data express decisive, low-IQR preferences for snow-balance algorithms (SNOBAL_SIMPLE_MELT preferred at 92% of basins with IQR mostly below 0.10) and for the dominant infiltration choice in most basins (VIC–Arno at 70%, with moderate IQR 0.10–0.20), but persistently high IQR continent-wide for deep baseflow regardless of which formulation is preferred. Second, where ambiguity persists, it is process-specific and physically interpretable. HMETS infiltration is preferred only rarely and with the broadest IQR distribution of the three infiltration options, and quickflow choices outside cropland-dominated lowlands carry residual posterior uncertainty. The strong landscape associations in Figure 4.6, with R^2 values up to 0.73 between IQR and basin attributes, demonstrate that the strength of structural preference, not just which algorithm is preferred, is systematically conditioned by the landscape. Together with the climate-driven predictability of structural variance share ($R^2 = 0.66$ from precipitation and temperature alone), this implies that, pending out-of-sample cross-validation, both the dominant uncertainty source and its identifiability at ungauged sites may be approximated from landscape attributes. The clearest practical conclusion concerns deep baseflow. Streamflow-driven calibration alone cannot resolve baseflow structural uncertainty, and reducing it would require data of a different kind, such as groundwater levels or storage anomalies.

The temporal decomposition adds a dimension that prior studies have not addressed at this scale. Clark et al. (2011) and Fenicia et al. (2008) showed that structural errors are regime-dependent, and Troin et al. (2022), using 180 discrete structures across 698 catchments, found that structural uncertainty varies across climate zones and flow metrics. Our results extend these findings temporally. At basins with prolonged summer

recessions, whether snow-influenced (gauge 01010000, Maine) or humid-eastern without snow (gauge 01668000, Virginia), structural uncertainty may expand four- to fivefold during summer–fall recession (Fig. 4.7a, c). At the semi-arid Spring River gauge (07186000, Missouri), seasonal expansion is muted (Fig. 4.7b). The continental synthesis (Fig. 4.7d) confirms that the expansion factor varies regionally. Ratios of summer–fall to winter–spring structural variance reach 4–6 across the Pacific Northwest, western mountains, and snow-influenced northern interior, with scattered occurrences in humid eastern uplands, and remain near 1–2 across the southern continental interior. The shared property of high-expansion basins is not snow but a sustained recession period during which slow-response algorithms can diverge; the shared property of low-expansion basins is a flashy or short recession that gives those algorithms little time to differentiate. In snow-influenced basins, the lag between snowfall, snowpack storage, melt, and runoff means that forcing errors are buffered and released through the snowpack, so part of what is attributed to structure during the recession may originate in forcing timing. The decomposition does not separate this lagged forcing contribution explicitly, which is a limitation in snow-dominated regimes. The operational consequence is that static uncertainty envelopes systematically understate structural uncertainty during summer recession in regions with sustained recession regimes, and overstate it during winter, errors most pronounced in the snow-influenced and humid-eastern domains. Ensemble systems propagating the full posterior could deliver seasonally appropriate estimates.

These findings, while supported by continental spatial evidence and detailed temporal evidence at three representative gauges with continental synthesis, are subject to methodological constraints that warrant note. The Gaussian homoscedastic likelihood provides a tractable starting point for inference (Schoups and Vrugt, 2010; Vrugt et al., 2008), but cannot adequately represent arid-region streamflow distributions (Hassan et al., 2022; Markstrom et al., 2016). Correlated structural errors may be absorbed into the parameter component, potentially underestimating structural uncertainty where it already dominates. Future work could address this with skew-normal, log-normal, or censored Gaussian likelihoods on a validation subset of arid catchments. Additionally, the continental structural-uncertainty share of 42% is conditional on two methodological choices, namely the uniform prior on weight-generating parameters after pie-sharing reparameterization and the 13 specific algorithmic options currently embedded in the blended model. Wider option sets or informative priors could yield different absolute magnitudes, though the relative spatial and seasonal patterns will likely remain. In particular, the blended model can express structural uncertainty only within the span of its embedded process options while the structural errors shared by all options are invisible to the decomposition and are absorbed by the parameter or residual components. Dedicated prior-sensitivity analysis remains a priority for

future work. We expect the prior to influence the absolute magnitude of the structural share more than its spatial and seasonal organization, since the latter is governed by hydroclimatic gradients that do not depend on the prior. The MANOVA linearity assumption is broadly satisfied (median residual $\sim 4\%$) but becomes less adequate in threshold-driven regimes where nonlinear attribution methods (Saltelli et al., 2008) may sharpen the decomposition. While Figure 4.7d already provides a continental view of seasonal expansion ratios, deeper archetype analysis, clustering basins by their seasonal fingerprint and identifying additional drivers of expansion magnitude beyond recession length, would be a useful extension. Benchmarking the blended model's structural uncertainty estimates against a traditional discrete multi-model ensemble on representative basins would also be useful, providing external validation that the continuously weighted and discrete formulations converge on comparable structural uncertainty magnitudes. Testing whether the uncertainty hierarchy shifts under climate change is a question of growing practical urgency.

Despite these limitations, the results offer continental-scale evidence for how hydrologic uncertainty should be understood and reduced. If structural uncertainty really does dominate the continental budget, the operative question shifts from how to better constrain or calibrate parameters to how to select the right model structure, and how confidently the data can support that choice. The probabilistic process-weight atlas (Fig. 4.4) provides a direct answer by identifying both which algorithms the data prefer and where those preferences are confident. This framework could serve as the basis for regionalized model selection in data-sparse regions (Samaniego et al., 2010), following the spirit of parameter transfer functions that link model behavior to landscape properties. The interpretable, process-based decomposition also provides information that purely data-driven approaches, such as Long Short-Term Memory (LSTM) networks, cannot currently provide. While LSTMs have demonstrated strong predictive skill at continental scale (Kratzert et al., 2019), they cannot attribute predictive spread to specific physical processes or guide targeted model improvement. Incorporating multi-variable observational constraints, satellite-derived evapotranspiration, snow water equivalent, or soil moisture, would further constrain structural and forcing parameters simultaneously (Dembelé et al., 2020; Széles et al., 2020). The result is an interpretable, landscape-specific, and seasonally resolved diagnostic that conveys the magnitude of predictive uncertainty together with its dominant source, the strength of structural support for the preferred algorithm, and the most effective route to reducing it.

4.4 Methods

We assemble a continental-scale Bayesian uncertainty analysis over 2,869 North American watersheds drawn from the HYSETS database. The methodology is thoroughly defined in Chapter 3. For each watershed, the blended hydrologic model (Chlumsky et al., 2021, 2024; Mai et al., 2020) represents structural uncertainty by computing process fluxes as weighted averages of multiple algorithmic options, while a stochastic perturbation scheme on precipitation and temperature represents forcing uncertainty. The full parameter vector (structural weights, hydrologic process parameters, and forcing perturbation parameters) is inferred jointly using the No-U-Turn Sampler (NUTS) (Hoffman and Gelman, 2014). Posterior ensembles are decomposed into structural, parameter, and forcing components via multivariate analysis of variance, evaluated probabilistically using the continuous ranked probability score, and related to catchment characteristics through correlations between posterior process weights and landscape attributes.

4.4.1 Study domain and data

The analysis is based on the HYSETS database (Arsenault et al., 2020), comprising 14,425 watersheds across North America with daily precipitation, minimum and maximum temperature, and observed streamflow. Following the screening procedure of Mai et al. (2022), we retained only watersheds with drainage areas greater than 300 km^2, more than five years of continuous daily streamflow records during the calibration period (January 1991 to December 2010), and a preliminary deterministic calibration Nash–Sutcliffe efficiency (NSE) of 0.5 or higher. The preliminary calibration used the Dynamically Dimensioned Search (DDS) algorithm (Tolson and Shoemaker, 2007) maximizing NSE with a budget of 1,000 model evaluations, following similar setup of Mai et al. (2022). Watersheds failing to achieve NSE $\geq$ 0.5 were excluded on the grounds that models unable to reproduce observed streamflow dynamics at this threshold cannot support meaningful uncertainty decomposition. Of the 3,316 watersheds analyzed by Mai et al. (2022), 447 did not meet our additional Bayesian convergence criteria (Gelman–Rubin $\hat{R} < 1.1$ for all parameters within 50,000 samples across three chains) and were excluded, yielding a final sample of 2,869 watersheds. This sample spans the full North American hydroclimatic gradient: drainage areas of $\sim$ 300–50,000 km^2, mean elevations from sea level to over 3,000 m, and mean annual precipitation from below 300 to over 3,000 mm yr^{-1}.

