Water Resources Research

RESEARCH ARTICLE
10.1029/2021WR031104

Probabilistic Predictions of Ecologically Relevant Hydrologic Indices Using a Hydrological Model

J. Sebastian Hernandez-Suarez[1,2] and A. Pouyan Nejadhashemi[1]

[1]Department of Biosystems and Agricultural Engineering, Michigan State University, East Lansing, MI, USA, [2]Department of Earth System Science, Stanford University, Stanford, CA, USA

Key Points:

- A calibration approach linked multi-objective optimization and Bayesian inference to improve the prediction of hydrologic indices
- The new approach was compared against Bayesian calibration using uniform priors for model parameters
- A higher precision in simulated hydrologic indices and large reductions of parametric uncertainty in streamflow predictions were observed

Correspondence to:

A. P. Nejadhashemi,
pouyan@msu.edu

Citation:

Hernandez-Suarez, J. S., & Nejadhashemi, A. P. (2022). Probabilistic predictions of ecologically relevant hydrologic indices using a hydrological model. *Water Resources Research*, *58*, e2021WR031104. https://doi.org/10.1029/2021WR031104

Received 11 APR 2022
Accepted 25 AUG 2022

Author Contributions:

Conceptualization: J. Sebastian Hernandez-Suarez, A. Pouyan Nejadhashemi
Methodology: J. Sebastian Hernandez-Suarez, A. Pouyan Nejadhashemi
Software: J. Sebastian Hernandez-Suarez
Supervision: A. Pouyan Nejadhashemi
Validation: J. Sebastian Hernandez-Suarez
Writing – original draft: J. Sebastian Hernandez-Suarez
Writing – review & editing: A. Pouyan Nejadhashemi

Abstract This study evaluated the effects of prior knowledge obtained from multi-objective optimization on the uncertainty quantification of simulated ecologically relevant hydrologic indices (ERHIs) using a hydrological model. Two experiments were formulated considering contrasting conditions in prior knowledge when implementing Bayesian parameter estimation. In the first experiment, we reproduced a common modeling practice in the absence of previous information where uniform priors are considered when calibrating hydrological and error model parameters. In the second experiment, near-optimal Pareto solutions from multi-objective calibration were used to build a multivariate prior distribution for estimating hydrological and error model parameters using Bayesian calibration. In both experiments, a likelihood function considering heteroscedasticity and autocorrelation effects was employed. The Unified Non-dominated Sorting Algorithm III was used for multi-objective calibration, whereas the multiple-try differential evolution adaptive metropolis algorithm was used for Markov Chain Monte Carlo sampling. The experiments were tested using a hydrologic model. We selected a subset of ERHIs comprised of 32 central tendency Indicators of Hydrologic Alteration and seven additional indices describing basic streamflow statistical properties. In general, using near-optimal Pareto parameter distributions as prior knowledge in Bayesian calibration reduced both the bias and variability ranges in ERHIs prediction. Furthermore, there was no significant loss in the reliability of streamflow predictions (from 2% to 6%, $p = 0.07$) when targeting the selected ERHIs, while improving precision (from 66% to 58%, $p < 0.05$) and reducing the bias (from 6% to 4%, $p < 0.05$). Moreover, parametric uncertainty substantially shrank by 72% when linking multi-objective calibration and Bayesian parameter estimation.

1. Introduction

Hydrologic signatures (a.k.a., hydrologic indices) are quantitative features that characterize the statistical properties of hydrologic time series (McMillan, 2020b). These signatures, which are most likely obtained from streamflow data, have received increasing attention due to their significance in understanding hydrologic and ecological processes (Carlisle et al., 2017; McMillan, 2020a). In hydrological modeling, streamflow signatures are typically used for model evaluation (Euser et al., 2013; Gupta et al., 2008; Jehn et al., 2019), model calibration (Shafii & Tolson, 2015), and for informing watershed management in ungauged and poorly gauged watersheds (Guo et al., 2021). Ecologically relevant hydrologic indices (ERHIs) are a subset of hydrologic signatures that can be obtained from hydrologic simulations to predict the biological condition of freshwater ecosystems in sites lacking streamflow or biological data (Hernandez-Suarez & Nejadhashemi, 2018; Mazor et al., 2018). Likewise, simulated ERHIs can be used to evaluate hydrologic and ecological alterations due to anthropogenic interventions or changes in climate and land use (Bejarano et al., 2019; S. K. McKay et al., 2019; Sengupta et al., 2018).

Using hydrological models to simulate ERHIs introduces uncertainty. However, uncertainty sources are not only limited to modeling uncertainties (i.e., inputs, structure, parameters, initial/boundary conditions, and measurement errors), but also include the signature computation method (Westerberg & McMillan, 2015; Westerberg et al., 2016), the number of records in the time series (i.e., record length), and non-stationarity effects (e.g., time period of record) (Kennard et al., 2010). For this reason, simulated ERHIs can result in large prediction errors, especially when hydrological models do not explicitly target those ERHIs during model calibration (Hallouin et al., 2020; Vigiak et al., 2018). Whether a calibration method accounts for uncertainties or not, it can be broadly classified into probabilistic and deterministic methods (Tasdighi et al., 2018).

On one side, probabilistic methods for model calibration consider different uncertainty sources and can be classified into informal and formal approaches (Schoups & Vrugt, 2010). The most important difference in these two

10.1029/2021WR031104

approaches resides in the likelihood function formulation (Beven & Binley, 2014). Informal methods use subjective measures (i.e., Limits of Acceptability) to identify those simulations that are a good fit to the observations (i.e., behavioral solutions), and then generate predictive distributions of the model parameters using sampling algorithms (Vrugt & Beven, 2018). Generally, the characteristics of model residual errors are treated implicitly and mapped onto the resulting parameter distributions (Beven & Smith, 2015). In this context, hydrologic signatures are typically used to further constrain the identification of behavioral solutions (Blazkova & Beven, 2009). Examples using informal methods and EHRIs can be found in Kiesel et al. (2017, 2020). Meanwhile, formal methods explicitly consider a model of residual errors to formulate the likelihood function, providing uncertainty estimates for the parameters of both hydrological and error models (McInerney et al., 2017; Smith et al., 2015). Under a Bayesian framework, different sources beyond parameter uncertainty can be explicitly addressed (Moges et al., 2021). However, the most common practice, which is also conceptually problematic, uses a lumped error model to account for those sources (Ammann et al., 2019). One of the major difficulties with hydrologic signatures has been the formulation of closed-form and tractable likelihood functions (Sadegh et al., 2015). Thus, these methods have been mainly implemented when the modeling objective is to predict the streamflow time series themselves (i.e., time-domain calibration). In recent years, the application of formal probabilistic calibration targeting the hydrologic signatures directly (i.e., signature-domain calibration) was introduced by implementing Approximate Bayesian Computation (ABC) methods. ABC methods do not require likelihood function evaluations and, instead, they sample Bayesian posterior distributions at the expense of a higher number of model evaluations (Fenicia et al., 2018; Kavetski et al., 2018; Sadegh & Vrugt, 2014).

On the other side, deterministic methods are mainly comprised of optimization methods using single or multiple objective functions to generate parameter sets that provide the best fit between observations and simulations (McInerney et al., 2018; Shafii and de Smedt, 2009). Single-objective methods are rather unreliable for decision-making since they do not address equifinality and identifiability issues (Beven, 2006). In addition, they only provide point estimates of model parameters and predictions and ignore the distributional properties of the model residual errors (Farmer & Vogel, 2016). In contrast, multi-objective methods provide ranges of solutions (i.e., Pareto-optimal solutions). However, the resulting Pareto-optimal distribution of model parameters and predictions do not necessarily correspond to probabilistic solutions suitable for uncertainty analysis (Reichert & Schuwirth, 2012; Tang et al., 2018). Hydrological models used for the replication of ERHIs are typically calibrated using deterministic methods (e.g., Hallouin et al., 2020; Hernandez-Suarez et al., 2021; Parker et al., 2019; Pool et al., 2017; Sengupta et al., 2018; Shrestha et al., 2014, 2016; Zhang et al., 2016). However, it is worth noting that the reported prediction errors for some of these studies are originated based on the distribution of the relative differences between observed and point predictions of ERHIs across multiple locations (Vigiak et al., 2018), multiple calibration trials (Pool et al., 2017; Vis et al., 2015), or Pareto-optimal solutions (Hernandez-Suarez et al., 2018). These distributions are valuable since they can be used in other modeling exercises as prior knowledge (e.g., modeling at ungauged watersheds, parameter regionalization), especially when using probabilistic methods (Almeida et al., 2013).

Improvements in ERHIs prediction have been mainly driven by the choice of objective functions on untransformed or transformed streamflows (Hallouin et al., 2020). These objective functions target specific flow conditions (i.e., extreme flows), regime facets (i.e., magnitude, duration, frequency, rate of change, timing), or hydrologic indices (Hallouin et al., 2020). As a result, several calibration strategies have been devised, some of them resulting in ensembles of model solutions (Hernandez-Suarez et al., 2018; Kiesel et al., 2020; Sengupta et al., 2018). Among these strategies, those using multi-objective calibration gained popularity since they consider tradeoffs among different targets (Efstratiadis & Koutsoyiannis, 2010; Kollat et al., 2012). However, these methods do not provide formal uncertainty estimates for model parameters and outputs. These estimates are relevant when predicting streamflows and ERHIs within the spatial domain of distributed or semi-distributed models. In addition, it is important to estimate uncertainty when developing and evaluating regionalization schemes based on hydrological modeling results (Addor et al., 2018; Almeida et al., 2016; Guo et al., 2021; Mazor et al., 2018; Moges et al., 2021; Prieto et al., 2019).

In time-domain Bayesian calibration, the likelihood function is typically evaluated on transformed time series to satisfy the assumptions of the selected error model. When using non-informative priors, the parameter estimation mainly relies on the likelihood function. Thus, many studies have focused on selecting the most appropriate function formulation (Gupta et al., 2022). However, these functions hardly extract independent pieces of information targeting multiple flow conditions or regime facets (Sadegh et al., 2018). The definition of meaningful prior distributions can be a complementary approach to addressing this limitation and reducing uncertainty in predictions (Hrachowitz et al., 2014). In addition, informative priors can incorporate effects that are not explicitly considered in the likelihood function, such as constraints in variables of interest or model parameters (Nijzink et al., 2018). Still, the choice of priors usually relies on expert knowledge and/or previous studies in the area of interest, which can result in a rather subjective task (Beven et al., 2011). As an alternative, Tang et al. (2018) showed that multi-objective calibration results can be used to define priors for error parameters in a more objective manner. Since this approach requires further exploration, we extended it to the definition of priors for model parameters in the context of ERHIs prediction.

The main goal of this study was to evaluate the effect of prior knowledge obtained from multi-objective calibration on the resulting posterior parameter distributions and ERHIs predictions when using a time-domain Bayesian calibration method. The main novelty of this study is the development of a methodology that links the advances in deterministic ERHIs prediction and uncertainty quantification. To the best of our knowledge, this is the first time that an evaluation of this kind has been performed to predict ERHIs. The objectives of this study were to (a) estimate the total uncertainty in predicting a set of ERHIs when targeting the overall streamflow time series, (b) compare the posterior model parameter distributions when using uniform versus Pareto-optimal priors, and (c) identify the extent of the changes in parameter estimation when using uniform versus Pareto-optimal priors. We performed this evaluation in an agriculture-dominated watershed in Michigan, US, using the Unified Non-dominated Sorting Algorithm III (U-NSGA-III) (Seada & Deb, 2016) for multi-objective calibration, the soil and water assessment tool (SWAT) as the hydrological model, and the multiple-try differential evolution adaptive metropolis (ZS) (MT-DREAM$_{(ZS)}$) algorithm (Laloy & Vrugt, 2012) for sampling the posterior distributions. For the likelihood function, we used a lumped residual errors model accounting for heteroscedasticity and autocorrelation (McInerney et al., 2017).

