AGU100 ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE

10.1002/2017WR021616

This article is a companion to Kavetski et al. [2018], https://doi.org/10.1002/2017WR020528.

Key Points:

- Bayesian signature-domain inference enables quantification of predictive streamflow uncertainty in the time-domain
- The use of signatures usually leads to a loss of information, and hence wider parameter posteriors and streamflow predictive distributions
- In some cases, careful selection of signatures can partially reduce the impact of deficiencies in the assumed model on its calibration

Correspondence to:
D. Kavetski,
dmitri.kavetski@adelaide.edu.au

Citation:
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*, 3958–3987. https://doi.org/10.1002/2017WR021616

Received 5 AUG 2017
Accepted 27 MAR 2018
Accepted article online 6 APR 2018
Published online 27 JUN 2018

Signature-Domain Calibration of Hydrological Models Using Approximate Bayesian Computation: Empirical Analysis of Fundamental Properties

Fabrizio Fenicia[1], Dmitri Kavetski[1,2], Peter Reichert[1], and Carlo Albert[1]

[1]Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland, [2]School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, Australia

Abstract This study investigates Bayesian signature-domain inference of hydrological models using Approximate Bayesian Computation (ABC) algorithms, and compares it to "traditional" time-domain inference. Our focus is on the quantification of predictive uncertainty in the streamflow time series and on understanding the information content of particular combinations of signatures. A combination of synthetic and real data experiments using conceptual rainfall-runoff models is employed. Synthetic experiments demonstrate: (i) the general consistency of signature and time-domain inferences, (ii) the ability to estimate streamflow error model parameters (reliably quantify streamflow uncertainty) even when calibrating in the signature domain, and (iii) the potential robustness of signature-domain inference when the (probabilistic) hydrological model is misspecified (e.g., by unaccounted timing errors). The experiments also suggest limitations of the signature-domain approach in terms of information loss when general (nonsufficient) statistics are used, and increased computational costs incurred by the ABC implementation. Real data experiments confirm the viability of Bayesian signature-domain inference and its general consistency with time-domain inference in terms of predictive uncertainty quantification. In addition, we demonstrate the utility of the flashiness index for the estimation of streamflow error parameters, and show that signatures based on the Flow Duration Curve alone are insufficient to calibrate parameters controlling streamflow dynamics. Overall, the study further establishes signature-domain inference (implemented using ABC) as a promising method for comparing the information content of hydrological signatures, for prediction under data-scarce conditions, and, under certain circumstances, for mitigating the impact of deficiencies in the formulation of the predictive model.

1. Introduction

Most if not all environmental models have parameters that cannot be measured directly and are estimated through calibration. Model calibration is therefore common to many areas of environmental sciences, including hydrology (e.g., Razavi & Tolson, 2013), ecology (e.g., Paillex et al., 2017), biology (e.g., Zhu et al., 2015), meteorology (e.g., Duan et al., 2017), and others.

In catchment-scale hydrological modeling, the goal of model calibration is typically to find model parameter values that lead to the closest match between simulated and observed catchment-scale responses. Model outputs used for calibration include streamflow (e.g., Beven & Binley, 1992; Gupta et al., 1998), groundwater levels (e.g., Freer et al., 2004; Seibert & McDonnell, 2002), concentrations of various substances including water isotopes (e.g., Fenicia et al., 2008; Son & Sivapalan, 2007), mineral solutes (e.g., Benettin et al., 2015; Hrachowitz et al., 2013), and herbicides/pesticides (e.g., Bertuzzo et al., 2013; Gassmann et al., 2013).

The most common approach to model calibration is to seek to match the observed and simulated time series directly, for example by constructing objective functions and performance metrics that operate directly on the observed and simulated streamflow time series. We refer to this approach as calibration in the "time domain." An alternative approach is to compare "signatures" of the observed and simulated time series. In the case of streamflow, these signatures may include quantiles of the streamflow distribution (values of the flow duration curve, FDC), the base flow index, the flashiness index and many others (e.g., Kavetski et al., 2018; Westerberg & McMillan, 2015). We refer to this approach as calibration in the "signature domain."

In hydrological modeling, signature-domain calibration is driven mainly by the following motivations:

Motivation 1. Calibration and prediction in data-poor settings. For example, when observed rainfall and streamflow time series are nonconcomitant, calibration in the signature domain can proceed by assuming the streamflow signatures are approximately constant in time (Montanari & Toth, 2007; Sadegh et al., 2015; Winsemius et al., 2006). In ungauged catchments, some streamflow signatures such as flow duration curves (FDCs) can be estimated by regionalization (e.g., Botter et al., 2008; Castellarin et al., 2007) and used for calibration (e.g., Yadav et al., 2007; Yu & Yang, 2000).

Motivation 2. Ability to emphasize or deemphasize the fitting of specific hydrograph features, in order to:

1. Tailor the calibration to specific modeling objectives, such as the timing and magnitude of flood peaks, low flow volumes, etc. For example, Jepsen et al. (2016) calibrated the Penn State Integrated Hydrologic Model (PIHM) to the shape of the base flow recession curve, as their focus was to capture the relationship between groundwater flow and subsurface water storage;
2. Improve confidence in the model realism, by directly forcing the model to reproduce hydrologically important characteristics of the hydrograph (e.g., Gupta et al., 2008; Sivapalan, 2006). For example, the base flow index reflects the partitioning of effective rainfall between fast and slow pathways, whereas the flashiness index reflects the dynamics of the flow. In terms of catchment hydrobiogeochemistry, signatures based on isotope ratios can be used as a proxy for the mean residence time (McGuire et al., 2005). The ability to reproduce these multiple distinct signatures (aspects) of catchment dynamics is indicative of the degree to which a model captures the processes it is intended to represent (see also Gupta et al., 2008); and
3. Make the calibration more robust to deficiencies in simple objective functions of streamflow time series, by using signatures that are less sensitive to features poorly reproduced by the hydrological and/or error models. This is a particularly intriguing aspect of calibration in the signature domain. For example, calibration to FDCs, which are insensitive to errors in the timing and magnitude of individual flood peaks, can in theory produce more robust parameter estimates than calibration to (squared) residuals of streamflow time series (e.g., Liu et al., 2011). Similarly, signatures such as the total runoff ratio and base flow index, which are based on spatial and/or temporal averages of the streamflow time series, can filter out noise and reduce the sensitivity of the objective function to data uncertainty (e.g., Sadegh et al., 2015; Westerberg & McMillan, 2015).

Like any modeling choice, signature-domain calibration has its downsides:

First, the use of signatures is usually associated with a loss of information. This loss of information may be undesired, and have a negative impact on parameter estimates and predictions. In principle, this problem can be avoided by using a set of sufficient statistics as signatures. However, proving that a set of statistics is sufficient for a given hydrological model, let alone deriving it, is a daunting task with no general guidelines; for this reason, practical interest is in "informative" rather than "strictly sufficient" statistics (e.g., see Kavetski et al., 2018, section 5.3.1). There are varying suggestions in the literature regarding the choice of streamflow signatures for the hydrological model calibration. For example, Westerberg et al. (2011) and Yu and Yang (2000) suggest that signatures based on the FDC can adequately constrain streamflow predictions; Yadav et al. (2007) suggest to use slope of the FDC, runoff ratio, and high pulse count.

Second, even when calibrating in the signature domain, the goal is seldom to predict the signatures themselves (e.g., Botter et al., 2008; Castellarin, 2014). Rather, the goal is usually to predict the underlying variables from which the signatures are calculated (e.g., Shafii & Tolson, 2015; Westerberg et al., 2011; Yilmaz et al., 2008)—and, importantly, to estimate the associated predictive uncertainty. Previous studies have shown that calibrating to streamflow signatures can result in reasonable simulations of streamflow time series (Westerberg et al., 2011; Yadav et al., 2007). But what about the *uncertainty* in these simulated time series? Can it be reliably estimated from the signatures themselves?

In this study, we investigate fundamental properties of Bayesian inference in the signature domain, including its ability to: (i) infer parameters of hydrological models and residual error models; (ii) produce reliable predictive distributions of streamflow time series despite accessing solely signature-domain data; and (iii) serve as a platform for comparing the information content and utility of different streamflow signatures.

Our empirical analyses are grounded in the Bayesian perspective on signature-domain calibration and its numerical implementation using Approximate Bayesian Computation (ABC) sampling algorithms, as detailed in Kavetski et al. (2018). This perspective differs in several important aspects from the majority of current applications of signature-domain inference and ABC algorithms in the hydrological literature (e.g., Nott et al., 2012; Vrugt & Sadegh, 2013). In addition, note that while our presentation focuses on hydrological model calibration and streamflow prediction, the same concepts apply to model calibration to other environmental data sets (e.g., groundwater levels, chemical concentrations) and their signatures.

In order to test the theoretical and practical properties of inference in the signature domain, we distinguish between two scenarios. In *Scenario* 1, the hydrological model is assumed to provide a correct description of the data, and the use of signatures is driven by Motivation 1 above. Under these circumstances, unless a complete set of sufficient statistics is available, signature-domain can be expected to provide *at best an approximation* to the time-domain inference, because (nonsufficient) signatures do not capture the full information content of the time series. In *Scenario* 2, the model does not provide a correct description of the data. Under these circumstances, the modeler may choose to deliberately use nominally nonsufficient signatures as per Motivation 2 above, i.e., to *improve* on the inference in the time domain.

Empirical analysis of Scenario 1 helps understand the theoretical relationship between signature and time domain inference. In particular, theoretical considerations (see Kavetski et al., 2018) suggest that inferences in the time and signature domain should produce consistent parameter and predictive distributions, with signature domain results representing an approximation of time domain results (to the extent that the set of signatures used is nonsufficient). These considerations lead to the following study objectives:

Objective 1. How does the correspondence between inferences in the signature versus time domains depend on the number and type of signatures? The expectation is that, if the selected signatures do not capture the entire information content of the original time series (i.e., if they are not "sufficient"), the parameter, and predictive distributions obtained in the signature domain will be wider than those obtained in the time domain. It can also be expected that, as more "informative" signatures are included, the differences between the results of signature domain versus time domain inferences will diminish.

Objective 2. How does data length affect the inference in the signature domain? It is well known that Bayesian posterior parameter distributions tend to tighten asymptotically as more data is included in the inference (Box & Tiao, 1973; Gelman et al., 2004), unless the data are noninformative with respect to the model setup (or, under some inference setups, if the data is incompatible with the prior, etc). Theoretically, such tightening should also occur if the inference is formulated in the signature domain. However, many signatures, such as the base flow and flashiness indices, are scalar-valued independently from the length of the underlying streamflow time series. Can such signatures support the tightening of the posterior as the streamflow data length increases, and, if so, how is this additional information transferred to the posterior distributions?

Empirical analysis of Scenario 2 helps interpret the differences between signature-domain and time-domain inference under nonideal conditions, where the data and/or model are known to be corrupted by particular types of errors. In this respect, the following objectives are investigated:

Objective 3. Are there cases where signature-domain calibration produces "better" results than time-domain calibration? For example, in the case of data and/or model deficiencies such as timing errors, can the calibration to signatures provide uncorrupted estimates of model parameter and streamflow predictions? How do we interpret such results in comparison with traditional time-domain inference?

Another important question, critical for practical applications, is the computational cost (feasibility) of signature-domain inference. This leads to our next objective, namely:

Objective 4. Is there a major difference in computational costs of signature-domain inference implemented using ABC algorithms and time-domain inference implemented using standard MCMC algorithms? Is there an intuitive reason for these differences?

Objectives 1–4 are investigated using synthetic data, so that the outcomes of the inference in the signatures domain can be compared to a known "ground truth." When investigating Objective 3 above, i.e., inference using a deficient model, the use of synthetic data avoids a speculative interpretation of the results.

The ultimate goal is to apply signature-domain inference using ABC in real data conditions. In order to corroborate the results of the synthetic analysis, we carry out a real data study representative of many practical applications. The real data study can help confirm or refute the outcomes of the synthetic case study, particularly with respect to how the correspondence between time-domain versus signature-domain inference is affected by the number and type of signatures (point 1). The real data study also yields insights into which signatures are most effective in constraining hydrograph predictions. In particular, we focus on signatures based on the FDCs, which are among the most frequently used for summarizing hydrographs (e.g., Castellarin et al., 2013; Sivapalan, 2006; Yilmaz et al., 2008), and explore our final objective in this work:

Objective 5. Are signatures based on the FDC alone sufficient to constrain streamflow predictions, as suggested in previous studies (Westerberg et al., 2011; Yu & Yang, 2000)? If not, which other signatures can be used in addition? How does the uncertainty in the FDC map to the uncertainty in the hydrograph? For example, does 20% error in the FDC correspond to the same error in the streamflow?

The paper is structured as follows. Section 2 presents the theory of the signature-domain and the time-domain inference implemented using the Bayesian framework. It also describes the selection of streamflow signatures and the Box-Cox Gaussian AR1 residual error model used in the case studies. Section 3 describes the case study experiments. Section 4 presents the case study results, followed by a discussion in section 5. Section 6 summarizes the key conclusions of the study and outlines directions for future research.

2. Theory

2.1. Bayesian Inference Framework

Suppose the observed streamflow time series $\tilde{\boldsymbol{q}}$ are treated as a sample from a probabilistic hydrological model $\boldsymbol{Q}(\boldsymbol{\theta},\boldsymbol{x})$,

$$\tilde{\boldsymbol{q}} \leftarrow \boldsymbol{Q}(\boldsymbol{\theta},\boldsymbol{x}) \tag{1}$$

where $\boldsymbol{\theta}$ are model parameters requiring calibration and $\boldsymbol{x}$ represents all required model forcings (e.g., precipitation, potential evaporation, etc).

Denoting the data used for calibration by $\tilde{\boldsymbol{y}}$, time-domain calibration is defined by

$$\tilde{\boldsymbol{y}}=\tilde{\boldsymbol{q}} \quad , \quad \boldsymbol{Y}(\boldsymbol{\theta},\boldsymbol{x})=\boldsymbol{Q}(\boldsymbol{\theta},\boldsymbol{x}) \tag{2}$$

and signature-domain calibration is defined by

$$\tilde{\boldsymbol{y}}=\boldsymbol{g}(\tilde{\boldsymbol{q}}) \quad , \quad \boldsymbol{Y}(\boldsymbol{\theta},\boldsymbol{x})=\boldsymbol{g}(\boldsymbol{Q}(\boldsymbol{\theta},\boldsymbol{x})) \tag{3}$$

where $\boldsymbol{g}(\boldsymbol{q})$ is a vector of signatures computed from streamflow time series $\boldsymbol{q}$.

The posterior distribution of model parameters, $p(\boldsymbol{\theta}|\tilde{\boldsymbol{y}},\boldsymbol{x})$, is given by Bayes' theorem,

$$p(\boldsymbol{\theta}|\tilde{\boldsymbol{y}},\boldsymbol{x})=\frac{p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta},\boldsymbol{x})p(\boldsymbol{\theta})}{p(\tilde{\boldsymbol{y}}|\boldsymbol{x})} \propto p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta},\boldsymbol{x})p(\boldsymbol{\theta}) \tag{4}$$

where $p(\boldsymbol{\theta})$ is the prior distribution of the model parameters, and $p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta},\boldsymbol{x})$ is the likelihood function defined by the probability density function (pdf) $p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{x})$ of the probability model $\boldsymbol{Y}(\boldsymbol{\theta},\boldsymbol{x})$, as given in equation (2) or equation (3), evaluated at the observed data and viewed as a function of parameters $\boldsymbol{\theta}$.

