The quest for hydrological signatures: Effects of data transformation on Bayesian inference of watershed models

Mojtaba Sadegh[1*], Morteza Shakeri Majd[2], Jairo Hernandez[1], Ali Torabi Haghighi[3]

[1]Department of Civil Engineering, Boise State University, Boise, ID, USA

[2]Department of Civil and Environmental Engineering, University of California, Irvine, CA, USA

[3]Water Resources and Environmental Engineering Research Unit, University of Oulu, Oulu, Finland

[*]Corresponding author: mojtabasadegh@boisestate.edu

ABSTRACT

Hydrological models contain parameters whose values cannot be directly measured in many field-scale projects, hence need to be meaningfully inferred through calibration against historical records. Much progress has been made in development of efficient search algorithms in order to find optimal parameter values and their underlying uncertainty distributions. Yet, relatively little is known about the effects of calibration data (or error residual) transformations on the identifiability of model parameters, reliability of model predictions, and convergence speed of the search algorithm. Such transformations are common in the model-data synthesis literature, but often appear ad-hoc without a comprehensive and comparative theoretical or practical justification. Effects of calibration

data transformations on the posterior parameter distribution and predictive capability of two parsimonious hydrological models are here analyzed. We are particularly concerned with convergence speed and reliability of parameters and predictive uncertainty estimates of hydrological models. Our results depict that calibration data transformations significantly influence parameter and predictive uncertainty estimates, sometimes distorting the information content of data. In particular, transformations which distort the temporal distribution of calibration data, such as flow duration curve, normal quantile transform, and Fourier transform, considerably deteriorate the identifiability of hydrological model parameters derived in a formal Bayesian framework with a residual-based likelihood function. Other transformations, such as wavelet, BoxCox and square root, while demonstrating some merits in identifying specific model parameters, would not consistently improve predictive capability of hydrological models in a single objective inverse problem. Multi-objective optimization schemes may present a more rigorous basis to extract several independent pieces of information from different data transformations. Finally, data transformations might offer a greater potential for model evaluation and selection than calibration in a single objective framework.

KEYWORDS: Data transformation, Hydrological signatures, Bayesian inference, MCMC, Parameter identifiability, Prediction reliability.

1 INTRODUCTION AND SCOPE

Rainfall-Runoff (RR) models are simplified representations of complex bio-geophysical processes in a watershed transforming climatic forcing into runoff that emanates from catchment outlet [Sadegh et al., 2015]. Such models are utilized for two general purposes, scientists scrutinize them to better understand hydrological behavior of watersheds, while engineers utilize them as predictive tools for decision making and policy planning [Jakeman et al., 2006; Pushpalatha et al., 2012]. Parameters of (conceptual) hydrological models are often not directly measurable, at the scale of interest, and need to be adjusted to optimally fit model simulations to measured data. In other words, they should be inferred indirectly through inverse modeling [Yapo et al., 1996]. Goodness of fit between simulations and observations is typically evaluated through an objective function that summarizes, as consistently as possible, the mismatch between modeled and observed system behavior.

There is an extensive and growing body of literature on optimization, parameter estimation, and uncertainty analysis methods [Thiemann et al., 2001; Reed et al., 2000, 2003; Tolson and Shoemaker, 2007; Vrugt et al., 2009; Guzman et al., 2015], but input, structural and measurement errors overburden the identifiability of unbiased parameter distributions [Vrugt and Sadegh, 2013]. Such errors lack an inherent probability distribution that can inform formulation of an explicit likelihood/objective function [Vrugt and Sadegh, 2013]. The commonly used likelihood/objective functions in the literature are based on some assumptions regarding error residuals, many of which are not borne out of post-processing

of the inverse modeling results [Sadegh and Vrugt, 2014]. In order to satisfy these assumptions, a common practice is to transform simulated and observed data or error residual time series prior to evaluating the objective function [Yapo et al., 1996; Bennett et al., 2013]. Data transformation is usually adopted either to satisfy predefined assumptions about the distribution of a time series of data [Dotto et al., 2014; Madadgar and Moradkhani, 2014], or to efficiently extract desired information from data that highlight different aspects of model behavior in a given domain [Bennett et al., 2013]. Desired information may emphasize a specific portion of the system response. Impacts of low flows in an inverse modeling practice, for example, could be accentuated through logarithmic, square root and inverse transformation of streamflows [Pushpalatha et al., 2012]. In other words, time series of simulations and measured flows can be filtered through a "context" prior to collation in the objective function aiming to highlight a particular behavior of catchment.

Signature-based analysis proposed by Gupta et al. [2008] suggests hydrological models should be calibrated against historical data presented in a "context" rather than their original (raw) format. Gupta et al. [2008] argue that "data" is not identical to "information". "Information" is one's understanding of the data presented through a filter of a "context". The appropriate "context" yields most relevant information of the observations which has "*clear and compelling diagnostic power*" to reconcile a model with measurements [Gupta et al., 2008]. This has been postulated as a potentially beneficial approach to calibrate models, and to ensure models are able to mimic key characteristics of the hydrological system of interest.

This paper aims to evaluate if data transformations can help us better understand hydrological behavior of watersheds, and consequently improve information extraction from available calibration data in a Bayesian framework. In this study, a range of commonly used transformations with widespread applications in hydrological modeling have been scrutinized. This includes square root, BoxCox, flow duration curve and normal quantile transformations, as well as spectral and wavelet spectral analysis. For each different transformation the posterior parameter distributions of hydrological models are estimated using a state of the art hybrid Markov Chain Monte Carlo (MCMC) algorithm [Sadegh et al., 2017] within a Bayesian framework. Posterior model parameter and prediction distributions, as well as convergence speed of MCMC simulation, are employed to assess the applicability of each transformation. Two lumped conceptual hydrological models, namely GR4J and HyMod, are used to analyze three catchments in the United States obtained from the MOPEX data set (available at ftp://hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data/). Our MCMC simulation formally maintains detailed balance, and we therefore argue that the convergence speed is a proxy statistic to determine the relative information content of the data. Overall, our results show while some transformations might offer promise to improve model parameter identifiability and predictive capability, a multi-objective framework is necessary to harness all available independent pieces of information obtained from different transformations. Moreover, some data transformations, such as flow duration curve, normal quantile, and Fourier transform, leave detrimental imprints on the inverse modeling results in a formal Bayesian context.

This paper is organized as follows: Section 2 describes the materials and methods used in this study, which details a general modeling framework, followed by a short description of Bayesian inference. This section continues with a concise explanation of data transformations used in this study, including square root, flow duration curve, normal quantile, BoxCox, spectral, and wavelet spectral analysis. Subsequently, the two conceptual rainfall runoff models, GR4J and HyMod, and case studies used in this analysis are briefly discussed. Section 3 presents the results of Bayesian analysis of these two hydrological models calibrated against different transformations of historical and synthetic records from three catchments in the U.S. Finally, section 5 presents the concluding remarks.

2 MATERIALS AND METHODS

In a common forward modeling practice, a set of describing equations and physical constraints are used to simulate the response of a system to a cohort of drivers. A rainfall-runoff model, M, for example, translates a set of inputs, $\tilde{\boldsymbol{I}}$, including precipitation, $\widetilde{\boldsymbol{P}}$, and potential evapotranspiration, $\widetilde{\boldsymbol{PET}}$, to streamflow discharge, $\boldsymbol{D}$, at catchment outlet given a parameter set, $\boldsymbol{\theta}$:

$$\boldsymbol{D} = M\left(\boldsymbol{\theta}, \tilde{\boldsymbol{I}}\right) + \boldsymbol{e}. \quad (1)$$

In this equation, $\boldsymbol{e}$ represents error residuals stemming from the difference between model hypothesis, M, and the underlying "true" unknown streamflow generating process, K, as

well as input, calibration data and parameter errors [Sadegh and Vrugt, 2013; Sadegh et al., 2017].

We wish to estimate model parameters, $\boldsymbol{\theta}$, through an optimization algorithm that tune them to minimize the discrepancy between model simulation, $\boldsymbol{D}$, and streamflow observation, $\widetilde{\boldsymbol{D}}$. Several optimization algorithms, such as Shuffled Complex Evolution (SCE-UA) of Duan et al. [1992], have been employed to find the optimal parameters of conceptual rainfall-runoff models. Such algorithms locate the model realization which best fits observed data without recourse to the underlying uncertainties of the modeling practice [Thiemann et al., 2001]. Bayesian inference, on the other hand, has shown great applicability and promise to characterize the underlying modeling uncertainties in the field of hydrology and beyond [Kuczera, 1999].

2.1. Bayesian Inference

Bayesian inference systematically incorporate new information, as they become available, to update the probability of a hypothesis. Bayes' theorem conveniently assumes model parameters are the sole source of uncertainty, and estimates posterior probability of a hypothesis, $p(\boldsymbol{\theta}, \widetilde{\boldsymbol{D}})$, through multiplication of prior probability, $p(\boldsymbol{\theta})$, likelihood value, $L(\boldsymbol{\theta}|\widetilde{\boldsymbol{D}})$, and inverse of evidence, $p(\widetilde{\boldsymbol{D}})$,

$$p(\boldsymbol{\theta}, \widetilde{\boldsymbol{D}}) = \frac{p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\widetilde{\boldsymbol{D}})}{p(\widetilde{\boldsymbol{D}})}. \quad (2)$$

In this equation, evidence, $p(\widetilde{\boldsymbol{D}})$, is a constant value, and can be conveniently excluded from the formula if estimation of posterior probability of the hypothesis is merely intended,

$$p\left(\boldsymbol{\theta}, \widetilde{\boldsymbol{D}}\right) \propto p(\boldsymbol{\theta}) L(\boldsymbol{\theta}|\widetilde{\boldsymbol{D}}). \quad (3)$$

In the field of environmental science, it is a common practice to assume a uniform prior distribution for model parameters, $p(\boldsymbol{\theta})$, in the absence of any useful information [Thiemann et al., 2001]. The second element of equation 3, likelihood function $L(\boldsymbol{\theta}|\widetilde{\boldsymbol{D}})$ characterizes the distance between model simulations and observations. There is a lot of discussion in the literature on how to formulate the likelihood function that sufficiently characterizes the probabilistic characteristics of forcing, model, and calibration data errors [Schoups and Vrugt, 2010]. However, it is noteworthy to acknowledge a simple likelihood function that assumes error residuals are uncorrelated, homoscedastic, Gaussian distributed with zero mean is probably most widely used in the field of hydrology and environmental science [Thyer et al., 2009; Sadegh et al., 2017]. Such a residual-based Gaussian likelihood function is formulated as,

$$L\left(\boldsymbol{\theta}\middle|\widetilde{\boldsymbol{D}}\right) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\tilde{\sigma}^2}} \exp\left\{-\frac{1}{2}\tilde{\sigma}^{-2}\left[\tilde{d}_t - d_t(\boldsymbol{\theta})\right]^2\right\}, \quad (4)$$

where $\tilde{\sigma}$ signifies measurement error, n denotes length of observational data, and $\tilde{d}_t$ and $d_t(\boldsymbol{\theta})$ represent observed and simulated streamflow at time step t.