2. Materials and Methods

We outlined two experiments using daily data to compare the performance of time-domain Bayesian calibration under contrasting prior knowledge conditions. We employed the same likelihood function regardless of the experiment. For each experiment, we defined two time periods with the same number of consecutive streamflow observations. In *Experiment 1*, we reproduced a common modeling practice when implementing Bayesian inference for model calibration for each time period. This practice uses uniform priors to infer model parameters that are usually built upon physically meaningful ranges for the parameters (Gupta et al., 2022). Meanwhile, *Experiment 2* was devised to evaluate the effect of prior knowledge about the calibration parameters on the modeling results when using Bayesian inference. In particular, we were interested in incorporating the information about the model parameters obtained from multi-objective calibration into the overall Bayesian calibration process. For this purpose, we obtained Pareto-optimal parameter distributions from one time period (*Period 1*) and used them as prior knowledge for calibrating the model using data for the other time period (*Period 2*). We compared the Pareto-optimal parameter distributions against the posterior distributions using Bayesian inference for *Period 1* with uniform priors. Likewise, we compared the predictive distributions for model parameters and ERHIs using non-uniform and uniform priors under *Period 2*. Finally, we assessed the reliability, precision, and bias of the streamflow and ERHIs predictions for each experiment.

2.1. Bayesian Parameter Estimation

In this study, we assumed that a streamflow observation $\tilde{Y}_t$ at time step t is linked to a deterministic hydrological model H with model parameters $\boldsymbol{\theta}_H$, and given forcing data $\tilde{\boldsymbol{X}}$, as follows,

$$\tilde{Y}_t \leftarrow H_t\left(\boldsymbol{\theta}_H, \tilde{\boldsymbol{X}}\right) + \varepsilon_t\left(\boldsymbol{\theta}_\varepsilon\right) \tag{1}$$

where, ε_t represents the raw residuals as an aggregated measure of predictive errors. We also assumed that the residuals follow a probability distribution with parameters $\boldsymbol{\theta}_\varepsilon$. Using the Bayes equation, the posterior probability distribution of hydrological and residual error model parameters can be obtained by conditioning the model to observations and the given forcing data (McInerney et al., 2017; Vrugt, 2016),

10.1029/2021WR031104

$$p\left(\theta_H, \theta_\varepsilon | \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}\right) \propto p\left(\tilde{\boldsymbol{Y}} | \theta_H, \theta_\varepsilon, \tilde{\boldsymbol{X}}\right) p(\theta_H, \theta_\varepsilon) \tag{2}$$

where, $p(\theta_H, \theta_\varepsilon)$ is the joint prior distribution of hydrological and residual error model parameters, and $L\left(\theta_H, \theta_\varepsilon | \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}\right) \equiv p\left(\tilde{\boldsymbol{Y}} | \theta_H, \theta_\varepsilon, \tilde{\boldsymbol{X}}\right)$ is the likelihood function. In the following sections, we present the model of residual errors and corresponding likelihood function (Section 2.1.1), the prior distributions we used (Section 2.1.2), and the sampling procedure to approximate the posterior distributions (Section 2.1.3).

2.1.1. Likelihood Function

The likelihood function summarizes the distance between the model simulations and observations and is built on top of the model of residual errors (Vrugt, 2016). The error model used here corresponds to a typical formulation in hydrological sciences that describes the total effect of all sources of error (Ammann et al., 2019). We followed a transformational strategy to account for heteroscedasticity and skewness in predictive errors (McInerney et al., 2017). For this purpose, we used the Box-Cox or power transformation with parameter λ (Box & Cox, 1964),

$$z\left[Y_t; \lambda\right] = \begin{cases} \left(Y_t^\lambda - 1\right) / \lambda & \text{if } \lambda \neq 0 \\ \log Y_t & \text{if } \lambda = 0 \end{cases} \tag{3}$$

Thus, the transformed residual η at time step t is obtained as follows,

$$\eta_t = z\left[\tilde{Y}_t, \lambda\right] - z\left[H_t\left(\theta_H, \tilde{\boldsymbol{X}}\right); \lambda\right] \tag{4}$$

Since errors in daily hydrological model outputs are usually highly autocorrelated, we used a first-order autoregressive (AR1) model to consider the temporal persistence of the (transformed) residual errors,

$$\eta_t = \phi\eta_{t-1} + W_t \tag{5}$$

where, ϕ is the autoregressive parameter and W_t is the disturbance or innovation. We assumed that innovations followed a truncated Gaussian distribution $\mathcal{TN}$ to avoid negative streamflow predictions with mean $\mu = 0$, standard deviation $\sigma = \sigma_W$, and lower bound $L_{W,t}$ (Fenicia et al., 2018),

$$W_t \sim \mathcal{TN}\left(0, \sigma_W, L_{W,t}\left(\theta_H, \tilde{\boldsymbol{X}}, \eta_{t-1}\right)\right) \tag{6}$$

Note that $L_{W,t}$ is defined such that $z\left[H_t\left(\theta_H, \tilde{\boldsymbol{X}}\right); \lambda\right] + \eta_t(\theta_\varepsilon) \geq z[0; \lambda]$, which makes it time-dependent. Assuming innovations are independent, the likelihood function is formulated as follows (Fenicia et al., 2018):

$$L\left(\theta_H, \theta_\varepsilon | \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}\right) = \prod_{t=1}^{N_t} \frac{f_N(W_t | 0, \sigma_W)}{1 - F_N\left(z[0; \lambda] - z\left[H_t\left(\theta_H, \tilde{\boldsymbol{X}}\right); \lambda\right] - \phi\eta_{t-1} | 0, \sigma_W\right)} \times \frac{\partial z\left[\tilde{Y}_t; \lambda\right]}{\partial Y} \tag{7}$$

where, $W_t = \eta_t - \phi\eta_{t-1}$, $f_N(v|\mu, \sigma)$ is the Gaussian probability distribution function with mean μ and standard deviation σ evaluated for v, $F_N(v|\mu, \sigma)$ is the corresponding cumulative distribution function (CDF), and N_t is the total number of observations. It is worth noting that for large N_t, which is the case for hydrologic time series, the likelihood is a very small number that can result in arithmetic underflow. Thus, it is common to work with the log-likelihood, $\mathcal{L}\left(\theta_H, \theta_\varepsilon | \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}\right)$, instead,

$$\begin{aligned} \mathcal{L}\left(\theta_H, \theta_\varepsilon | \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}\right) \cong & -\frac{N_t}{2}\log 2\pi - N_t \log \sigma_W - \frac{1}{2\sigma_W^2}\sum_{t=2}^{N_t}(\eta_t - \phi\eta_{t-1})^2 + \sum_{t=1}^{N_t}\log\frac{\partial z\left[\tilde{Y}_t; \lambda\right]}{\partial Y} \\ & - \sum_{t=2}^{N_t}\log\left\{1 - F_N\left(z[0; \lambda] - z\left[H_t\left(\theta_H, \tilde{\boldsymbol{X}}\right); \lambda\right] - \phi\eta_{t-1} | 0, \sigma_W\right)\right\} \end{aligned} \tag{8}$$

The complete set of error model parameters is $\theta_\varepsilon = \{\lambda, \phi, \sigma_W\}$. In this study, we fixed the values of λ and ϕ to 0.2 and 0.8, respectively, following Evin et al. (2014) and McInerney et al. (2017) recommendations for reducing parameter interactions during model calibration.

2.1.2. Prior Distributions

2.1.2.1. Experiment 1—Non-Informative Priors

In this experiment, we used uniform distributions that defined the feasible parameter space by providing the minimum and maximum values for the hydrological and error model parameters. Upper and lower limits for the hydrological model parameters were determined from the literature and previous modeling exercises in the study area (see Section 2.4.3), whereas σ_W limits were defined from an initial screening of modeling results. The initial states for the chains used by the Markov chain Monte Carlo (MCMC) sampling algorithm for Bayesian inference (see Section 2.1.3) were drawn using Latin-Hypercube Sampling (M. D. McKay et al., 1979) subject to the aforementioned parameter limits.

2.1.2.2. Experiment 2—Multi-Objective Model Calibration

A constrained, performance-based, multi-objective calibration targeting a set of ERHIs was executed to obtain near-optimal Pareto distributions of model parameters. We implemented the recently developed evolutionary multi-objective optimization algorithm U-NSGA-III (Seada & Deb, 2016). The calibration consisted in minimizing six objective functions $f(\boldsymbol{\theta}_H)$ derived from performance metrics $P_m(\boldsymbol{\theta}_H)$. Each $f(\boldsymbol{\theta}_H)$is computed on transformed or untransformed streamflow values to accentuate different flow conditions or regime facets,

$$f_j(\boldsymbol{\theta}_H) = 1 - P_{m_j}(\boldsymbol{\theta}_H) \tag{9}$$

where, $j = 1, 2, \ldots, 6$. The performance metrics used in this study for calibration were the Kling-Gupta Efficiency (Gupta et al., 2009) computed on untransformed and inverse-transformed values (KGE and $\mathrm{KGE_{inv}}$, respectively), the relative Index of Agreement d_{rel} (Krause et al., 2005), and the coefficient of determination computed on untransformed, inverse-, and square-root-transformed values (R^2, R^2_{inv}, and R^2_{sqrt}, respectively). An optimization constraint, which must not be greater than 0, was defined to limit all targeted ERHIs to not exceed a predefined acceptability threshold τ in terms of relative error $e_{\mathrm{rel}}(\boldsymbol{\theta}_H)$,

$$e_{\mathrm{rel}_i}(\boldsymbol{\theta}_H) = \frac{I_i\left(H\left(\boldsymbol{\theta}_H, \tilde{\boldsymbol{X}}\right)\right) - I_i\left(\tilde{\boldsymbol{Y}}\right)}{I_i\left(\tilde{\boldsymbol{Y}}\right)} \tag{10}$$

where, I_i is the ith ERHI evaluated for the simulation $H\left(\boldsymbol{\theta}_H, \tilde{\boldsymbol{X}}\right)$ and observations $\tilde{\boldsymbol{Y}}$. The constraint $CV(\boldsymbol{\theta}_H)$ was formulated to penalize high relative errors and for separating feasible from unfeasible solutions. An unfeasible solution results in ERHI values outside the predefined acceptability threshold,

$$\mathrm{CV}(\boldsymbol{\theta}_H) = \sum_{i=1}^{m} k_i(\boldsymbol{\theta}_H)\left[1 + w_i\left(\frac{|e_{\mathrm{rel}_i}(\boldsymbol{\theta}_H)|}{\tau} - 1\right)\right]$$

$$k_i(\boldsymbol{\theta}_H) = \begin{cases} 0 & \text{if } \dfrac{|e_{\mathrm{rel}_i}(\boldsymbol{\theta}_H)|}{\tau} - 1 \leq 0 \\ 1 & \text{Otherwise} \end{cases}$$

$$w_i = \frac{1}{g \times h_i} \tag{11}$$

where, m is the total number of ERHIs, w_i is the weighting factor for the ith ERHI, g is the number of ERHI categories, and h_i is the total number of ERHIs in the category that contains the ith ERHI. The ERHIs used in this study and their categories are presented in Section 2.4.4. The value of τ used in this study was 0.3 (Hernandez-Suarez et al., 2018), which means that all the targeted ERHIs must attain relative errors within $\pm 30\%$. Once the near-optimal Pareto solutions were obtained, we computed σ_W for each individual solution using Equations 3–5. A multivariate kernel distribution was generated to have a non-parametric representation of the joint distribution of parameters $\boldsymbol{\theta}_H$ and σ_W using the resulting near-optimal Pareto parameter sets. For this purpose, we employed the *mvksdensity* function in MATLAB R2019b using a Gaussian kernel. An initial vector of bandwidths was defined using the Silverman's rule of thumb (Silverman, 1986),

$$b_k = \sigma_k \left[\frac{4}{(d+2)n} \right]^{1/(d+4)} \quad (12)$$

where, d is the number of dimensions (i.e., number of hydrological and error model parameters), n is the number of observations (i.e., Pareto-optimal solutions), $k = 1, 2, \ldots, d$, and σ_k is the standard deviation of the kth variate (i.e., parameter). This vector of bandwidths was further refined by maximizing the agreement between the marginal empirical CDF of each parameter and the corresponding marginal CDF obtained from the multivariate kernel density distribution using a five-fold cross-validation and a genetic algorithm. For this purpose, we used the *ga* function in MATLAB R2019b. The lower and upper bounds for the optimization search were defined from multiplying b_k by 0.1 and 1.1, respectively. The objective function to minimize was the cross-validated average sum of squared errors between the empirical and estimated marginal CDFs. We employed the default *ga* settings: population size of 200 individuals, uniform random initial population, stochastic uniform selection, 10 elite individuals, scattered crossover for recombination using a 0.8 fraction, and adaptive mutation. The resulting optimized multivariate kernel distribution was then used as the prior distribution $p(\boldsymbol{\theta}_H, \boldsymbol{\theta}_\varepsilon)$ under this experiment. It is worth noting that the initial states for the chains used by the MCMC sampling algorithm were directly drawn from $p(\boldsymbol{\theta}_H, \boldsymbol{\theta}_\varepsilon)$.