The next step is to specify the structure of the probabilistic model $\boldsymbol{Q}(\boldsymbol{\theta},\boldsymbol{x})$, which defines $\boldsymbol{Y}(\boldsymbol{\theta},\boldsymbol{x})$.

2.2. Description of Predictive Uncertainty in Streamflow Space

Formulating the probabilistic model $\boldsymbol{Q}(\boldsymbol{\theta},\boldsymbol{x})$ requires describing uncertainties involved in the streamflow modeling process (e.g., input, output, and structural uncertainty) and their effect on the model output (simulated streamflow). This study uses a traditional formulation where a deterministic hydrological model $\boldsymbol{h}(\boldsymbol{\theta}_h,\boldsymbol{x})$ is supplemented with a random residual error model $\mathcal{E}(\boldsymbol{\theta}_{\mathcal{E}})$ to describe the total effect of all sources of error (e.g., McInerney et al., 2017). However, Bayesian signature-domain inference can be applied to more complicated stochastic models, including those where uncertainties are described using (potentially multiple) terms internal to the model structure, e.g., when attempting to distinguish uncertainty in the inputs from uncertainty in distinct model components, improve the characterization of structural errors, and so forth (e.g., Albert et al., 2016; Del Giudice et al., 2016; Renard et al., 2011; and others).

The heteroscedasticity and skewness of predictive errors is described using the Box-Cox transformation,

$$z[\mathbf{Q}(\boldsymbol{\theta},\mathbf{x});\lambda]=z[\mathbf{h}(\boldsymbol{\theta}_h,\mathbf{x});\lambda]+\boldsymbol{\mathcal{E}}(\boldsymbol{\theta}_{\mathcal{E}}) \tag{5}$$

from which the probabilistic model $\mathbf{Q}(\boldsymbol{\theta},\mathbf{x})$ can be expressed directly as

$$\mathbf{Q}(\boldsymbol{\theta},\mathbf{x})=z^{-1}[z[\mathbf{h}(\boldsymbol{\theta}_h,\mathbf{x});\lambda]+\boldsymbol{\mathcal{E}}(\boldsymbol{\theta}_{\mathcal{E}});\lambda] \tag{6}$$

where $z[q;\lambda]$ denotes the Box-Cox transformation with parameter λ (e.g., Bates & Campbell, 2001),

$$z[q;\lambda]=\begin{cases}(q^{\lambda}-1)/\lambda & \text{if } \lambda\neq 0\\ \log q & \text{if } \lambda=0\end{cases} \tag{7}$$

The temporal persistence of residual errors is described using a first-order autoregressive (AR1) model,

$$\mathcal{E}_t=\phi\varepsilon_{t-1}+W_t \tag{8}$$

where ϕ is the autoregressive parameter and W_t is the innovation (disturbance).

In order to avoid negative streamflow predictions, the innovations are assumed to follow a truncated Gaussian distribution, which for $\lambda>0$ corresponds to

$$W_t\sim\mathcal{TN}\left(0,\sigma_W,L_{W,t}(\boldsymbol{\theta},\mathbf{x},\varepsilon_{t-1})\right) \tag{9}$$

with parameters $\mu=0$, $\sigma=\sigma_W$ and the lower bound $L_{W,t}$ defined such that $\mathbf{Q}(\boldsymbol{\theta},\mathbf{x})\geq\mathbf{0}$ (which makes it dependent on the particular model realization).

The complete set of error model parameters is $\boldsymbol{\theta}_{\mathcal{E}}=(\lambda,\phi,\sigma_W)$. In this work, we fix λ and ϕ to values given in Table 1 and only calibrate σ_W. The fixing of ϕ is motivated by the findings of Evin et al. (2014) to reduce parameter interactions, and the fixing of λ follows the recommendations of McInerney et al. (2017).

2.3. Hydrological Signatures

The following signatures are considered:

1. Streamflow values corresponding to individual quantiles of the marginal streamflow distribution, given by selected FDC values. The FDC signature corresponding to streamflow quantile X is defined as

$$g_{qX}[\mathbf{q}_{1:N_T}]=q \text{ such that } \#[\mathbf{q}_{1:N_T}<q]/N_T=X/100 \tag{10}$$

where #[ ] denotes the count function. For example, g_{q50} is the median streamflow, g_{q95} is the 95% highest streamflow, and so forth. Here, we consider the 10th, 50th, 75th, and 95th FDC percentiles.

2. The "entire" FDC is used as a (vector) signature,

$$\mathbf{g}_{FDC}[\mathbf{q}_{1:N_T}]=\left\{g_{q\varsigma_i}[\mathbf{q}_{1:N_T}];i=1,\cdots,N_{FDC}\right\} \tag{11}$$

with quantiles points ς selected to provide higher resolution for higher flows (Westerberg et al., 2011),

$$\frac{\varsigma_i}{100}=1-\left(\frac{i+1}{N_{\mathrm{FDC}}+2}\right)^2, \text{for } i=1,\cdots,N_{\mathrm{FDC}} \tag{12}$$

We set $N_{FDC}=20$, in which case the most extreme quantiles within $\mathbf{g}_{FDC}$ are g_{q9} and g_{q99}.

3. The base flow index g_b (e.g., Eckhardt, 2008), defined as the fraction of base flow $q^{(b)}$ within the total streamflow q,

$$g_b[\mathbf{q}_{1:N_T}]=\sum_{t=1}^{N_T}q_t^{(b)}/\sum_{t=1}^{N_T}q_t \tag{13}$$

The base flow was estimated from streamflow using a low-pass filter (Lyne & Hollick, 1979; see also Eckhardt, 2008, equation (5)),

Table 1
Parameters of the Single-Reservoir Hydrological Model and the Error Model in Experiment 1

	Single-reservoir hydrological model (monthly time step)				Residual error model		
	k (mm$^{1-\alpha}$/t$_m$)	c_e	α	m (mm)	σ_W (mm$^\lambda$/t$_m^\lambda$)	λ	ϕ
Synthetic data generation	10^{-3}	1.0	2.0	0.5	2.0	0.5	0.5
Calibration	10^{-4}–10^{-1}	0.1–3	2.0	0.5	10^{-6}–6.0	0.5	0.5

Note. The row labeled "synthetic data generation" lists the parameter values used to generate the synthetic observations in Experiment 1. The row labeled "calibration" lists the bounds of calibrated parameters and the values of fixed parameters. The symbol t$_m$ in the units refers to the monthly time step (4 weeks).

$$q_t^{(b)}=\min\left(q_t, \vartheta_b q_{t-1}^{(b)}+\frac{1-\vartheta_b}{2}(q_{t-1}+q_t)\right) \tag{14}$$

Following Eckhardt (2008), a single forward filter pass is applied, with filtering parameter ϑ_b set to 0.925.

4. The flashiness index g_f (Baker et al., 2004), defined by scaled changes in streamflow,

$$g_f[\boldsymbol{q}_{1:N_T}]=\sum_{t=2}^{N_T}|q_t-q_{t-1}| \Bigg/ \sum_{t=2}^{N_T} q_t \tag{15}$$

and used to describe the "responsiveness" of a catchment.

2.4. Implementation of Signature-Domain Inference

In this work, signature-domain inference is defined as the calibration of the probabilistic model $\boldsymbol{Q}(\boldsymbol{\theta},\boldsymbol{x})$ to the streamflow signatures $\boldsymbol{g}$ from section 2.3. For this choice of signatures, all of which are nonlinear and noninvertible functions of $\boldsymbol{q}$, deriving the probability density function (pdf) $p(\boldsymbol{g}(\boldsymbol{q})|\boldsymbol{\theta},\boldsymbol{x})$—and hence the likelihood function $p(\boldsymbol{g}(\tilde{\boldsymbol{q}})|\boldsymbol{\theta},\boldsymbol{x})$—in closed form is difficult or impossible. By "closed form" we mean, loosely speaking, a well-defined procedure (equation or algorithm) for calculating a mathematical quantity of interest to high precision, using generally accepted "existing" functions as building blocks, i.e., without having to design and implement new approximations (e.g., http://mathworld.wolfram.com/Closed-FormSolution.html). The lack of a (known) closed form expression for the likelihood function makes the latter impossible to evaluate directly, unless a numerical approximation is developed. In contrast, we note that *sampling* from the random function $\boldsymbol{g}(\boldsymbol{Q}(\boldsymbol{\theta},\boldsymbol{x}))$ is relatively straightforward (e.g., see steps 1–4 in section 2.6). This situation is well suited for sampling the posterior $p(\boldsymbol{\theta}|\boldsymbol{g}(\tilde{\boldsymbol{q}}),\boldsymbol{x})$ using an Approximate Bayesian Computation (ABC) approach; see Kavetski et al. (2018) for theoretical background. In this work, we use the ABC algorithm described in section 2.4.1, with the distance metric given in section 2.4.2.

2.4.1. ABC Algorithm

The archetypal form of ABC algorithms is given by the following formulation (Weiss & von Haeseler, 1998):

1. Draw a sample $\boldsymbol{\theta}^{(i)}$ from the prior $p(\boldsymbol{\theta})$;
2. Draw a sample from the probability model $\boldsymbol{Y}(\boldsymbol{\theta},\boldsymbol{x})$ given the parameters $\boldsymbol{\theta}^{(i)}$ from Step 1, i.e., $\boldsymbol{y}^{(i)} \leftarrow \boldsymbol{Y}(\boldsymbol{\theta}^{(i)},\boldsymbol{x})$. In our case, samples from $\boldsymbol{g}(\boldsymbol{Q}(\boldsymbol{\theta},\boldsymbol{x}))$ are generated as described in section 2.6;
3. Accept $\boldsymbol{\theta}^{(i)}$ if $\rho(\tilde{\boldsymbol{y}}, \boldsymbol{y}^{(i)}) \leq \tau_\rho$ where ρ is a distance metric such as the RMSE or similar, and τ_ρ is a small tolerance; and
4. Repeat Steps 1–3 for $i=1,2,\ldots,N_{sam}$, where N_{sam} is the number of samples required.

The basic ABC algorithm is generally too computationally inefficient, especially when $\boldsymbol{y}$ and $\boldsymbol{\theta}$ are high dimensional. This study employs the SABC algorithm (Albert et al., 2014), which combines two algorithmic enhancements: (i) sampling $\boldsymbol{\theta}$ values in Step 1 from a Markov chain rather than from the prior (Marjoram et al., 2003); and (ii) tightening the ABC tolerance τ_ρ as the sampling progresses (Toni et al., 2009), similar to Simulated Annealing in optimization (Press, 1992). SABC evolves a population of particles $(\boldsymbol{\theta}^{(i)};\boldsymbol{y}^{(i)})$ defined in the joint parameter-output space, according to the sequence of distributions

$$p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})\times\exp\left(-\rho(\tilde{\boldsymbol{y}},\boldsymbol{y})/\tau_{\rho,k}\right) \tag{16}$$

for a tightening sequence of tolerances $\tau_{\rho,k+1} < \tau_{\rho,k}$. The particles move according to a Markovian jump distribution, with the jump variance and Metropolis acceptance test controlled by properties of the particle distribution and the current tolerance. The tolerance is tightened according to an adaptive "annealing" schedule based on the mean distance of the particles from the data $\tilde{\boldsymbol{y}}$, using the thermodynamic principle of entropy production to minimize the number of "wasted" iterations (Albert et al., 2014).

In our case studies, the SABC algorithm is configured to return 5,000 samples from a total of 2×10^6 iterations. The relatively high number of SABC iterations is employed to enable convergence to sufficiently small values of the tolerance τ_ρ, so that the choice of distance metric and tolerance has little to no effect on the ABC posterior samples (see in Kavetski et al., 2018, sections 2.4 and 5.2.2 for a discussion).

2.4.2. ABC Distance Metric

We define the ABC distance metric ρ to allow for the use of sets of observed and simulated signatures, $\tilde{\boldsymbol{g}}=\{\tilde{\boldsymbol{g}}_1,\tilde{\boldsymbol{g}}_2,..,\tilde{\boldsymbol{g}}_{N_g}\}$ and $\boldsymbol{g}=\{\boldsymbol{g}_1,\boldsymbol{g}_2,..,\boldsymbol{g}_{N_g}\}$ respectively, where an individual signature can potentially be vector-valued itself (e.g., as the FDC signature $\boldsymbol{g}_{FDC}$ in equation (11)). In this work, the largest distance over the individual signature metrics in the set is used,

$$\rho(\tilde{\boldsymbol{g}},\boldsymbol{g})=\max\left[\xi(\tilde{\boldsymbol{g}}_1,\boldsymbol{g}_1),\xi(\tilde{\boldsymbol{g}}_2,\boldsymbol{g}_2),\ldots,\xi(\tilde{\boldsymbol{g}}_{N_g},\boldsymbol{g}_{N_g})\right] \tag{17}$$

where ξ denotes an auxiliary distance function defined for an individual (potentially vector-valued) signature $\boldsymbol{g}_k=\{\boldsymbol{g}_{k,1},\boldsymbol{g}_{k,2},..,\boldsymbol{g}_{k,n_k}\}$. The auxiliary distance function is specified as the average relative difference over the elements of the signature,

$$\xi(\tilde{\boldsymbol{g}}_k,\boldsymbol{g}_k)=\frac{1}{n_k}\sum_{i=1}^{n_k}\left|\frac{\tilde{g}_{k,i}-g_{k,i}}{\tilde{g}_{k,i}}\right| \tag{18}$$

The definition of ρ in equations (17) and (18) is convenient because it gives equal weighting to (scaled) distances of individual signatures, such that a high-dimensional individual signature (such as the FDC, represented by a vector of its quantiles) will be prevented from dominating the overall distance metric merely by the virtue of comprising more elements than a lower-dimensional signature (such as the base flow index, which is a scalar).

Our attention to the specification of the distance metric is motivated by practical convergence aspects, in particular: (i) early sampling stages of the SABC algorithm, when the tolerance is still very coarse; and (ii) the matching of low quantiles of the FDC signature $\boldsymbol{g}_{FDC}$, where even minor discrepancies in streamflow magnitude correspond to large relative errors and can considerably delay ABC convergence (for this reason equation (18) is based on the average, rather than on the largest, error in the FDC curve).

The objectives of this study are particularly demanding with respect to ABC convergence and numerical accuracy. A lax ABC tolerance introduces potentially large approximation errors that manifest in wider posteriors (e.g., see Kavetski et al., 2018, sections 2.4.1 and 5.2). Such numerical artifacts would obscure the comparisons of signature and time-domain inferences, in particular with respect to the information content of various signatures. To assess the convergence of the SABC algorithm, we carried out a sensitivity analysis by changing the distance metric in equation (17) to be defined by the arithmetic average (rather than the maximum) of the individual distances ξ. Near-identical inferences and predictions were obtained in all SABC experiments reported here and in the companion paper Kavetski et al. (2018), indicating that SABC approximation errors in these studies are likely to be minor to negligible.