Catchment attributes were compiled from multiple sources. These include seven land cover fractions (crop, forest, grass, shrub, urban, water, wetland), six topographic indices

(elevation, slope, aspect, drainage area, Gravelius compactness, perimeter), geographic coordinates, the snow/ice fraction, and three climate indicators (Knoben et al., 2018) (aridity, seasonality, fraction of precipitation as snow), extracted from HYSETS and supplemented from the CAMELS attribute set (Addor et al., 2017) where watersheds overlapped.

4.4.2 Blended hydrologic model and forcing perturbation

The blended model (Mai et al., 2020) addresses structural uncertainty by computing each process flux as a weighted average of multiple algorithmic options. Rather than committing a priori to a single algorithm for, say, infiltration or snowmelt, the model evaluates several established formulations in parallel and combines their outputs through non-negative weights that sum to unity. The constraint guarantees that the blended flux remains bounded by the individual algorithm outputs, preserving physical consistency, while the weights themselves become inferred quantities, turning the choice of process representation into a learnable feature rather than a fixed modeling decision. This weighting strategy is related in spirit to automatic relevance determination in Bayesian machine learning (Neal, 1996), in which continuous hyperparameters govern the effective contribution of competing model components. The key differences here are that the weights act on physically interpretable process algorithms, are constrained to the simplex, and are inferred jointly with the process and forcing parameters rather than treated as tunable hyperparameters.

The model is implemented within the Raven hydrologic modeling framework (Craig et al., 2020) and encompasses five process groups summarized in Table 4.1. This yields 13 process algorithms governed by 8 weight-generating parameters, plus 35 hydrologic process parameters and 6 forcing perturbation parameters, for a total of 49 inferred parameters per watershed. The model setup is identical to that used by Mai et al. (2020, 2022) and Chlumsky et al. (2024).

Meteorological forcing uncertainty was represented through a stochastic perturbation scheme applied independently to precipitation, minimum temperature, and maximum temperature (McMillan et al., 2011; Tang et al., 2023). Each variable is governed by two parameters, a scale parameter controlling perturbation magnitude and a bias parameter adjusting systematic offset, yielding six forcing parameters in total. Precipitation was perturbed multiplicatively using a log-normal formulation that preserves its non-negative, right-skewed distribution; temperature was perturbed additively, consistent with the approximately symmetric distribution of temperature anomalies.

Inter-variable correlations across precipitation and temperature were preserved through joint sampling using the empirical Pearson correlation matrix estimated from standardized

Table 4.1: Process algorithms in the blended hydrologic model. Each process group contains multiple algorithm options whose weights are jointly inferred. Code identifiers are those used in the Raven framework configuration files.

Process group	Code identifier	Algorithm descriptive name
Infiltration	INF_HMETS	HMETS empirical infiltration
	INF_VIC_ARNO	VIC/ARNO saturation-excess scheme
	INF_HBV	HBV soil-moisture-based infiltration
Quickflow / interflow	BASE_LINEAR_ANALYTIC	Linear reservoir (analytic solution)
	BASE_VIC	VIC nonlinear baseflow (ARNO)
	BASE_TOPMODEL	TOPMODEL exponential subsurface flow
Evapotranspiration	SOILEVAP_ALL	ET drawn uniformly from full soil column
	SOILEVAP_TOPMODEL	ET weighted by topographic wetness index
Deep baseflow	BASE_LINEAR_ANALYTIC	Linear baseflow from deep store
	BASE_POWER_LAW	Nonlinear (power-law) baseflow
Snow balance	SNOBAL_HMETS	HMETS degree-day melt with refreezing
	SNOBAL_SIMPLE_MELT	Simple degree-day melt (no refreezing)
	SNOBAL_HBV	HBV degree-day melt with refreezing index

historical anomalies. Deterministic seeding based on the perturbation parameter values ensured that the same perturbation time series is reproduced for any given parameter configuration, which is essential during Bayesian inference when the sampler's rejection step may cause re-evaluation of identical parameter sets.

To enforce physical parameter bounds required by the Raven framework, a smooth transformation based on the hyperbolic tangent function (tanh) was applied to map unbounded NUTS proposals into the admissible parameter space while preserving the differentiability required for gradient-based sampling. Unlike rejection-based approaches, which introduce inefficiencies through repeated proposal rejection, this deterministic transformation ensures that every NUTS proposal maps to a valid parameter set. The total computational burden across all watersheds was $2{,}869 \times 3 \times 50{,}000 = 430.35$ million forward model simulations, performed on the Digital Research Alliance of Canada high-performance computing clusters.

4.4.3 Uncertainty decomposition and evaluation

For each watershed, 5,000 posterior parameter samples (drawn from the converged chains, which satisfy the Gelman–Rubin criterion, while keeping the continental-scale propagation

computationally feasible) were propagated through the blended model to generate an ensemble of simulated hydrographs over the simulation period (January 1991 to December 2010, following a two-year warm-up). The ensemble was reshaped into a three-dimensional tensor with dimensions [samples × years × days], enabling efficient variance decomposition for each Julian calendar day independently. Because each of the 49 parameters is pre-assigned to one of three functional groups based on its role in the model (structural weights, hydrologic process parameters, and forcing perturbation parameters), the total spread in simulated streamflow can be attributed to these sources via a multivariate analysis of variance (MANOVA).

The continuous ranked probability score (CRPS) was computed for each watershed as the mean over all daily timesteps of the integrated squared difference between the posterior predictive cumulative distribution function and the observation (Hersbach, 2000). To enable comparison across watersheds with different flow magnitudes, the CRPS was divided by the mean observed streamflow at each site, yielding a dimensionless scaled CRPS. Lower values indicate better probabilistic prediction skill.

To characterize structural identifiability, we computed at each watershed the IQR of the posterior weight distribution for each algorithm. The dominant algorithm at a basin was identified as the one with the maximum posterior weight in its process group; for each algorithm, we then mapped the IQR distribution across all basins where it was the dominant choice (Fig. 4.4), and quantified the relationship between IQR and 16 basin attributes through Pearson R^2 (Fig. 4.6). Attributes spanning more than two orders of magnitude (drainage area, elevation, slope, and all land cover fractions) were log-transformed. Statistical significance was assessed using the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995) to control the false discovery rate at $q < 0.05$ across all $16 \times 13 = 208$ simultaneous hypothesis tests. All statistical tests were two-tailed. Sample size: $n = 2{,}869$ unless otherwise stated.

Data availability

The HYSETS database is publicly available at `https://doi.org/10.17605/OSF.IO/RPC3W`. The blended model setups and xSSA results from Mai et al. (2022) are at `https://doi.org/10.5281/zenodo.5730428`.

Code availability

The Raven hydrological modeling framework is open-source at https://raven.uwaterloo.ca/. Bayesian inference was performed using the LaplacesDemon R package (LaplacesDemon Core Team, 2023). Custom R scripts for NUTS inference and MANOVA decomposition will be deposited in a permanent repository upon publication.

Funding

This work was funded by Natural Resources Canada's Floodplain Hazard and Mapping Program grant (FHIMP23-27P2), NSERC Discovery grants held by co-authors Craig (RGPIN-2017-03920; RGPIN-2023-03459) and Tolson (RGPIN-2022-03890), and by funding from co-author Craig's NSERC Tier II Canada Research Chair in Hydrologic Modeling and Analysis (CRC-2020-00176).