2.1.3. Sampling Algorithm

In this study, the MT-DREAM$_{(ZS)}$ algorithm (Laloy & Vrugt, 2012) was used to efficiently explore the posterior distribution of hydrological and error model parameters. MT-DREAM$_{(ZS)}$ is an adaptive MCMC algorithm using multiple-try sampling, snooker updating, and an archive of past states to improve the convergence of computationally intensive and high-dimensional models (Vrugt, 2016). This method belongs to the differential evolution adaptive metropolis (DREAM) multi-chain family of algorithms for Bayesian inference. The original DREAM algorithm automatically adjusts the scale and orientation of the proposal distribution used for posterior inference. In addition, it employs subspace sampling and outlier chain correction while maintaining balance and ergodicity (Vrugt et al., 2009). DREAM$_{(ZS)}$ introduced the use of past samples (from an external archive) into the jump distribution. As a result, the sampling procedure requires a smaller number of chains to explore the target distribution, outliers do not need a forceful treatment, and chains can run in a distributed manner (Vrugt, 2016). To ensure convergence (given the violation of Markovian principles introduced by adaptive Metropolis samplers), the adaptation rate decreases with the number of generations (Ter Braak & Vrugt, 2008). MT-DREAM$_{(ZS)}$ introduced multiple-try sampling in each of the chains, creating m different proposals in each chain that can be evaluated in parallel, which is more efficient than running DREAM with large chain numbers (Laloy & Vrugt, 2012).

Convergence to a stationary posterior distribution was monitored using the multivariate $\hat{R}$-statistic proposed by Gelman and Rubin (1992). The multivariate $\hat{R}$-statistic, which is computed using the last 50% samples of each parallel chain, is used to evaluate whether the between- and within-covariance matrices of these chains are similar. When these matrices are very similar, the $\hat{R}$-statistic is close to unity. An $\hat{R}$-statistic below 1.2 is used in practice to declare convergence (Gelman et al., 2013).

2.2. Generation of Predictive Distributions of ERHIs

Predictive distributions were generated by propagating the posterior probability distributions for the model and error parameters through the hydrological model and ERHIs computation methods. For the i-th individual set of parameters $\boldsymbol{\theta}^i = \left(\boldsymbol{\theta}^i_H, \boldsymbol{\theta}^i_\varepsilon\right)$, a prediction of a given ERHI was obtained as follows (Fenicia et al., 2018; McInerney et al., 2017):

1. Sample W_t using Equation 6.
2. Obtain η^i_t using Equation 5. For $t = 1$, η^i_t is directly sampled and step 1 is ignored:

$$\eta^i_1 \leftarrow f_N\left(0, \sigma^i_\eta\right) \quad (13)$$

where, $\sigma^2_\eta = \sigma^2_W / \left(1 - \phi^2\right)$

3. Compute the streamflow prediction $\boldsymbol{Y}$ at time step t for the given $\boldsymbol{\theta}^i$ as follows:

$$Y_t^i(\boldsymbol{\theta}^i) = z^{-1}\left(z\left[H_t(\boldsymbol{\theta}_H^i, \tilde{\boldsymbol{X}}); \lambda^i\right] + \eta_t^i\right) \tag{14}$$

   where, z^{-1} is the inverse Box-Cox transformation.
4. Once Y_t^i is obtained for the N_t time steps, the j-th ERHI is computed using Equation 15:

$$ERHI_j^i = I_j\left(\boldsymbol{Y}^i(\boldsymbol{\theta}^i)\right) \tag{15}$$

The parameter sets were taken from the last 20% posterior samples obtained by the MT-DREAM$_{(ZS)}$ algorithm.

2.3. Performance Evaluation

We computed three measures to quantify reliability, precision, and bias for evaluating the performance of the Bayesian parameter estimation under each experiment. A prediction is reliable when the observations can be considered samples of the predictive distribution (i.e., observations consistently fall within the prediction bounds). The reliability measure that was employed in this study represents the average absolute difference between the predictive quantile-quantile (PQQ) plot and a 1:1 line representing the CDF of a standard uniform distribution $\boldsymbol{U}(0,1)$ (McInerney et al., 2017). Regarding precision, we determined the average coefficient of variation of the predicted streamflows using the observations as a proxy to the average streamflow at each time step (McInerney et al., 2017). The precision measure represents the width of the prediction bounds. The following equation was used to compute the precision:

$$\text{Precision}\left[Y, \tilde{Y}\right] = \frac{1}{N_t}\sum_{t=1}^{N_t}\frac{\text{std}(Y_t)}{\tilde{Y}_t} \tag{16}$$

where, $\text{std}(Y_t)$ is the standard deviation of streamflow predictions at time step t. Finally, we computed the absolute volumetric bias to evaluate the long-term water balance error of the predictions. For this purpose, we used the mean prediction value at each time step $\overline{Y}_t$:

$$\text{Bias}\left[Y, \tilde{Y}\right] = \left|\frac{\sum_{t=1}^{N_t}\tilde{Y}_t - \sum_{t=1}^{N_t}\overline{Y}_t}{\sum_{t=1}^{N_t}\tilde{Y}_t}\right| \tag{17}$$

The reliability, precision, and bias measures used herein targeted the overall streamflow predictions; for the ERHIs predictions, we directly compared the resulting predictive distributions against the ERHIs obtained from the observations. For the latter, we visually inspected whether the ERHIs from observations fell within the corresponding predictive bounds. Also, we verified whether the median relative errors felt within the nominal $\pm 30\%$ relative error range. This error range has been reported in previous studies as a reference value for uncertainty in ERHIs due to data length effects when working with 15-year time series (Kennard et al., 2010). In addition, we computed the coefficient of variation of each ERHI distribution using the observed ERHI as a reference and obtained the average relative error between each predicted and observed ERHI.

We used generalized least-square (GLS) estimation with an Autoregressive model with lag 1, or AR(1), to evaluate the differences in reliability, precision, and bias of streamflow predictions by comparing observed quantiles, coefficients of variation, and the average predictions at a daily time step. To this end, the differences between the two experiments for each time period were used as response variables to fit a simple regression model with intercept. The differences are considered significant if the reported p-value for the intercept parameter is less than 0.05 (i.e., 95% confidence interval for the intercept does not include zero).

10.1029/2021WR031104

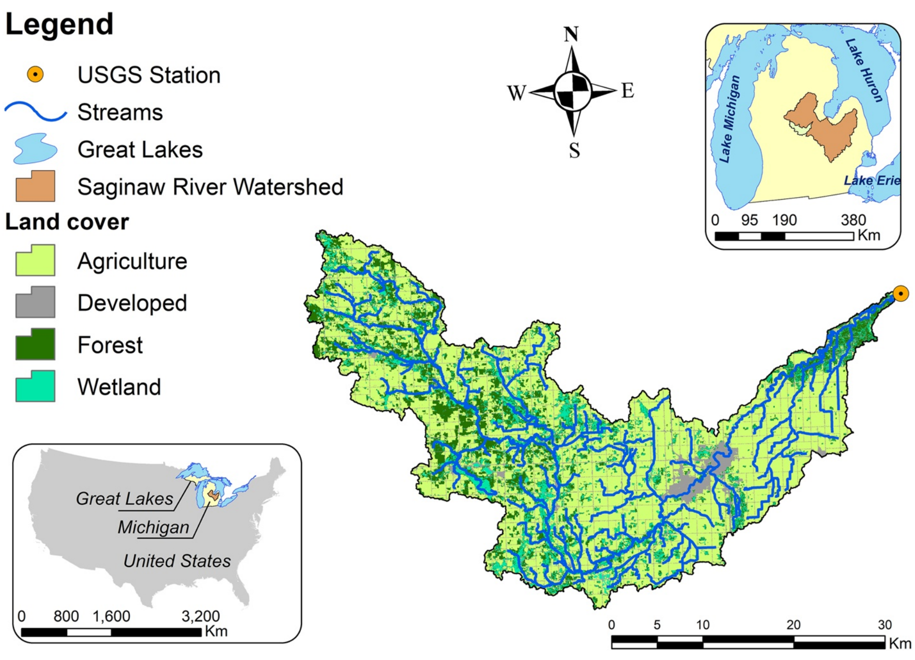

Figure 1. Study area location and major land uses.

2.4. Case Study

2.4.1. Study Area and Model

We executed the calibration experiments in the Honeyoey Creek-Pine Creek Watershed located in east-central Michigan, US (Figure 1). This watershed has a drainage area of 1,010 km^2, and its land use is predominantly agriculture, covering about 50% of the total area, followed by forests (~24%) and wetlands (~16%). Developed areas account for less than 4% of the total watershed area (Hernandez-Suarez et al., 2018). SWAT 2012, Rev. 622 was used as the deterministic hydrological model in Equation 1 for predicting streamflow at the watershed outlet. SWAT is a process-based model widely used to simulate daily water quantity and quality time series at the watershed scale (Arnold et al., 2012). In SWAT, a watershed is divided into subwatersheds, which are comprised of hydrologic response units (HRUs). An HRU is a homogeneous land unit concerning land use/cover, soil type, and slope. In this study, the Honeyoey Creek-Pine Creek Watershed was divided into 250 subwatersheds, each one comprised by a single HRU representing the dominant land use, soil type, and slope conditions (Einheuser et al., 2012). SWAT was used to simulate daily streamflows from 2003 to 2014 (*Period 1*) and from 1983 to 1994 (*Period 2*). Warm-up periods of 2 years (1981–1982 and 2001–2002) were used to reduce the effect of initial conditions in both periods. Potential evapotranspiration was estimated using the Penman-Monteith equation (Monteith, 1965). Surface runoff was obtained using the soil conservation service (SCS) curve number method (USDA-SCS, 1972), and the selected routing method was the variable storage coefficient routine developed by Williams (1969).

2.4.2. Data

Input data included the 30-m resolution National Elevation Data set provided by the US Geological Survey (USGS, 2018), the 30-m resolution Cropland Data Layer provided by the National Agricultural Statistics Service of the US Department of Agriculture (USDA-NASS, 2012), soil properties extracted from the Soil Survey Geographic Database (SSURGO) from the USDA Natural Resources Conservation Service (USDA-NRCS, 2020), and daily precipitation and maximum and minimum air temperature time series from 1981 to 2014 collected at two land stations from the National Centers for Environmental Information of the National Oceanic and Atmospheric Administration (NOAA-NCEI, 2020). Missing values and remaining input weather data such as solar radiation, wind speed, and relative humidity were estimated using the SWAT built-in WXGEN stochastic weather generator (Neitsch et al., 2011). Daily observed streamflow records were obtained at the watershed outlet for the period of study from the Pine River Near Midland USGS gauging station 04155500 (USGS, 2020).