2.5. Implementation of Time-Domain Inference

For our choice of probability model $\boldsymbol{Q}(\boldsymbol{\theta},\boldsymbol{x})$ in section 2.2, the likelihood function $p(\tilde{\boldsymbol{q}}|\boldsymbol{\theta},\boldsymbol{x})$ for time-domain inference is readily available in closed form and is computationally fast to evaluate once the deterministic hydrological model $\boldsymbol{h}(\boldsymbol{\theta}_h,\boldsymbol{x})$ has been computed (see section 2.5.1). Our implementation of time-domain inference is hence based on constructing the posterior $p(\boldsymbol{\theta}|\tilde{\boldsymbol{q}},\boldsymbol{x})$ in closed form, and exploring this posterior using a combination of optimization and MCMC sampling (section 2.5.2).

2.5.1. Likelihood Function

The likelihood function corresponding to the probability model in equations (5–9) is approximately

$$p(\tilde{\boldsymbol{q}}|\boldsymbol{\theta},\boldsymbol{x})=\prod_{t=2}^{N_t}\frac{f_{\mathcal{N}}\left(\tilde{\varepsilon}_t-\phi\tilde{\varepsilon}_{t-1};0,\sigma_W\right)}{1-F_{\mathcal{N}}\left(z[0;\lambda]-z[h_t(\boldsymbol{\theta}_h,\boldsymbol{x});\lambda]-\phi\tilde{\varepsilon}_{t-1};0,\sigma_W\right)}\times\frac{\partial z[\tilde{q}_t;\lambda]}{\partial q} \tag{19}$$

where $f_{\mathcal{N}}(v;\mu,\sigma)$ is the Gaussian pdf with mean μ and standard deviation σ and $F_{\mathcal{N}}(v;\mu,\sigma)$ is the corresponding Gaussian cumulative distribution function (cdf). The denominator in equation (19) reflects the truncated nature of the error model, because we have $\tilde{\boldsymbol{q}} \geq 0$ and $\boldsymbol{h} \geq 0$. For simplicity, we have omitted the probability density term corresponding to $t=1$ (McInerney et al., 2017; Priestley, 1981).

The residuals $\tilde{\varepsilon}$ in equation (19) are computed as follows

$$\tilde{\varepsilon}=z[\tilde{\boldsymbol{q}}(\boldsymbol{\theta},\boldsymbol{x});\lambda]-z[\boldsymbol{h}(\boldsymbol{\theta}_h,\boldsymbol{x});\lambda] \tag{20}$$

Note that equation (19), which gives the likelihood function $p(\tilde{\boldsymbol{q}}|\boldsymbol{\theta},\boldsymbol{x})$ as a closed form expression, is used only in the time-domain inference. The signature-domain inference implemented using ABC does *not* require access to the likelihood function $p(\boldsymbol{g}(\tilde{\boldsymbol{q}})|\boldsymbol{\theta},\boldsymbol{x})$ and does *not* make use of equation (19); instead the ABC approach employs an indirect "sampling-based" approximation of $p(\boldsymbol{g}(\tilde{\boldsymbol{q}})|\boldsymbol{\theta},\boldsymbol{x})$ (see section 2.4).

2.5.2. Parameter Optimization and Sampling Algorithms

In the time-domain inference, we estimate the most likely parameter values by maximizing the posterior distribution in equation (4) with the likelihood in equation (19), and explore the entire posterior using MCMC sampling. Optimization was carried out using a multistart quasi-Newton algorithm (e.g., Kavetski & Clark, 2010). Twenty independent quasi-Newton searches were initiated from random initial seeds in the feasible parameter space. The quasi-Newton algorithm used trust-region safeguards to stabilize convergence to the nearest optimum. The posterior distributions were explored using the MCMC sampling strategy described by Thyer et al. (2009) with a total of 50,000 model runs and five parallel chains. During the first 10,000 samples, the jump distribution was tuned one parameter at a time. During the next 10,000 samples, the jump distribution was tuned by scaling its entire covariance matrix. The jump distribution was then fixed and 30,000 samples collected. The first 25,000 samples were treated as a burn-in and discarded from the analysis (Gelman et al., 2004) and the final 5,000 samples were used to analyze and report the parameter distributions.

2.6. Generation of Predictive Distributions

In both the signature and time-domain inferences, predictive distributions of streamflow time series and signatures are generated by propagating posterior parameter uncertainty through the probabilistic model in equation (5). For a given parameter set $\boldsymbol{\theta}^{(i)}=(\boldsymbol{\theta}_h^{(i)},\boldsymbol{\theta}_{\mathcal{E}}^{(i)})$ and input data $\boldsymbol{x}$, samples from the probability model in equation (5) are generated as follows:

1. Run the deterministic model to compute $\boldsymbol{h}^{(i)}=\boldsymbol{h}(\boldsymbol{\theta}_h^{(i)},\boldsymbol{x})$;
2. Sample a random vector $\boldsymbol{\varepsilon}^{(i)} \leftarrow \mathcal{E}(\boldsymbol{\theta}_{\mathcal{E}}^{(i)})$ from the truncated Gaussian AR1 process in equations (8) and (9);
3. Compute $\boldsymbol{q}^{(i)}$ from $\boldsymbol{h}^{(i)}$ and $\boldsymbol{\varepsilon}^{(i)}$ using equation (6); and
4. Compute the signatures $\boldsymbol{y}^{(i)}=\boldsymbol{g}(\boldsymbol{q}^{(i)})$. This step is only needed if sampling/predicting signatures.

When generating predictions, the parameter set $\boldsymbol{\theta}^{(i)}$ is taken from the set of posterior samples generated earlier during the respective parameter calibrations (sections 2.4 and 2.5).

In addition, within the signature-domain inference, samples from the probability model are used when calibrating the model parameters using the ABC approach (Step 2 of the algorithm in section 2.4.1). In this case, the parameter set $\boldsymbol{\theta}^{(i)}$ can be either drawn from the prior (Step 1 in the basic ABC algorithm), or generated as part of internal approximations within the more complex SABC algorithm (Albert et al., 2014).

3. Case Study Description

3.1. Study Area and Data

The case study experiments are based on the Lacmalac catchment in south-east Australia (Australian Bureau of Meteorology, http://www.bom.gov.au/water/hrs, Gauge 410057), as it has a relatively long data record and its pronounced seasonal behavior spans a range of hydrological regimes. The catchment has an area of 673 km^2 and is fed by the Goobarragandra River. The catchment elevation drops from around 1,330 to 270 m over the river's course length of about 56 km (http://www.bonzle.com). Precipitation, potential evaporation, and streamflow time series are available from 1957 to 2006. Over this period, average annual precipitation is 1,085 mm, and average annual streamflow is 345 mm. The mean monthly temperatures range from a high of around 30°C in summer (January) to a low of around 3°C in winter (July) (http://weather.mla.

com.au/climate-history/nsw/goobarragandra). The majority of the catchment is expected to be snow-free the entire year, though some areas at the higher elevations may experience episodic light snowfall.

Figure 1 shows the monthly averages of precipitation, potential evaporation, and streamflow. In summer potential evaporation is about three times higher than precipitation, whereas in winter it is about three times lower. These variations translate into a high seasonality in streamflow; for example, the runoff coefficient varies from about 0.2 in summer to about 0.5 in winter. The catchment exhibits appreciable ephemerality, with streamflow below 1 mm/d comprising 65% of the daily record.

In addition to seasonality and ephemerality, the Lacmalac catchment exhibits appreciable interannual variability in its streamflow characteristics, including in the runoff coefficient, flashiness index, and so forth. Our selection of data periods for the real data study (section 3.4.2) excludes some years where particularly strong interannual variability could not be handled using the selected hydrological and residual error models, a modeling limitation that we consider beyond the scope of this work. This is further noted in section 5.4, where we discuss the limitations of the current case study and outline directions for future investigations.

3.2. Hydrological Models

Two hydrological models, representing the deterministic term ***h*** in equation (5), are used. The first model is a single reservoir model, used with monthly data in Experiment 1 (section 3.4.1). The second model is a smoothed version of the hydrological model HyMod, used with daily data in Experiment 2 (section 3.4.2). Both models are implemented within the SUPERFLEX framework (Fenicia et al., 2011; Kavetski & Fenicia, 2011) and solved using the implicit Euler time stepping scheme (Kavetski & Clark, 2010). All models are forced by precipitation $r(t)$ and potential evaporation $e_p(t)$ and return, as responses, deterministic estimates of streamflow $h(t)$ and actual evaporation $e_a(t)$.

Single nonlinear reservoir model, with storage $s(t)$ and mass balance equation

$$\frac{\mathrm{d}s}{\mathrm{d}t}=r-e_a(s,e_p)-\hbar(s) \tag{21}$$

The reservoir storage-discharge relation is nonlinear,

$$\hbar(s;k,\alpha)=ks^{\alpha} \tag{22}$$

The actual evaporation flux is given by

$$e_a(s,e_p;c_e,m)=c_e e_p(1-\exp(-s/m)) \tag{23}$$

where m is a smoothing parameter that controls the transition from $e_a=0$ when $s=0$ to $e_a\approx c_e e_p$ when $s\gg m$.

We apply the reservoir model with 2 calibration parameters: k and c_e. The values of α and m are fixed a priori to 2.0 [-] and 0.5 mm, respectively (see Table 1) and are not fitted. The calibrated response is the flow volume over the (monthly) time step, $h_t=\int_0^{\Delta t}\hbar(t+\varsigma)d\varsigma$, approximated by the implicit Euler scheme.

HyMod. HyMod is a conceptual rainfall-runoff model with five reservoirs described by states s_u, s_{1f}, s_{2f}, $s_{3f,}$ and s_s (see Vrugt et al., 2003, Figure 7), and five calibration parameters, namely s_{uMax}, β, α, k_f, and k_s (e.g., Boyle, 2001; Vrugt et al., 2003; Wagener et al., 2001).

An "unsaturated" reservoir, with storage $s_u(t)$, controls the amount of precipitation that is converted to flow, $q_u(t)$, according to $q_u/r=1-(1-\bar{s}_u)^{\beta}$, where $\bar{s}_u=s_u/s_{uMax}$. The evaporation also takes place from the unsaturated reservoir and is given by $e_a/e_p=\bar{s}_u(1+m)/(\bar{s}_u+m)$, where the smoothing parameter m controls the transition from $e_a=0$ when $\bar{s}_u=0$ to $e_a\approx e_p$ when $\bar{s}_u\gg m$. The value of m is fixed a priori to 10^{-2} (Table 2). Our implementation of

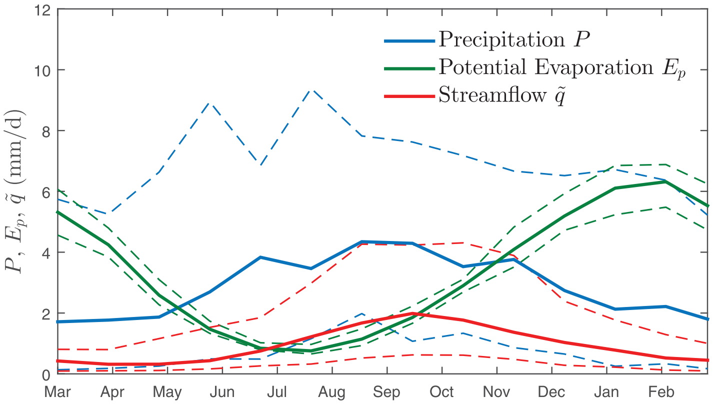


Figure 1. Monthly averages of precipitation, potential evaporation, and streamflow in the Lacmalac catchment (Australia), calculated over the period 1957–2006. Solid lines indicate average values; dashed lines indicate 90% probability limits. A strong seasonality can be noted, with winter characterized by higher precipitation, lower potential evaporation, and higher streamflow than summer.

Table 2
Parameters of the HyMod Model and the Error Model in Experiment 2

Hydrological model HyMod (daily time step)						Residual error model		
S_{uMax} (mm)	β	α	k_f (1/t_d)	k_s (1/t_d)	m (mm)	σ_W (mm$^\lambda$/t_d^λ)	λ	ϕ
1.0–10^3	0.1–10	0.0–1.0	0.1–2.0	10^{-3}–10^{-1}	10^{-2}	10^{-6}–2.0	0.2	0.8

Note. Calibration bounds and fixed parameter values are listed. The symbol t_d in the units refers to the daily time step.

HyMod employs a smooth evaporation relationship in order to improve its numerical behavior (Kavetski & Kuczera, 2007).

The flow from the unsaturated zone $q_u(t)$ is partitioned into two pathways intended to represent "quick" and "slow" flows, respectively. The "quick" pathway receives a flow fraction αq_u and routs it through a series of three linear reservoirs with the storage-discharge relation $q_{if}=k_f s_{if}$ for $i=1, 2, 3$, where s_{if} and q_{if} are, respectively, the storage and outflow from the ith reservoir. The "slow" pathway receives the remaining flow fraction, $(1-\alpha)q_u$, and routs it through a single linear reservoir with the storage-discharge relation $q_s=k_s s_s$, where s_s is the storage and q_s is the outflow.

The total streamflow $\hbar(t)$ is given by the sum of "quick" and "slow" flows, $\hbar(t)=q_{3f}(t)+q_s(t)$. The calibrated response h_t is the flow volume over the (daily) time step, given by the implicit Euler approximation.

3.3. Comparison Criteria

The signature-domain and time-domain inferences are compared in terms of the following criteria:

1. Posterior parameter distribution. We use histograms to compare the marginal posteriors from the signature-domain and time-domain inferences. In addition, in Experiment 1.2, we use the standard deviation of the parameter samples to quantify posterior uncertainty.
2. Precision of streamflow predictions. We compare the width of streamflow predictions obtained using signatures to the width of streamflow predictions obtained using the time-domain inference.
3. Statistical reliability of streamflow predictions. We compare the extent to which the predictive distributions capture the actual uncertainty in the model predictions. Statistical reliability is assessed using PQQ plots (Thyer et al., 2009, Figure 7). A PQQ plot checks the assumption that observed data $\tilde{\mathbf{y}}$ is consistent with being a sample from a given predictive distribution $p(\mathbf{y}|\cdot)$. The plot is constructed as follows. For each time step t, the nonexceedance probability of the observed data, $p_t=p(y_t < \tilde{y}_t|\cdot)$, within the predictive distribution $p(y_t|\cdot)$ is calculated. Under the assumption that $\tilde{\mathbf{y}}$ is a realization from $p(\mathbf{y}|\cdot)$, the distribution of p_t should be approximately uniform, and their empirical cdf approximately linear. The PQQ plot displays the cdf of p_t; departures from linearity suggest that the corresponding predictive distribution is statistically unreliable in the sense used in the forecasting literature (Gneiting et al., 2007).

In addition to inspecting predictive distributions of streamflow, Experiment 2 also consider predictive distributions of the FDC, calculated as the collection of FDCs of individual streamflow time series sampled from the predictive distribution. The predictive FDC distribution provides an additional posterior diagnostic to assess the quality of streamflow predictions.

In all case study experiments, predictive distributions and performance criteria are shown for the validation period only.

3.4. Case Study Experiments

3.4.1. Experiment 1 (Synthetic Monthly Data, Single Reservoir Model)

The objective of Experiment 1 is to investigate the behavior of the signature-domain inference under controlled synthetic conditions. Four experiments, labeled Experiments 1.1–1.4, are carried out.