In a common Bayesian inference problem, we wish to estimate the posterior distribution of model parameters, $p\left(\boldsymbol{\theta}, \widetilde{\boldsymbol{D}}\right)$, through equation 3. Analytical solution to the Bayes' equation is, however, not always warranted. Hence, a numerical solution is often required to estimate the posterior distribution of model parameters given the observed data. Markov Chain

Monte Carlo (MCMC) simulation has emerged as a powerful statistical tool to sample from high dimensional, multimodal, complex distributions [Andrieu and Thomas, 2008], and can be readily applied to solve the Bayes' rule. In this paper, we employ the state of the art hybrid-evolution MCMC algorithm of Sadegh et al., 2017 (Algorithm 1) that utilizes adaptive proposal distributions, namely Adaptive Metropolis (AM), Differential Evolution (DE), and snooker update, to search the prior space [Gilks et al., 1994; Storn and Price, 1995, 1997; Roberts and Sahu, 1997; Haario et al., 1999, 2001; ter Braak, 2006; Roberts and Rosenthal, 2009].

In words, this algorithm starts as random search with $20 * d$ samples (d: number of model parameters) selected through Latin Hypercube Sampling (LHS). These samples then drive the RR model and calculate the associated likelihood value. Next step is to randomly assign these samples to $N = 2 * d$ complexes, in which the best sample (highest likelihood) is selected as the starting point for Markov chains. Then, parallel (with a probability of 90%) and snooker (with a probability of 10%) direction updates are employed to propose new samples, acceptance/rejection of which is determined through Metropolis ratio. To diversify the jump algorithm, we assign one chain to use Adaptive Metropolis and the rest to employ Differential Evolution. Finally, Gelman-Rubin $\tilde{\mathcal{R}}$ statistics is used to monitor the convergence of Markov chains [Gelman and Rubin, 1992]. For more information regarding the MCMC algorithm, interested readers are referred to Algorithm 1 of Sadegh et al. [2017].

Convergence speed of MCMC simulation provides a proxy diagnostic measure for the information content of data and the simplicity to extract this information. Filtering data through a context which potentially extracts more pieces of independent information from the data provides the MCMC algorithm a higher power to reconcile model with data, and hence increases the convergence speed of the MCMC simulation. Our MCMC simulation formally maintains detailed balance, and we therefore argue that the convergence speed is a proxy statistic to determine the relative information content of data. In order to demonstrate the effects of data transformations on model parameter identifiability and predictive capability, as well as convergence speed of MCMC simulation, two hydrological models, namely GR4J and HyMod, are calibrated against data from three watersheds in the U.S., details of which are provided later in this section. In the following paragraphs, different data transformations and hydrological models, used in this study, are explained in detail.

2.2 Data Transformations

In principle, there are infinite number of data transformations. In this paper, we restrict our attention to six commonly used class of transformations in the field of hydrology, namely square root, BoxCox, flow duration curve, normal quantile transform, spectral analysis and wavelet spectral analysis. In this section, we first demonstrate how a data transformation affects calibration of a simple hydrological model, and then describe different data transformations in detail.

As an intuitive example, consider a very simple, three parameters "abc" hydrological model as following [Fiering, 1967; Mantovan and Todini, 2006]:

$$\begin{cases} d_t = (1-a-b) \times p_t + c \times s_{t-1} \\ s_t = (1-c) \times s_{t-1} + a \times p_t \end{cases}, \quad (5)$$

in which p_t, d_t and s_t represent precipitation, discharge and aquifer storage at time t, respectively. In this model, discharge consists of a direct runoff component (as a result of precipitation) and a baseflow component (as a result of aquifer storage). Aquifer storage at each time step, t, depends on storage at the previous time step as well as precipitation at the current time. We use a simple and yet frequently used quadratic objective function of Sum of Squared Residuals (SSR $= \sum(d_t - \bar{d}_t)^2$), to calibrate this hydrological model. In this equation, d_t and $\bar{d}_t$ represent simulated and measured discharge at time t, respectively. The idea is to study model response sensitivity to the aquifer storage parameter (c), as a result of an extreme precipitation event. In a hypothetical rainfall time series, consider an extreme precipitation that mainly contributes to direct runoff, and results in an extreme flow discharge of 200 m^3/s. Also, suppose a long period of dryness which makes the effect of precipitation on the discharge negligible and consider a hypothetical discharge of 10 m^3/s, primarily due to baseflow.

A change of 10% in the discharge during the extreme precipitation event (e.g., observation of 200 m^3/s and simulation of 180 m^3/s) would induce a change of 400 units in the objective function, while a change of 10% in the discharge during the dry period (e.g., observation of 10 m^3/s and simulation of 9 m^3/s) would result in a 1 unit change in the objective function. During the extreme precipitation event, the dominant parameter is the direct

runoff parameter, while in the dryness period the baseflow parameter is dominant. In this example, model would be insensitive to the parameter representing baseflow (c), because of the trivial impact of this parameter on the objective function comparing to the direct runoff parameter.

Given this forcing, a square root transformation will alleviate the dominant effect of extreme event on the objective function; as a change of 10% in the discharge during the extreme precipitation (dominated by direct runoff) would prompt a change of 0.53 units in the objective function; while a change of 10% in discharge during the dry period (dominated by baseflow) would generate a 0.03 unit change in the objective function. In other words, square root transformation gives higher importance to lower discharge values comparing to the original data, and enhance model sensitivity to the parameter representing baseflow. In the following sections, different data transformations are explained, and consequently applied to two conceptual models of streamflow prediction.

2.2.1 Square root transformation

Square root transformation is a simple square root of the discharge data, which help decrease the dominant effects of extreme peak events in the calibration process, and is denoted by SQRT hereafter in this paper.

2.2.2 Flow duration curve transformation

Flow duration curve (FDC) is a cumulative distribution function that depicts the relationship between the magnitude of a certain flow event and its frequency. The FDC

transformation exhibits the percentage of time that discharge equals or exceeds a particular value, and associates each discharge value to an exceedance probability [Vogel and Fennessey, 1994].

According to Vogel and Fennessey, 1994, "*FDC provides a simple, yet comprehensive, graphical view of the overall historical variability associated with stream-flow in a river basin*". However, removing temporal order of events can be regarded as a potential drawback of this transformation. In the field of hydrology and water resources, FDCs are widely used for planning purposes [Mitchell, 1957; Searcy, 1959; Male and Ogawa, 1984; Gordon et al. 1992]. Blazkova and Beven [2009] utilized information summary of FDCs to test a rainfall-runoff model performance.

Computing FDC involves the following steps:

1) Sort discharge values in a descending order and rank them 1 to n (sample size).

2) Compute exceedance probability associated with each discharge value:

$$P = 100 \times \frac{i_t}{n+1}, \quad (6)$$

in which, i_t is rank of discharge d_t.

2.2.3 Normal quantile transformation

Normal Quantile Transform (NQT), also referred to as Inverse Normal Score, is a strong statistical method to alter the form of a cumulative distribution, $p(\boldsymbol{D} \leq d_t)$, in the field of

hydrology [Moran, 1970; Kelly and Krzysztofowicz, 1997; Hosking and Wallis, 1998]. NQT is usually denoted as [Kelly and Krzysztofowicz, 1997; Montanari and Brath, 2004]:

$$NP(t) = Q^{-1}[p(\boldsymbol{D} \leq d_t)] \quad (7)$$

In which $NP(t)$ is the modified Gaussian distribution, $p(\boldsymbol{D} \leq d_t)$ is the marginal cumulative distribution of variable d_t, and Q^{-1} represents the inverse of the standard normal distribution. We follow Montanari and Brath [2004] in implementing NQT:

1) Compute cumulative frequency $F(d_t)$ based on Weibull plotting position:

$$F(d_t) = \frac{i_t}{n+1}. \quad (8)$$

2) Compute normal quantile $NP(t)$ for each frequency $F(d_t)$, and assign it to the associated d_t.

2.2.4 BoxCox transformation

BoxCox transformation is a strong tool to eliminate the heteroscedasticity of data and stabilize the variance of a time series [Sorooshian and Dracup, 1980]. Heteroscedastic model residuals are prevalent in hydrological modeling, in contrast to the homoscedasticity assumption of the statistical inference approaches. To solve this problem, and to alleviate the assumption of Gaussianity, it is assumed that there is a BoxCox transformation with an unknown parameter λ that generates a set of more likely normally distributed values with N constant variances. BoxCox transformation is defined for positive values as [Box and Cox, 1964]:

$$d_t^{(\lambda)} = \begin{cases} \frac{(d_t^{\lambda}-1)}{\lambda}, & \text{when} \quad \lambda \neq 0 \\ \log(d_t), & \text{when} \quad \lambda = 0 \end{cases} . \tag{9}$$

Our investigation showed that choice of $\lambda = 0.3$ is proper to stabilize the total error variance and reduce heteroscedasticity, confirming previous studies of Misirli et al. [2003]; and Vrugt et al. [2006].

2.2.5 Spectral analysis

One can adjust model parameters so that spectral properties of model simulations fit spectral properties of measurements, instead of fitting in the temporal domain [Pauwels and De Lannoy, 2011]. It is specifically useful in the case of scarce data or ungauged basins, as spectral properties of a process can be obtained through analyzing scarce, old and non-overlapping data [Montanari and Toth 2007]. Calibration in the spectral domain has been implemented by several researchers in the field of hydrology [Montanari et al., 2000; Montanari, 2003; Montanari and Toth, 2007; Quets et al., 2010; Pauwels and De Lannoy, 2011].

Spectral properties can be computed through Fourier transform, $F(d_t)$, of time series of observations and simulations. In case of constant time step measurements and no missing data, one can use a Fast Fourier Transform (FFT) to decompose a time series to its spectral components. In our study, we use FFT to change the time domain into frequency domain.