Chapter 5

Conclusions

5.1 Summary

A long-standing working assumption among hydrologic modellers has been that parameter uncertainty dominates the predictive spread of process-based models. The results of this thesis indicate that this assumption does not hold in general. Across 2,869 North American watersheds, structural uncertainty, the imperfection of the model structure itself, accounts for the largest share of predictive variance, exceeding both parameter and forcing uncertainty at the continental median. Reaching that finding required two methodological advances, an efficient inference framework (Chapter 2) and a well-posed decomposition (Chapter 3), that the continental analysis of Chapter 4 then put to use. The four objectives of the thesis followed directly from two practical constraints on Bayesian uncertainty analysis, the computational cost of joint inference and the lumped form of its predictive output, and the chapters address them in turn.

The first objective was to exploit the conditional structure between hydrologic and error model parameters and to develop a machine-learning-assisted Bayesian framework that would materially reduce the sampling cost of joint inference. Chapter 2 addresses this objective. The Surrogate-aided Heteroscedastic Autocorrelated (SHA) error model trains a support vector regression surrogate offline on roughly one thousand hydrologic parameter realizations and deploys it inside the No-U-Turn Sampler to predict error model parameters at every MCMC step rather than sampling them. The result is a roughly fifty percent reduction in the number of samples and the wall-clock time required for convergence, with no loss in probabilistic skill. Computational efficiency at this point in

the inference pipeline is what makes previously infeasible analyses practical, including the continental-scale application of Chapter 4.

The second objective was to formulate structural, parameter, and forcing uncertainty as components of a single well-posed Bayesian inference and to develop a variance decomposition that resolves their contributions across the hydrological year. Chapter 3 addresses this objective. By encoding structural uncertainty as parameter uncertainty in the weight-generating coefficients of a blended hydrologic model, the framework circumvents the ill-posed simultaneous estimation of separate input and structural error models that has constrained prior decomposition attempts. A multivariate analysis of variance applied to the posterior ensemble for every Julian calendar day yields a continuous time-varying budget. On the MOPEX12 benchmark sample, parameter uncertainty dominates the annual mean budget (median $\approx 53\%$), but structural uncertainty triples during summer–fall recession when competing baseflow algorithms diverge most strongly. These results show that predictive uncertainty is not a static property of a model but varies seasonally with flow regime.

The third objective was to deploy the combined framework at continental scale and to translate the resulting uncertainty budgets into practical guidance. Chapter 4 addresses both. Across 2,869 North American watersheds, structural uncertainty (continental median 42% of predictive variance) exceeds parameter uncertainty (36%) and forcing uncertainty (19%), in contrast to the parameter-dominated hierarchy found for the smaller MOPEX12 sample in Chapter 3. The dominant source varies regionally in a pattern that is highly predictable from climate alone, with roughly two-thirds of the spatial variance in structural share explained by mean annual precipitation, precipitation variability, and temperature seasonality, while topographic factors explain comparatively little. Structural uncertainty also expands four- to six-fold during summer–fall recession at basins with prolonged recessions, while remaining muted at semi-arid basins with flashy hydrographs. In practical terms, additional streamflow observations yield the largest gains in humid, parameter-dominated catchments. Model and process refinement yields more than additional gauges in structurally dominated regions. Improved precipitation data is most effective in forested or orographic catchments where forcing dominates. Because the dominant source is predictable from hydroclimatic attributes alone, this guidance can in principle be extended to ungauged basins.

5.2 Contributions

The principal contributions of this thesis can be summarized as follows.

- A surrogate-aided Bayesian calibration framework that exploits the conditional dependence of error model parameters on hydrologic parameters to halve the MCMC sampling cost of joint inference, with preserved or improved probabilistic skill (Chapter 2).
- A unified Bayesian framework that, by reformulating structural uncertainty as uncertainty in blended-model weight-generating parameters, jointly infers all three uncertainty sources in a single well-posed problem and produces, via per-Julian-day MANOVA, what is to the author's knowledge the first continuous seasonal budget of hydrologic uncertainty reported in the literature (Chapter 3).
- The largest multi-source uncertainty decomposition study to date, spanning 2,869 North American watersheds, two orders of magnitude larger than any previously published sample, made tractable by the efficiency gains of the surrogate-aided framework (Chapter 4).
- Empirical evidence that structural uncertainty (not parameter uncertainty) dominates the continental hydrologic uncertainty budget, that climate alone explains roughly two-thirds of its spatial variation, and that its seasonal expansion is governed by recession regime, findings that run counter to a long-standing default assumption of the field (Chapter 4).

5.3 Limitations and Future Research

Several limitations bound these conclusions. The decomposition relies on a linear MANOVA approximation between posterior parameters and simulated streamflow. In nonlinear, threshold-driven regimes such as semi-arid catchments, nonlinear attribution methods (random-forest importance, neural-network surrogates) may yield sharper decompositions. The Gaussian homoscedastic likelihood is similarly restrictive in flashy regimes, and skew-normal, log-normal, and censored Gaussian alternatives merit systematic evaluation. The analyses are also confined to North America and to lumped configurations. Extension to global samples and to semi-distributed setups would test whether the climate-based predictability of the budget generalizes. Three extensions stand out. The first is integration with climate-change scenarios to test how the budget shifts as regimes evolve. The second is the use of multi-variable observational constraints (evapotranspiration, snow water equivalent, soil moisture, gravimetric storage) to constrain the three parameter blocks

simultaneously. The third is the coupling of the surrogate-aided framework with deep-learning hydrologic models, where the conditional decomposition logic developed here could be transferred to data-driven model classes.

By making Bayesian uncertainty analysis efficient enough for continental application and interpretable enough to attribute predictive spread to its underlying sources, this work links uncertainty quantification to the practical question of where to invest in model improvement. For a given watershed, the framework indicates, by climate, by season, and by process, which source dominates predictive uncertainty and which intervention would most effectively reduce it.

References

N. Addor, A. J. Newman, N. Mizukami, and M. P. Clark. The CAMELS data set: catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, 21(10):5293–5313, 2017. doi: 10.5194/hess-21-5293-2017.

M. I. Ahmed, K. Shook, A. Pietroniro, T. Stadnyk, J. W. Pomeroy, C. Pers, and D. Gustafsson. Implementing a parsimonious variable contributing area algorithm for the prairie pothole region in the HYPE modelling framework. *Environmental Modelling & Software*, 167:105769, 2023.

J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

N. K. Ajami, Q. Duan, and S. Sorooshian. An integrated hydrologic Bayesian multimodel combination framework: confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resources Research*, 43(1):W01403, 2007. doi: 10.1029/2005WR004745.

Rezgar Arabzadeh, Jonathan Romero-Cuellar, James R. Craig, Bryan A. Tolson, and Robert Chlumsky. A surrogate-aided approach for accelerated Bayesian calibration of hydrologic models. *Environmental Modelling and Software*, 198:106894, 2026. doi: 10.1016/j.envsoft.2026.106894.

R. Arsenault, F. Brissette, J.-L. Martel, M. Troin, G. Lévesque, J. Davidson-Chaput, M. Castañeda Gonzalez, A. Ameli, and A. Poulin. A comprehensive, multisource database for hydrometeorological modeling of 14,425 North American watersheds. *Scientific Data*, 7:243, 2020. doi: 10.1038/s41597-020-00583-2.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

S. Bergström and A. Forsman. Development of a conceptual deterministic rainfall-runoff model. *Nordic Hydrology*, 4:240–253, 1973.

K. Beven. A manifesto for the equifinality thesis. *Journal of Hydrology*, 320(1–2):18–36, 2006. doi: 10.1016/j.jhydrol.2005.07.007.

K. Beven. *Environmental Modelling: An Uncertain Future?* Routledge, 2009.

K. Beven. *Rainfall-Runoff Modelling: The Primer*. Wiley-Blackwell, 2nd edition, 2012.

K. Beven and J. Freer. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of Hydrology*, 249(1–4):11–29, 2001. doi: 10.1016/S0022-1694(01)00421-8.

Keith J. Beven and Andrew M. Binley. The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, 6(3):279–298, 1992. doi: 10.1002/hyp.3360060305.

Keith J. Beven, P. J. Smith, and J. E. Freer. So just why would a modeller choose to be incoherent? *Journal of Hydrology*, 354(1–4):15–32, 2008. doi: 10.1016/j.jhydrol.2008.02.007.