Experiment 1.1: Effect of number and type of signatures. We compare the posterior parameter distributions and predictive streamflow distributions obtained using the following sets of signatures:

$$\mathbf{g}^{(1)}=\{g_{q50}\} \tag{24}$$

$$\boldsymbol{g}^{(2)}=\{g_{q10}, g_{q50}, g_{q75}, g_{q95}\} \tag{25}$$

$$\boldsymbol{g}^{(3)}=\{g_{q10}, g_{q50}, g_{q75}, g_{q95}, g_{f}\} \tag{26}$$

$$\boldsymbol{g}^{(4)}=\{g_{q10}, g_{q50}, g_{q75}, g_{q95}, g_{b}\} \tag{27}$$

A particular focus is on the ability to estimate parameters of the error model, as this is necessary to reliably capture predictive uncertainty.

Although informative of the hydrological regime, the sets of signatures in equations (24–27) are unlikely to represent the complete set of sufficient signatures for the (monthly) nonlinear reservoir model. Hence we expect a certain loss of information when these signatures are used in lieu of the streamflow time series (e.g., see Kavetski et al., 2018, sections 2.2 and 5.3.1).

Experiment 1.2: Effect of data length. We compare the posterior parameter distributions inferred from 25, 50, 100, and 500 years of data. All analyses in this experiment use the signature set $\boldsymbol{g}^{(3)}$. Provided the data is informative with respect to the inferred model and the posterior distribution has a finite variance, we expect the posterior standard deviation of any parameter θ in our inference setups to obey the (asymptotic) square-root scaling law with respect to the data length N_T (e.g., Box & Tiao, 1973; Gelman et al., 2004),

$$\text{sdev}[\theta] \sim \sqrt{N_T} \tag{28}$$

This applicability of this relationship to signature-domain inference is investigated by computing the standard deviation of the posterior parameter samples and plotting it against N_T on a log-log scale.

Experiment 1.3: Effect of deficiencies in the probability model. To investigate this aspect, we additionally corrupt the synthetic observed data with timing errors. Timing errors are likely to be common in hydrological time series (especially rainfall) but are seldom if ever represented in the error models (and hence in the likelihood functions). In this study, a crude form of timing errors is introduced by shifting the streamflow time series by 2 time steps forward. The signature-domain inference is carried out using signature set $\boldsymbol{g}^{(3)}$ (FDC quantiles plus flashiness index). Note that the error model in equations (5–9) does not represent any timing errors, but the signatures comprising $\boldsymbol{g}^{(3)}$ are insensitive to timing shifts.

In this experiment, we undertake two tests: (i) compare the posterior parameter distributions from signature-domain versus time-domain inference, and (ii) evaluate the predictive streamflow distributions from signature-domain versus time-domain inference against both the shifted and unshifted synthetic streamflow time series.

Experiment 1.4: Convergence and computational cost comparison. We use box plots of posterior parameter samples to assess and compare the convergence of SABC sampling (in the signature-domain inference based on signature set $\boldsymbol{g}^{(3)}$) and traditional MCMC sampling (in the time-domain inferences) over the course of 2×10^6 iterations. The MCMC approximation of the posterior at a given number of iterations is represented by 5,000 samples extracted by thinning the preceding chain. The SABC approximation is taken as the 5,000 parameter sets comprising the SABC population after a given number of iterations. In addition, to provide further insights into ABC convergence, we report the distance metric values of the final SABC samples, and illustrate the lack of sensitivity of the SABC approximation to the choice of distance metric (section 2.4.2).

Experiments 1.1–1.4 use the single-reservoir model (section 3.2). The time series data are given by monthly averages of rainfall, PET, and streamflow, generated synthetically based on approximate calibration of the single-reservoir model to the observed time series. Using synthetic data allows the generation of sufficiently long-time series to investigate the dependence of the inference on the data length in Experiment 1.2. We use 2 years for warm up, 50 years for calibration, and 50 years for validation (except for Experiment 1.2, see above). The signatures are computed directly from the monthly streamflow data.

The procedure used to generate synthetic forcing time series (precipitation and evaporation) is described in Appendix A. The synthetic streamflow time series are generated using the single reservoir model and Box-Cox-transformed Gaussian AR1 noise, with the reference parameter values given in Table 1. In order to make the synthetic experiment more realistic, the reference parameter values are chosen close to the values obtained by calibrating the hydrological models to the observed (monthly) data.

In Experiments 1.1–1.4, the signature and time-domain inferences are applied to the same three calibration parameters: the single-reservoir hydrological model parameters c_e and $\log k$, which control the water balance and the shape of the hydrograph, respectively, and the error model parameter σ_W, which reflects the magnitude of the error. The remaining parameters are kept fixed at their reference values (Table 1). By using the same model setup and inferring the same parameters (and by monitoring the numerical convergence of SABC and MCMC), we can attribute changes in the estimated parameters and predictions solely to the switch from calibration to streamflow time series to calibration to streamflow signatures. Given our usage of uniform (noninformative) priors, we can also interpret differences in the width of parameter posteriors as evidence of differences in the information content of the calibration data (time series versus various signature sets).

3.4.2. Experiment 2 (Real Daily Data, HyMod)

The objectives of real-data Experiment 2 are as follows: (i) corroborate the findings of synthetic Experiment 1 using real data, including a comparison of signature and time-domain inferences in terms of streamflow uncertainty quantification, and (ii) investigate the information content of FDCs and the base flow/flashiness indices. These objectives are investigated using daily scale data and the HyMod model (section 3.2).

The following sets of signatures are used:

$$\boldsymbol{g}^{(5)}=\{\boldsymbol{g}_{FDC}\} \tag{29}$$

$$\boldsymbol{g}^{(6)}=\{\boldsymbol{g}_{FDC}, g_f\} \tag{30}$$

$$\boldsymbol{g}^{(7)}=\{\boldsymbol{g}_{FDC}, g_b\} \tag{31}$$

The signatures in equations (29)–(31) are unlikely to represent the complete set of sufficient signatures for the (daily) hydrological model HyMod; a certain loss of information when these signatures are used in lieu of the streamflow time series is hence expected a priori.

In the hydrological model HyMod, we calibrate all parameters except the evaporation smoothing parameter m(see section 3.2.1). In the residual error model, only σ_W is calibrated. See Table 2 for details. The signatures are computed from the daily streamflow data.

Experiment 2 uses 1 year for warm up, 4 years for calibration and 4 years for validation. More specifically, the period 1 March 1995 to 1 March 1996 is used for warm up, the period 1 March 1996 – 1 March 2000 is used for calibration, and the period 1 March 2002 to 1 March 2006 is used for validation. This selection of calibration and validation periods provides a challenging case study due to the seasonality of the Lacmalac catchment, but excludes years where the interannual variability is too strong to be captured using the simple conceptual hydrological model HyMod and the Box-Cox Gaussian AR1 residual error model (see sections 3.1 and 5.4 for further discussion of this limitation).

4. Results

4.1. Experiment 1 (Synthetic Monthly Data, Single Reservoir Model)

4.1.1 Experiment 1.1: Effect of Number and Type of Signatures

Figure 2 shows the parameter distributions inferred using the signatures sets $\boldsymbol{g}^{(1)}$, $\boldsymbol{g}^{(2)}$, $\boldsymbol{g}^{(3)}$, and $\boldsymbol{g}^{(4)}$, and compares them to parameter distributions inferred from the streamflow time series. The reference parameter values used to generate the synthetic data are also shown. The following results can be noted:

1. When calibrating to g_{q50} alone (signature set $\boldsymbol{g}^{(1)}$, row 1), the signature-domain inference is generally poor: the posterior parameter distributions capture the "known" reference parameter values but are exceedingly wide. In contrast, the time-domain inference produces sharp posteriors centered near the reference parameter values;
2. Once four additional FDC quantiles are added to the signature set (rows 2 onward), the parameter posteriors inferred from the signatures achieve much closer visual agreement with the reference parameter values, in the sense that the posterior distributions are approximately centered on the reference parameter values. For example, in row 2, the posterior of reservoir parameter c_e inferred using the signatures sharpens to a range of about 0.9 to 1.2, and captures well the reference value of 1.0. The posterior of the error parameter σ_W also captures the reference value of 2.0;

3. The posterior distributions from the signatures-domain inference are always wider than the distributions from time-domain inference. In particular, when using a single signature g_{q50} (row 1), the posteriors span almost the entire prior range, and are 30–100 times wider than the posteriors from the time-domain inference. Even when using signature set $\boldsymbol{g}^{(4)}$ (row 4), the signature-domain posteriors are 3–5 times wider than time-domain posteriors;
4. As more signatures are added, the signature-domain posteriors initially become much narrower (e.g., row 2 versus row 1), and then stay approximately constant (e.g., row 3 versus row 2 and row 3 versus row 4); and
5. As more signatures are added, the signature-domain posteriors become closer to the time-domain posteriors. Once again, the strongest change occurs when switching from using a single FDC quantile ($\boldsymbol{g}^{(1)}$) to using four FDC quantiles ($\boldsymbol{g}^{(2)}$); the effect of additional signatures is less pronounced.

Figure 3 compares the predictive distributions obtained using the signature-domain versus time-domain inferences. The following results can be noted:

1. The predictive distributions obtained from the signature-domain inference become increasingly precise, approaching those obtained from the time-domain inference. The streamflow predictions obtained from the signature-domain calibration using signature sets $\boldsymbol{g}^{(1)}$ and $\boldsymbol{g}^{(2)}$ are visibly less precise than the predictions obtained using the time-domain calibration. When signature sets $\boldsymbol{g}^{(3)}$ and $\boldsymbol{g}^{(4)}$ are used, the distributions are virtually indistinguishable; and

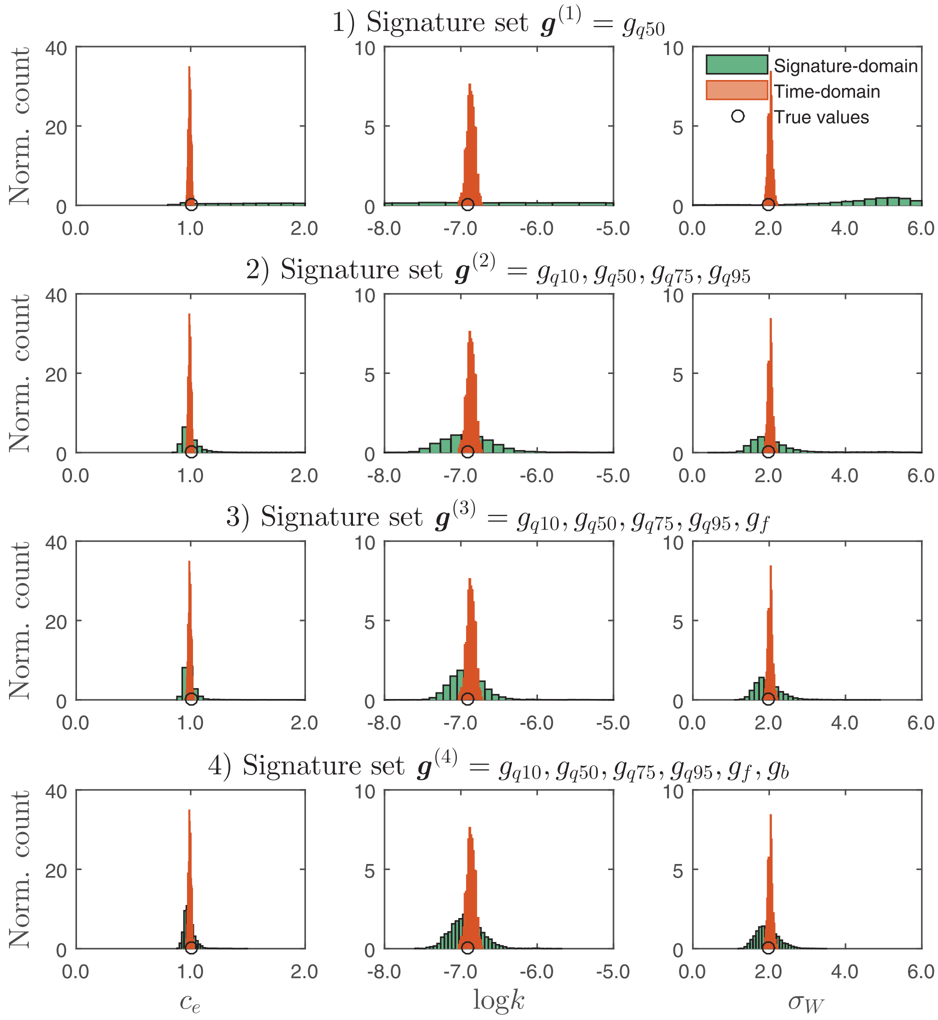


Figure 2. (Experiment 1.1). Effect of additional signatures on the posteriors of hydrological and error model parameters. As more signatures are included in the inference, the posteriors become narrower and closer to the posteriors inferred from time-domain calibration.

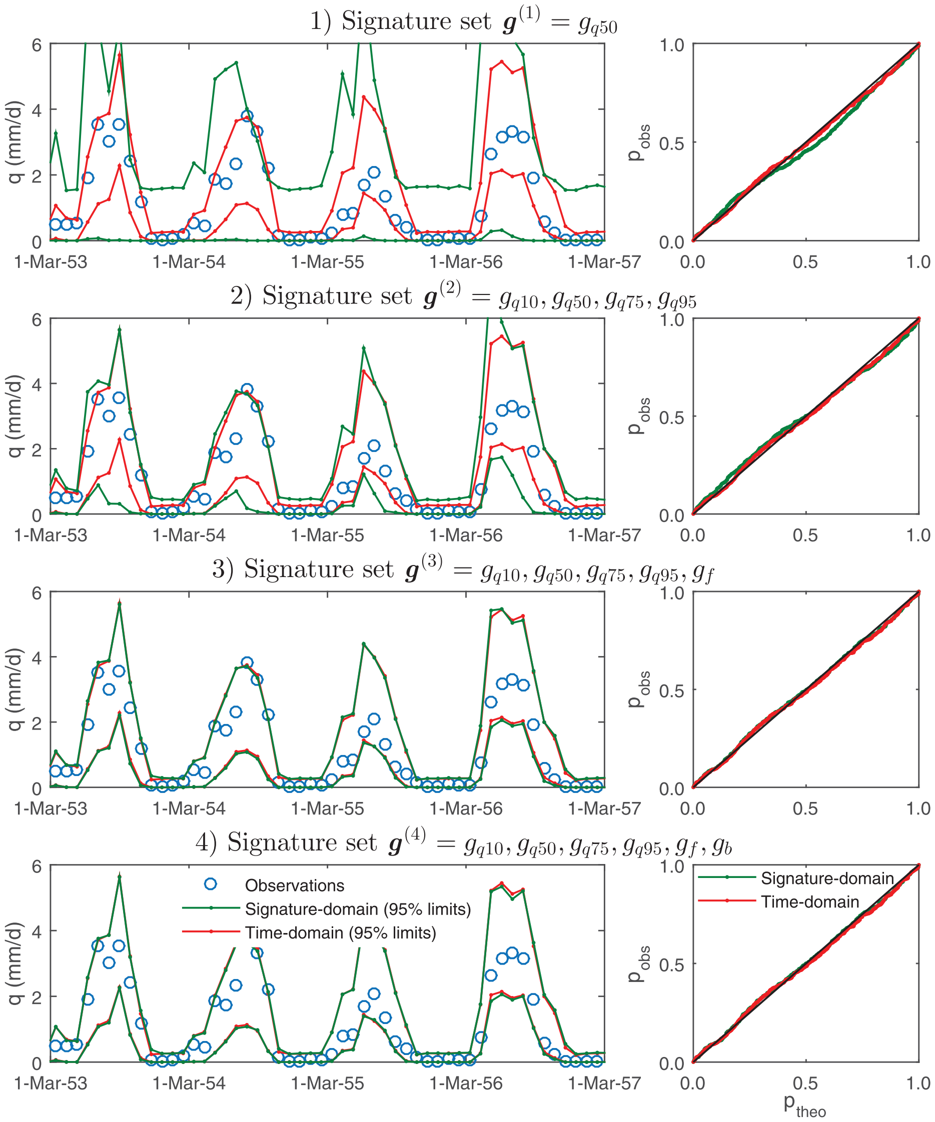


Figure 3. (Experiment 1.1). Streamflow predictions corresponding to the parameter posteriors from Figure 2. The *x* axis refers to dates within the synthetically generated data set. (left column) 95% prediction limits; (right column) the corresponding PQQ plots. As more signatures are included, the predictive distributions obtained using signatures become narrower and closer to those obtained from the streamflow time series. The PQQ plots are always close to a straight line, i.e., despite differences in precision, all predictive distributions shown here are statistically reliable.