Table 1
Model Calibration Parameters and Ranges

Parameter	Description	Calibration range
Biomix	Biological mixing efficiency	[0, 1]
CN2[a]	Initial soil conservation service (SCS) runoff number for moisture condition II	[−0.25, 0.25]
Canmx	Maximum canopy storage (mm H_2O)	[0, 100]
Esco	Plant uptake compensation factor	[0, 1]
Epco	Soil evaporation compensation factor	[0, 1]
Alpha bf	Baseflow alpha factor (days^{-1})	(0, 1]
Gw delay	Groundwater delay time (days)	[0, 500]
Gwqmn	Threshold depth of water in the shallow aquifer required for return flow to occur (mm H_2O)	[0, 5,000]
Gw revap	Groundwater "revap" coefficient	[0.02, 0.2]
Revapmn	Threshold depth of water in the shallow aquifer for "revap" or percolation to the deep aquifer to occur (mm H_2O)	[0, 1,000]
Rchrg dp	Deep aquifer percolation fraction	[0, 1]
Ch n2	Manning's n value for the main channel	(0, 0.3]
Ch k2	Effective hydraulic conductivity in main channel alluvium (mm h^{-1})	(0, 500]
Sol awc[a]	Available water capacity of the soil layer (mm H_2O mm^{-1} soil)	[−0.25, 0.25]
Surlag	Surface runoff lag coefficient	[1, 24]

[a]These parameters were adjusted by perturbing the initial default spatially-varying values for each HRU by a global fraction. HRU, hydrologic response units.

2.4.3. Calibration Parameters

The set of θ_H to be estimated was comprised of 15 SWAT model parameters. Maximum and minimum limits for each parameter were defined following the model documentation (Neitsch et al., 2011) and previous studies (Herman et al., 2018; Hernandez-Suarez et al., 2018). Model parameters were adjusted during the calibration process either by replacing the original value with a new one or by perturbing the original value by a fraction. Most of the parameters were assumed to have the same value in every HRU. Only the curve number for moisture condition II (CN2) and the available water capacity of the soil layer (Sol awc) were assumed to spatially change and were adjusted by perturbing the initial default values by a global fraction. These global fractions were calibrated instead of estimating CN2 and Sol awc at each individual HRU. Calibration model parameters and their calibration ranges are presented in Table 1.

2.4.4. Ecologically Relevant Hydrologic Indices

The selection of hydrologic indices depends on the ecohydrological application objectives. For instance, some studies target non-redundant hydrologic metrics for streamflow classification (Eng et al., 2017), others target specific hydrologic indices relevant to the condition of specific biological communities (George et al., 2021). Our goal in this study was to calibrate the hydrological model targeting a balanced representation of the streamflow regime, which is of interest when defining environmental flows (Poff et al., 2010). Therefore, we selected 32 indices of hydrologic alteration (IHA) (The Nature Conservancy, 2009), describing the central tendency of several streamflow regime characteristics. These indices were originally classified into five categories representing distinct regime facets such as magnitude, duration, frequency, timing, and rate of change of flows (Richter et al., 1997). IHA have been used in previous SWAT model applications to evaluate the effects of climate and land use changes on hydrological extremes (Chen et al., 2019). In addition, we considered seven indices proposed by Archfield et al. (2014) (a.k.a., Magnificent seven) describing basic properties of streamflow time series such as central tendency, variability, skewness, kurtosis, autocorrelation, and seasonality. The list of 39 ERHIs is presented in Table 2.

2.4.5. Experiments Set Up

The U-NSGA-III algorithm was implemented using the *pymoo* library in Python 3.7 (Blank & Deb, 2020). We developed a Python interface to modify SWAT's input text files to link *pymoo* and SWAT. The multi-objective optimization algorithm was executed for 1,000 generations, using 100 well-spaced reference directions obtained

Table 2
List of ERHIs Used in This Study

Category	Index[a]	Description
Magnitude of monthly water conditions	MA12—MA23	Mean monthly flows from January to December ($m^3\ s^{-1}$)
Magnitude and duration of annual extreme water conditions	DL1—DL5	Annual minimum with 1-, 3-, 7-, 30-, and 90-day moving average flow ($m^3\ s^{-1}$)
	DH1—DH5	Annual maximum with 1-, 3-, 7-, 30-, and 90-day moving average flow ($m^3\ s^{-1}$)
	ML17	Baseflow index based on the 7-day minimum flow
Timing of annual extreme water conditions	TL1	Julian day of annual minimum
	TH1	Julian day of annual maximum
Frequency and duration of high and low pulses	FL1	Mean low flow pulse count per water year ($year^{-1}$)
	DL16	Mean low flow pulse duration (days)
	FH1	Mean high flow pulse count per water year with a threshold equal to the 75th percentile of the entire flow record ($year^{-1}$)
	DH15	Mean high flow pulse duration with a threshold equal to the 75th percentile of the entire flow record (days)
Rate and frequency of water condition changes	RA1	Rise rate ($m^3\ s^{-1}\ d^{-1}$)
	RA3	Fall rate ($m^3\ s^{-1}\ d^{-1}$)
	RA8	Reversals ($year^{-1}$)
Magnificent seven	MAG1—MAG4	First four L–moments (mean, coefficient of variation, skewness, and kurtosis)
	MAG5	Autoregressive lag-one AR(1) correlation coefficient
	MAG6—MAG7	Amplitude and phase of the seasonal signal

[a]Index abbreviations for Indicators of Hydrologic Alteration (IHA) as presented by Olden and Poff (2003).

with the Riesz s-Energy method (Blank et al., 2021). This resulted in a total number of 100,000 model evaluations. Other U-NSGA-III parameters included the crossover probability, the distribution index for the Simulated Binary Crossover operator, the mutation probability, and the distribution index for the polynomial mutation operator, defined as 0.9, 10, 1/15 (i.e., the inverse of the number of the hydrological model calibration parameters), and 20, respectively. The MT-DREAM$_{(ZS)}$ algorithm was executed in MATLAB R2019b using the MT-DREAM$_{(ZS)}$ package developed by Vrugt (2016). The SWAT interface in Python was linked with MATLAB to compute the log-likelihood function. MT-DREAM$_{(ZS)}$ was executed using three Markov chains and five multi-try proposals for 10,000 generations (for a total of 150,000 model evaluations). Additional MT-DREAM$_{(ZS)}$ parameters were assigned the default values reported by Vrugt (2016). The calibration experiments were executed using up to 20 threads in parallel on a machine equipped with two Intel® Xeon® CPU E5-2640 Processors at 2.5 GHz with 64 GB RAM running Ubuntu 16.04.7 LTS.

3. Results and Discussion

3.1. Convergence of Multi-Objective and Bayesian Calibration Experiments

Convergence of the U-NSGA-III algorithm (*Experiment 2*, *Period 1*) was monitored using the Hypervolume Indicator (Auger et al., 2009), which started to show a steady behavior after 800 generations (i.e., 80,000 model evaluations). The first feasible solution (i.e., model simulation with all the selected ERHIs within ±30% relative error) was found after 4,800 model evaluations. The total computation time for the multi-objective calibration was 32.43 hr. Regarding the Bayesian calibration experiments, the MT-DREAM$_{(ZS)}$ algorithm converged after 5,400 generations (i.e., 81,000 model evaluations) for the *Experiment 1*, *Period 1*; 8,400 generations (i.e., 126,000 model evaluations) for the *Experiment 1, Period 2*; and 6,200 generations (i.e., 93,000 model evaluations) for the *Experiment 2, Period 2*. The total computation times for the Bayesian calibration experiments, which were simultaneously executed in the same machine, were 118.74 hr for *Experiment 1, Period 1*; 120.19 hr for *Experiment 1, Period 2*; and 118.97 hr for *Experiment 2, Period 2*. It is worth noting that Bayesian calibration using MCMC sampling had 50% more model evaluations than the multi-objective calibration. According to the results, Bayesian calibration convergence using prior knowledge was 26% faster than the non-informative case. This

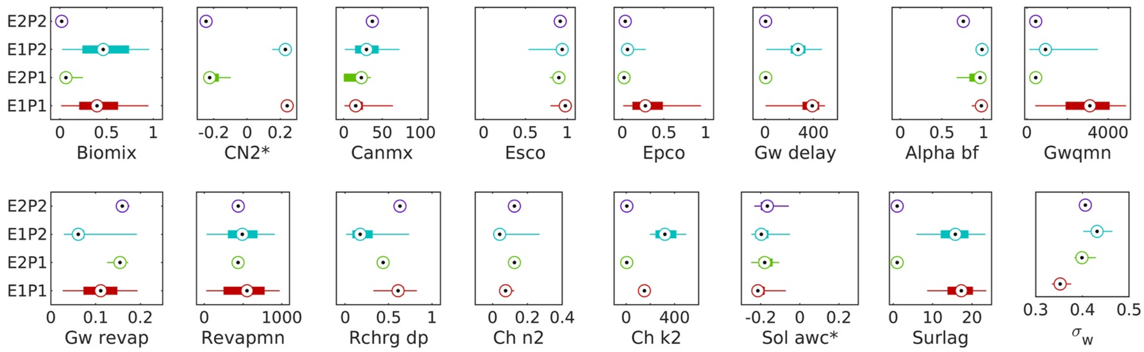

Figure 2. Distribution of model parameters obtained from multi-objective calibration (*Experiment 2, Period 1*—E2P1) and Bayesian parameter estimation (*Experiment 1, Period 1*—E1P1; *Experiment 1, Period 2*—E1P2; *Experiment 2, Period 2*—E2P2). Box and whisker plots represent the 50% and 95% confidence limits, respectively; points represent median parameter values. Parameter descriptions are reported in Table 1. *These parameters were calibrated using global multipliers.

faster convergence can be partially attributed to the further constrained search space for the informative case through the prior distribution.

3.2. Comparison Between Posterior Parameter Distributions Using Non-Informative Priors and Pareto-Optimal Results

The distribution of the hydrological and error model parameters considered in this study are presented in Figure 2 for each experiment and calibration period. *Under Experiment 1*, those distributions that were similar for both calibration periods corresponded mostly to parameters related to surface runoff generation and initial losses. Streamflow simulations using SWAT are usually very sensitive to most of these parameters, such as CN2, alpha bf, Sol awc, Esco, and Canmx (Khorashadi Zadeh et al., 2017). Since we selected two periods with the same length, and a similar number of high-frequency events and seasonality, it is not surprising that these parameters behaved consistently. Other parameters that usually present a low sensitivity and identifiability, such as Biomix, Revapmn, and Surlag, resulted in wide posterior parameter ranges for both periods. Meanwhile, distributions that showed a very distinct behavior corresponded mostly to soil- and groundwater-related parameters: Epco, Gwqmn, Gw revap (which represents water movement from the shallow aquifer to the overlying unsaturated zone), Rchrg dp, and Ch k2 (associated with transmission losses in the main channel). The standard deviation of the transformed autocorrelated residuals σ_W for *period 2* was greater than *period 1*. Therefore, results showed a higher total uncertainty in streamflow during *period 2* compared to *period 1*. These differences in hydrological and error model parameter distributions can be explained by the time-varying nature of model parameters driven by non-stationarity effects from input weather data and changes in land use (Xiong et al., 2019).

Regarding *Experiment 2*, parameter distributions were more consistent across the two calibration periods compared with *Experiment 1*. This behavior revealed the strong influence of the prior distribution on the Bayesian calibration results for *period 2*. However, some parameters showed important differences, including Canmx (related to surface water interception) and Alpha bf (related to the shape of recession curves). These differences were originated from the contribution of new data under *period 2*. Excepting the global multiplier for Sol awc, the posterior parameter distributions from Bayesian calibration (i.e., *period 2*) were narrower compared to the multi-objective calibration distributions (i.e., *period 1*) for *Experiment 2*. This reduction in parametric uncertainty resulted from the assimilation of a greater amount of information (i.e., *Experiment 2, period 2* ultimately used information from both *periods 1* and *2*).