G. Blöschl, M. F. Bierkens, A. Chambel, C. Cudennec, G. Destouni, A. Fiori, et al. Twenty-three unsolved problems in hydrology (UPH)—a community perspective. *Hydrological Sciences Journal*, 64(10):1141–1158, 2019. doi: 10.1080/02626667.2019.1620507.

R. Chlumsky, J. Mai, J. R. Craig, and B. A. Tolson. Simultaneous calibration of hydrologic model structure and parameters using a blended model. *Water Resources Research*, 57(5):e2020WR029229, 2021. doi: 10.1029/2020WR029229.

R. Chlumsky, J. Mai, J. R. Craig, and B. A. Tolson. Advancement of a blended hydrologic model for robust model performance. *Journal of Hydrologic Engineering*, 29(5):04024033, 2024. doi: 10.1061/JHYEFF.HEENG-6246.

Martyn P. Clark, Dmitri Kavetski, and Fabrizio Fenicia. Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, 47(9):W09301, 2011. doi: 10.1029/2010WR009827.

Martyn P. Clark et al. A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resources Research*, 51(4):2498–2514, 2015. doi: 10.1002/2015WR017198.

Martyn P. Clark et al. The evolution of process-based hydrologic models: Historical challenges and the collective quest for physical realism. *Hydrology and Earth System Sciences*, 21(8):3427–3440, 2017. doi: 10.5194/hess-21-3427-2017.

James R. Craig et al. Flexible watershed simulation with the Raven hydrological modelling framework. *Environmental Modelling and Software*, 129:104728, 2020. doi: 10.1016/j.envsoft.2020.104728.

D. Del Giudice, C. Albert, J. Rieckermann, and P. Reichert. Describing the catchment-averaged precipitation as a stochastic process improves parameter and input estimation. *Water Resources Research*, 52(4):3162–3186, 2016. doi: 10.1002/2015WR017871.

M. Dembelé, N. Ceperley, S. J. Zwart, E. Salvadore, G. Mariethoz, and B. Schaefli. Potential of satellite and reanalysis evaporation datasets for hydrological modelling under various model calibration strategies. *Advances in Water Resources*, 143:103667, 2020. doi: 10.1016/j.advwatres.2020.103667.

Q. Duan et al. Model parameter estimation experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. *Journal of Hydrology*, 320(1–2):3–17, 2006. doi: 10.1016/j.jhydrol.2005.07.031.

G. Evin, D. Kavetski, M. Thyer, and G. Kuczera. Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration. *Water Resources Research*, 49(7):4518–4524, 2013. doi: 10.1002/wrcr.20284.

G. Evin, M. Thyer, D. Kavetski, D. McInerney, and G. Kuczera. Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity. *Water Resources Research*, 50(3):2350–2375, 2014.

F. Fenicia, H. H. Savenije, P. Matgen, and L. Pfister. A comparison of alternative multiobjective calibration strategies for hydrological modeling. *Water Resources Research*, 43(3), 2007. doi: 10.1029/2006WR005098.

F. Fenicia, H. H. G. Savenije, P. Matgen, and L. Pfister. Understanding catchment behavior through stepwise model concept improvement. *Water Resources Research*, 44(1): W01402, 2008. doi: 10.1029/2006WR005563.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992. doi: 10.1214/ss/1177011136.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2 edition, 2004. doi: 10.1201/9780429258480.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. CRC Press, 3 edition, 2013. doi: 10.1201/b16018.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, 69(2): 243–268, 2007.

H. V. Gupta, S. Sorooshian, and P. O. Yapo. Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information. *Water Resources Research*, 34(4):751–763, 1998. doi: 10.1029/97WR03495.

H. V. Gupta, H. Kling, K. K. Yilmaz, and G. F. Martinez. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. *Journal of Hydrology*, 377(1–2):80–91, 2009. doi: 10.1016/j.jhydrol.2009.08.003.

H. V. Gupta, M. P. Clark, J. A. Vrugt, G. Abramowitz, and M. Ye. Towards a comprehensive assessment of model structural adequacy. *Water Resources Research*, 48(8), 2012.

H. V. Gupta, H. Kling, K. K. Yilmaz, and G. F. Martinez. Towards a comprehensive assessment of model structural adequacy. *Water Resources Research*, 50(7):6248–6272, 2014a. doi: 10.1029/2011WR011044.

H. V. Gupta, C. Perrin, G. Blöschl, A. Montanari, R. Kumar, M. Clark, and V. Andréassian. Large-sample hydrology: A need to balance depth with breadth. *Hydrology and Earth System Sciences*, 18(2):463–477, 2014b. doi: 10.5194/hess-18-463-2014.

H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001. doi: 10.2307/3318737.

S. M. T. Hassan, G. R. Ghimire, and G. Bhatt. Uncertainty quantification in watershed hydrology: which method to use? *Journal of Hydrology*, 616:128703, 2022. doi: 10.1016/j.jhydrol.2022.128703.

H. Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000. doi: 10.1175/1520-0434(2000)015⟨0559:DOTCRP⟩2.0.CO;2.

M. D. Hoffman and A. Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

M. Hrachowitz, H. H. G. Savenije, G. Blöschl, J. J. McDonnell, M. Sivapalan, J. W. Pomeroy, et al. A decade of predictions in ungauged basins (PUB)—A review. *Hydrological Sciences Journal*, 58(6):1198–1255, 2013. doi: 10.1080/02626667.2013.803183.

X. Jiang, H. V. Gupta, Z. Liang, and B. Li. Toward improved probabilistic predictions for flood forecasts generated using deterministic models. *Water Resources Research*, 55 (11):9519–9543, 2019. doi: 10.1029/2019WR025477.

D. Kavetski, G. Kuczera, and S. W. Franks. Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water Resources Research*, 42(3):W03408, 2006. doi: 10.1029/2005WR004376.

S. Khatami, M. C. Peel, T. J. Peterson, and A. W. Western. Equifinality and flux mapping: a new approach to model evaluation and process representation under uncertainty. *Water Resources Research*, 55:8922–8941, 2019. doi: 10.1029/2018WR023750.

W. J. M. Knoben, R. A. Woods, and J. E. Freer. A quantitative hydrological climate classification evaluated with independent streamflow data. *Water Resources Research*, 54:5088–5109, 2018. doi: 10.1029/2018WR022913.

W. J. M. Knoben, J. E. Freer, and R. A. Woods. Technical note: inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, 23(10):4323–4331, 2019. doi: 10.5194/hess-23-4323-2019.

F. Kratzert, D. Klotz, G. Shalev, G. Klambauer, S. Hochreiter, and G. Nearing. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, 23:5089–5110, 2019. doi: 10.5194/hess-23-5089-2019.

F. Kratzert, G. Nearing, N. Addor, T. Erickson, M. Gauch, O. Gilon, et al. Caravan—A global community dataset for large-sample hydrology. *Scientific Data*, 10:61, 2023. doi: 10.1038/s41597-023-01975-w.

W. Krzanowski. *Principles of Multivariate Analysis*. Oxford University Press, 2000.

G. Kuczera and E. Parent. Monte Carlo assessment of parameter uncertainty in conceptual catchment models: The Metropolis algorithm. *Journal of Hydrology*, 211(1–4):69–85, 1998. doi: 10.1016/S0022-1694(98)00198-X.

G. Kuczera et al. A limited-memory acceleration strategy for MCMC sampling. *Water Resources Research*, 46(7), 2010. doi: 10.1029/2009WR008985.

E. Laloy and J. A. Vrugt. High-dimensional posterior exploration of hydrologic models using multiple-try DREAM(ZS) and high-performance computing. *Water Resources Research*, 48(1), 2012. doi: 10.1029/2011WR010608.

LaplacesDemon Core Team. *LaplacesDemon: complete environment for Bayesian inference*, 2023. URL https://cran.r-project.org/package=LaplacesDemon. R package version 16.1.6.

S.-Y. Liong and C. Sivapragasam. Flood stage forecasting with support vector machines. *JAWRA Journal of the American Water Resources Association*, 38(1):173–186, 2002. doi: 10.1111/j.1752-1688.2002.tb01544.x.

Y. Liu and H. V. Gupta. Uncertainty in hydrologic modeling: toward an integrated data assimilation framework. *Water Resources Research*, 43(7):W07401, 2007. doi: 10.1029/2006WR005756.