2. The PQQ plots for all predictive distributions are close to the diagonal line, indicating that despite having different precisions the streamflow predictions from the signature-domain and time-domain inferences are both statistically reliable.

Figures 2 and 3 show that the use of signature set $\boldsymbol{g}^{(4)}$ did not provide noticeable improvement over using signature set $\boldsymbol{g}^{(3)}$. For this reason, experiments 1.2–1.4 use the signature set $\boldsymbol{g}^{(3)}$.

4.1.2. Experiment 1.2: Effect of Data Length

Figure 4 shows the effect of data length on the posterior distributions inferred using signature set $\boldsymbol{g}^{(3)}$ and the posterior distributions inferred using the streamflow time series.

The left plots of Figure 4 show box plots of the posterior distributions. It can be seen that:

1. Both posteriors remain generally close to the true parameter values used to generate the synthetic data;
2. Both posterior distributions become narrower as the data length increases; and
3. The signature-domain posteriors are about 3–5 times wider than time-domain posteriors.

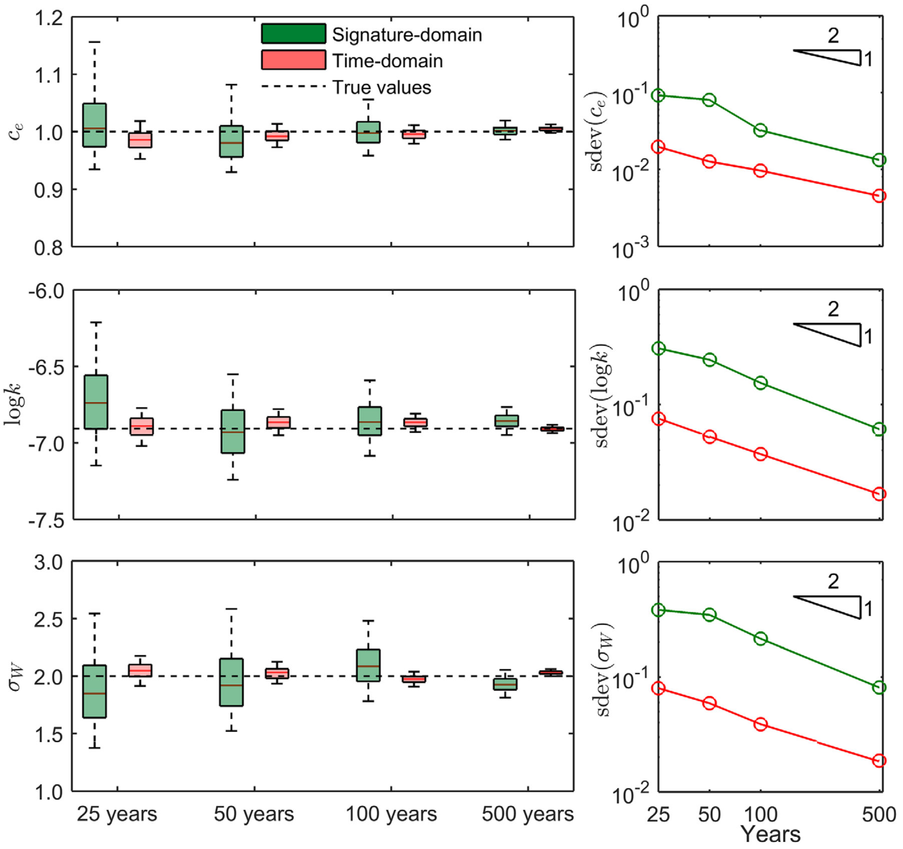


Figure 4. (Experiment 1.2). Effect of calibration data length on posterior parameter distributions. In both the signature-domain and time-domain inferences, posterior uncertainty decreases as more data are added, with the distributions converging to the "true" parameter values used to generate the synthetic data. The red line indicates the median, the boxes indicate the 25th and 75th percentiles, and the whiskers indicate the 5th and 95th percentiles.

The right plots of Figure 4 show the posterior standard deviation of the parameters plotted against the data length. It can be seen that:

1. For both signature-domain and time-domain inferences, the relationship between the posterior standard deviation of the parameters and data length follows a straight line on log-log scale, with a slope of $^1/_2$, as expected from theory (section 3.4.1). The relationship is particularly evident for parameters $\log k$ and σ_W. For parameter c_e, the relationship deviates from linearity for the lower values of N_T, but asymptotes toward the linear decreasing trend as N_T increases;
2. As follows from point 1 above, the ratio of standard deviations of posteriors from the signature-domain versus time-domain inferences remains (approximately) constant as the data length increases.

Overall, these results indicate that parameter distributions are sensitive to data length, both in the case of signature-domain and time-domain inferences.

Note that the posterior parameter distributions become narrower, while still encompassing the "true" parameter set. However, the posterior medians do not coincide with the true values. Such variability is to be expected when calibrating to a single realization of observed data, and is exactly why uncertainty quantification is important when undertaking parameter estimation.

4.1.3. Experiment 1.3: Effect of Deficiencies in the Probability Model

Figure 5 compares the parameter posteriors obtained in the case where the data are consistent with the probability model (data do not contain timing error, left-hand plots) versus the case where the data contains an error unaccounted for by the probability model (data contains timing error, right-hand plots).

1. The time-domain parameter posteriors differ considerably in the two cases, with effectively no overlap even allowing for posterior uncertainty. The posteriors of parameters $\log k$, c_e and σ_W estimated from

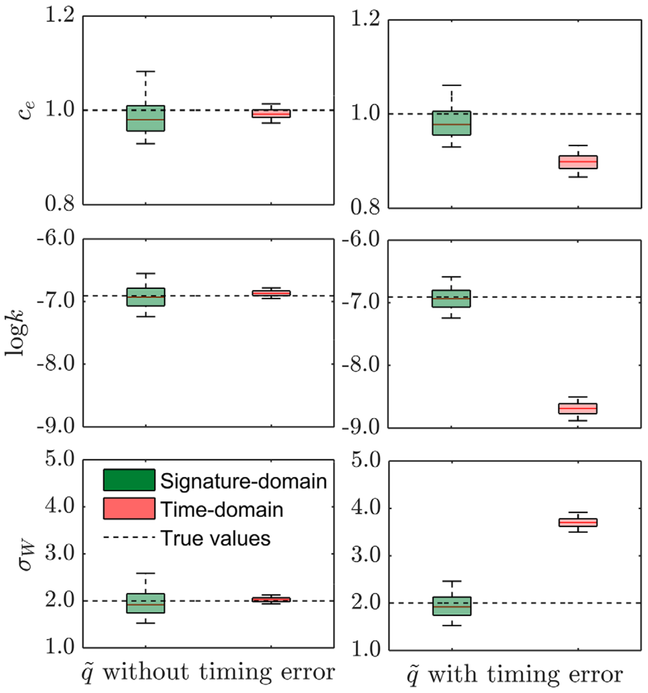


Figure 5. (Experiment 1.3). (right column) Posterior distributions obtained in the presence of timing errors unaccounted for in the likelihood model, (left column) compared to distributions inferred from data without timing errors. Signature-domain inference using signatures that are insensitive to timing errors can produce more robust parameter estimates than the time-domain inference. The red line indicates the median, the boxes indicate the 25th and 75th percentiles, and the whiskers indicate the 5th and 95th percentiles.

time-shifted data do not include the reference parameter values. The shift is particularly pronounced for parameter *k*, the value of which drops by a factor of about 60, corresponding to flatter recessions. The estimate of the error model parameter σ_W increases by a factor of 2, in recognition of the larger model misfit when fitting the time-shifted data; and

2. In contrast, parameter posteriors inferred from the data signatures are identical, irrespective of whether fitting to the original or time-shifted data sets, and are centered on the reference parameter values.

Figure 6a shows the predictive streamflow distributions corresponding to the parameter posteriors estimated from the time-shifted data. It can be seen that:

1. When calibrating to time-shifted data, the streamflow predictions obtained by signature-domain calibration are much narrower than those obtained from time-domain calibration.

The increased predictive streamflow uncertainty for the time-domain inference arises due to a flatter simulated hydrograph, as given by the lower value of parameter *k* in Figure 5, and is reflected in higher values of the error parameter σ_W, again as seen in Figure 5. Although not shown, signature-domain calibration generates the same streamflow predictions when calibrating to time-shifted data or to the original data without timing errors (Figure 3, row 3), because the inferred parameter values are identical in both cases. These streamflow time series are also identical to the streamflow predictions obtained by time-domain calibration to the original data (Figure 3, row 3).

Figures 6b and 6c show an assessment of the reliability of the streamflow distributions in predicting the original data (no timing errors) and the time-shifted data.

1. Plot b shows a PQQ plot analysis of the streamflow predictions in plot a against the time-shifted "observed" data used in the calibration. The PQQ plot of predictions generated using the signature-domain inference is S-shaped, corresponding to a general underestimation of predictive uncertainty. In contrast, the PQQ plot of predictions generated using time-domain inference is much closer to a diagonal line, with a clearly milder S-shaped discrepancy;
2. Plot c shows a PQQ plot analysis of the predictive distributions in plot a, but this time against the underlying data without timing error. In this case, the predictions obtained using signature-domain calibration are characterized by a near-diagonal PQQ plot, with only minor discrepancies. In contrast, the predictive distributions obtained using time-domain inference produce a mildly S-shaped PQQ plot, once again corresponding to underestimated uncertainty.

The findings in Figures 6b and 6c can be corroborated by inspection of Figure 6a. In particular, many streamflow observations fall outside the prediction limits estimated from the signature-domain inference; the PQQ plot is hence strongly S-shaped. In contrast, the prediction limits estimated from time-domain inference are wider and hence capture more observations, which results in a clearly more diagonal PQQ plot.

In summary: if the modeler is interested in reproducing the time-shifted data (i.e., the calibration data), signature-domain inference is *less* reliable than time-domain inference. Conversely, if the interest is in reproducing the underlying original data (without timing errors), signature-domain inference is *more* reliable than time-domain inference. This distinction will be discussed in section 5.1.4.

4.1.4. Experiment 1.4: Computational Convergence and Cost Comparison

Figure 7 shows how the model parameter distributions obtained with signature-domain and time-domain inferences change as a function of the number of iterations.

The posteriors from signature-domain inference implemented using SABC (left column) show a converging pattern and stabilize after approximately 10^6 iterations. The posteriors from time-domain inference

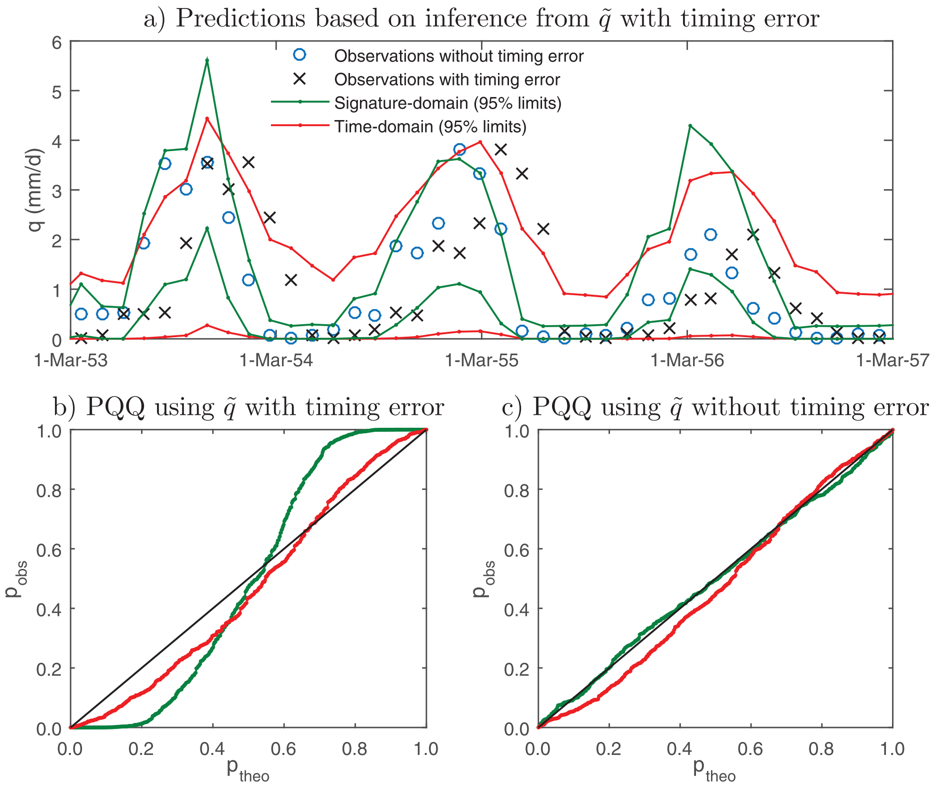


Figure 6. (Experiment 1.3). Streamflow predictions obtained using signatures-domain versus time-domain inferences in the presence of unaccounted timing errors in the streamflow data. The *x* axis in plot a refers to dates within the synthetically generated data set. Despite calibration to data with timing errors, signature-domain inference produces *unreliable* predictions of the data with timing errors, but *reliable* predictions of the data without timing errors. In contrast, time-domain inference with a deficient likelihood function partially compensates for unaccounted timing errors, and produces unreliable predictions of both data sets.

implemented using MCMC stabilize already after 25,000 iterations and fluctuate mildly around the stationary distribution. Therefore, in this experiment, signature-domain inference is around 40 times more expensive than time-domain inference. The computational efficiency of the ABC algorithm and its practicality in the context of hydrological model inference are discussed in section 5.3.

Figure 8 compares the SABC results obtained using the max-function distance metric to the SABC results obtained using an alternative distance function where the arithmetic average of the individual signature distances is used. The SABC approximations of the parameter posteriors obtained under these different numerical settings are virtually indistinguishable. This result illustrates a key property of ABC: provided the tolerance is driven to sufficiently small values, the choice of distance metric has little impact on the results (e.g., see Kavetski et al., 2018, section 2.4 for a theoretical exposition). Achieving this lack of sensitivity provides stronger empirical evidence of ABC convergence and general robustness of the results.