In general, *Experiment 2* distributions were drastically narrower than *Experiment 1* distributions. Bayesian calibration with uniform non-informative priors (especially *Experiment 1, period 1*) resulted in poorly-informative posteriors. These posteriors included distributions for Biomix, groundwater parameters such as Gwqmn, Gw revap, and Revapmn, and Surlag. Poorly-informative posteriors are related to the equifinality problem, indicating that different parameter combinations yield similar outputs (Beven, 2006). This may apply to the groundwater parameters interacting with each other and Surlag, which represents water storage. Likewise, non-informative posteriors are indicative of a low sensitivity of the modeling outputs to changes in those particular parameters, which can explain Biomix behavior. Nevertheless, *Experiment 1*, *period 1* attained the lowest σ_W, which can offset a lower total uncertainty in streamflow predictions.

The reduction in parametric uncertainty attained by *Experiment 2* compared to *Experiment 1* was mainly driven by the constraints to ERHIs simulation accuracy, introduced by the proposed multi-objective calibration strategy. Another effect of the ERHIs constraints was the difference in the actual range of values for certain parameters. For instance, *Experiment 2* results indicated that the global multiplier perturbing initial CN2 values was around −25%, whereas *Experiment 1* results indicated the opposite (i.e., around +25%). Higher CN2 makes the watershed more impervious, resulting in higher runoff generation (and therefore higher streamflows). Similarly, while *Experiment 2* resulted in Surlag values close to unity, *Experiment 1* resulted in Surlag values mostly between 10 and 20. Surlag controls the fraction of total available water that enters the main channel on a daily basis; higher Surlag values result in higher fractions. Likewise, *Experiment 2* generated Gw delay close to zero, whereas *Experiment 1* yielded values greater than 200 days. Gw delay represents the time that water spends in the vadose zone before becoming shallow aquifer recharge. Gw delay ultimately affects groundwater contributions to the main channel, which also impacts low-flow dynamics.

3.3. Performance of Uncertainty Quantification of Daily Streamflows

Uncertainty quantification performance of streamflow predictions is presented in Figure 3 for each experiment and calibration period. As expected, the streamflow prediction bounds resulted in a lower parametric uncertainty for *Experiment 2*, which was consistent with the narrower parameter distributions presented in Figure 2 and the effects of the ERHIs' optimization constraints. In fact, parametric uncertainty for *Experiment 2*, *Period 2* (i.e., Bayesian parameter estimation using multi-objective calibration prior) was drastically lower compared to the other cases (~72% narrower compared to *Experiment 1* for the same period). Regarding total uncertainty, the hydrographs in Figure 3 (first column from left to right) reveal wider limits for low flows in *Experiment 1* (i.e., uniform priors), and wider limits for high flows in *Experiment 2*.

Scatter plots comparing average streamflow predictions with observations are presented in Figure 3, the second column. In these plots, we also report the corresponding root-mean-square error (RMSE). In both time periods, *Experiment 2* showed a lower dispersion than *Experiment 1*, with very similar RMSE values. The difference was greater in *Period 2*, which emphasizes the effect of the ERHIs constraining through the Pareto-optimal prior distributions. It is worth noting that the scatter plot for *Experiment 2*, *Period 1* shows streamflow over-prediction. For this reason, absolute differences in high flows are more noticeable in Figure 3b (first column) compared to the other cases, where streamflows were mostly underpredicted.

In terms of reliability, PQQ plots indicated that *Experiment 1* (Figure 3 row a, third column) slightly over-estimated predictive uncertainty during *Period 1* (curve above the 1:1 diagonal at the lower-left corner and below the same line at the upper-right corner). Meanwhile, Pareto-optimal solutions (*Experiment 2*, Figure 3 row b, third column) over-predicted streamflows under the same calibration period (curve was mostly below the 1:1 diagonal). The latter occurred because total uncertainty was estimated using the σ_w obtained from the residuals of each Pareto-optimal solution. Since the multi-objective calibration strategy did not consider the additive error term included in the Bayesian calibration framework, the streamflow predictions resulted in higher values by explicitly adding the residual innovations to the Pareto-optimal simulations. Regarding *Period 2*, PQQ plots indicated excellent reliability for the streamflow predictions in both experiments (Figure 3 rows c and d, third column). When analyzing the scatter and PQQ plots together, the skewed nature of the predictions over time is evident. While the total uncertainty bound is quite wide, most predictions concentrate around lower values.

When statistically comparing the observed quantiles from both experiments with a 95% confidence level, we did not find evidence of a significant difference in the reliability measure under *Period 2* ($p = 0.070$). For *Period 1*, the reliability was significantly different for both experiments ($p = 1.7 \times 10^{-5}$), with *Experiment 1* presenting a better result overall (i.e., 10% vs. 13%, see Figure 3, fourth column). Regarding precision, no evidence of a significant difference was found between the two experiments under *Period 1* ($p = 0.11$). Meanwhile, *Experiment 2* resulted in a higher precision than *Experiment 1* under *Period 2* (58% vs. 66%, $p = 1.3 \times 10^{-8}$). Regarding bias in the long-term water balance, no significant difference was found for *Period 1* ($p = 0.40$), whereas for *Period 2*, *Experiment 2* attained a significantly lower bias compared with *Experiment 1* (4% vs. 6%, $p = 0.0059$). In other

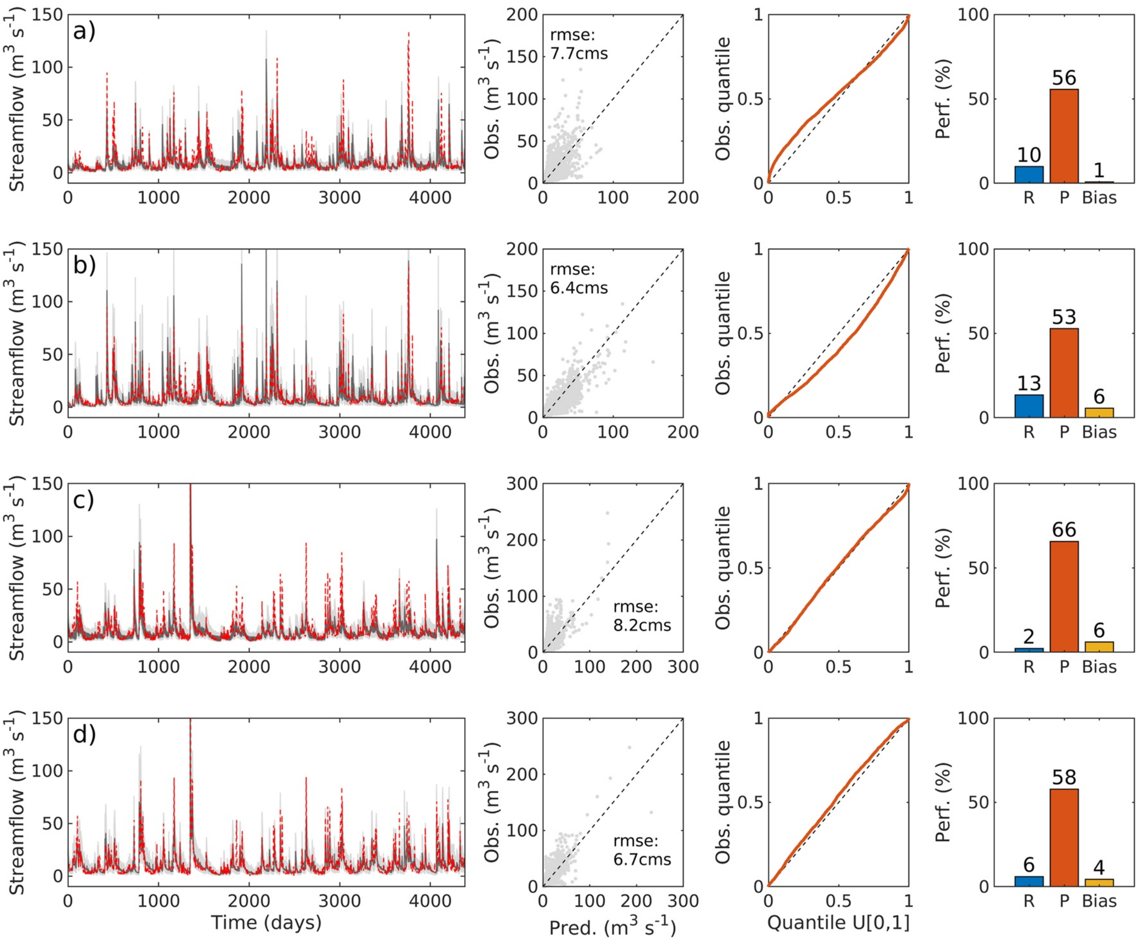

Figure 3. Uncertainty quantification performance using multi-objective calibration and Bayesian parameter estimation. The hydrographs (first column from left to right) represent the 95% prediction bounds for streamflow; light gray is for total uncertainty, dark gray is for parametric uncertainty, red line are observations. The second column presents scatter plots comparing the average model predictions and observations, reporting the root-mean-square error (rmse). The third column presents the corresponding quantile-quantile plots (PQQ) using a standard uniform distribution. The fourth column presents the overall performance indices for reliability (R), precision (P), and Bias. (a) *Experiment 1, Period 1*; (b) *Experiment 2, Period 1;* (c) *Experiment 1, Period 2;* (d) *Experiment 2, Period 2*.

words, Bayesian calibration using multi-objective priors resulted in lower total uncertainty in streamflow predictions with lower bias in long-term water balance and no significant loss in reliability.

3.4. Performance of Uncertainty Quantification of ERHIs

Figure 4 shows the distribution of the relative errors for the 32 IHA and Magnificent seven indices selected in this study and reported in Table 2. This figure presents the parametric predictive distributions for the multi-objective calibration under *Period 1* because σ_w was not directly calibrated in this case. As expected, all Pareto-optimal predictions fell within the ±30% relative error range due to the optimization constraints in ERHIs accuracy. However, note that only about 50% of the ERHIs computed on the observed streamflows fell within the Pareto-optimal predictive distributions. The situation was not better for the non-informative Bayesian calibration predictions under the same period. Meanwhile, the non-informative case under *Period 2* contained 64% of the observed ERHIs, against 56% for *Experiment 2* (see columns 3 and 4 in Table 3). Furthermore, all the predictive distributions for indices under the "frequency and duration of high and low pulses" category did not contain observed ERHIs under *Period 2*. The same situation was observed for DL1 (i.e., annual daily minimum), RA8 (i.e., reversals), and MAG3 (i.e., L-skewness). Distributions obtained with the Bayesian calibration using prior knowledge were particularly limited in the prediction of low-flow-related ERHIs across all the streamflow regime facets (i.e., magnitude, duration, frequency, rate of change, and timing). This limitation is likely a consequence