J. Mai, J. R. Craig, and B. A. Tolson. Simultaneously determining global sensitivities of model parameters and model structure. *Hydrology and Earth System Sciences*, 24(12): 5835–5858, 2020. doi: 10.5194/hess-24-5835-2020.

J. Mai, B. A. Tolson, J. R. Craig, R. Arsenault, et al. The sensitivity of simulated streamflow to individual hydrologic processes across North America. *Nature Communications*, 13:455, 2022. doi: 10.1038/s41467-022-28010-7.

P. Mantovan and E. Todini. Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology. *Journal of Hydrology*, 330(1–2):368–381, 2006. doi: 10.1016/j.jhydrol.2006.04.046.

S. L. Markstrom, L. E. Hay, and M. P. Clark. Towards simplification of hydrologic modeling: identification of dominant processes. *Hydrology and Earth System Sciences*, 20: 4655–4671, 2016. doi: 10.5194/hess-20-4655-2016.

L. Marshall, D. Nott, and A. Sharma. Hydrological model selection: A Bayesian alternative. *Water Resources Research*, 41(10), 2005. doi: 10.1029/2004WR003719.

D. McInerney, M. Thyer, D. Kavetski, J. Lerat, and G. Kuczera. Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. *Water Resources Research*, 53(3):2199–2239, 2017. doi: 10.1002/2016WR019168.

H. McMillan, J. Freer, F. Pappenberger, T. Krueger, and M. Clark. Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions. *Hydrological Processes*, 24(10):1270–1284, 2010. doi: 10.1002/hyp.7587.

H. McMillan, B. Jackson, M. Clark, D. Kavetski, and R. Woods. Rainfall uncertainty in hydrological modelling: an evaluation of multiplicative error models. *Journal of Hydrology*, 400(1–2):83–94, 2011. doi: 10.1016/j.jhydrol.2011.01.026.

H. McMillan, T. Krueger, and J. Freer. Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality. *Hydrological Processes*, 26(26): 4078–4111, 2012. doi: 10.1002/hyp.9384.

P. A. Mendoza, M. P. Clark, M. Barlage, B. Rajagopalan, L. Samaniego, G. Abramowitz, and H. Gupta. Are we unnecessarily constraining the agility of complex process-based models? *Water Resources Research*, 51:716–728, 2015. doi: 10.1002/2014WR015820.

E. Moges, Y. Demissie, L. Larsen, and F. Yassin. Review: Sources of hydrological model uncertainties. *Water*, 13(1):28, 2021. doi: 10.3390/w13010028.

A. Montanari. What do we mean by 'uncertainty'? The need for a consistent terminology about uncertainty assessment in hydrology. *Hydrological Processes*, 21(6):841–845, 2007. doi: 10.1002/hyp.6623.

A. Montanari and D. Koutsoyiannis. Towards a Bayesian total error analysis of conceptual rainfall-runoff models: characterising model error using storm-dependent parameters. *Journal of Hydrology*, 337(3–4):246–262, 2007. doi: 10.1016/j.jhydrol.2007.01.050.

A. Montanari, G. Young, H. H. G. Savenije, D. Hughes, T. Wagener, L. L. Ren, D. Koutsoyiannis, C. Cudennec, E. Toth, S. Grimaldi, et al. "Panta Rhei—Everything Flows": change in hydrology and society—the IAHS scientific decade 2013–2022. *Hydrological Sciences Journal*, 58(6):1256–1275, 2013. doi: 10.1080/02626667.2013.809088.

R. D. Moore. Application of a conceptual streamflow model in a glacierized drainage basin. *Journal of Hydrology*, 150(1):151–168, 1993. doi: 10.1016/0022-1694(93)90159-7.

Radford M. Neal. *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer, New York, 1996. doi: 10.1007/978-1-4612-0745-0.

A. J. Newman, M. P. Clark, K. Sampson, A. Wood, L. E. Hay, A. Bock, R. J. Viger, D. Blodgett, L. Brekke, J. R. Arnold, T. Hopson, and Q. Duan. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set

characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, 19:209–223, 2015. doi: 10.5194/hess-19-209-2015.

B. Nijssen and D. P. Lettenmaier. Effect of precipitation sampling error on simulated hydrological fluxes and states: anticipating the Global Precipitation Measurement satellites. *Journal of Geophysical Research: Atmospheres*, 109(D2):D02103, 2004. doi: 10.1029/2003JD003497.

V. M. H. Ong, D. J. Nott, M.-N. Tran, S. A. Sisson, and C. C. Drovandi. Likelihood-free inference in high dimensions with synthetic likelihood. *Computational Statistics and Data Analysis*, 128:271–291, 2018. doi: 10.1016/j.csda.2018.07.008.

Z. Pan, P. Liu, S. Gao, J. Xia, J. Chen, and L. Cheng. Improving hydrological projection performance under contrasting climatic conditions using spatial coherence through a hierarchical Bayesian regression framework. *Hydrology and Earth System Sciences*, 23 (8):3405–3421, 2019. doi: 10.5194/hess-23-3405-2019.

K. Peng, D. B. Wright, Y. Derin, G. A. Alexander, A. Pradhan, D. Zoccatelli, and J. Tan. Advancing ensemble streamflow prediction through satellite-based precipitation product and model parameter uncertainty quantification. *Journal of Advances in Modeling Earth Systems*, 17(10):e2025MS004953, 2025.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL https://www.R-project.org/.

O. Rakovec, N. Mizukami, R. Kumar, A. J. Newman, S. Thober, A. W. Wood, and L. Samaniego. Diagnostic evaluation of large-domain hydrologic models calibrated across the contiguous United States. *Journal of Geophysical Research: Atmospheres*, 124:13991–14007, 2019. doi: 10.1029/2019JD030767.

S. Razavi, B. A. Tolson, and D. H. Burn. Review of surrogate modeling in water resources. *Water Resources Research*, 48(7):W07401, 2012. doi: 10.1029/2011WR011527.

B. Renard, D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, 46(5), 2010. doi: 10.1029/2009WR008328.

B. Renard, D. Kavetski, E. Leblois, M. Thyer, G. Kuczera, and S. W. Franks. Toward a reliable decomposition of predictive uncertainty in hydrological modeling: characterizing rainfall errors using conditional simulation. *Water Resources Research*, 47(11):W11516, 2011. doi: 10.1029/2011WR010643.

J. Romero-Cuellar, R. Arabzadeh, J. R. Craig, B. A. Tolson, and J. Mai. A multi-model evaluation of probabilistic streamflow predictions via residual error modelling. *Journal of Hydrology*, 635:131152, 2024. doi: 10.1016/j.jhydrol.2024.131152.

P. Salamon and L. Feyen. Disentangling uncertainty in hydrological modeling through multiplicative error models and sequential data assimilation. *Water Resources Research*, 46(12):W12501, 2010. doi: 10.1029/2009WR009022.

A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. *Global Sensitivity Analysis: The Primer*. John Wiley & Sons, 2008. doi: 10.1002/9780470725184.

S. Samadi, D. L. Tufford, and G. J. Carbone. Estimating hydrologic model uncertainty in the presence of complex residual error structures. *Stochastic Environmental Research and Risk Assessment*, 32:1259–1281, 2018. doi: 10.1007/s00477-017-1489-6.

L. Samaniego, R. Kumar, and S. Attinger. Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, 46:W05523, 2010. doi: 10.1029/2008WR007327.

J. Schaake, S. Cong, and Q. Duan. US MOPEX data set. Technical Report UCRL-JRNL-221228, Lawrence Livermore National Laboratory, 2006.

G. Schoups and J. A. Vrugt. A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resources Research*, 46(10), 2010. doi: 10.1029/2009WR008933.

M. Shafii, B. Tolson, and L. S. Matott. Uncertainty-based multi-criteria calibration of rainfall–runoff models: A comparative study. *Stochastic Environmental Research and Risk Assessment*, 28(6):1493–1510, 2014. doi: 10.1007/s00477-014-0855-x.