Fourier amplitude spectra of a variable is estimated through FFT, which is in turn defined as:

$$X_k = \sum_{t=0}^{N-1} d_t \omega^{t \times k} \qquad k = 0,1,2,\dots,n-1$$
$$\omega = e^{\frac{-2\pi i}{n}} \tag{10}$$

in which, X_k is the Fourier amplitude of the kth harmonic wave. To keep at least two samples for each wavelength, highest harmonic (k) in this formula is chosen to be less than $\frac{n}{2}$ [Quets et al., 2010; Shannon, 1949]. We calibrate hydrological models in the spectral domain following suggestions of Quets et al [2010]; and Pauwels and De Lannoy [2011] using Fourier amplitude spectra (denoted by FFT).

Spectral density of a process can also be approximated as square of the Fourier amplitude spectra. When the requirement of constant sampling time step is not met in the available data, spectral density can be defined as the Fourier transform of the autocorrelation function of the time series [Pauwels and De Lannoy, 2011]:

$$S_k = F\big(R_\tau(d)\big); \quad R_\tau(rd) = E[d(t)d(t-\tau)], \tag{11}$$

in which, S_k is the spectral density of the kth harmonic wave, F represents the Fourier transform, and $R_\tau(d)$ is the autocorrelation function of the variable d. For more information refer to Quets et al. [2010]; and Pauwels and De Lannoy [2011]. Results of transformation of equation 11 are not presented in this paper, as they consistently deteriorate model performance for all watersheds.

2.2.6 Wavelet spectral analysis

Wavelet analysis provides time localized power spectra for each frequency/scale [Lane 2007], as opposed to spectral analysis which yields a power spectrum for each frequency and loses time localization. Wavelet transform is able to decompose a signal into scaled and translated versions of a mother wavelet which are defined in a three-dimensional space: time, scale/frequency and power [Lafrenière and Sharp, 2003]. As stated by Dhanya and Kumar [2011], "*localization property of wavelets, utilizing different dilation and translation parameters, helps in capturing most of the statistical properties of the observed data.*" Wavelet spectral analysis has become a very attractive option in the field of hydrology to identify and analyze scale variability, long term oscillation and cycles of hydrological data [Torrence and Compo, 1998; Lafrenière and Sharp, 2003; Labat et al., 2005; Lane, 2007; Schaefli et al., 2007; Dhanya and Kumar, 2011].

Wavelet analysis is a useful tool in hydrological studies, since it enables us to discriminate between measurements and model simulations through time localization and scale/frequency decomposition of data series [Lane, 2007]. Continuous wavelet transformation of a discrete time series, d_t, with respect to a pre-specified wavelet function $g(t)$ is defined as [Torrence and Compo, 1998]:

$$W_t(s) = \sum_{t'=0}^{n-1} d_{t'}\, g^* \left[\frac{(t'-t)}{s}\delta t\right], \qquad (12)$$

where, g^* represents a complex conjugate of wavelet function g, and s, t and δt denote scale, time and measurement resolution, respectively. The so-called convolution (equation 12) should be repeated n times for each scale, where n is the dimension of the time series. It is worth mentioning that wavelet transform can be implemented much faster in the Fourier space than the original domain. For more information refer to Torrence and Compo [1998].

Wavelet functions can be generally characterized as follows [Torrence and Compo, 1998]:

1. Orthogonal or nonorthogonal wavelets: Nonorthogonal wavelet functions are appropriate for time series analysis, and can be utilized for both continuous and discrete waveforms. Orthogonal wavelets, on the other hand, can only be used for discrete wavelet transforms.
2. Complex or real-valued wavelet functions: Real wavelets only provide information on the power of a signal, whereas complex wavelets offer information on both power and phase of a signal, which is more informative for analyzing signal oscillations.
3. Width: The balance between width in time and frequency spaces determines the resolution of a wavelet function. There is a trade-off between resolution in time and frequency; a narrow wavelet function in time has a high time resolution and a low frequency resolution, whereas a broad wavelet function is vice versa.
4. Shape: Shape of wavelet function should represent structure, form and properties of the time series.

One should consider different characteristics of wavelet functions and select an appropriate function for the studied data set. However, study of Torrence and Compo [1998] suggests choice of wavelet does not drastically affect the qualitative results of wavelet analysis. Two forms of wavelets namely Morlet wavelet (complex) [Lafrenière and Sharp, 2003; Labat et al., 2005] and Mexican Hat wavelet (real) [Lane, 2007] are most commonly in use in hydrological studies. These two wavelet functions will be utilized in this study.

It is a common practice to use a set of scales, instead of one single scale. This helps generate a more comprehensive representation of the data set. Choice of scales for a wavelet transformation is arbitrary, and different suggestions can be found in the literature in this regard [Torrence and Compo, 1998; Lane, 2007]. A rough estimate of minimum scale for Morlet wavelet is $s_{\min} = 2.06\delta t$ and maximum scale is $s_{\max} < \frac{n}{2\epsilon}\delta t$ [Lane 2007]. Here, δt and ϵ denote measurement resolution and cone of influence, respectively. ϵ is here set to $\sqrt{2s}$ for both Morlet and Mexican Hat wavelets [Torrence and Compo, 1998]. Cone of influence represents the edge effect of data transformation. Due to inherent cyclic assumption of data in wavelet transformation, it does not yield accurate transformed values for the edge parts of a time series. In order to alleviate the impacts of edge effect errors, we followed Torrence and Compo [1998] to pad zeros to both sides of time series before applying the transformation. Added values are cut off afterwards.

In order to satisfy the computational efficiency, one can choose a set of integer scale values between the minimum and maximum thresholds [Lane, 2007]. Our investigations show a scale set of $\{2, 4, 8, 16, 32\}$ seems appropriate to represent almost all features of our data set. In this scale set, the maximum value is carefully selected, since larger scale values require higher computational time with only marginal impact on model analysis. Each scale here provides a wavelet spectra series that complies with the length of the raw data set. In order to compare measurements with simulation results in our objective function, all the wavelet spectra with different scales are put in one vector of length $m * n$, in which n is the length of measurement data and m is the number of scales (in our study $m = 6$). Note that wavelet analysis considerably increases the computational effort. Next, we discuss the conceptual hydrological models used in this study.

2.3 Rainfall-Runoff Models

Rainfall-Runoff (RR) models can be generally categorized into two classes of conceptual and physically-based models. In this study, we calibrate two lumped conceptual hydrological models, widely used in the literature, namely HyMod and GR4J against daily data from French Broad, Skykomish, and Rogue river watersheds, located in the U.S. The two hydrological models are briefly explained in the following paragraphs.

2.3.1 HyMod model

HyMod is a five-parameter parsimonious conceptual RR model, developed by Boyle [2001]. HyMod has been extensively used in the field of hydrological science. This model consists of a nonlinear soil moisture storage compartment and two parallel series of linear

reservoirs, forming a total of five storage compartments. Soil moisture storage processes evapotranspiration and excess rainfall, and feed the parallel slow (one storage compartment) and quick (three serial storage compartments) routing reservoirs. Input forcing of this model include mean areal precipitation and potential evapotranspiration, and its outputs are simulated streamflow discharge and evapotranspiration. For the sake of brevity, we refrain from explaining the model in more details and refer interested readers to Boyle [2001]; Wagener et al. [2001]; and Gharari et al. [2013] for more information. A schematic representation of HyMod is provided in Figure 1, and a short description of its parameters along with their feasible ranges are provided in Table 1.

2.3.2 GR4J model

Perrin [2000] introduced GR4J as a parsimonious RR model with a great flexibility to simulate a wide range of watersheds' response to mean areal precipitation and potential evapotranspiration forcing. Perrin et al. [2003] argue that a model complexity of three to five parameters for rainfall-runoff transformation is adequate to obtain a satisfactory performance at a daily time step. GR4J consists of two storage compartments and a few transfer functions that are governed by a total of four tunable parameters and two fixed scalars. GR4J satisfies evapotranspirative demand first, and then routes net rainfall through a storage compartment (production store) that feeds two unit-hydrograph (UH) based transfer functions. One UH then feeds the routing store, whereas the other directly routes water to the catchment outlet. For a schematic representation of the GR4J model see Figure 1, and for a detailed description refer to Perrin et al. [2001]; and Perrin et al. [2003]. Table 2 provides a short description of the GR4J parameters along with their feasible ranges.

2.3.3 Snow module

A snow module is used in this study to classify precipitation into snow and rainfall categories based on observed temperature, before driving RR models. Precipitation accumulates as snow, if mean daily temperature is less than a freezing threshold, T_0, and pass as rainfall when temperature is above the freezing threshold ($T_0 = 0$ degree Celsius, in this study). Snow melts based on a tunable degree-day factor parameter, and cumulative sum of snowmelt and rainfall drives the RR model(s). For more details regarding the snow module, refer to AghaKouchak and Habib [2010] and AghaKouchak et al. [2013].

2.4 Case Studies

In this study, we analyze three watersheds from the MOPEX dataset with different hydrological behavior, including French Broad river catchment near Newport, Tennessee, U.S. (USGS ID: 03455000), Skykomish river catchment near Gold Bar, Washington, U.S. (USGS ID: 12134500), and Rogue river at Raygold near Central Point, Oregon, U.S. (USGS ID: 14359000). Historical data consists of daily mean areal precipitation (mm/day), potential evapotranspiration (mm/day), maximum and minimum daily temperature (℃) and streamflow (m^3/s). Table 2 presents a concise description of these watersheds' hydrological characteristics. We calibrate GR4J and HyMod RR models against five years of observed data (01/01/1949-12/31/1953) from each watershed, and evaluate the calibrated models against five years of out of sample independent observations (01/01/1954-12/31/1958).

Bayesian inference was repeated for each model and watershed with all the aforementioned data transformations. In a quest for sufficient hydrological signatures, we transform observed and simulated streamflow data prior to likelihood evaluation. To further investigate the influence of data transformation on Bayesian inference, we also evaluate the performance of each transformation for synthetic data. We generate time series of synthetic data with known randomly selected parameter sets for each model and each watershed. These synthetic observations are in turn corrupted with a red noise (random error series drawn from a Gaussian distribution centered at zero with a variance of 20% of original flow values) to represent real-world errors. Similar time periods as that of the real-world observations are used for Bayesian analysis of the synthetic data.