10.1029/2021WR031104

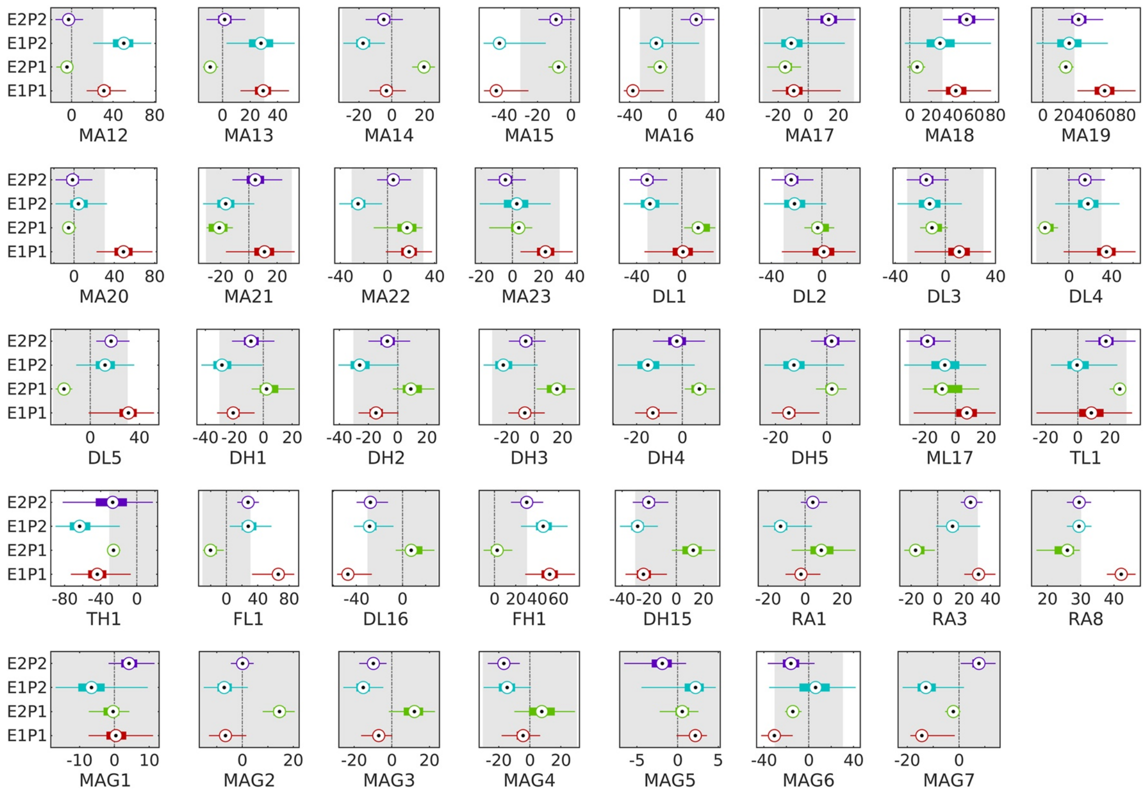

Figure 4. Distribution of relative errors of the selected ecologically relevant hydrologic indices (ERHIs) using multi-objective calibration (*Experiment 2, Period 1*—E2P1) and Bayesian parameter estimation (*Experiment 1, Period 1*—E1P1; *Experiment 1, Period 2*—E1P2; *Experiment 2, Period 2*—E2P2). Box and whiskers represent the 50% and 95% confidence limits, respectively; points represent median relative error values; the vertical dotted line represents the zero axis, the gray area represent the nominal ±30% ERHI uncertainty. Index abbreviations are reported in Table 2.

of structural inadequacies in the SWAT model (Tan et al., 2020). For instance, end-of-summer monthly flows (i.e., July-August months) were drastically over-predicted due to a lagged simulation in the timing of annual minima (i.e., a delayed transition toward the fall season). Therefore, modeling inaccuracies affect intermediate- and base-flows (e.g., oversimplified unsaturated zone model structure). Regarding the frequency and duration of extreme pulses, temporal and spatial uncertainty in input weather data with missing events are likely the main factor driving poor ERHIs predictions (Tan et al., 2020). Other factors affecting model performance include limitations in the snow melt and evapotranspiration modules (Duan et al., 2018; Valipour et al., 2017). Still, it is worth noting that Bayesian calibration using prior knowledge resulted in narrower predictive distributions compared with the non-informative case under *Period 2* (see columns 7 and 8 in Table 3), which explains the tradeoff between precision and reliability.

The aforementioned limitations in the accuracy of ERHIs predictive distributions were not surprising since we were trying to fit multiple streamflow facets simultaneously. Moreover, total uncertainty propagation through ERHIs computation affects precision and impacts the overall central tendency of the predictive distributions. Instead, our main interest was to limit the bias and the overall dispersion within the nominal ±30% relative error range. When comparing the median ERHIs relative errors against the acceptability threshold, we found that both experiments under *Period 2* yielded 90% of the median ERHIs within the acceptability threshold. For *Experiment 2*, ERHIs with the median outside the acceptability range include MA18-19 (i.e., monthly flows for July and August, summer flows), DL1, and FH1 (i.e., frequency of high flow pulses). For *Experiment 1*, ERHIs with the median outside the acceptability range include MA12 and MA15 (i.e., monthly flows for January and April), TH1 (i.e., Julian day of annual maximum), and FH1.

In general, *Experiment 2* exhibited a better performance in ERHIs prediction compared with *Experiment 1* because the overall precision had an average increase of ~32%. Regarding bias, ~59% ERHIs exhibited lower (better) values for *Experiment 2* compared with the non-informative case. It is worth noting that ERHIs predictions

Table 3
Performance of Predictive Distributions of ERHIs Obtained Under Period 2

ERHI category	ERHI	Dist. contains observed ERHI		Median within ±30% range		Precision (%)		Bias (%)	
		Exp. 1	Exp. 2	Exp. 1	Exp. 2	Exp. 1	Exp. 2	Exp. 1	Exp. 2
Magnitude of monthly water conditions	MA12		✓		✓	14.4	6.8	49.6	−2.7
	MA13		✓	✓	✓	12.8	7.2	27.9	1.8
	MA14		✓	✓	✓	6.4	5.8	−17.4	−4.7
	MA15		✓		✓	7.7	5.7	−41.8	−8.8
	MA16	✓		✓	✓	12.6	8.1	−13.3	22.7
	MA17	✓	✓	✓	✓	13.1	8.4	−9.6	13.9
	MA18	✓		✓		20.2	12.1	29.9	53.5
	MA19	✓		✓		17.5	10.9	26.1	34.9
	MA20	✓	✓	✓	✓	12.9	9.4	5.2	−1.0
	MA21	✓	✓	✓	✓	9.3	9.0	−15.9	4.9
	MA22		✓	✓	✓	9.2	7.4	−24.5	5.0
	MA23	✓	✓	✓	✓	11.7	6.3	2.2	−4.3
Magnitude and duration of annual extreme water conditions (mean daily flow)	DL1			✓		12.3	8.3	−28.0	−30.5
	DL2	✓		✓	✓	12.6	8.3	−21.6	−24.0
	DL3	✓	✓	✓	✓	13.1	8.4	−12.5	−14.9
	DL4	✓	✓	✓	✓	15.0	8.8	17.9	15.2
	DL5	✓		✓	✓	11.9	7.0	12.2	17.3
	DH1		✓	✓	✓	10.6	7.6	−27.3	−8.0
	DH2	✓	✓	✓	✓	10.6	7.3	−24.8	−6.7
	DH3	✓	✓	✓	✓	9.9	6.7	−21.3	−5.9
	DH4	✓	✓	✓	✓	8.1	5.8	−14.4	−2.2
	DH5	✓	✓	✓	✓	7.1	4.5	−12.4	2.0
	ML17	✓		✓	✓	13.7	7.3	−6.6	−18.1
Timing of annual extreme water conditions	TL1	✓		✓	✓	11.2	8.1	0.8	18.0
	TH1		✓		✓	34.7	33.5	−59.2	−27.0
Frequency and duration of high and low pulses	FL1			✓	✓	13.8	7.2	29.0	27.6
	DL16			✓	✓	8.8	6.9	−27.6	−27.8
	FH1					11.2	8.0	46.0	30.5
	DH15			✓	✓	7.2	6.7	−28.5	−20.1
Rate and frequency of water condition changes	RA1	✓	✓	✓	✓	6.6	3.6	−12.6	4.3
	RA3	✓		✓	✓	8.1	4.2	12.1	25.1
	RA8			✓	✓	1.9	1.8	29.5	29.5
Magnificent seven	MAG1	✓	✓	✓	✓	6.4	3.4	−6.2	4.3
	MAG2	✓	✓	✓	✓	4.4	2.3	−6.9	0.1
	MAG3			✓	✓	5.4	3.7	−15.2	−9.8
	MAG4	✓		✓	✓	7.8	5.2	−14.7	−16.7
	MAG5	✓	✓	✓	✓	2.5	2.0	1.7	−2.1
	MAG6	✓	✓	✓	✓	20.0	10.7	4.9	−15.7
	MAG7	✓		✓	✓	5.9	3.4	−12.2	7.5

Note. Reliability was evaluated by identifying whether the distributions contained the ERHIs from observations, and whether the median of the distributions was within the ±30% relative error range. Exp. 1 = *Experiment 1*; Exp. 2 = *Experiment 2*. ERHI, ecologically relevant hydrologic indices.

behaved differently depending on the calibration approach, as revealed by the differences in some posterior model parameter distributions and related measures (especially bias) such as the magnitude of monthly flows (i.e., MA12—MA22), the duration of high flow conditions (i.e., DH1—DH5), and timing of extreme flows (i.e., TL1, TH1). For instance, *Experiment 1* increased runoff generation and transmission losses and regulated groundwater return to the main channels. Meanwhile, *Experiment 2* decreased runoff generation, increased water storage, and increased groundwater and lateral flow contributions. As a result, non-informative Bayesian calibration generated a lower bias for low flow ERHIs than *Experiment 2*. Likewise, *Experiment 2* resulted in a lower bias for high flow ERHIs compared with *Experiment 1*. The main factors contributing to these differences were the constraints inherited through the informative prior distribution and the likelihood function (particularly, the transformational approach selected to address heteroscedasticity). Deciding for a preferred calibration option ultimately requires a better understanding and assessment of the internal modeling processes using observations for variables of other water cycle components such as groundwater levels, soil moisture, and evapotranspiration. Also, a better characterization of other uncertainty sources is required.

4. Conclusions

In this study, we successfully linked multi-objective calibration with Bayesian parameter estimation for ERHIs uncertainty quantification. We achieved this by using a multivariate prior distribution of model parameters based on near-optimal Pareto solutions. The connection allowed the transfer of predefined ERHIs accuracy constraints into the overall Bayesian inference framework. The proposed method can be applied to any hydrological model, study area, and subsets of hydrologic signatures. The main advantage of the developed strategy is the use of multiple sources of information contained within a single continuously measured quantity for improving streamflow regimes prediction. Other benefits—compared with Bayesian calibration using uniform priors—included: (a) 26% faster convergence to a stationary multivariate posterior distribution of model parameters using the MT-DREAM$_{(ZS)}$ algorithm, (b) drastic reduction (72%) of parametric uncertainty in streamflow predictions, (c) lower average coefficient of variance (a.k.a., precision, 58% vs. 66%, $p < 0.05$) in streamflow predictive uncertainty with lower bias (4% vs. 6%, $p < 0.05$) in the long-term water balance and no significant loss in reliability (6% vs. 2%, $p = 0.07$), and (d) 32% increase in precision of ERHIs' prediction.

A disadvantage of the proposed two-stage framework is the increased computational cost since results of multi-objective model calibration must be obtained before executing the Bayesian parameter estimation algorithm. Therefore, further work is needed to develop a more efficient multi-objective Bayesian calibration framework, including the use of surrogate models. Meanwhile, it is worth noting that using prior knowledge had an important effect on how the hydrological model internally simulated surface runoff, interflow, and baseflow, as reflected by differences in related parameter posteriors. For example, when using uniform priors, the chosen likelihood function favored the representation of low flows at the expense of high flows. Meanwhile, multi-objective calibration priors resulted in the opposite outcomes, yielding better results for high flows and behaving rather poorly for low flows. Therefore, further work is needed for understanding the tradeoffs between priors and likelihood functions for improving streamflow and ERHIs prediction. Ultimately, deciding on a particular modeling path depends on a better characterization of uncertainty sources (e.g., weather data, land use change, non-stationarity effects) and the incorporation of additional information, including remote sensing products, for other water cycle components (e.g., evapotranspiration, soil moisture, groundwater levels, leaf area index). Also, persistent limitations in ERHIs prediction for low flows, rate of change, and frequency and duration of high and low pulses require the revision of modeling workflows and structures to describe the ecohydrological behavior of freshwater ecosystems better. In particular, additional work is needed to evaluate the impacts of different potential evapotranspiration methods and surface-groundwater interaction routines in the prediction of ERHIs.