T. Smith, L. Marshall, and A. Sharma. Modeling residual hydrologic errors with Bayesian inference. *Journal of Hydrology*, 528:29–37, 2015. doi: 10.1016/j.jhydrol.2015.05.051.

A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14:199–222, 2004. doi: 10.1023/B:STCO.0000035301.49549.88.

Statisticat LLC. *LaplacesDemon: Complete Environment for Bayesian Inference*, 2021. URL https://github.com/LaplacesDemonR/LaplacesDemon. R package version 16.1.6.

J. R. Stedinger, R. M. Vogel, S. U. Lee, and R. Batchelder. Appraisal of the generalized likelihood uncertainty estimation (GLUE) method. *Water Resources Research*, 44(12), 2008. doi: 10.1029/2008WR006822.

B. Széles, J. Parajka, P. Hogan, R. Silasari, L. Pavlin, P. Strauss, and G. Blöschl. The added value of different data types for calibrating and testing a hydrologic model in a small catchment. *Water Resources Research*, 56:e2019WR026153, 2020. doi: 10.1029/2019WR026153.

G. Tang, M. P. Clark, W. J. M. Knoben, H. Liu, S. Gharari, L. Arnal, H. E. Beck, A. W. Wood, A. J. Newman, and S. M. Papalexiou. The impact of meteorological forcing uncertainty on hydrological modeling: a global analysis of cryosphere basins. *Water Resources Research*, 59(10):e2022WR033767, 2023. doi: 10.1029/2022WR033767.

A. Thiboult, F. Anctil, and M.-A. Boucher. Accounting for three sources of uncertainty in ensemble hydrological forecasting. *Hydrology and Earth System Sciences*, 20(5):1809–1825, 2016. doi: 10.5194/hess-20-1809-2016.

M. Thyer, B. Renard, D. Kavetski, G. Kuczera, S. W. Franks, and S. Srikanthan. Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis. *Water Resources Research*, 45(12), 2009. doi: 10.1029/2008WR006825.

B. A. Tolson and C. A. Shoemaker. Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, 43:W01413, 2007. doi: 10.1029/2005WR004723.

M. Troin, J.-L. Martel, R. Arsenault, and F. Brissette. Large-sample study of uncertainty of hydrological model components over North America. *Journal of Hydrology*, 609:127766, 2022. doi: 10.1016/j.jhydrol.2022.127766.

W.-P. Tsai, D. Feng, M. Pan, H. Beck, K. Lawson, Y. Yang, et al. From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nature Communications*, 12(1):5988, 2021. doi: 10.1038/s41467-021-26107-z.

A. Valéry, V. Andréassian, and C. Perrin. 'As simple as possible but not simpler': What is useful in a temperature-based snow-accounting routine? Part 2–Sensitivity analysis of the CEMANeige snow accounting routine on 380 catchments. *Journal of Hydrology*, 517:1176–1187, 2014. doi: 10.1016/j.jhydrol.2014.04.059.

J. A. Vrugt. Distribution-based model evaluation and diagnostics: Elicitability, propriety, and scoring rules for hydrograph functionals. *Water Resources Research*, 60(6): e2023WR036710, 2024. doi: 10.1029/2023WR036710.

J. A. Vrugt, C. J. F. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson. Treatment of input uncertainty in hydrologic modeling: doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resources Research*, 44:W00B09, 2008. doi: 10.1029/2007WR006720.

J. A. Vrugt, C. J. Ter Braak, H. V. Gupta, and B. A. Robinson. Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? *Stochastic Environmental Research and Risk Assessment*, 23(7):1011–1026, 2009. doi: 10.1007/s00477-008-0274-y.

J. A. Vrugt, D. Y. de Oliveira, G. Schoups, and C. G. Diks. On the use of distribution-adaptive likelihood functions: Generalized and universal likelihood functions, scoring rules and multi-criteria ranking. *Journal of Hydrology*, 615:128542, 2022. doi: 10.1016/j.jhydrol.2022.128542.

I. K. Westerberg, J.-L. Guerrero, J. Seibert, K. J. Beven, and S. Halldin. Stage-discharge uncertainty derived with a non-stationary rating curve in the Choluteca River, Honduras. *Hydrological Processes*, 25(4):603–613, 2011. doi: 10.1002/hyp.7848.

F. Woldemeskel, D. McInerney, J. Lerat, M. Thyer, D. Kavetski, D. Shin, et al. Evaluating post-processing approaches for monthly and seasonal streamflow forecasts. *Hydrology and Earth System Sciences*, 22(12):6257–6278, 2018. doi: 10.5194/hess-22-6257-2018.

C. S. Wongchuig, R. C. D. Paiva, V. Siqueira, and W. Collischonn. Hydrological reanalysis across the 20th century: a case study of the Amazon Basin. *Journal of Hydrology*, 570: 755–773, 2019. doi: 10.1016/j.jhydrol.2019.01.025.

Appendices

Appendix 1.A: SVR Training Sample Size Selection

To determine an appropriate training sample size for the SVR model, a series of experiments are conducted aimed at empirically identifying the point at which additional samples no longer provide meaningful improvements in predictive performance. As shown in Figure 5.1, the SVR mean absolute error (MAE) exhibits a clear decaying trend as the sample size increases. The curve begins to flatten at roughly 1,000 samples, indicating that further increases in sample size yield diminishing returns. Based on this behavior, a training sample size of approximately 1,000 is selected for the SVR model. This threshold is specific to the GR4J–CEMANeige configuration and the MOPEX12 records, and models with higher-dimensional parameter spaces or different residual characteristics may require larger training sets. In these cases, the stabilization of the MAE curve should be re-verified for each new application.

This analysis is repeated across all MOPEX12 case studies and for all error-model parameters. In every case, a similar pattern is observed, with the MAE consistently decreasing for larger sample sizes but stabilizing around the same threshold of roughly 1,000 samples. This consistent behavior across basins and parameters reinforces the choice of 1,000 training samples as a practical and effective reference point for the SVR surrogate construction.

Appendix 1.B: Residual Diagnostics for Gaussian Assumption

The suitability of the Gaussian heteroscedastic–autocorrelated error formulation is evaluated further, as the focused configuration in this study is the SHA method. Figure 5.2

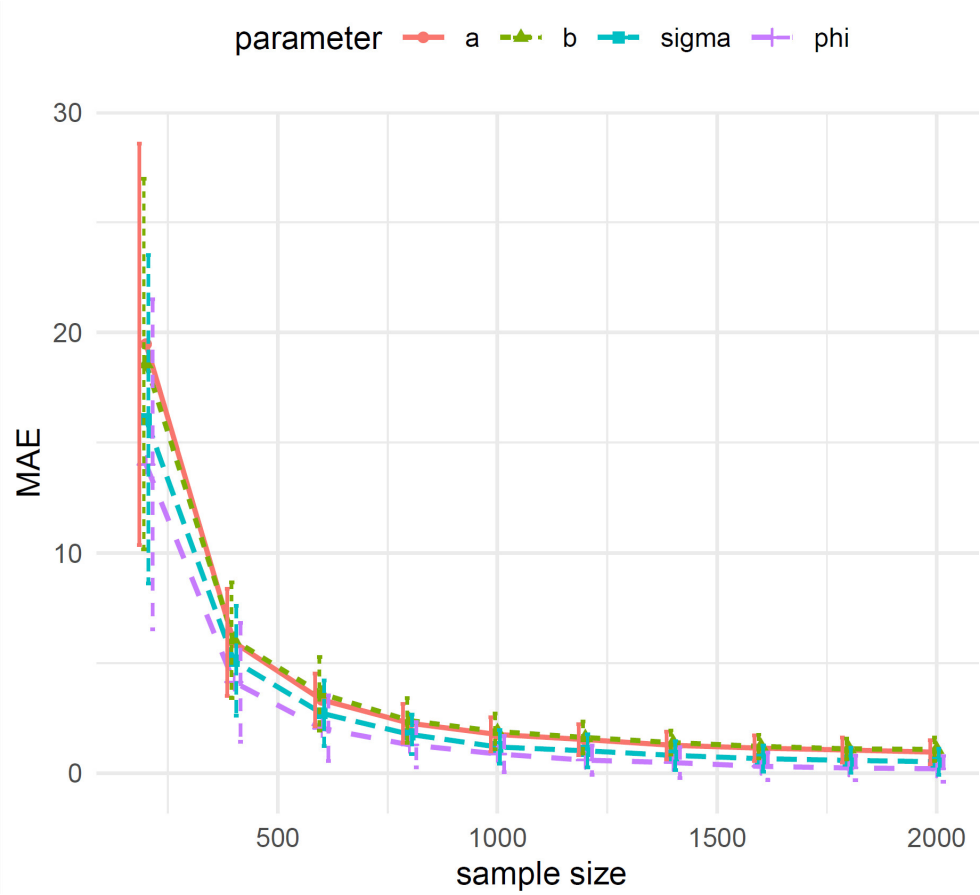


Figure 5.1: Mean Absolute Error (MAE) computed for the SVR model on the calibration data. The vertical error bars indicate the variability of the metric across the MOPEX12 watersheds, with each line representing a different variable.