3 RESULTS AND DISCUSSION

In this section, we present results of numerical simulation of French Broad, Skykomish, and Rogue river basins with GR4J and HyMod models. Posterior distribution of model parameters for each watershed is inferred using a state of the art hybrid-MCMC algorithm with 35,000 function evaluations, through calibration against synthetic and real-world observed data. Hybrid-MCMC algorithm of Sadegh et al., 2017 maintains detailed balance, and equilibrium state of this algorithm accurately represents the posterior distribution of model parameters. Last 20% draws of Markov chains, satisfying the convergence criteria, were used to form the posterior distributions, which are in turn used for predictive analysis. The so-derived posterior distributions of model parameters contain enough information to characterize model predictive uncertainty ranges. As discussed earlier, Bayesian inference attribute the modeling uncertainties to parameters, and hence the estimated predictive

uncertainty ranges include only the influence of parameter uncertainty. Assuming the remaining unexplained uncertainty is additive, we then perturb each model simulation run with a residual error series, $\varepsilon \sim N(0, \sigma_v^2)$, to obtain total predictive uncertainty ranges. RMSE of best model realization is used as an intuitive approximation of measurement error variance, σ_v^2. Note that we drive RR model with posterior parameter sets and observed forcing to obtain model simulation runs.

In this study, we used a likelihood function based on a simple quadratic objective function of Sum of Square Residuals (equation 3) to minimize the influence of complex objective function assumptions on the posterior results; and to ensure that results are maximally reflecting the impacts of data transformations. We acknowledge that the underlying assumptions of our likelihood function (residuals being uncorrelated, homoscedastic, Gaussian distributed with mean zero) influence the posterior results, but this is minimal compared to other choices. More complex objective functions would interact with data transformations, leaving detection of the pure impacts of data transformations very difficult, if not impossible. In the following paragraphs, the results of parameter inference against synthetic and real data are described in detail.

Figure 2 depicts the posterior distribution of maximum capacity of the production store parameter ($S1_{\max}$) of GR4J calibrated against synthetic data of French Broad river catchment, using different data transformations. The “true” parameter value (used to generate the synthetic data) is portrayed in each plot with a red square. Each plot shows

the results of one data transformation, as described in the title. Most conspicuous in this figure is the distorted posterior distributions of $S1_{max}$ in cases of FDC and FFT transformations. Indeed, FDC's results show multimodality in posterior $S1_{max}$ parameter distribution, and FFT derived parameters are limited to a few samples. Latter observation might suggest that MCMC has reached premature convergence for the FFT transformed data, but a closer look (not presented herein) shows that Markov chains representing $S1_{max}$ parameter indeed have explored the entire prior space and converged to a few points without much diversity. We attribute this behavior to the impact of FFT transformation in distorting the information content of data, which in turn is manifested in rejecting almost all the proposed samples.

It is also noticeable in Figure 2 that the posterior distributions of $S1_{max}$ parameter derived with no transformation (None), as well as BoxCox, SQRT and NQT transformations encompass "true" parameter value, demonstrated with red box in each plot. It is interesting that the mode of distribution, which represents the parameter with highest likelihood, coincides with the true parameter value using NQT transformation. This is a desired behavior, but further investigation of the results of NQT transformation reveals that the overall influence of this transformation is disruptive on the calibration results. We will revisit this issue later in this section. SQRT transformation and original data (no transformation) are both able to nicely recover the "true" posterior distribution of this parameter, with the most likely parameter (mode of the distribution) closely positioned around the "true" parameter value. Posterior distributions derived by both wavelet transformations (Mexican Hat, MexH, and Morlet, Morl, wavelets), on the contrary, fail to

encompass the “true” parameter value. It is noteworthy, however, that posterior distributions of these transformations are very tight, partially explaining why wavelet transformation did not recover the “true” parameter value. Indeed, both wavelet transformations, in the employed format, significantly reduce the variance of the posterior parameter distribution for all watersheds and both models.

Similar conclusions can be drawn for the HyMod model’s results. Figure 3 presents the residence time of quick flow reservoir (R_q) parameter of the HyMod model, calibrated against five years of synthetic data from the Skykomish river catchment in Washington, U.S. For this parameter, results of the original data (no transformation), as well as BoxCox and SQRT transformations nicely converge to the “true” parameter value, with distribution mode coinciding with the “true” parameter. NQT transformation also returns a desirable posterior distribution, whereas FDC and FFT transformations distort the parameter distributions, as also observed in Figure 2. Expectedly, Mexican Hat wavelet transformation derives a tight distribution without encompassing the “true” parameter value. Morlet wavelet transformation, however, shows a surprising behavior in yielding a bimodal posterior distribution with its mode coinciding with the “true” parameter value.

As previously discussed in the illustrative example, data transformations can affect the identifiabilty of model parameters. SQRT transformation, for example, enhance the weight of low flow values (baseflow, for example) in the objective function, and improve the identifiability of associated model parameters. Figure 4 shows the marginal distribution of

553 the residence time of slow flow reservoir (R_s) parameter of the HYMOD model calibrated
554 against synthetic data from the Rogue river basin in Oregon, U.S. SQRT transformation
555 helps identify the "true" parameter value of R_s, while mode of posterior distribution of R_s
556 derived with the original data and other transformations do not coincide with the "true"
557 parameter value. R_s governs baseflow generation in HyMod, impacts of which are
558 accentuated using the SQRT transformation.

559

560 To further analyze the impacts of data transformation on the predictive performance of
561 GR4J and HyMod models, Table 3 presents 95% predictive uncertainty ranges due to
562 parameter and total uncertainty of the GR4J model for the Skykomish river catchment. This
563 table also details the associated coverage of synthetic streamflow observations in the 95%
564 parameter and total predictive uncertainty ranges, as well as root mean square error
565 (RMSE) associated with best parameter set, and mean RMSE of model predictions driven
566 by posterior parameters. These results also depict the poor performance of FFT and FDC
567 transformations in terms of predictive performance of the calibrated GR4J model. Best and
568 mean RMSE values for both transformations in calibration and evaluation periods are
569 significantly higher compared to the original data results. In more details, best RMSE value
570 of FDC and FFT transformations manifest values of 118.30 and 94.30 m^3/s, respectively,
571 compared to that of 78.82 m^3/s for the original data in the calibration period. Coverage of
572 synthetic observed streamflow in 95% predictive uncertainty ranges is also poor for both
573 transformations. FDC and FFT's total predictive uncertainty ranges cover 87.02% and
574 90.42% of observations, whereas the original data results cover 93.54% of observed data.

575 Note, however, that the total predictive uncertainty ranges for these two transformations
576 are also marginally smaller than that of the original data (None).

577

578 Mexican Hat and Morlet wavelet transformations return comparable results in terms of best
579 and mean RMSE in both calibration and evaluation periods to that of the original data. This
580 moderates the concerns regarding the posterior parameter distribution of the wavelet
581 transformed data not coinciding with the "true" parameter values seen in Figures 2-4.
582 However, in terms of predictive uncertainty ranges due to parameter uncertainties, both
583 wavelet transformations return much tighter uncertainty ranges compared to the original
584 data results in both calibration and evaluation periods (2.97 [2.97] m^3/s for Mexican Hat
585 and 7.78 [8.01] m^3/s for Morlet wavelets, as opposed to that of 12.37 [12.46] m^3/s for
586 original data; evaluation period statistics are presented within brackets). This is
587 accompanied with lower coverage of observed data (2.41% [1.92%] for Mexican Hat and
588 6.52% [5.59%] for Morlet wavelets, as opposed to that of 8.93% [7.56%] for original data;
589 evaluation period statistics are presented within brackets), which shadows the superiority
590 of tighter predictive uncertainty ranges of the wavelet transformed analysis. NQT
591 transformation, contrary to the results of Figures 2-4, show a much higher 95% predictive
592 uncertainty range due to parameter impacts compared to that of the original data (37.56
593 [39.83] m^3/s versus 12.37 [12.46] m^3/s), which is associated with higher coverage in the
594 predictive uncertainty range. Higher best and mean RMSE values for NQT analysis
595 compared to that of the original data (84.29 [88.02] m^3/s versus 78.82 [80.41] m^3/s),
596 however, conclusively manifest the disruptive impacts of NQT transformation on Bayesian
597 inference. Finally, SQRT and BoxCox transformations show a rather similar behavior to

that of the original data in terms of predictive uncertainty ranges, and coverage of observed data within the 95% uncertainty range. RMSE values associated with the BoxCox and SQRT transformations are expectedly marginally higher than that of the original data, since they highlight sections of the streamflow distribution that might not necessarily hold highest weight for RMSE computation. Likelihood function of equation 3 intrinsically minimizes RMSE, and hence Bayesian inference with this objective function and original data yields the most preferred RMSEs.

We now focus our attention on the results of analyzing real-world data from these three watersheds. Overall, we notice that the effects of data transformations on the calibration procedure are magnified using real-world historical data. For example, NQT transformation's results for synthetic data in terms of predictive uncertainty spread, associated coverage of streamflow observation, RMSE, and parameter identifiability do not greatly diverge from those obtained from the original data; but the NQT derived model performance deteriorates greatly for real-world historical data. This is manifested in increasing the 95% predictive uncertainty spread, as well as best and mean RMSE of model predictions. Figure 5 presents the convergence speed of MCMC simulation of GR4J model for the Rogue river catchment in terms of minimum RMSE in a moving window of 50 Markov chain samples. We argue that convergence speed of MCMC simulation can be one proxy indicator, among others, of the quantity/quality of information extracted from the data. Our MCMC simulation formally maintains detailed balance and we therefore argue that the convergence speed is a proxy statistic to determine the relative information content of the calibration data. The speculation behind data transformation is that it changes the

context in which information is presented and transferred, and it potentially unveils multiple independent pieces of information in the calibration data which helps the model inference procedure. Figure 5, however, reveals that MCMC simulation with original data (black solid line, "None" in the legend) converges to its stable state faster than any transformation-based analysis. This rejects the underlying hypothesis that data transformation can improve model inference procedure through unveiling extra independent pieces of information in the calibration data in a formal Bayesian framework.

Figure 5 also confirms our previous findings that FDC and FFT transformations distort information extraction from calibration data, leaving the model performance significantly inferior to that of the original data. FDC transformation not only converges to a higher RMSE than other transformations and result in higher uncertainty spread, but also shows an erratic behavior in its convergence when calibration is repeated several times (results not shown herein). This behavior can be interpreted by time disaggregation property of FDC that prompts a substantial loss of information in the calibration procedure. Indeed, in a synthetic case study if the temporal distribution of a time series is distorted, its FDC transformed series will be identical to that of the original series. Recent literature has shown a lot of interest in using FDC as one important hydrological signature for model inference and analysis purposes [Westerberg et al., 2011; Euser et al., 2013; Vrugt and Sadegh, 2013; Sadegh et al., 2016], but our results confirm the argument of Vogel and Fennessey [1995] that FDC is inadequate as main descriptor of hydrological behavior of watersheds. FFT transformation, on the other hand, changes the time domain to frequency domain, and associates each frequency with an amplitude spectrum. The calibration algorithm, then,

644 compares the amplitude spectra of measurements and simulations in order to infer
645 parameter values. Our results both for real and synthetic data and for both models show
646 that FFT transformation degrades model performance in the time domain, which
647 controverts the findings of Quets et al. [2010]. We also attribute this behavior to the loss
648 of information associated with temporal disaggregation in FFT transformation, as observed
649 for FDC. We repeated calibration endeavor with square root of FFT and FFT
650 transformation of autocorrelation coefficients, both of which significantly deteriorated
651 model performance in terms of RMSE and coverage of observation points in 95%
652 predictive uncertainty range (not showed herein).