Note that the convergence of an ABC approximation to the posterior does not necessarily imply that the achieved ABC tolerance value is negligible in an absolute sense. For example, a separate analysis similar to the one reported in Figure 3 of Kavetski et al. (2018) revealed that, in Experiment 1.4, the smallest value of ρ across all the SABC samples corresponds to a 1.4% discrepancy in the signature values of observed and simulated stream flows, but the median is appreciably higher (7.5% discrepancy), and the largest discrepancy is as high as 50%. These discrepancies relate to the difficulty in matching the low quantiles of the FDC, where even minor changes in streamflow magnitude correspond to large relative errors. In contrast, in Experiment 1 (where the median streamflow was used as the sole signature), the smallest discrepancy (value of ρ) is as low as 10^{-8}, the median discrepancy is 0.1% and the largest discrepancy is 5%.

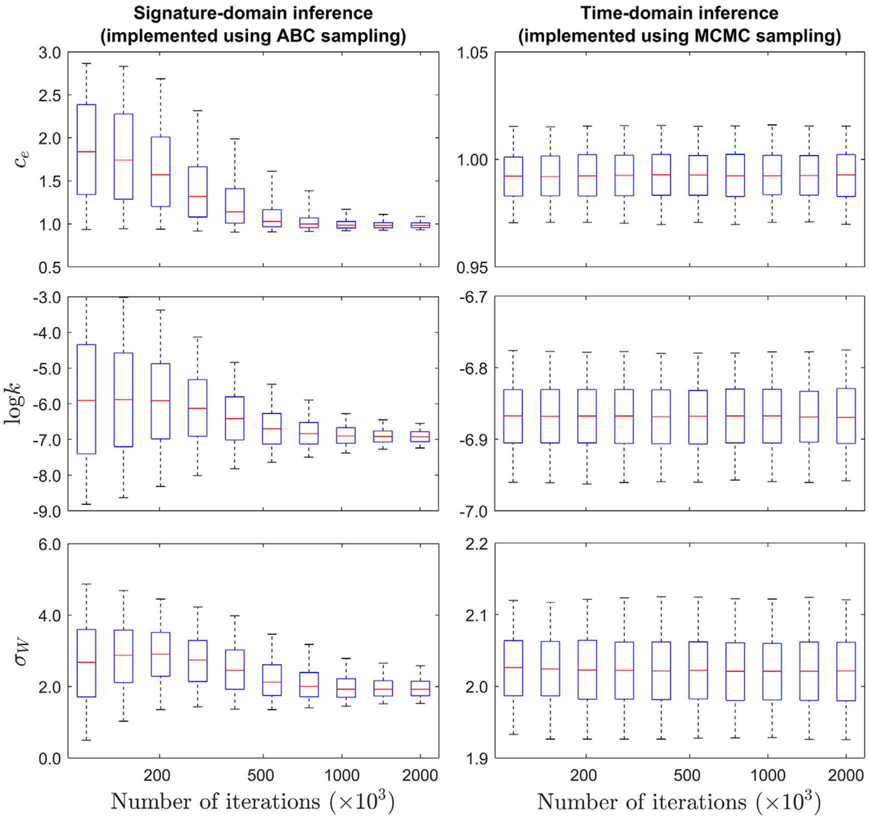


Figure 7. (Experiment 1.4). Evolution of parameter distributions throughout the sampling process, as a function of the number of iterations (hydrological model runs). (left column) The signature-domain inference (signature set $\boldsymbol{g}^{(3)}$) stabilizes after approximately 10^6 iterations, whereas (right column) the time-domain inference stabilizes already after around 25,000 iterations.

4.2. Experiment 2 (Real Daily Data, HyMod)

Figure 9 shows the predictive distributions obtained in Experiment 2, where HyMod is calibrated to daily data. In a given row of plots, the inference based on a specific set of signatures is compared to the time-domain inference (same in all rows).

Row 1 shows the inference based on FDC quantiles alone. The left plot shows the streamflow predictive distributions. The distributions obtained from signature-domain inference are much wider than those obtained from time-domain inference. For example, in the summer period, the lower limit of the predictive distributions obtained from signatures is always close to zero, while the upper limit is twice higher than the upper

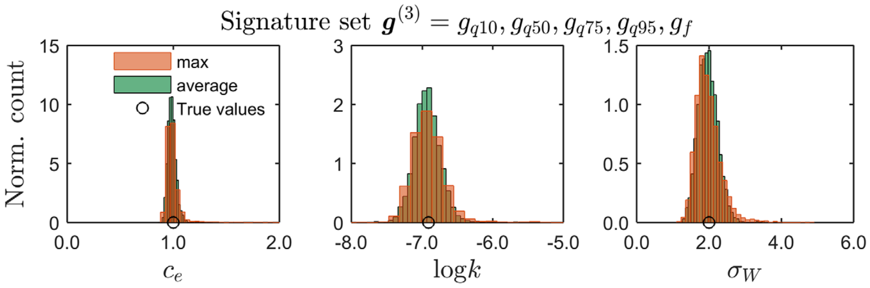
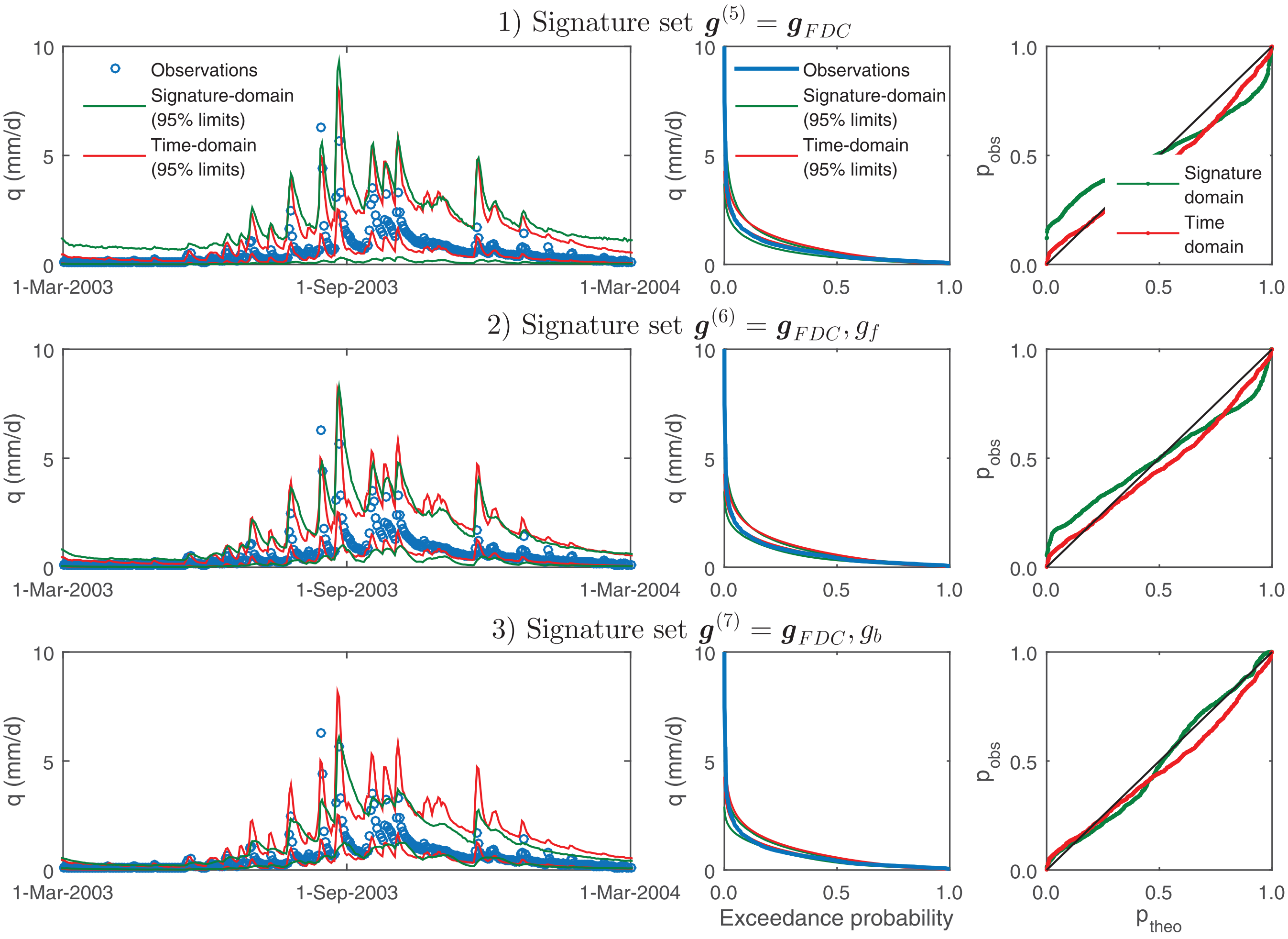


Figure 8. (Experiment 1.4). Comparison of parameter posteriors of signature-domain inference implemented using SABC for two choices of the ABC distance metric function; “average” refers to replacing the “max” function in equation (17) with the arithmetic average. Low sensitivity to the choice of distance metric is indicative that a sufficiently tight ABC tolerance has been achieved and that SABC sampling has converged.



1) Signature set $\boldsymbol{g}^{(5)} = \boldsymbol{g}_{FDC}$

2) Signature set $\boldsymbol{g}^{(6)} = \boldsymbol{g}_{FDC}, g_f$

3) Signature set $\boldsymbol{g}^{(7)} = \boldsymbol{g}_{FDC}, g_b$

Figure 9. (Experiment 2). HyMod predictions of daily streamflow in the Lacmalac catchment. Comparison of streamflow predictions obtained using signature-domain versus time-domain calibrations (the latter are identical in all rows). Similarly to the monthly simulations in Figure 3, daily simulations benefit from the inclusion of the flashiness index in the set of calibration signatures. However, the inclusion of the base flow index results in missed peaks.

limit of the distributions estimated from the streamflow time series. The middle plot shows the predictive distributions of the FDC. The FDC appears well represented, by both the signature-domain and time-domain inferences. The uncertainty in the FDC is much smaller than the uncertainty in the streamflow time series. The right plot compares the PQQ plots of the streamflow predictions. The PQQ plot corresponding to the signature-domain inference is less straight than the PQQ plot corresponding to the time-domain inference, indicating lower reliability.

Row 2 shows the inference based on FDC and flashiness index. The left plot shows that the streamflow distributions obtained using signature-domain calibration are similar to those obtained using time-domain calibration. The middle plot shows the predictive distributions of the FDC. Also in this case, the FDC appears well represented, by both the signature and time-domain inferences. The right plot shows the PQQ plots of the streamflow predictions. Compared to row 1, the PQQ plot obtained with signatures in row 2 is closer to the diagonal, although less than the PQQ obtained using the streamflow time series.

Row 3 shows the inference based on FDC and the base flow index. The left plot shows that the streamflow predictive distributions obtained using signature-domain calibration miss most observed peaks. In contrast, the middle plot shows that the observed FDC appears well captured by the model predictions. The PQQ plots in the right plot exhibit less deviation from the diagonal in the case of signature-domain calibration compared to time-domain calibration. Despite deficiencies in capturing peak flows, the predictive distribution obtained from signature-domain calibration appears quite reliable.

5. Discussion

5.1. Properties Investigated Using Synthetic Data (Experiment 1)

5.1.1. Effect of Number and Type of Signatures

In Experiment 1.1, based on synthetic data, the time-domain inference represents the ground truth, which the signature-domain inference seeks to approximate. This experiment offers the following points for discussion:

1. The streamflow-domain inference is generally consistent with the time-domain inference. As long as enough signatures are used (e.g., FDC quantiles plus the flashiness index), the signature-domain parameter posteriors have similar central values to the time-domain posteriors, though the former are somewhat wider (Figure 2, rows 3 and 4). Despite being wider, the signature-domain parameter distributions generate predictive streamflow distributions that are indistinguishable from their time-domain counterparts (Figure 3, rows 3 and 4);
2. The parameter posteriors and predictive streamflow distributions estimated using signature-domain calibration are always wider than those estimated using time-domain calibration (Figures 2 and 3). This result is in agreement with the theoretical expectation that conditioning an inference on general (non-sufficient) signatures in lieu of the full time series leads to a loss of information; and
3. Posterior and predictive distributions clearly depend on the set of signatures used (Figures 2 and 3). For example, using the FDC median as the sole signature produces extremely wide predictive distributions. This result highlights that signature-domain inference does require some minimal set of signatures before useful results can be obtained.

The general insights from this experiment can be summarized as follows:

1. Bayesian signature-domain inference can serve as a platform for assessing the information content of particular choices of signatures. Specifically, it can test the effectiveness of different combinations of signatures to serve as "summary statistics" of the time series for calibrating model parameters. This analysis can be carried out by comparing signature-domain inference to a time-domain inference (if available), or by comparing signature-domain inferences with different sets of signatures;
2. The use of signatures generally leads to a loss of information, unless the complete set of sufficient statistics for the calibrated model is used. As discussed in section 5.3.1 of Kavetski et al. (2018), it may be difficult or impossible to formally derive such a set of sufficient statistics for nontrivial probability models of observed streamflow time series, especially for catchments with strong seasonality and other nonstationary behavior (Sadegh et al., 2015; Westra et al., 2014). Sets of *approximately* sufficient statistics are hence of interest—we see from Figure 2 that although signature set $\boldsymbol{g}^{(1)}$ is far from sufficient, set $\boldsymbol{g}^{(3)}$ is arguably quite close to sufficiency for parameter c_e of the monthly model. In addition Figure 3 suggests that $\boldsymbol{g}^{(3)}$ is close to sufficiency for the practical purpose of using this relatively simple combination of a single-reservoir model and the Box-Cox Gaussian AR1 residual error model to make (monthly) probabilistic streamflow predictions;
3. The inference setups in the signature and in the time domain are mutually consistent, and give similar results when an appropriate selection of signatures is used. Under these circumstances, the set of signatures appears to capture enough information content of the original streamflow time series to produce a similar inference. However, important differences can arise when the probability model assumed to describe the observed data is deficient (see section 5.1.4 for details); and
4. Further work is needed to understand the extent to which volatile signatures such as low and/or high quantiles of the FDC are indicative of hydrological consistency, and to establish their impact on the computational efficiency of ABC.

The signature set found suitable in this experiment, $\boldsymbol{g}^{(3)}$, is clearly specific to the particular choice of data and model, as will be elaborated further in section 5.2.

5.1.2. Ability to Estimate Streamflow Error Parameters and Predictive Uncertainty From Signatures

An important distinction between our work and previous ABC applications in hydrology is that we explicitly infer the residual error variance, σ_W^2. In previous applications, the error model was either absent (e.g., Nott et al., 2012; Sadegh & Vrugt, 2013; Vrugt & Sadegh, 2013), or used with predefined variance values (Sadegh & Vrugt, 2013, Figure 15). Our results show that with an appropriate selection of signatures, e.g., FDC quantiles and flashiness index, the error variance can be estimated as part of calibration (see Experiment 1). In other words, it is possible to estimate a probabilistic model of streamflow even when calibrating to streamflow signatures. The design of signatures suitable for estimating error model parameters is an important direction for future research. For example, the structure of the flashiness index, defined by differences between consecutive streamflow points, is similar to the data error estimator used by Vrugt et al. (2005). The work of Westerberg and McMillan (2015) on the classification of hydrological signatures according to their ability to filter out the uncertainty in the original data is another useful step in this direction.

5.1.3. Effect of Data Length

Experiment 1.2 shows that, similar to time-domain posteriors, the signature-domain parameter posteriors become increasingly precise as the length of data used for calibration is increased, and converge asymptotically to the reference values used to generate the synthetic data.