presents the residual characteristics for the Guadalupe watershed based on samples drawn from the joint posterior distribution. The histogram of residuals (panel a) shows close alignment between the empirical distribution and the fitted Gaussian density, with only minor deviations near zero/low flow values, indicating that the Gaussian assumption provides a reasonable approximation for the residual behavior. The autocorrelation function (panel b) demonstrates that the AR(1) term effectively captures serial dependence, as no significant correlation remains beyond lag 1. In addition, residuals plotted against simulated flow quantiles (panel c) reveal a weak trend and nearly constant variance across flow magnitudes, confirming that the variance stabilization achieved through the heteroscedastic component is appropriate for this case. These diagnostics collectively support the adequacy of the Gaussian residual assumption for the SHA formulation in the analyzed basins.

Although the diagnostics confirm that a Gaussian distribution is reasonable for the tested watersheds, the surrogate-aided framework is not restricted to this residual structure. The surrogate model learns the deterministic mapping between hydrologic parameters (θ) and error-model parameters (ω) and can therefore accommodate alternative error model formulations, such as the Student-t distribution tested in the THA experiment. The improved convergence and predictive performance of SHA compared to THA, GHA, and G (Figure 2.3) primarily arise from the surrogate-based decoupling of θ and ω during inference rather than from the Gaussian assumption itself. Nevertheless, extending the surrogate to probabilistic or adaptive forms that propagate emulator uncertainty could further enhance robustness, particularly for catchments exhibiting non-Gaussian residual behavior or non-stationary conditions.

Appendix 2.A: Blended-model parameter list

Table 5.1 lists the full set of 51 Raven parameters inferred via the NUTS sampler in this study, grouped by their physical role and the corresponding uncertainty class.

Appendix 2.B: Aitchison effect size for compositional comparison

For each of the 12 USGS gauges, we compare the output compositions of three model configurations: blended model v1 (Chlumsky et al., 2021), blended model v2 (Chlumsky et al.,

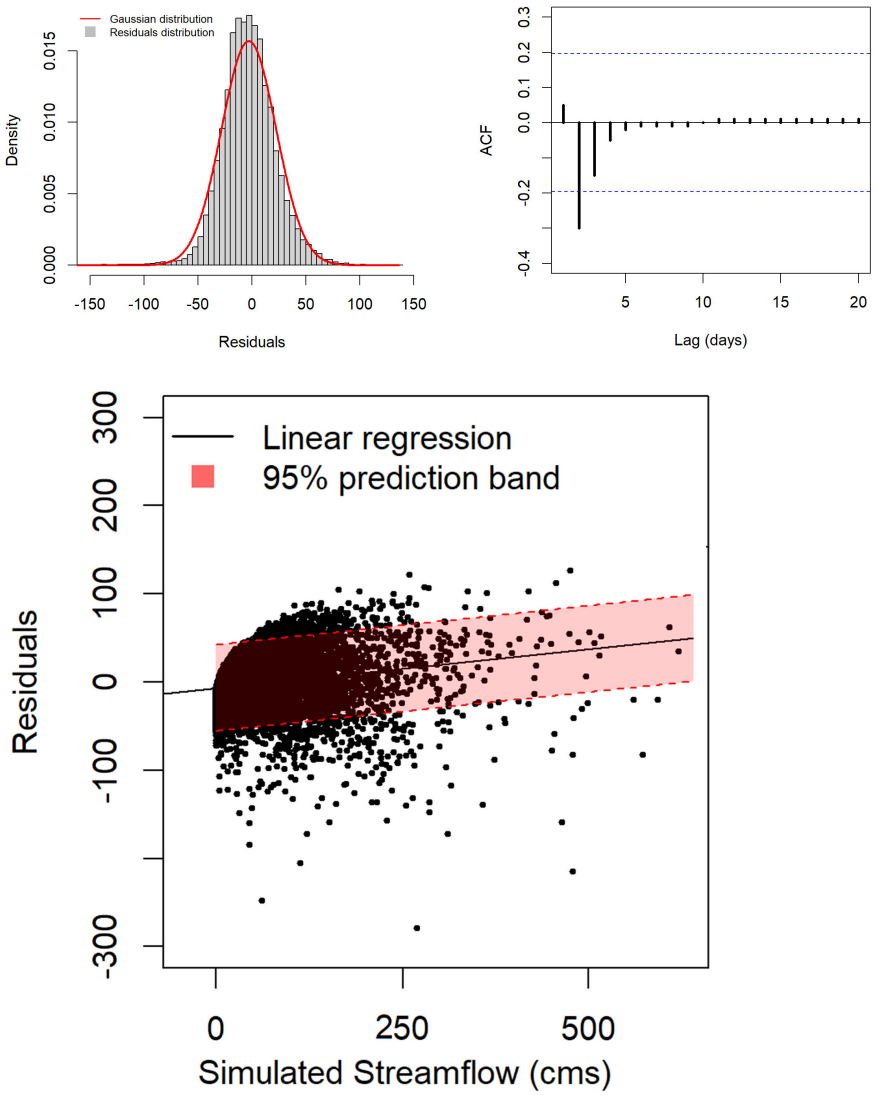


Figure 5.2: Residual diagnostics for a sample drawn from the joint posterior distribution of the Guadalupe, TX watershed. (a) Residual distribution, (b) residual autocorrelation, and (c) residuals versus simulated streamflow quantiles, illustrating heteroscedasticity.

Table 5.1: List of Raven parameters inferred via the MCMC sampler in Chapter 3.