653

654 Poor performance of NQT transformation is most conspicuous in Figure 5, in which RMSE
655 decreases to around 50 m^3/s in the begining of the MCMC search, but then increases as
656 search algorithm progresses, and converges to about 70 m^3/s. This is much higher than
657 the best RMSE value of roughly 45 m^3/s derived from model inference with the original
658 data. Such behavior is explained by distortion of information content of calibration data
659 through NQT transformation, which shadows parameter identifiability and degrades model
660 performance. The SQRT and BoxCox transformations, in agreement with the results of
661 synthetic analyses, both expectedly converge to a slightly higher RMSE than that of the
662 original data. Most important note in case of these two transformations, in this figure, is
663 that convergence is marginally slower for the transformed data compared to the original
664 data. This is a rather undesirable behavior. For the wavelet transformations, Mexican Hat
665 wavelet converges to the vicinity of RMSE obtained from the original data, while Morlet

wavelet converges to a relatively higher RMSE. Both wavelet transformations slow down the convergence speed of MCMC simulation.

To further demonstrate the influence of data transformations on model inference, Table 4 details the model performance in the calibration and evaluation periods for the GR4J model calibrated against real-world historical data from the Rogue river catchment. Expectedly, inferred model realization with original data outperform all data transformation runs in terms of best and mean posterior RMSE in the calibration and evaluation periods, with one exception in the evaluation period in which Mexican Hat wavelet results marginally outperform the original data in terms of both best and mean posterior RMSE. This is a rather interesting result showing the consistency of model performance derived from wavelet analysis (with Mexican Hat mother wavelet) over the calibration and evaluation periods. Indeed, wavelet analysis shows a great promise, compared to other data transformations in this study, to improve model inference practice. However, given the results of synthetic analysis (Figures 2-4), in which posterior parameter distributions from the wavelet transformation analysis do not encapsulate the "true" parameter value, we suggest retaining a level of caution in employing wavelet transformed data as sole source of information in the model data synthesis practice. It is also noticed in the calibration period that SQRT and BoxCox results are marginally superior to that of original data in covering a higher percentage of observation data in the 95% parameter predictive uncertainty ranges (10.24% and 10.13% for SQRT and BoxCox, respectively, versus 5.37% for the original data). Moreover, SQRT's results are superior to that of BoxCox, since a higher rate of coverage (10.24% for SQRT versus 10.13% BoxCox) is obtained in

689 a smaller predictive uncertainty spread (7.84 m^3/s vs 9.13 m^3/s). The superiority of
690 SQRT and BoxCox transformations over original data, however, does not hold in the
691 evaluation period, as results of the original data analysis outperform their SQRT and
692 BoxCox transformed counterparts in coverage of observed streamflow in the 95%
693 parameter predictive uncertainty range (8.38% for the original data, versus 7.17% for
694 SQRT and 5.31% for BoxCox). FDC, FFT and NQT transformations, as expected,
695 consistently deteriorate model performance in terms of predictive uncertainty ranges and
696 associated coverage of observational data, as well as best and mean posterior RMSE.

697

698 In general, results of this study show that transformation of calibration data imposes a
699 significant impact on identifiability of hydrological model parameters, and its
700 corresponding predictive performance. Data manipulation may not significantly help
701 extract information more readily from the data, and the major influencing source of
702 information for the calibration remains the original (raw) data. Some data transformations
703 such as SQRT and BoxCox might improve model predictive performance in the calibration
704 period, but this superiority does not extend to the evaluation period. Other transformations,
705 such as wavelet families, reduce posterior parameter uncertainty ranges at the expense of
706 not encapsulating the "true" parameter value (as in synthetic studies). Finally, some data
707 transformations, such as FDC and NQT lose important pieces of information during the
708 data manipulation process, returning significantly inferior results than that of the original
709 data. Hence, choice of data transformation needs an extra attention.

710

711 Note that spectral, and wavelet spectral analysis have some interesting characteristics
712 which can be helpful for hydrological modeling. Spectral analysis can be used for
713 calibrating ungauged basins where only scarce and old data is available. In this case,
714 spectral density of measurements, which can also be calculated using data from similar
715 basins, is compared with the spectral density of simulations [Montanari and Toth, 2007;
716 Quets et al., 2010]. Moreover, the ability of wavelet transformation to decompose a time
717 series to time and scale/frequency localized components enables the modeler to diagnose
718 model inaccuracies, insufficiencies and errors based on different scales, as well as the
719 temporal scale in which mismatch between measurements and simulations occurred.
720 However, one should notice that decomposing a time series to several components and
721 compare the constituent components in the likelihood function increases the computational
722 cost of the parameter estimation algorithm.

723

724 The commonly used likelihood/objective functions in the field of hydrology aggregate
725 discrepancies at different time steps into one single value, and make it impossible to take
726 full advantage of the wavelet analysis. One suggestion is to develop new objective
727 functions that keep the frequency and time localization. Lane [2007] suggested that the
728 most appropriate way to employ wavelet analysis might be using a set of objective
729 functions to choose a cluster of “plausible” model predictions, and then use the wavelet-
730 based objective functions to act as a diagnostic measure to help improve the model, and
731 also help realize the impacts of each parameter set on model prediction. So future research
732 should focus on developing one/multiple objective function(s) which can use the

733 localization feature of wavelet spectral analysis to constrain model parameters, and capture
734 a more complete set of watershed characteristics.

735

736 To sum up, the choice of data transformation requires attention. The framework in which
737 hydrological signatures (acquired through data transformation) are used for model
738 inference also plays a pivotal role on the parameter identifiability and model predictive
739 performance. While our previous research signifies the importance of FDC, as one
740 hydrological signature in a cohort of metrics, in the approximate Bayesian framework
741 [Vrugt and Sadegh, 2013; Sadegh and Vrugt, 2013; Sadegh and Vrugt, 2014; Sadegh et al.,
742 2015], our current study shows FDC transformation of data when used within the Bayesian
743 context with a residual-based likelihood function is indeed disruptive to the model
744 calibration process. Moreover, some data transformations in a single objective formal
745 Bayesian context introduce parameter multimodality, and distort the marginal posterior
746 parameter distributions. We suggest using data transformations within a multi-objective
747 calibration framework that can separately capture and retain different pieces of information
748 for model analysis. Finally, some data transformations might suggest more potential for
749 model evaluation and analysis, than model calibration.

750 5 CONCLUSIONS

751 In this paper, we analyzed the impacts of data transformations on the posterior parameter
752 distributions, and convergence rate of the MCMC simulation. This latter analysis is used
753 to assess how readily the information contained in the discharge time series can be
754 extracted for parameter estimation purposes. Two relatively simple hydrological models

were used to illustrate our results. Our analysis shows that calibrating a hydrological model in time domain, depending on the calibration framework, is generally superior to the spectral or wavelet spectral domains. Our results also convincingly demonstrate that transformations that disaggregate the temporal order of the calibration data are not recommended, as they deteriorate the model performance. One such transformation is FDC, which is finding increasing use in hydrological modeling as a hydrological signature. Sole use of this hydrological signature is not productive. Square root and BoxCox transformations can help solve the parameter identifiability problem, especially those parameters that hold low sensitivity in a quadratic objective function. It is most productive to use data transformations in a multi-objective optimization framework that can extract and retain multiple pieces of information. Finally, some data transformations might suggest more potential for model evaluation and analysis, rather than calibration.

ACKNOWLEDGEMENTS:

We obtained hydrological data for the watersheds studied herein from the MOPEX dataset freely available to public at: ftp://hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data/. MATLAB codes for the hydrological models and MCMC algorithm can be obtained from the corresponding author.

REFERENCES

Aghakouchak, A. and Habib, E.: Application of a conceptual hydrologic model in teaching hydrologic processes, International Journal of Engineering Education, 26, 2010.

AghaKouchak, A., Nakhjiri, N., and Habib, E.: An educational model for ensemble streamflow simulation and uncertainty analysis, Hydrology and Earth System Sciences, 17, 2013.

Andrieu, C., and Thoms, J., A tutorial on adaptive MCMC, Statistics and Computing, 18(4), 343–373, 2008.

Bennett N.D., Croke B.F., Guariso G., Guillaume J.H., Hamilton S.H., Jakeman A.J., Marsili-Libelli S., Newham L.T., Norton J.P., Perrin C., Pierce S.A.: Characterising performance of environmental models. Environmental Modelling & Software, 40, 1-20, 2013.

Blazkova, S. and Beven K.: A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic, Water Resour. Res., 45, W00B16, doi:10.1029/2007WR006726, 2009.

Box, G. E. P., and Cox, D. R.: The analysis of transformations, J. Roy. Statist. Soc., Series B, 26(2), 211-252, 1964.

Boyle, D.P.: Multicriteria calibration of hydrologic models, University of Arizona, 2001.

Dhanya, C. T. and Kumar D. N.: Predictive uncertainty of chaotic daily streamflow using ensemble wavelet networks approach, Water Resour. Res., 47, W06507, doi:10.1029/2010WR010173, 2011.

Dotto, C.B.S., Kleidorfer, M., Deletic, A., Rauch, W., and McCarthy, D.T., Impacts of measured data uncertainty on urban stormwater models. Journal of hydrology, 508, 28-42, 2014.

808 Duan, Q., Sorooshian, S., and Gupta, V.: Effective and efficient global optimization for
809 conceptual rainfall-runoff models, Water Resour. Res., 28(4), 1015–1031, 1992,
810 doi:10.1029/91WR02985.

811

812 Euser, T., Winsemius, H.C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., and Savenije,
813 H.H.G.: A framework to assess the realism of model structures using hydrological
814 signatures, Hydrology and Earth System Sciences, 17, 1893-1912, 2013, doi:10.5194/hess-
815 17-1893-2013

816

817 Fiering, M. B.: Streamflow Synthesis, Harvard University Press, Cambridge, MA, 1967.

818

819 Gharari, S., Hrachowitz, M., Fenicia, F., and Savenije, H.: An approach to identify time
820 consistent model parameters: sub-period calibration, Hydrology and Earth System
821 Sciences, 17, 149–161, 2013.