Data Availability Statement

The codes and model files used in this study are available at: https://github.com/jshernandezs/swat-pytools. Model calibration outputs and scripts for reproducing the figures presented in this article are available at: http://www.hydroshare.org/resource/02b1eaa825614f02a3ad07b94eac1e95.

Acknowledgments

This work was supported in part by the USDA National Institute of Food and Agriculture, Hatch project 1019654. J. Sebastian Hernandez-Suarez is grateful to Fulbright and the Colombian Ministry of Science, Technology, and Innovation for the financial support of his doctoral studies. We thank Dr. Jasper Vrugt for providing the MT-DREAM package in MATLAB.

References

Addor, N., Nearing, G., Prieto, C., Newman, A. J., Le Vine, N., & Clark, M. P. (2018). A ranking of hydrological signatures based on their predictability in space. *Water Resources Research*, *54*(11), 8792–8812. https://doi.org/10.1029/2018WR022606

Almeida, S., Bulygina, N., McIntyre, N., Wagener, T., & Buytaert, W. (2013). Improving parameter priors for data-scarce estimation problems. *Water Resources Research*, *49*(9), 6090–6095. https://doi.org/10.1002/wrcr.20437

Almeida, S., Le Vine, N., McIntyre, N., Wagener, T., & Buytaert, W. (2016). Accounting for dependencies in regionalized signatures for predictions in ungauged catchments. *Hydrology and Earth System Sciences*, *20*(2), 887–901. https://doi.org/10.5194/hess-20-887-2016

Ammann, L., Fenicia, F., & Reichert, P. (2019). A likelihood framework for deterministic hydrological models and the importance of non-stationary autocorrelation. *Hydrology and Earth System Sciences*, *23*(4), 2147–2172. https://doi.org/10.5194/hess-23-2147-2019

Archfield, S. A., Kennen, J. G., Carlisle, D. M., & Wolock, D. M. (2014). An objective and parsimonious approach for classifying natural flow regimes at a continental scale. *River Research and Applications*, *30*(9), 1166–1183. https://doi.org/10.1002/rra.2710

Arnold, J. G., Moriasi, D. N., Gassman, P. W., Abbaspour, K. C., White, M. J., Srinivasan, R., et al. (2012). Swat: Model use, calibration, and validation. *Transactions of the ASABE*, *55*(4), 1491–1508. https://doi.org/10.13031/2013.42256

Auger, A., Bader, J., Brockhoff, D., & Zitzler, E. (2009). Theory of the hypervolume indicator: Optimal μ-distributions and the choice of the reference point. In *Proceedings of the tenth ACM SIGEVO workshop on foundations of genetic algorithms—FOGA '09* (pp. 87–102). https://doi.org/10.1145/1527125.1527138

Bejarano, M. D., Sordo-Ward, A., Gabriel-Martin, I., & Garrote, L. (2019). Tradeoff between economic and environmental costs and benefits of hydropower production at run-of-river-diversion schemes under different environmental flows scenarios. *Journal of Hydrology*, *572*, 790–804. https://doi.org/10.1016/j.jhydrol.2019.03.048

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

Beven, K., & Binley, A. (2014). GLUE: 20 years on. *Hydrological Processes*, *28*(24), 5897–5918. https://doi.org/10.1002/hyp.10082

Beven, K., & Smith, P. (2015). Concepts of information content and likelihood in parameter calibration for hydrological simulation models. *Journal of Hydrologic Engineering*, *20*(1), 1–15. https://doi.org/10.1061/(asce)he.1943-5584.0000991

Beven, K., Smith, P. J., & Wood, A. (2011). On the color and spin of epistemic error (and what we might do about it). *Hydrology and Earth System Sciences*, *15*(10), 3123–3133. https://doi.org/10.5194/hess-15-3123-2011

Blank, J., & Deb, K. (2020). Pymoo: Multi-objective optimization in python. *IEEE Access*, *8*, 89497–89509. https://doi.org/10.1109/ACCESS.2020.2990567

Blank, J., Deb, K., Dhebar, Y., Bandaru, S., & Seada, H. (2021). Generating well-spaced points on a unit simplex for evolutionary many-objective optimization. *IEEE Transactions on Evolutionary Computation*, *25*(1), 48–60. https://doi.org/10.1109/TEVC.2020.2992387

Blazkova, S., & Beven, K. (2009). A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic. *Water Resources Research*, *45*(12), 1–12. https://doi.org/10.1029/2007WR006726

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B*, *26*(2), 211–252. https://doi.org/10.2307/2287791

Carlisle, D. M., Grantham, T. E., Eng, K., & Wolock, D. M. (2017). Biological relevance of streamflow metrics: Regional and national perspectives. *Freshwater Science*, *36*(4), 927–940. https://doi.org/10.1086/694913

Chen, Q., Chen, H., Wang, J., Zhao, Y., Chen, J., & Xu, C. (2019). Impacts of climate change and land-use change on hydrological extremes in the Jinsha River Basin. *Water*, *11*(7), 1398. https://doi.org/10.3390/w11071398

Duan, Y., Liu, T., Meng, F., Luo, M., Frankl, A., de Maeyer, P., et al. (2018). Inclusion of modified snow melting and flood processes in the SWAT model. *Water*, *10*(12), 1715. https://doi.org/10.3390/w10121715

Efstratiadis, A., & Koutsoyiannis, D. (2010). One decade of multi-objective calibration approaches in hydrological modelling: A review. *Hydrological Sciences Journal*, *55*(1), 58–78. https://doi.org/10.1080/02626660903526292

Einheuser, M. D., Nejadhashemi, A. P., Sowa, S. P., Wang, L., Hamaamin, Y. A., & Woznicki, S. A. (2012). Modeling the effects of conservation practices on stream health. *Science of the Total Environment*, *435–436*, 380–391. https://doi.org/10.1016/j.scitotenv.2012.07.033

Eng, K., Grantham, T. E., Carlisle, D. M., & Wolock, D. M. (2017). Predictability and selection of hydrologic metrics in riverine ecohydrology. *Freshwater Science*, *36*(4), 915–926. https://doi.org/10.1086/694912

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., & Savenije, H. H. G. (2013). A framework to assess the realism of model structures using hydrological signatures. *Hydrology and Earth System Sciences*, *17*(5), 1893–1912. https://doi.org/10.5194/hess-17-1893-2013

Evin, G., Thyer, M., Kavetski, D., McInerney, D., & Kuczera, G. (2014). Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity. *Water Resources Research*, *50*(3), 2350–2375. https://doi.org/10.1002/2013WR014185

Farmer, W. H., & Vogel, R. M. (2016). On the deterministic and stochastic use of hydrologic models. *Water Resources Research*, *52*(7), 5619–5633. https://doi.org/10.1002/2016WR019129

Fenicia, F., Kavetski, D., Reichert, P., & Albert, C. (2018). Signature-domain calibration of hydrological models using approximate Bayesian computation: Empirical analysis of fundamental properties. *Water Resources Research*, *54*(6), 3958–3987. https://doi.org/10.1002/2017WR021616

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 15–51. https://doi.org/10.1214/ss/1177011136

George, R., McManamay, R., Perry, D., Sabo, J., & Ruddell, B. L. (2021). Indicators of hydro-ecological alteration for the rivers of the United States. *Ecological Indicators*, *120*, 106908. https://doi.org/10.1016/j.ecolind.2020.106908

Guo, Y., Zhang, Y., Zhang, L., & Wang, Z. (2021). Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: A comprehensive review. *Wiley Interdisciplinary Reviews: Water*, *8*(1), 1–32. https://doi.org/10.1002/wat2.1487

Gupta, A., Govindaraju, R. S., Morbidelli, R., & Corradini, C. (2022). The role of prior probabilities on parameter estimation in hydrological models. *Water Resources Research*, *58*(5), e2021WR031291. https://doi.org/10.1029/2021WR031291

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

Gupta, H. V., Wagener, T., & Liu, Y. (2008). Reconciling theory with observations: Elements of a diagnostic approach to model evaluation. *Hydrological Processes*, *22*(18), 3802–3813. https://doi.org/10.1002/hyp.6989

Hallouin, T., Bruen, M., & O'Loughlin, F. E. (2020). Calibration of hydrological models for ecologically relevant streamflow predictions: A trade-off between fitting well to data and estimating consistent parameter sets? *Hydrology and Earth System Sciences*, *24*(3), 1031–1054. https://doi.org/10.5194/hess-24-1031-2020

Herman, M. R., Nejadhashemi, A. P., Abouali, M., Hernandez-Suarez, J. S., Daneshvar, F., Zhang, Z., et al. (2018). Evaluating the role of evapotranspiration remote sensing data in improving hydrological modeling predictability. *Journal of Hydrology*, *556*, 39–49. https://doi.org/10.1016/j.jhydrol.2017.11.009

Hernandez-Suarez, J. S., & Nejadhashemi, A. P. (2018). A review of macroinvertebrate- and fish-based stream health modeling techniques. *Ecohydrology*, *11*(8), e2022. https://doi.org/10.1002/eco.2022

Hernandez-Suarez, J. S., Nejadhashemi, A. P., & Deb, K. (2021). A novel multi-objective model calibration method for ecohydrological applications. *Environmental Modelling & Software*, *144*, 105161. https://doi.org/10.1016/j.envsoft.2021.105161

Hernandez-Suarez, J. S., Nejadhashemi, A. P., Kropp, I. M., Abouali, M., Zhang, Z., & Deb, K. (2018). Evaluation of the impacts of hydrologic model calibration methods on predictability of ecologically-relevant hydrologic indices. *Journal of Hydrology*, *564*, 758–772. https://doi.org/10.1016/j.jhydrol.2018.07.056

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., et al. (2014). Process consistency in models: The importance of system signatures, expert knowledge, and process complexity. *Water Resources Research*, *50*(9), 7445–7469. https://doi.org/10.1002/2014WR015484

Jehn, F. U., Chamorro, A., Houska, T., & Breuer, L. (2019). Trade-offs between parameter constraints and model realism: A case study. *Scientific Reports*, *9*(1), 1–12. https://doi.org/10.1038/s41598-019-46963-6

Kavetski, D., Fenicia, F., Reichert, P., & Albert, C. (2018). Signature-domain calibration of hydrological models using approximate Bayesian computation: Theory and comparison to existing applications. *Water Resources Research*, *54*(6), 4059–4083. https://doi.org/10.1002/2017WR020528

Kennard, M. J., Mackay, S. J., Pusey, B. J., Olden, J. D., & Marsh, N. (2010). Quantifying uncertainty in estimation of hydrologic metrics for ecohydrological studies. *River Research and Applications*, *26*(2), 137–156. https://doi.org/10.1002/rra.1249

Khorashadi Zadeh, F., Nossent, J., Sarrazin, F., Pianosi, F., van Griensven, A., Wagener, T., & Bauwens, W. (2017). Comparison of variance-based and moment-independent global sensitivity analysis approaches by application to the SWAT model. *Environmental Modelling & Software*, *91*, 210–222. https://doi.org/10.1016/j.envsoft.2017.02.001

Kiesel, J., Guse, B., Pfannerstill, M., Kakouei, K., Jähnig, S. C., & Fohrer, N. (2017). Improving hydrological model optimization for riverine species. *Ecological Indicators*, *80*, 376–385. https://doi.org/10.1016/j.ecolind.2017.04.032

Kiesel, J., Kakouei, K., Guse, B., Fohrer, N., & Jähnig, S. C. (2020). When is a hydrological model sufficiently calibrated to depict flow preferences of riverine species? *Ecohydrology*, *13*(3), 1–15. https://doi.org/10.1002/eco.2193

Kollat, J. B., Reed, P. M., & Wagener, T. (2012). When are multiobjective calibration trade-offs in hydrologic models meaningful? *Water Resources Research*, *48*(3), 1–19. https://doi.org/10.1029/2011WR011534

Krause, P., Boyle, D. P., & Base, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences*, (5), 89–97. https://doi.org/10.5194/adgeo-5-89-2005