Regardless of the type of inference—signature-domain or time-domain—the reduction in posterior uncertainty with respect to data length is seen to follow the well-known square-root scaling law (section 3.4.1). Interestingly, this convergence held even though the total number of "evaluation points" associated with the signatures remained constant: four FDC quantiles and the flashiness index, irrespective of the length of the calibration time series. It is interesting to consider the mathematics of how signatures transfer the information about data length into the posterior parameter distributions. We suggest that this behavior is due to the combination of the use of a probability model of the signatures that is derived by propagating the uncertainty of the time series and the use of signatures that integrate the properties of the time series (all signatures considered in this work have this property). Under these conditions, as $N_T \rightarrow \infty$, $p(\boldsymbol{g}(\boldsymbol{q})|\boldsymbol{\theta}, \boldsymbol{x})$ derived from $p(\boldsymbol{q}|\boldsymbol{\theta}, \boldsymbol{x})$ becomes tighter. In turn, the tightening of the likelihood function $p(\boldsymbol{g}(\tilde{\boldsymbol{q}})|\boldsymbol{\theta}, \boldsymbol{x})$ with respect to $\boldsymbol{g}(\tilde{\boldsymbol{q}})$ typically translates into a tightening of the posterior $p(\boldsymbol{\theta}|\boldsymbol{g}(\tilde{\boldsymbol{q}}), \boldsymbol{x})$ with respect to $\boldsymbol{\theta}$. As such, the information from the streamflow time series "flows" into the signatures and the posterior gets narrower. Another intuitive way to interpret this behavior is that, as the data length increases, signatures such as flow quantiles and flashiness/base flow indices tend to stabilize, and their sensitivity to particular realizations of the probability model decreases. In turn, this translates into tighter posteriors.

Experiment 1.2 (section 4.1.2) further corroborates the findings of Experiment 1.1 (section 4.1.1) regarding the loss of information when calibrating to signatures instead of time series. In Figure 4, the signature-domain parameter posteriors are consistently 3–5 times wider than the time-domain parameter posteriors. This constant ratio of posterior standard deviations, irrespective of calibration data length, may be indicative of the relative loss of information by the particular signatures.

5.1.4. Effect of Deficiencies in the Probability Model

The theoretical analysis in section 2 and the empirical investigation using synthetic data (sections 4.1.1–4.1.2) indicates the general correspondence of inferences based on signatures-domain versus time-domain calibrations. However, Experiment 1.3 (section 4.1.3) suggests that the choice of signatures and distance metric can be used to tamper with this correspondence. This finding is of interest when some of the assumptions underlying the probability model (and hence the likelihood function) are violated. For example, when calibrating to data corrupted by a substantial timing error, the inferences obtained using certain signatures become substantially different from inferences obtained using the time series, even if the same probability model is used in the calibration. These differences in posterior estimates can be attributed to the complete insensitivity of the selected signatures, namely the FDC quantiles and the base flow index, to timing shifts. In contrast, inferences based on the squares-type probability models are highly sensitive to such timing shifts.

The differences in parameter estimates obtained using particular signatures in the presence of unaccounted errors are of clear interest and represent one of the motivations for using signatures (see section 1). However, there are important subtleties. For example, the parameter posteriors obtained using the signature-domain inference appears preferable in this example where the correct result is known. However, reliability analysis of the predictive streamflow distributions against the (time-shifted) time series data—which is all the modeler will have available in practice—could favor the time-domain inference. This apparent "improvement" is due to the combination of the hydrological model parameters partially compensating for the time shift and the error model parameter *correctly* capturing the resulting larger predictive uncertainty.

Relevant questions then include the interpretation of the time shift, as well as the modeling objectives. In particular, we distinguish the following scenarios:

1. The time shift represents an error in the data. For example, it is common for meteorological services to report daily precipitation averages at 6:00 A.M., but report streamflow averages at midnight, thus creating a mismatch of 6 (or 18) h between the time series. If such shifts are present in the data, but not in the model, the (insensitive) behavior of the signature-domain inference could be considered superior to the (compensatory) behavior of the time-domain inference, because the genuine data the modeler is trying to predict is the underlying unshifted data. A more complex example of timing errors can arise due to the travel of a storm cell across a sparsely gauged large catchment; and
2. The inability to simulate the time shift represents a structural error in the hydrological model. For example, lumped rainfall-runoff models without appropriate routing components might incur major timing errors when applied to large catchments. In this case, the interest is clearly in predicting the "time-shifted" data. A hydrologist interested in improving the model representation will undoubtedly prefer to avoid parameter values compensating for structural deficiencies, yet unless the modeler is already aware of this particular defect in their model, there may be little apparent evidence of this compensation. An operational forecaster, interested first and foremost in setting up a given model to produce the most reliable predictions for their catchment of interest, may be justified in preferring the "compensatory" inference based on the streamflow time series, at least until the model deficiency can be identified and its correction shown to improve specific predictive criteria of interest.

Both types of scenarios—or combinations of scenarios—are likely common in hydrological applications, making it difficult to conclusively favor the results of either inference approach without further context or statement of modeling objectives. Timing errors in observed data could manifest more strongly in small catchments (especially if relying on rain gauges located outside the catchment), whereas routing errors in the hydrological model could be more pronounced when modeling large catchments. Therefore, although the use of particular signatures can in principle mitigate against deficiencies in the probability model, making practical use of such behavior is tricky, especially if the interest is primarily in the streamflow predictions rather than in parameter estimation per se. Achieving the former may require an interpretation of the nature of the modeling deficiencies, in which case it would be preferable to use this understanding to directly improve the probability model itself rather than to try to "trick" the inference into ignoring these deficiencies.

The analysis in Experiment 4 is intended to provide only a basic illustration of the effect of model deficiencies on time-domain and signature-domain inference. In practice, such deficiencies are likely to be much more complex than a simple uniform shift of the entire time series of streamflow, and their impact will depend substantially on the hydrological model and the time scale of application. We also emphasize that the ability to mitigate against potential deficiencies by "throwing away" information conflicting with the model specification is a property of working in the signature domain, rather than a property of the ABC algorithm. Deeper understanding of these effects clearly requires substantial further research.

5.2. Properties Investigated Using Real Data (Experiment 2)

5.2.1. Comparison With Synthetic Findings

Experiment 2, which uses real data, corroborates the earlier findings made in Experiment 1, which uses synthetic data, in the following respects:

1. When an appropriate signature set is used, such as a combination of FDC quantiles and the flashiness index, signature-domain inference is generally consistent with time-domain inference, in the sense that predictions are similar, but less precise. The loss of precision is indicative of the degree to which signatures sets $\boldsymbol{g}^{(5)}$, $\boldsymbol{g}^{(6)}$, and $\boldsymbol{g}^{(7)}$ are not sufficient for the HyMod model;
2. The flashiness index supports the estimation of the residual error variance, and hence the quantification of predictive uncertainty in the streamflow time series, even when streamflow time series are not used in the calibration; and
3. The addition of signatures to the calibration set, in particular the addition of the flashiness index to the FDC quantiles, results in a tightening of the predictive distribution. This can be attributed to the additional information in the flashiness index contributing to a reduction in posterior parameter uncertainty, especially in the parameters controlling hydrograph dynamics and streamflow uncertainty. While we would not consider $\boldsymbol{g}^{(6)}$ to be sufficient, it appears to be a useful signature set that can appreciably inform the inference of the HyMod model.

Real data experiments also point to some important differences from the synthetic experiments. In synthetic Experiment 1.1, signature-domain calibration generated predictive distributions that always contained the predictive distributions from time-domain calibration, and always contained predictive distributions estimated using fewer signatures. In real-data Experiment 2, this is not always the case. For example, addition of the base flow index to the FDC quantiles signature set results in predictive distributions that no longer capture the observed hydrograph peaks, in contrast to the predictive distributions from signature-domain inference to the FDCs alone and from the time-domain inference (Figure 8, row 3). A detailed interpretation of these differences is complicated by the fact that, in going from the synthetic to real experiments, we have changed both the hydrological model *and* the time scale of application. That said, it is likely that, in the real-data Experiment 2, deficiencies in the probability model result in tradeoffs in the ability of the model to fit particular signatures and, depending on the signature set selected, favor fitting some aspects of the hydrograph at the expense of others. These tradeoffs are reminiscent of tradeoffs in model performance that motivated multiobjective optimization and diagnostics (e.g., Gupta et al., 1998) in the reliability versus precision versus bias metrics achievable by residual error models depending on the value of the Box-Cox power parameter (McInerney et al., 2017). These tradeoffs are clearly not present in synthetic Experiment 1.1 where the "correct" probability model of the data is used in the calibration.

5.2.2. Information Content of FDC

The use of real data in Experiment 2 helps us understand the extent to which signatures derived from the FDC capture the parameter-related information in the original streamflow time series, and which additional signatures can complement such information. The FDC is arguably the most common summary of the hydrograph (see Castellarin et al., 2013 for a review). For example, in the studies of Yu and Yang (2000) and Westerberg et al. (2011), model calibration was based solely on FDC characteristics.

In contrast to Yu and Yang (2000) and Westerberg et al. (2011), who suggested that FDCs can be successfully used as an alternative to time-domain hydrological model calibration, we reach an opposite conclusion. Streamflow predictive distributions from the signature-domain inference are considerably wider than streamflow predictive distributions obtained from the time-domain inference (Figure 9, rows 1, left plots). In terms of statistical reliability, the streamflow predictive distributions obtained from the FDC signature are significantly worse than those obtained from the streamflow time series (Figure 9, rows 1, right plots).

It is interesting to note that the uncertainty in FDC space is about an order of magnitude smaller in than in the streamflow time series space (Figure 9, rows 1, left and central plots). The reduction of uncertainty occurs because very different hydrographs can map to very similar FDCs, and illustrates the "uncertainty filtering" effect of the FDC (Westerberg & McMillan, 2015). However, the downside of this filtering behavior is that streamflow predictions that capture the FDC relatively well can be quite poor, both in terms of statistical reliability and in terms of reproducing other characteristics of observed streamflow.

Experiment 2 suggests that model calibration using FDC characteristics alone may be insufficient to constrain parameter estimates and produce reliable streamflow predictions.

5.2.3. Added Value of Base Flow Index, Flashiness Index, or Other Signatures

Given the limitations of FDC-only signature calibration discussed in section 5.2.2, it is of interest to consider additional signatures to complement the information content of the FDC.

Experiment 2, where the HyMod model was calibrated to daily streamflow data, suggests that complementing the FDC quantiles with the flashiness index improves both the reliability and precision of the streamflow predictions, which become similar to the predictions obtained using the time-domain calibration (Figure 9, row 2, left and right plots). However, complementing the FDC quantiles with the base flow index did not produce comparable improvements: most peaks are still missed (Figure 9, rows 3, left plot) even if the streamflow predictions are overall quite reliable (Figure 9, row 3, right plot).

It can be seen that signature-domain inference benefits from supplementing the FDC quantiles with the flashiness index. Both reliability and precision are improved, with no tradeoffs between these performance criteria. The added value of the flashiness index could be explained by its ability to capture the timing aspects of the streamflow dynamics, which are omitted by the FDC. The addition of base flow and flashiness indices improves the streamflow predictions while maintaining the FDC fit, which is consistent with the earlier observation in section 6.2.2 that notably different hydrographs can have similar or near-identical FDCs.

5.3. Practicalities: Computational Cost of the Inference

This study highlights several appealing features of signature-domain inference. However, the computational cost of signature-domain inference can be substantially higher than the cost of time-domain inference. For example, in Experiment 1.4, ABC sampling from the signature-domain posterior required approximately 40 times more hydrological model runs than MCMC sampling from the time-domain inference.

The higher computational cost of ABC sampling should not be surprising given that $\boldsymbol{Y}(\boldsymbol{\theta},\boldsymbol{x})$ is a probabilistic model. Determining parameters $\boldsymbol{\theta}$ that "match" the observations $\tilde{\boldsymbol{y}}$, as required by ABC algorithms, depends not only on the parameters $\boldsymbol{\theta}$ themselves, but just as importantly on the random realization. To match the observations, ABC algorithms not only need to sample "good" parameter values, but also have the probability model generate a "lucky" random sample of outputs that matches the observations (see Kavetski et al., 2018, section 5.1.2). This is computationally expensive, especially when trying to achieve small relative errors in signatures such as the low quantiles of the FDC. In contrast, traditional MCMC sampling from the posterior (employed in this work to implement the time-domain inference) exploits the ability to evaluate the posterior density in computationally fast closed form, and does not require matching observed data.

We stress that the underlying cause of elevated computational costs is not the use of signatures per se, but the difficulty in deriving the probability density of the signature-domain predictions, $p(\boldsymbol{g}(\boldsymbol{q})|\boldsymbol{\theta},\boldsymbol{x})$ (Kavetski et al., 2018). If this pdf were obtainable in closed form, or if a suitably accurate approximation could be constructed from the samples of $\boldsymbol{Y}(\boldsymbol{\theta},\boldsymbol{x})=\boldsymbol{g}(\boldsymbol{Q}(\boldsymbol{\theta},\boldsymbol{x}))$ (e.g., Jennings et al., 2010; Lockart et al., 2015; Wood, 2010), the likelihood function $p(\boldsymbol{g}(\tilde{\boldsymbol{q}})|\boldsymbol{\theta},\boldsymbol{x})$ could be evaluated directly. Under these circumstances, provided the likelihood expressions were suitably fast to evaluate, ABC would no longer be necessary and signature-domain inference could be implemented using the much faster "traditional" MCMC techniques.

The SABC algorithm, motivated by principles of thermodynamics and simulated annealing (Albert et al., 2014), is one of many algorithms available for implementing the ABC approach. For example, in the hydrological literature, differential evolution principles and continuous acceptance kernels have been exploited to develop the DREAM-ABC algorithm (Sadegh & Vrugt, 2014). Previous studies have demonstrated the computational efficiency of DREAM-ABC when applied to inference setups with a *deterministic* model and a coarse ABC tolerance τ_ρ; testing of DREAM-ABC for ABC-based inference of a *probabilistic* model, especially as the tolerance τ_ρ is tightened, is hence of interest (see Kavetski et al., 2018, sections 3.2 and 5.1.3).

The computational cost of signature-domain inference using ABC will generally depend on the number or type of signatures used, on the model/data setup, and on the ABC numerical settings. Inference costs, irrespective of the setup (e.g., signature or time-domain), increase with model complexity, and we therefore have restricted this study to parsimonious models with few parameters. ABC costs may prove impractical for computationally expensive models, such as large-scale distributed hydrological models. This is a well-recognized and inherent limitation of applying demanding analysis techniques to expensive models (e.g., Hill et al., 2016). Parallel computation offers opportunities for mitigating these costs; e.g., the SABC algorithm can be organized to evolve individual particles in parallel, or, similar to standard MCMC, evolve multiple populations of particles (naïve parallelization). Relaxing the ABC tolerance is a riskier avenue for computational cost-cutting, as it can introduce large approximation errors; these errors would manifest in a loss of information (wider estimates of the posteriors) and introduce an undesirable dependence on the choice of distance metric function. Overall, ABC sampling is unlikely to be computationally competitive with traditional MCMC sampling when calibrating deterministic hydrological models with exogenous residual error models for which the likelihood function has a computationally fast closed form expression. Hence, computational cost considerations will generally favor the time-domain inference—unless specific circumstances, including those described in sections 1 and 5.1.4, require the use of signatures.