822

823 Gelman, A., and Rubin D. B.: Inference from iterative simulation using multiple
824 sequences, Statistical Science, 7, 457–472, 1992.

825

826 Gilks, W. R., Roberts, G.O. and George E. I., Adaptive direction sampling, The statistician,
827 179–189, 1994.

828

829 Gordon, N. D., McMahon, T. A. and Finlayson, B. L.: Stream hydrology-an introduction
830 for ecologists. John Wiley & Sons, New York, N.Y., 373-377, 1992.

831

832 Gupta H. V., Wagener T. and Liu Y.: Reconciling theory with observations: elements of a
833 diagnostic approach to model evaluation, Hydrological Processes, 22, 3802–3813, 2008,
834 doi: 10.1002/hyp.6989.

835

836 Guzman J.A., Shirmohammadi A., Sadeghi A., Wang X., Chu M.L., Jha M.K., Parajuli P.,
837 Harmel D., Khare Y., and Hernandez, J.E. : Uncertainty considerations in calibration and
838 validation of hydrologic and water quality models. Transactions of the American Society
839 of Agricultural and Biological Engineers, 58(6), 1745-1762, 2015.

840

841 Haario, H., Saksman, E., and Tamminen, J. : Adaptive proposal distribution for random
842 walk metropolis algorithm, Computational Statistics,14(3), 375–396, 1999.

Haario, H., Saksman, E. , and Tamminen, J. : An adaptive metropolis algorithm, Bernoulli, 223–242, 2001.

Hosking, J. R. M. and Wallis, J. R.: The effect of intersite dependence on regional flood frequency analysis, Water Resources Research, 24, 588– 600, 1988.

Jakeman, A.J., Letcher, R.A., Norton, J.P.: Ten iterative steps in development and evaluation of environmental models. Environmental Modelling and Software, 21(5), 602-614, 2006.

Kelly, K. S. and Krzysztofowicz, R.: A bivariate meta-Gaussian density for use in hydrology, Stochastic Hydrology and Hydraulics, 11, 17-31, 1997.

Kuczera, G.: Comprehensive at-site flood frequency analysis using Monte Carlo Bayesian inference, Water Resources Research, 35(5), 1551–1557, 1999, doi:10.1029/1999WR900012.

Labat D., Ronchail, J. and Guyot, J. L.: Recent advances in wavelet analyses: Part 2—Amazon, Parana, Orinoco and Congo discharges time scale variability, J. Hydrol., 314, 289–311, doi:10.1016/j.jhydrol.2005.04.004, 2005.

Lafrenière M. and Sharp M.: Wavelet analysis of inter-annual variability in the runoff regimes of glacial and nival stream catchments, Bow Lake, Alberta., Hydrol. Process., 17: 1093–1118, 2003.

Lane S. N.: Assessment of rainfall-runoff models based upon wavelet analysis, Hydrol. Process. 21, 586–607, DOI: 10.1002/hyp.6249, 2007.

Madadgar, S., and Moradkhani, H.: Improved Bayesian multimodeling: Integration of copulas and Bayesian model averaging. Water Resources Research, 50(12), 9586-9603, 2014.

Male, J. W. and Ogawa, H.: Tradeoffs in water quality management, J. Water Resour. Plann. Manage., 110(4), 434-444, 1984.

878

879 Mantovan, P. and Todini, E.: Hydrological forecasting uncertainty assessment:
880 Incoherence of the GLUE methodology, J. Hydrol., 330(1-2), 368-381, ISSN 0022-1694,
881 10.1016/j.jhydrol.2006.04.046, 2006.

882

883 Misirli F, Gupta H. V., Sorooshian S. and Thiemann M.: Bayesian recursive estimation of
884 parameter and output uncertainty for watershed models. In: Duan et al (eds) Calibration of
885 watershed models, Water Sci. Appl. Ser., AGU, Washington, 6, 113–124, 2003.

886

887 Mitchell, W. D.: Flow duration curves of Illinois streams, Illinois Dept. of Public Works
888 and Buildings, Division of Waterways, 1957.

889

890 Montanari, A., Rosso, R. and Taqqu, M. S.: A seasonal fractional ARIMA model applied
891 to the Nile River monthly flows at Aswan, Water Resour. Res., 36, 1249– 1259, 2000.

892

893 Montanari, A.: Long range dependence in hydrology, in Theory and Applications of Long-
894 Range Dependence, edited by P. Doukhan, G. Oppenheim, and M. S. Taqqu, Springer,
895 New York, 461– 472, 2003.

896

897 Montanari, A. and Brath, A.: A stochastic approach for assessing the uncertainty of rainfall-
898 runoff simulations, Water Resour. Res., 40, W01106, doi:10.1029/2003WR002540, 2004.

899

900 Montanari, A. and Toth E.: Calibration of hydrological models in the spectral domain: An
901 opportunity for scarcely gauged basins?, Water Resour. Res., 43, W05434,
902 doi:10.1029/2006WR005184, 2007.

903

904 Moran, P. A. P.: Simulation and evaluation of complex water system operations, Water
905 Resour. Res., 6, 1737–1742, 1970.

906

907 Pauwels, V. R. N. and De Lannoy, G. J. M.: Multivariate calibration of a water and energy
908 balance model in the spectral domain, Water Resour. Res., 47, W07523,
909 doi:10.1029/2010WR010292, 2011.

910

911 Perrin, C.: Vers une amélioration d'un modele pluie-débit au travers d'une approche
912 comparative, Ph.D. thesis, Ph. D. Thesis, INP Grenoble/Cemagref Antony, France, 2000.

Perrin, C., Michel, C., and Andréassian, V.: Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments, Journal of Hydrology, 242, 275–301, 2001.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 279, 275–289, 2003.

Pushpalatha, R., Perrin, C., Le Moine, N. and Andréassian, V.: A review of efficiency criteria suitable for evaluating low-flow simulations. Journal of Hydrology, 420, 171-182, 2012.

Quets J. J., De Lannoy, G. J. M. and Pauwels, V. R. N.: Comparison of spectral and time domain calibration methods for precipitation-discharge processes, Hydrol. Process. 24, 1048–1062, DOI: 10.1002/hyp.7546, 2010.

Reed, P., Minsker B. S. and Goldberg, D. E.: Designing a competent simple genetic algorithm for search and optimization, Water Resour. Res., 36(12), 3757–3761, 2000.

Reed, P., Minsker, B. S., and Goldberg, D. E.: Simplifying multiobjective optimization: An automated design methodology for the nondominated sorted genetic algorithm-II, Water Resour. Res., 39(7), 1196, doi:10.1029/2002WR001483, 2003.

Roberts, G. O., and J. S. Rosenthal: Examples of adaptive mcmc, Journal of Computational and Graphical Statistics, 18(2), 349–367, 2009.

Roberts, G. O., and Sahu, S. K.: Updating schemes, correlation structure, blocking and parameterization for the gibbs sampler, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 59(2), 291–317, 1997.

Sadegh, M., and Vrugt, J. A.: Bridging the gap between GLUE and formal statistical approaches: approximate Bayesian computation. Hydrology and Earth System Sciences, 17(12), 2013.

947 Sadegh, M., and Vrugt, J. A.: Approximate bayesian computation using Markov chain
948 Monte Carlo simulation: DREAM$_{(ABC)}$. Water Resources Research, 50(8), 6767-6787,
949 2014.

950

951 Sadegh, M., Vrugt, J. A., Xu, C., and Volpi, E.: The stationarity paradigm revisited:
952 Hypothesis testing using diagnostics, summary metrics, and DREAM$_{(ABC)}$. Water
953 Resources Research, 51(11), 9207-9231, 2015.

954

955 Sadegh, M., Vrugt, J. A., Gupta, H. V., and Xu, C.: The soil water characteristic as new
956 class of closed-form parametric expressions for the flow duration curve. Journal of
957 Hydrology, 535, 438-456, 2016.

958

959 Sadegh, M., Ragno, E., and AghaKouchak, A.: Multivariate Copula Analysis Toolbox
960 (MvCAT): Describing dependence and underlying uncertainty using a Bayesian
961 framework. Water Resources Research, 53, doi:10.1002/2016WR020242.

962

963 Schaefli B., Maraun D. and Holschneider, M.: What drives high flow events in the Swiss
964 Alps? Recent developments in wavelet spectral analysis and their application to hydrology,
965 Adv. in Water Resour., 30, 2511–2525, doi:10.1016/j.advwatres.2007.06.004, 2007.

966

967 Schoups, G., and Vrugt, J. A.: A formal likelihood function for parameter and predictive
968 inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian
969 errors. Water Resources Research, 46(10), 2010.

970

971 Searcy, J. K.: Flow-duration curves, Water Supply Paper 1542-A, U.S. Geological Survey,
972 Reston, Virginia, 1959.

973

974 Shannon C. E.: Communication in the presence of noise. Proceedings of the Institute of
975 Radio Engineers, 13(1): 10–21, 1949.

976

977 Sorooshian, S. and Dracup, J. A.: Stochastic parameter estimation procedures for
978 hydrologie rainfall-runoff models: Correlated and heteroscedastic error cases, Water
979 Resour. Res., 16(2), 430–442, doi:10.1029/WR016i002p00430, 1980.

980

981 Storn, R., and Price, K. : Differential evolution-a simple and efficient adaptive scheme
982 for global optimization over continuous spaces, vol. 3, ICSI Berkeley, 1995.

Storn, R., and Price, K.: Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces, Journal of global optimization, 11(4), 341–359, 1997.

ter Braak, C. J. : A markov chain monte carlo version of the genetic algorithm differential evolution: easy bayesian computing for real parameter spaces, Statistics and Computing, 16(3), 239–249, 2006.

Thiemann, M., Trosset, M., Gupta, H. and Sorooshian, S.: Bayesian recursive parameter estimation for hydrologic models, Water Resour. Res., 37(10), 2521–2536, doi:10.1029/2000WR900405, 2001.

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W., & Srikanthan, S.: 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.

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

Torrence C. and Compo G. P.: A practical guide to wavelet analysis. Bulletin of the American Meteorological Society, 79, 61–78, 1998.

Vogel R. M. and Fennessey N. M.: Flow-Duration Curves. I: New Interpretations and Confidence Interval, J. Water Resour. Plann. Manage., 120(4), 485-504, 1994.