#	Parameter	Low	High	Type	Uncertainty
1	RAIN_CORRECTION	0.8	1.2	Gauge	Forcings
2	SNOW_CORRECTION	0.8	1.2	Gauge	Forcings
3	RELATIVE_ERROR (Pr)	0.1	0.5	Gauge	Forcings
4	RELATIVE_BIAS (Pr)	−0.2	0.2	Gauge	Forcings
5	RELATIVE_ERROR (Tmin)	0.1	0.5	Gauge	Forcings
6	RELATIVE_BIAS (Tmin)	−0.2	0.2	Gauge	Forcings
7	RELATIVE_ERROR (Tmax)	0.1	0.5	Gauge	Forcings
8	RELATIVE_BIAS (Tmax)	−0.2	0.2	Gauge	Forcings
9	HMETS_RUNOFF_COEFF	0	100	Land Use	Parameters
10	B_EXP TOPSOIL	0.1	3	Soil	Parameters
11	HBV_BETA TOPSOIL	0.5	3	Soil	Parameters
12	BASEFLOW_COEFF TOPSOIL	−5	−1	Soil	Parameters
13	MAX_BASEFLOW_RATE TOPSOIL	0	100	Soil	Parameters
14	BASEFLOW_N TOPSOIL	0.5	2	Soil	Parameters
15	TOPMODEL_LAMBDA	5	10	Terrain Classes	Parameters
16	PET_CORRECTION TOPSOIL	0	3	Soil	Parameters
17	SAT_WILT TOPSOIL	0	0.05	Soil	Parameters
18	FIELD_CAPACITY TOPSOIL	0	0.45	Soil	Parameters
19	BASEFLOW_COEFF PHREATIC	−5	−2	Soil	Parameters
20	BASEFLOW_N PHREATIC	0.5	2	Soil	Parameters
21	SNOW_SWI_MIN	0	0.1	Global	Parameters
22	SNOW_SWI_MAX	0.01	0.3	Global	Parameters
23	SWI_REDUCT_COEFF	0.005	0.1	Global	Parameters
24	DD_REFREEZE_TEMP	−5	2	Land Use	Parameters
25	REFREEZE_EXP	0	1	Land Use	Parameters
26	REFREEZE_FACTOR	0	5	Land Use	Parameters
27	SNOW_SWI	0	0.4	Global	Parameters
28	GAMMA_SHAPE	0.3	20	Land Use	Parameters
29	GAMMA_SCALE	0.01	5	Land Use	Parameters
30	GAMMA_SHAPE2	0.5	13	Land Use	Parameters
31	GAMMA_SCALE2	0.15	1.5	Land Use	Parameters
32	MIN_MELT_FACTOR	1.5	3	Land Use	Parameters
33	MAX_MELT_FACTOR	0	5	Land Use	Parameters
34	DD_MELT_TEMP	−1	1	Land Use	Parameters
35	DD_AGGRADATION	0.01	0.2	Land Use	Parameters
36	PERC_COEFF TOPSOIL	0.00001	0.02	Soil	Parameters
37	thickness TOPSOIL (m)	0	0.5	Soil Profiles	Parameters
38	thickness PHREATIC (m)	0	2	Soil Profiles	Parameters
39	RAINSNOW_TEMP	−3	3	Global	Parameters
40	RAINSNOW_DELTA	0.5	4	Global	Parameters
41	PERC_COEFF PHREATIC	0	0.02	Soil	Parameters
42	URBAN_IMPERVIOUSNESS_FRACTION	0.3	0.8	Land Use Classes	Parameters
43	SIGMA	0.001	100	Error Model	Parameters
44	PAR_W11_PIE	0	1	Infiltration weight	Structure
45	PAR_W12_PIE	0	1	Infiltration weight	Structure
46	PAR_W21_PIE	0	1	Quickflow weight	Structure
47	PAR_W22_PIE	0	1	Quickflow weight	Structure
48	PAR_W31_PIE	0	1	Evaporation weight	Structure
49	PAR_W41_PIE	0	1	Baseflow weight	Structure
50	PAR_W51_PIE	0	1	Snow Balance weight	Structure
51	PAR_W52_PIE	0	1	Snow Balance weight	Structure

2024), and blended model v2 with an added linear quickflow component (BASE_LINEAR). We conduct a formal statistical comparison of model-derived compositions to test whether the structural decomposition of model outputs differs significantly between configurations. Because the model outputs form a constrained three-part composition at each station, we perform all analyses in compositional (log-ratio) space.

We quantify differences between model configurations using the Aitchison distance (Aitchison, 1982) in centred log-ratio (CLR) space:

$$d_A = \left\| \mathbb{E}[\mathrm{clr}(A_1)] - \mathbb{E}[\mathrm{clr}(A_2)] \right\|. \tag{5.1}$$

This represents the Euclidean distance between mean log-ratio compositions of two model configurations and captures how the internal allocation of model components changes, independent of absolute magnitude.

Statistical inference is performed in isometric log-ratio space using a multivariate Hotelling framework applied to paired differences between model configurations:

$$F \sim \frac{n-p}{p\,(n-1)}\, n\, \boldsymbol{Y}^{\top} S^{-1} \boldsymbol{Y}, \qquad p = 2, \tag{5.2}$$

where $\boldsymbol{Y}$ is the mean paired difference in ILR coordinates and S is the sample covariance matrix.

To account for multiple comparisons across the 12 stations, p-values are adjusted using the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995), yielding FDR-corrected q-values.

Two stations (7378500, 3364000) exhibit strong and statistically robust compositional differences between BM v1 and both BM v2 and BM v2 + BASE_LINEAR ($q < 0.01$). Three additional stations (8167500, 1608500, 3179000) show marginal differences under blended model v2, with one becoming statistically significant under the BASE_LINEAR configuration. The remaining stations show no statistically significant differences beyond sampling variability. Overall, blended model v2 and especially BASE_LINEAR introduce localized but not widespread changes relative to blended model v1.

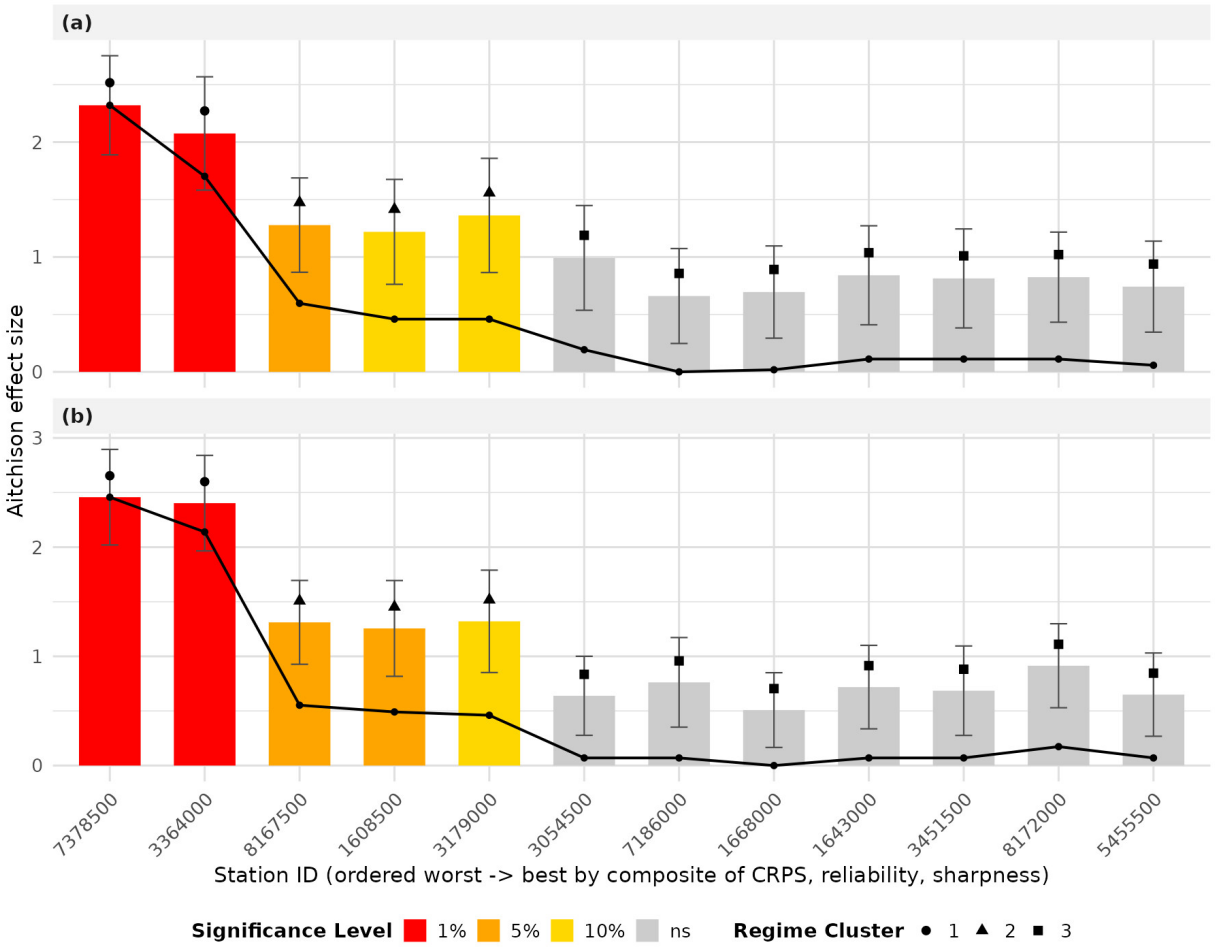


Figure 5.3: Compositional comparison of the model used in this study against (a) blended model v2 and (b) blended model v2 + BASE_LINEAR, across 12 USGS gauges. Bar height: Aitchison effect size (centroid shift in CLR). Colour: FDR-adjusted significance level. Black line: rescaled $-\log_{10} q$. Symbols: hierarchical clustering.