Laloy, E., & Vrugt, J. A. (2012). High-dimensional posterior exploration of hydrologic models using multiple-try DREAM (ZS) and high-performance computing. *Water Resources Research*, *48*(1), 1–18. https://doi.org/10.1029/2011WR010608

Mazor, R. D., May, J. T., Sengupta, A., McCune, K. S., Bledsoe, B. P., & Stein, E. D. (2018). Tools for managing hydrologic alteration on a regional scale: Setting targets to protect stream health. *Freshwater Biology*, *63*(8), 786–803. https://doi.org/10.1111/fwb.13062

McInerney, D., Thyer, M., Kavetski, D., Bennett, B., Lerat, J., Gibbs, M., & Kuczera, G. (2018). A simplified approach to produce probabilistic hydrological model predictions. *Environmental Modelling & Software*, *109*, 306–314. https://doi.org/10.1016/j.envsoft.2018.07.001

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

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, *21*(2), 239. https://doi.org/10.2307/1268522

McKay, S. K., Theiling, C. H., & Dougherty, M. P. (2019). Comparing outcomes from competing models assessing environmental flows in the Minnesota River Basin. *Ecological Engineering X*, *4*, 100014. https://doi.org/10.1016/j.ecoena.2019.100014

McMillan, H. K. (2020a). Linking hydrologic signatures to hydrologic processes: A review. *Hydrological Processes*, *34*(6), 1393–1409. https://doi.org/10.1002/hyp.13632

McMillan, H. K. (2020b). A review of hydrologic signatures and their applications. *Wiley Interdisciplinary Reviews: Water*, *8*(1), e1499. https://doi.org/10.1002/wat2.1499

Moges, E., Demissie, Y., Larsen, L., & Yassin, F. (2021). Review: Sources of hydrological model uncertainties and advances in their analysis. *Water*, *13*(1), 1–23. https://doi.org/10.3390/w13010028

Monteith, J. L. (1965). Evaporation and environment. *Symposia of the Society for Experimental Biology*, *19*, 205–234.

Neitsch, S. L., Arnold, J. G., Kiniry, J. R., & Williams, J. R. (2011). Soil and water assessment tool theoretical documentation version 2009 Texas water resources institute. Retrieved from http://swat.tamu.edu/media/99192/swat2009-theory.pdf

Nijzink, R. C., Almeida, S., Pechlivanidis, I. G., Capell, R., Gustafssons, D., Arheimer, B., et al. (2018). Constraining conceptual hydrological models with multiple information sources. *Water Resources Research*, *54*(10), 8332–8362. https://doi.org/10.1029/2017WR021895

NOAA-NCEI (2020). Climate data online data tools. Retrieved from https://www.ncdc.noaa.gov/cdo-web/datatools/

Olden, J. D., & Poff, N. L. (2003). Redundancy and the choice of hydrologic indices for characterizing streamflow regimes. *River Research and Applications*, *19*(2), 101–121. https://doi.org/10.1002/rra.700

Parker, S. R., Adams, S. K., Lammers, R. W., Stein, E. D., & Bledsoe, B. P. (2019). Targeted hydrologic model calibration to improve prediction of ecologically-relevant flow metrics. *Journal of Hydrology*, *573*, 546–556. https://doi.org/10.1016/j.jhydrol.2019.03.081

Poff, N. L., Richter, B. D., Arthington, A. H., Bunn, S. E., Naiman, R. J., Kendy, E., et al. (2010). The ecological limits of hydrologic alteration (ELOHA): A new framework for developing regional environmental flow standards. *Freshwater Biology*, *55*(1), 147–170. https://doi.org/10.1111/j.1365-2427.2009.02204.x

Pool, S., Vis, M. J. P. P., Knight, R. R., & Seibert, J. (2017). Streamflow characteristics from modeled runoff time series—Importance of calibration criteria selection. *Hydrology and Earth System Sciences*, *21*(11), 5443–5457. https://doi.org/10.5194/hess-21-5443-2017

Prieto, C., Le Vine, N., Kavetski, D., García, E., & Medina, R. (2019). Flow prediction in ungauged catchments using probabilistic random forests regionalization and new statistical adequacy tests. *Water Resources Research*, *55*(5), 4364–4392. https://doi.org/10.1029/2018WR023254

Reichert, P., & Schuwirth, N. (2012). Linking statistical bias description to multiobjective model calibration. *Water Resources Research*, *48*(9), 1–20. https://doi.org/10.1029/2011WR011391

Richter, B. D., Baumgartner, J., Wigington, R., & Braun, D. (1997). How much water does a river need? *Freshwater Biology*, *37*(1), 231–249. https://doi.org/10.1046/j.1365-2427.1997.00153.x

Sadegh, M., Shakeri Majd, M., Hernandez, J., & Haghighi, A. T. (2018). The quest for hydrological signatures: Effects of data transformation on Bayesian inference of watershed models. *Water Resources Management*, *32*(5), 1867–1881. https://doi.org/10.1007/s11269-018-1908-6

Sadegh, M., & Vrugt, J. A. (2014). Approximate Bayesian computation using Markov chain Monte Carlo simulation. *Water Resources Research*, *10*(2), 6767–6787. https://doi.org/10.1002/2014WR015386.Received

Sadegh, M., Vrugt, J. A., Xu, C., & Volpi, E. (2015). The stationarity paradigm revisited: Hypothesis testing using diagnostics, summary metrics, and DREAM (ABC). *Water Resources Research*, *51*(11), 9207–9231. https://doi.org/10.1002/2014WR016805

Schoups, G., & Vrugt, J. A. (2010). A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resources Research*, *46*(10), 1–17. https://doi.org/10.1029/2009WR008933

Seada, H., & Deb, K. (2016). A unified evolutionary optimization procedure for single, multiple, and many objectives. *IEEE Transactions on Evolutionary Computation*, *20*(3), 358–369. https://doi.org/10.1109/TEVC.2015.2459718

Sengupta, A., Adams, S. K., Bledsoe, B. P., Stein, E. D., McCune, K. S., & Mazor, R. D. (2018). Tools for managing hydrologic alteration on a regional scale: Estimating changes in flow characteristics at ungauged sites. *Freshwater Biology*, *63*(8), 769–785. https://doi.org/10.1111/fwb.13074

Shafii, M., & de Smedt, F. (2009). Multi-objective calibration of a distributed hydrological model (WetSpa) using a genetic algorithm. *Hydrology and Earth System Sciences*, *13*(11), 2137–2149. https://doi.org/10.5194/hess-13-2137-2009

Shafii, M., & Tolson, B. A. (2015). Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives. *Water Resources Research*, *51*(5), 3796–3814. https://doi.org/10.1002/2014WR016520

Shrestha, R. R., Peters, D. L., & Schnorbus, M. A. (2014). Evaluating the ability of a hydrologic model to replicate hydro-ecologically relevant indicators. *Hydrological Processes*, *28*(14), 4294–4310. https://doi.org/10.1002/hyp.9997

Shrestha, R. R., Schnorbus, M. A., & Peters, D. L. (2016). Assessment of a hydrologic model's reliability in simulating flow regime alterations in a changing climate. *Hydrological Processes*, *30*(15), 2628–2643. https://doi.org/10.1002/hyp.10812

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall/CRC. https://doi.org/10.1201/9781315140919

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

Tan, M. L., Gassman, P. W., Yang, X., & Haywood, J. (2020). A review of SWAT applications, performance and future needs for simulation of hydro-climatic extremes. In *Advances in water resources* (Vol. 143). Elsevier Ltd. https://doi.org/10.1016/j.advwatres.2020.103662

Tang, Y., Marshall, L., Sharma, A., & Ajami, H. (2018). A Bayesian alternative for multi-objective ecohydrological model specification. *Journal of Hydrology*, *556*, 25–38. https://doi.org/10.1016/j.jhydrol.2017.07.040

Tasdighi, A., Arabi, M., & Harmel, D. (2018). A probabilistic appraisal of rainfall-runoff modeling approaches within SWAT in mixed land use watersheds. *Journal of Hydrology*, *564*, 476–489. https://doi.org/10.1016/j.jhydrol.2018.07.035

Ter Braak, C. J. F., & Vrugt, J. A. (2008). Differential evolution Markov chain with snooker updater and fewer chains. *Statistics and Computing*, *18*(4), 435–446. https://doi.org/10.1007/s11222-008-9104-9

The Nature Conservancy. (2009). Indicators of hydrologic alteration version 7.1 user's manual. *Nature Conservancy*.

USDA-NASS. (2012). CropScape - NASS CDL program. Retrieved from https://nassgeodata.gmu.edu/CropScape/

USDA-NRCS. (2020). Web soil survey. Retrieved from https://websoilsurvey.sc.egov.usda.gov/App/WebSoilSurvey.aspx

USDA-SCS. (1972). National engineering handbook, section 4: Hydrology.

USGS. (2018). The national map: Elevation. Retrieved from https://nationalmap.gov/elevation.html

USGS. (2020). National water information system: Web interface. Retrieved from https://waterdata.usgs.gov/nwis

Valipour, M., Gholami Sefidkouhi, M. A., & Raeini−Sarjaz, M. (2017). Selecting the best model to estimate potential evapotranspiration with respect to climate change and magnitudes of extreme events. *Agricultural Water Management*, *180*, 50–60. https://doi.org/10.1016/j.agwat.2016.08.025

Vigiak, O., Lutz, S., Mentzafou, A., Chiogna, G., Tuo, Y., Majone, B., et al. (2018). Uncertainty of modelled flow regime for flow-ecological assessment in Southern Europe. *Science of the Total Environment*, *615*, 1028–1047. https://doi.org/10.1016/j.scitotenv.2017.09.295

Vis, M., Knight, R., Pool, S., Wolfe, W., & Seibert, J. (2015). Model calibration criteria for estimating ecological flow characteristics. *Water*, *7*(5), 2358–2381. https://doi.org/10.3390/w7052358

Vrugt, J. A. (2016). Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. *Environmental Modelling & Software*, *75*, 273–316. https://doi.org/10.1016/j.envsoft.2015.08.013

Vrugt, J. A., & Beven, K. J. (2018). Embracing equifinality with efficiency: Limits of Acceptability sampling using the DREAM(LOA) algorithm. *Journal of Hydrology*, *559*, 954–971. https://doi.org/10.1016/j.jhydrol.2018.02.026

Vrugt, J. A., Ter Braak, C. J. F., Diks, C. G. H., Robinson, B. A., Hyman, J. M., & Higdon, D. (2009). Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *International Journal of Nonlinear Sciences and Numerical Stimulation*, *10*(3), 273–290. https://doi.org/10.1515/IJNSNS.2009.10.3.273

Westerberg, I. K., & McMillan, H. K. (2015). Uncertainty in hydrological signatures. *Hydrology and Earth System Sciences*, *19*(9), 3951–3968. https://doi.org/10.5194/hess-19-3951-2015

Westerberg, I. K., Wagener, T., Coxon, G., McMillan, H. K., Castellarin, A., Montanari, A., & Freer, J. (2016). Uncertainty in hydrological signatures for gauged and ungauged catchments. *Water Resources Research*, *52*(3), 1847–1865. https://doi.org/10.1002/2015WR017635

Williams, J. R. (1969). Flood routing with variable travel time or variable storage coefficients. *Transactions of the ASAE*, *12*(1), 100–103. https://doi.org/10.13031/2013.38772

Xiong, M., Liu, P., Cheng, L., Deng, C., Gui, Z., Zhang, X., & Liu, Y. (2019). Identifying time-varying hydrological model parameters to improve simulation efficiency by the ensemble Kalman filter: A joint assimilation of streamflow and actual evapotranspiration. *Journal of Hydrology*, *568*, 758–768. https://doi.org/10.1016/j.jhydrol.2018.11.038

Zhang, Y., Shao, Q., Zhang, S., Zhai, X., & She, D. (2016). Multi-metric calibration of hydrological model to capture overall flow regimes. *Journal of Hydrology*, *539*, 525–538. https://doi.org/10.1016/j.jhydrol.2016.05.053