Vogel, R. M., and Fennessey, N. M.: Flow duration curves II: a review of applications in water resources planning. JAWRA Journal of the American Water Resources Association, 31(6), 1029-1039, 1995.

Vrugt, J. A., Gupta, H. V., Sorooshian, S., Wagener, T. and Bouten, W.: Application of stochastic parameter optimization to the Sacramento Soil Moisture Accounting model, J. Hydrol., 325(1–4), 288–307, doi:10.1016/j.hydrol.2005.10.041, 2006.

1017 Vrugt, J. A., ter Braak, C. J. F., Diks, C. G. H., Higdon, D., Robinson, B. A. and Hyman,
1018 J. M.: Accelerating Markov chain Monte Carlo simulation by differential evolution with
1019 self-adaptive randomized subspace sampling, I. J. Nonlin. Sci. Num., 10, 271–288, 2009.

1020

1021 Vrugt, J. A., and Sadegh, M.: Toward diagnostic model calibration and evaluation:
1022 Approximate Bayesian computation. Water Resources Research, 49(7), 4335-4345, 2013.

1023

1024 Wagener, T., Boyle D. P., Lees, M. J., Wheater, H. S., Gupta, H. V. and Sorooshian S.: A
1025 framework for development and application of hydrological models, Hydrol. Earth Syst.
1026 Sci., 5(1), 13– 26, 2001.

1027

1028 Westerberg, I.K., Guerrero, J.L., Younger, P.M., Beven, K.J., Seibert, J., Halldin, S., Freer,
1029 J.E., and Xu, C.Y.: Calibration of hydrological models using flow-duration
1030 curves. Hydrology and Earth System Sciences, 15(7), p.2205, 2011.

1031

1032 Yapo, P.O., Gupta, H.V. and Sorooshian, S.: Automatic calibration of conceptual rainfall-
1033 runoff models: sensitivity to calibration data. Journal of Hydrology, 181(1-4), 23-48, 1996.

Table 1. Parameter description of HYMOD and GR4J models, Snow Module, and their description and prior range.

	Description	**Minimum**	**Maximum**	**Units**
HyMod				
C_{max}	Maximum storage in watershed	1	500	mm
b_{exp}	Spatial variability of soil moisture storage	0.1	2.00	-
α	Distribution factor between two reservoirs	0.1	0.99	-
R_s	Residence time of slow flow reservoir	0.001	0.1	days
R_q	Residence time of quick flow reservoir	0.1	0.99	days
GR4J				
$S1_{max}$	Maximum storage in watershed	0	1500	mm
Exch	Spatial variability of soil moisture storage	-10	10	mm
$S2_{max}$	Distribution factor between two reservoirs	1	500	mm
UHB	Residence time slow flow reservoir	0.5	8	days
Snow Module				
DD	Degree-day factor	0.01	7	mm/℃

Table 2. Description of watersheds used in this study, including their USGS ID, catchment area (km^2), mean annual precipitation (P, mm), mean annual potential evapotranspiration (PET, mm), and mean annual runoff ratio (RR, -).

Watershed	USGS ID	Area (km^2)	P (mm)	PET (mm)	RR (-)
French Broad	03455000	4812.20	1402.64	771.25	0.40
Skykomish	12134500	1385.64	2701.65	668.82	0.66
Rogue	14359000	5317.25	1052.09	849.71	0.48

Table 3. Parameter and total predictive uncertainty spread (PUS and TPUS, respectively, m^3/s) of 95% streamflow prediction ranges, and associated coverage of observation points, as well as best and mean RMSE (m^3/s) of the posterior samples for the GR4J model calibrated against synthetic data from the Skykomish river basin.

	PUS (m^3/s)	**TPUS (m^3/s)**	**Coverage in PUS** (%)	**Coverage in TPUS** (%)	**Best RMSE (m^3/s)**	**Mean RMSE (m^3/s)**
Calibration period						
None	12.37	225.29	8.93	93.54	78.82	78.93
FDC	1.19	219.46	0.60	87.02	118.30	118.50
BoxCox	10.22	206.34	6.13	93.43	81.97	84.25
SQRT	9.66	213.74	5.70	94.41	80.44	81.45
NQT	37.56	198.68	16.98	92.17	84.29	91.73
MexH wavelet	2.97	219.65	2.41	93.70	78.94	79.09
Morl wavelet	7.78	229.34	6.52	92.99	78.95	79.38
FFT	32.49	218.33	13.03	90.42	94.30	101.74
Evaluation period						
None	12.46	228.94	7.56	93.04	80.41	80.95
FDC	1.13	220.44	0.16	86.36	113.96	114.26
BoxCox	10.61	208.96	6.13	91.62	85.18	88.13
SQRT	9.87	217.10	5.53	92.55	83.23	84.60
NQT	39.83	200.40	17.96	90.69	88.02	96.96
MexH wavelet	2.97	223.49	1.92	93.04	81.17	81.43
Morl wavelet	8.01	233.21	5.59	93.59	80.77	80.95
FFT	33.69	222.7	13.2	89.98	94.39	102.7

Table 4. Parameter and total predictive uncertainty spread (PUS and TPUS, respectively, m^3/s) of 95% streamflow prediction ranges, and associated coverage of observation points, as well as best and mean RMSE (m^3/s) of the posterior samples for the GR4J model calibrated against real-world historical data from the Rogue river basin.

	PUS (m^3/s)	**TPUS (m^3/s)**	**Coverage in PUS** (%)	**Coverage in TPUS** (%)	**Best RMSE (m^3/s)**	**Mean RMSE (m^3/s)**
Calibration period						
None	7.00	154.81	5.37	95.56	47.98	48.06
FDC	2.14	159.01	2.57	94.30	55.68	56.13
BoxCox	9.13	171.82	10.13	94.36	51.33	52.23
SQRT	7.84	162.85	10.24	94.74	49.54	50.32
NQT	35.73	155.85	14.07	93.76	63.49	67.66
Mexh wavelet	5.27	155.16	2.85	95.24	49.80	50.47
Morl wavelet	3.56	159.81	1.53	95.24	58.65	60.3
FFT	10.94	192.04	7.28	94.74	70.91	72.76
Evaluation period						
None	7.11	156.50	8.38	94.96	55.84	57.31
FDC	2.20	160.65	2.46	93.32	68.20	69.08
BoxCox	10.03	172.78	5.31	91.35	65.76	67.72
SQRT	8.70	164.30	7.17	91.84	61.84	63.92
NQT	38.69	159.72	15.39	97.43	60.82	65.16
Mexh wavelet	6.17	157.24	5.31	96.00	53.46	53.93
Morl wavelet	3.92	162.67	1.7	96.88	56.72	57.87
FFT	13.17	192.53	8.11	92.77	83.41	86.46

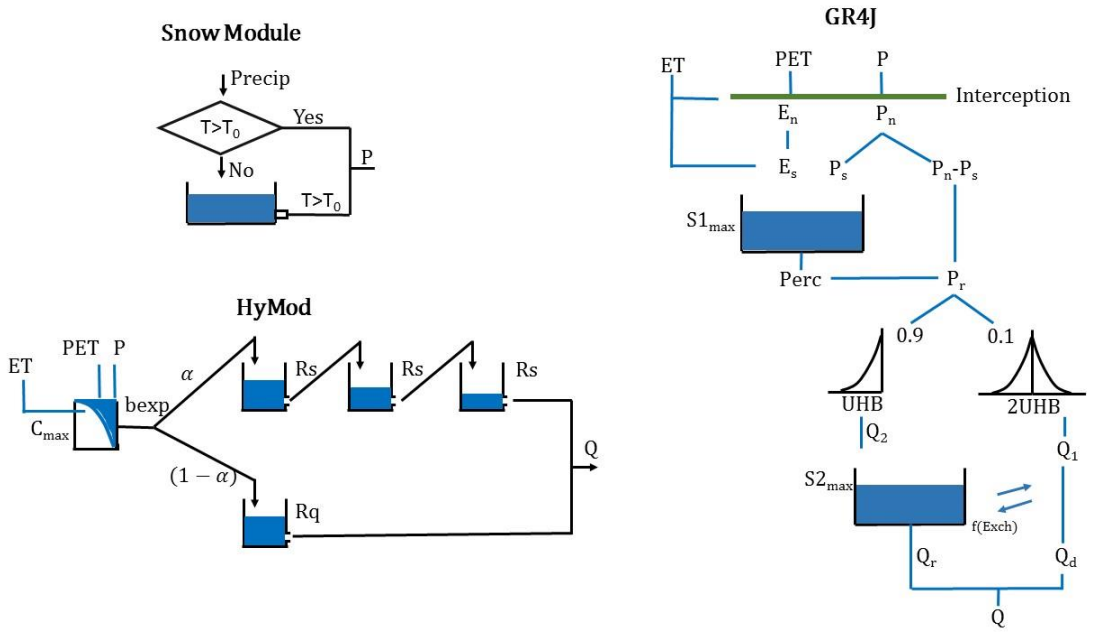

Figure 1. Schematic representation of Snow Module, HyMod and GR4J conceptual RR models.

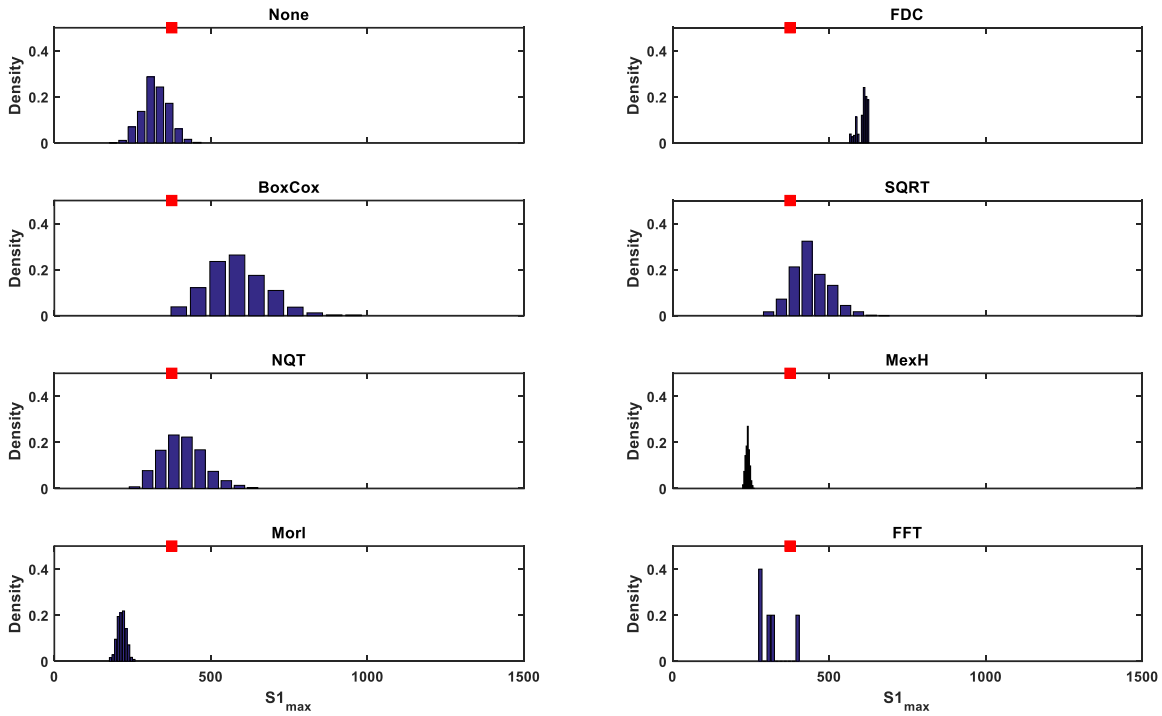

Figure 2. Posterior distribution of first parameter ($S1_{max}$, maximum capacity of the production store) of the GR4J model, inferred through MCMC simulation with different transformations of synthetic data from the French Broad river catchment. Red square on top of each plot shows the "true" parameter value used to generate the synthetic data. Each plot represents results of a data transformation, as described in the plot's title.

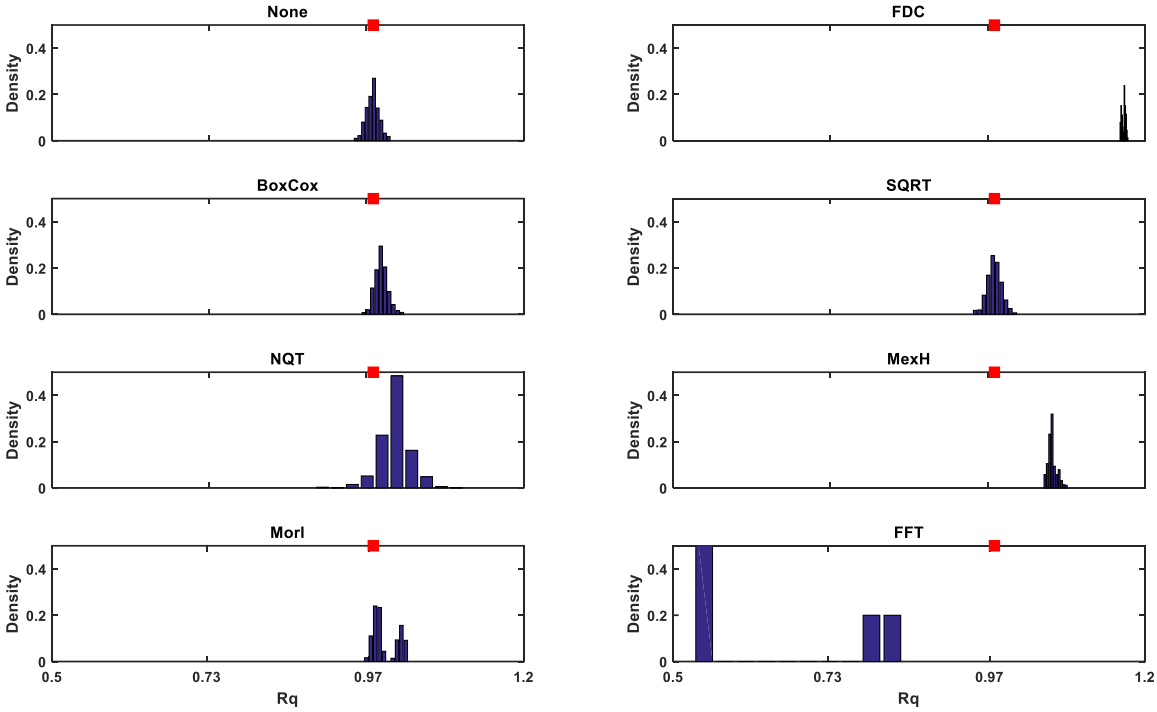

Figure 3. Posterior distribution of fifth parameter (R_q, residence time of quick flow reservoir) of the HyMod model, inferred by MCMC simulation using different transformations of synthetic data from the Skykomish river catchment. Red square on top of each plot shows the "true" parameter value used to generate the synthetic data. Each plot represents results of a data transformation, as described in the plot's title.

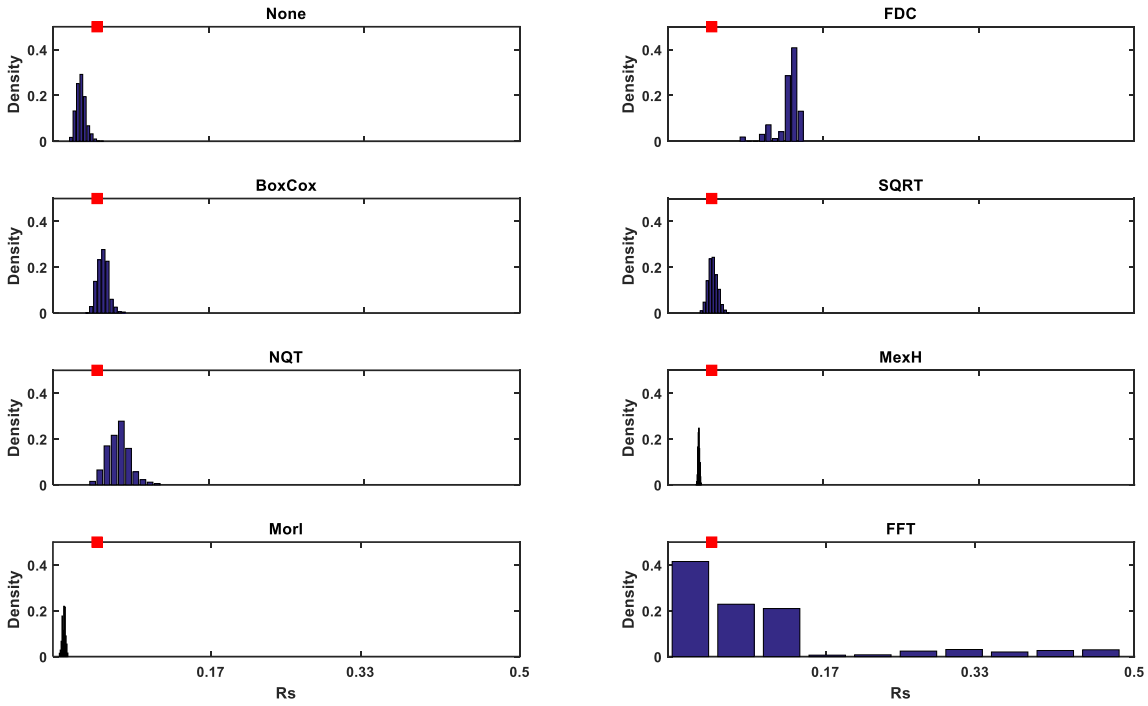

Figure 4. Posterior distribution of fourth parameter (R_s, residence time of slow flow reservoir) of the HyMod model, inferred by MCMC simulation using different transformations of synthetic data from the Rogue river catchment. Red square on top of each plot shows the "true" parameter value used to generate the synthetic data. Each plot represents results of a data transformation, as described in the plot's title.

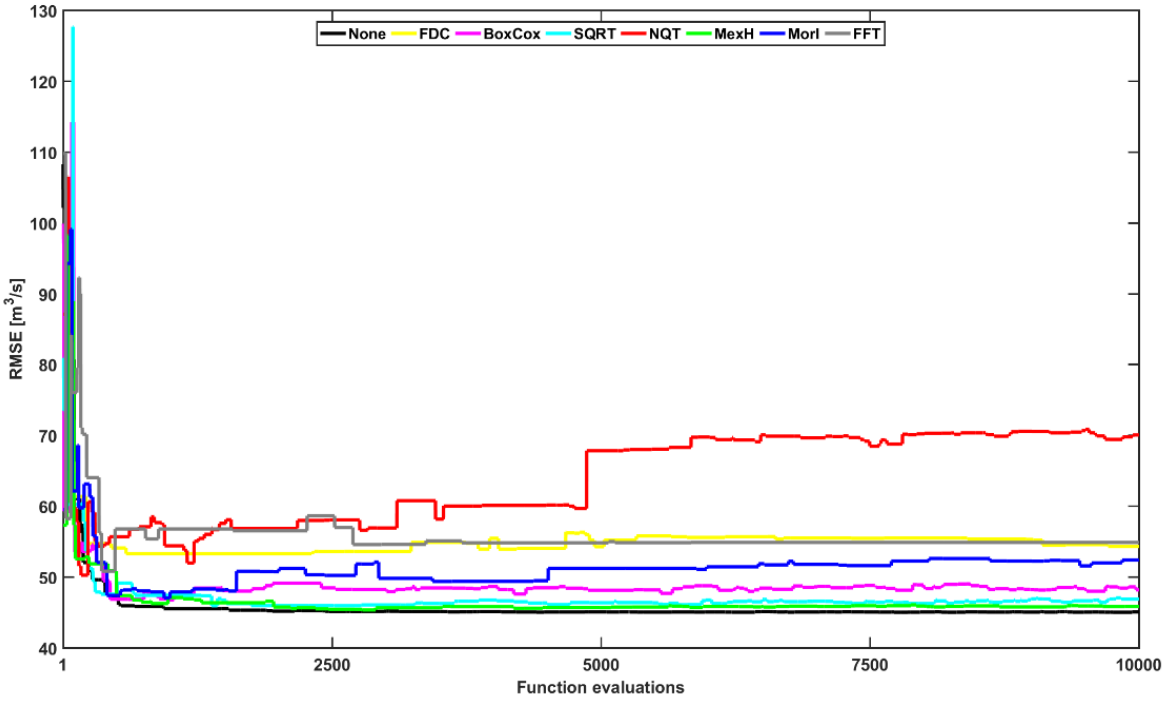

Figure 5. Evolution of moving minimum RMSE values derived from MCMC samples for the GR4J model analyzing the Rogue river catchment. Minimum RMSE in a sliding window of 50 Markov chain samples is presented in this figure, which is a proxy for the convergence speed of MCMC simulation. We only present, in this figure, the first 10,000 iterations of the total 35,000 iterations used for inference, as Markov chains reach their stable state in this range.