5.4. Limitations of the Study

This work focused on clarifying several important empirical properties of signature-domain inference using ABC, drawing insights from a set of synthetic and real data experiments. Given that we explored only a limited number of hydrological calibration scenarios, there are fertile grounds for further research.

1. We restricted the experiments to scenarios where data and model uncertainties are described using fairly simple Box-Cox Gaussian AR1 error models. One of the advantages of ABC lies in its ability to deal with much more complex stochastic models, where the stochastic terms are embedded within the model

structure, and the associated likelihood functions that are hence expensive or prohibitive to derive and evaluate even in the time domain (e.g., Albert et al., 2016; Lockart et al., 2015). Whether there is an advantage in using such complex probabilistic models in practical hydrological modeling, and whether ABC algorithms can provide a computationally feasible implementation, remains to be investigated;

2. We focused primarily on scenarios where the probability model (here, hydrological model plus residual error model) is plausible. In these cases, the time-domain inference is arguably the best achievable outcome of the inference process. In the case where the probability model is deficient, which is surely common in practice, it appears possible to at least partially mitigate the impact of these deficiencies by using signatures that are insensitive to these defects. However, further in-depth theoretical and empirical analysis is required to establish more conclusively the conditions and extent to which signature-domain inference can result in better predictions than the time-domain inference;
3. We considered only two hydrological models and a single catchment with daily and monthly data. Although we found the FDC and flashiness index to be a good combination in the real-data experiment, we expect the information content ("degree of sufficiency") of signatures to depend on the hydrological model (including the error model) and the time scale of application. It is clearly necessary to explore a much wider range and combinations of signatures, models, and data sets. Signatures of interest include other traditional hydrological signatures such as the runoff coefficient, as well as parametric descriptors such as the parameters of a fitted Flow Duration Curve (Sadegh et al., 2016). Empirical studies using catchments with diverse hydrological regimes, especially including arid and semi-arid conditions (e.g., Smith et al., 2010; Ye et al., 1997), and investigations of multiple time scales of application, e.g., from sub-hourly to monthly and yearly, are also of clear practical interest. Importantly, we recommend ensuring a controlled approach for the testing of signatures, e.g., to avoid the interpretative limitations encountered in our empirical study where we changed both the hydrological model and the time scale of application when going from the synthetic to the real case study;
4. We considered either purely time-domain or purely signature-domain inference setups. Inference setups that exploit a combination of time series *and* signatures are of interest, as well as setups where signatures are used to inform the calibration of individual model components (e.g., see the snowmelt calibration study by He et al., (2015) and the base flow calibration study by Su et al., (2016)). From the Bayesian inference perspective, a question that would need to be resolved is the specification of the inference in a way that does not "use the same data twice" (if both streamflow time series and their signatures are used); and
5. We sidestepped the important topic of dealing with strong variability and/or nonstationarity in the data, catchment, and/or model (see sections 3.1 and 3.4.2). For example, when nonstationarity in catchment conditions manifests in major changes in signatures, signature-domain inference can become problematic, in the same way as a time-domain inference. The treatment of hydrological nonstationarity is a major research challenge in its own right, certainly related to the question of model development, data collection, inference, diagnostic inspection, and application (e.g., Montanari et al., 2013; Sadegh et al., 2015; Vaze et al., 2010; Westra et al., 2014); it clearly warrants dedicated investigation.

These research questions will be pursued in future work.

6. Conclusions

This paper investigated Bayesian signature-domain calibration of hydrological models and its implementation using Approximate Bayesian Computation (ABC). Following a brief review of theoretical background along the perspectives presented by Kavetski et al. (2018), the empirical behavior of signature-domain inference using ABC was explored using a series of synthetic and real data experiments. The experiments employed deterministic hydrological models with additive residual error models, for which the time-domain likelihood function is available in computationally fast closed form. However, the signature-domain likelihood function was not available in closed form, which motivated the ABC implementation. The empirical analyses had a particular focus on comparing signature and time-domain inferences in terms of uncertainty quantification in the calibrated parameters and predicted streamflow, and in terms of computational costs.

The synthetic experiments suggest that:

1. Bayesian signature-domain inference (implemented using ABC) and Bayesian time-domain inference (implemented using traditional MCMC) provide consistent results, in the sense that: (i) posterior

parameter distributions and predictive distributions obtained using the time-domain inference are always enveloped by the distributions obtained using the signature-domain inference; (ii) with a suitable selection of signatures—here, four quantiles of the Flow Duration curve (FDC) and the flashiness index—the predictive distributions obtained from the two approaches are close to identical, suggesting common hydrological signatures can be relatively close to sufficiency, at least for relatively simple hydrological and residual error models; (iii) with a suitable selection of signatures (in particular, including the flashiness index), signature-domain inference is able to infer the standard deviation of the streamflow residual error model even when calibrating solely to streamflow signatures; and (iv) the posterior parameter uncertainty in signature-domain inference follows the same square-root scaling law with respect to the number of observations as the posterior parameter uncertainty in time-domain inference, even when calibrating to a set of signatures of fixed dimension;

2. Bayesian signature-domain inference provides a systematic platform for testing signatures as summary statistics of the original streamflow time series, for calibrating model parameters and for generating predictions. This analysis can proceed directly by comparing signature versus time-domain inferences (if the latter is available), or indirectly by comparing signature-domain inferences with different choices of signature sets. Signatures are generally associated with information loss, unless they represent the complete set of sufficient statistics for the assumed probability model. Such a set is difficult or impossible to derive for practical hydrological models. Therefore, identifying and understanding which signatures are most informative for hydrological parameter estimation, including for parameters controlling the uncertainty of the predictions, requires more research;
3. In some cases, the use of signatures can reduce the impact of particular deficiencies in the probability model on the estimated parameters; e.g., calibrating to the FDC can mitigate against unaccounted timing errors. This behavior is a consequence of working in the signature domain, rather than a property of the ABC algorithm. The use of signatures to improve the robustness of model calibration is of major practical interest; that said, achieving this effect is not straightforward and, in the absence of an understanding of the nature of the deficiency, can even become counterproductive; and
4. ABC sampling is computationally demanding. In our experiments, signature-domain inference using ABC sampling was about 40 times slower to converge than time-domain inference using traditional MCMC sampling. For this reason, ABC may be reserved for inference setups where the hydrological models are (relatively) fast computationally but their likelihood function is not readily accessible—examples of such setups include signature-domain inference motivated by lack of time series data, signature domain inference motivated by mitigation against deficiencies in the model (subject to the caveats in point 3 above), and general inference using complex stochastic models.

Experiments using streamflow data from the Lacmalac catchment in Australia yield the following insights:

1. Provided a suitable combination of signature is found, Bayesian signature-domain inference can produce probabilistic streamflow predictions comparable to those estimated using time-domain inference. The set of signatures including 20 FDC quantiles and the flashiness index provided good performance, with only a moderate loss of precision in the predicted daily streamflow time series. This finding corroborates the results of the synthetic study, and provides practical evidence of the ability of Bayesian signature-domain inference implemented using ABC to quantify uncertainties in the streamflow time series even when calibrating solely to a set of signatures;
2. In real data conditions, where the probability model is imperfect, there may be tradeoffs between the ability of the model to reproduce different signatures, and consequent tradeoffs in the ability of the model to fit different aspects of underlying streamflow dynamics.
3. The Flow Duration curve on its own does not convey sufficient information about the hydrograph to produce reliable and precise streamflow predictions; it is beneficial to include signatures such as the flashiness index that can provide information about flow dynamics. This finding appears to be intuitive and differs from the conclusions of previous studies; it shows the value of the Bayesian implementation of signature-domain inference as a platform to compare different signatures in terms of their ability to capture the information content of the corresponding time series.

The ability of Bayesian signature-domain inference (implemented using ABC) to provide reliable estimates of predictive streamflow uncertainty—even when calibrating to signatures alone—makes it attractive for several hydrological applications, including prediction in sparsely gauged and ungauged locations, where streamflow time series might not be available but signatures might be estimated through regionalization. The potential ability of signature-domain inference to mitigate against deficiencies in the assumed probability model is another promising area for investigation. While the current study focused on model calibration to streamflow data, the Bayesian inference methods and ABC numerical sampling approaches described in this work are general and applicable to the modeling of other environmental data sets, including groundwater levels, chemical concentrations, and so forth.

Appendix A: Synthetic Forcing Data

The synthetic precipitation and potential evaporation data were generated in two stages:

Stage 1: Calibrate a time series model to the observed data. This ensures that the synthetic data broadly resembles real hydrological data;
Stage 2: Generate realizations from the time series model calibrated in Stage 1.

The technical details of Stage 1 and 2 are listed next.

A1. Stage 1: Model Calibration

The time series model is calibrated as follows:

1. IN: Daily data *x*;
2. Calculate "monthly-averaged daily-step" data $\bar{\boldsymbol{x}}$ by applying a moving average filter to daily data $\boldsymbol{x}$, with a window width N_m of 28 days,

$$\bar{x}_t = \text{ave}\left(x_t, x_{t-1}, \ldots, x_{t-N_m+1}\right) \tag{A1}$$

3. Remove skewness from $\bar{x}$ by applying a square root transformation

$$\psi_t = \sqrt{\bar{x}_t} \tag{A2}$$

4. Detrend ψ_t as follows (Kantelhardt et al., 2006):
5. Calculate the mean $\bar{\psi}_t$ and standard deviation $\bar{\bar{\psi}}_t$ of the time series for each day of the calendar year (here, the multiple samples are given by multiple years in the record, so if we have 20 years of data then we have 20 samples for each calendar day).

$$\bar{\psi}_t = \text{ave}[\psi_t, \psi_{t+365}, \psi_{t+2\times365}, \ldots] \tag{A3}$$

$$\bar{\bar{\psi}}_t = \text{sdev}[\psi_t, \psi_{t+365}, \psi_{t+2\times365}, \ldots] \tag{A4}$$

6. To reduce the periodic seasonal trend, calculate the departures ("anomalies") δ_t of ψ_t from $\bar{\psi}_t$, for each calendar day:

$$\delta_t = \psi_t - \bar{\psi}_t \tag{A5}$$

7. To reduce remaining trends, standardize the departures δ_t by $\bar{\bar{\psi}}_t$, for each calendar day:

$$\xi_t = \frac{\delta_t}{\bar{\bar{\psi}}_t} \tag{A6}$$

8. Model ξ_t as an AR1 process at a monthly time step:

$$\xi_t = \gamma \xi_{t-N_m} + u_t \tag{A7}$$

with the innovations u_t assumed to follow a truncated Gaussian distribution $\mathcal{TN}\left(0, \sigma_u^2, L_{u,t}\right)$, with the lower bound $L_{u,t}$ defined such that $\psi_t > 0$,

$$L_{u,t} = -\left(\gamma \xi_{t-N_m} + \frac{\bar{\psi}_t}{\bar{\bar{\psi}}_t}\right) \tag{A8}$$

The parameters γ and σ_u of the AR1 process in equation (A7) were estimated from the observed data using the method of Maximum Likelihood.

Note that the procedure in steps 1–5 is only used to generate data for the synthetic experiments and is not intended as a rigorous stochastic model of precipitation or potential evaporation.

A2. Stage 2: Data Generation

Realizations of the synthetic precipitation and evaporation time series were generated as follows:

1. Sample a time series realization $\{\xi_t^{(i)}, t=1, \ldots, N_t\}$ from the AR1 process with parameters (γ, σ_u);
2. Calculate $\{\bar{x}_t^{(i)}, t=1, \ldots, N_t\}$ by applying equations (A2–A6) in reverse order.
3. Repeat for $i=1, \ldots, N_{rep}$, where N_{rep} is the required number of synthetic replicates.

In order to undertake Experiment 1.2, we generated a single long-time series of 500 years (6,000 monthly time steps). In experiments 1, 2, and 4, we used a 102-year long subset of this series (1,331 monthly time steps). The data partitioning into warm up, calibration, and validation periods are described in section 3.4.1.

Acknowledgments

We thank Editors Alberto Montanari, Scott Mackay, and the Associate Editor for handling our manuscript. We are grateful to Mojtaba Sadegh, Saman Razavi, Mahyar Shafii, and two anonymous reviewers for their insightful comments and constructive criticisms, which helped us substantially improve the study and its presentation. The contributions of the second author were partially supported by the Australian Research Council Linkage Grant LP140100978. The hydrological time series and the ABC algorithm used in the case study experiments are available on request from the first author (FF). The SABC algorithm, including parallelization facilities, is available in a Python library from https://github.com/eth-cscs/abcpy.

References

Albert, C., Kunsch, H., & Scheidegger, A. (2014). A simulated annealing approach to approximate Bayes computations. *Statistics and Computing*, *25*(6), 1217–1232. https://doi.org/10.1007/s11222-014-9507-8

Albert, C., Ulzega, S., & Stoop, R. (2016). Boosting Bayesian parameter inference of nonlinear stochastic differential equation models by Hamiltonian scale separation. *Physical Review E*, *93*(4), 043313.

Baker, D. B., Richards, R. P., Loftus, T. T., & Kramer, J. W. (2004). A new flashiness index: Characteristics and applications to midwestern rivers and streams. *Journal of the American Water Resources Association*, *40*(2), 503–522. https://doi.org/10.1111/j.1752-1688.2004.tb01046.x

Bates, B. C., & Campbell, E. P. (2001). A Markov chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modeling. *Water Resources Research*, *37*(4), 937–947.

Benettin, P., Bailey, S. W., Campbell, J. L., Green, M. B., Rinaldo, A., Likens, G. E., et al. (2015). Linking water age and solute dynamics in streamflow at the Hubbard Brook Experimental Forest, NH, USA. *Water Resources Research*, *51*, 9256–9272. https://doi.org/10.1002/2015WR017552

Bertuzzo, E., Thomet, M., Botter, G., & Rinaldo, A. (2013). Catchment-scale herbicides transport: Theory and application. *Advances in Water Resources*, *52*, 232–242. https://doi.org/10.1016/j.advwatres.2012.11.007

Beven, K., & Binley, A. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, *6*(3), 279–298.

Botter, G., Zanardo, S., Porporato, A., Rodriguez-Iturbe, I., & Rinaldo, A. (2008). Ecohydrological model of flow duration curves and annual minima. *Water Resources Research*, *44*, W08418. https://doi.org/10.1029/2008WR006814

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis* (588 s. pp.). Reading, MA: Addison-Wesley.

Boyle, D. P. (2001). *Multicriteria calibration of hydrological models* (PhD thesis). Tucson: Department of Hydrology and Water Resources, University of Arizona.

Castellarin, A. (2014). Regional prediction of flow-duration curves using a three-dimensional kriging. *Journal of Hydrology*, *513*, 179–191. https://doi.org/10.1016/j.jhydrol.2014.03.050

Castellarin, A., Botter, G., Hughes, D. A., Liu, S., Ouarda, T. B. M. J., Parajka, J., et al. (2013). Prediction of flow duration curves in ungauged basins. In G. Blöschl et al. (Eds.), *Runoff prediction in ungauged basins*. Cambridge, UK: Cambridge University Press.

Castellarin, A., Camorani, G., & Brath, A. (2007). Predicting annual and long-term flow-duration curves in ungauged basins. *Advances in Water Resources*, *30*(4), 937–953. https://doi.org/10.1016/j.advwatres.2006.08.006

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

Duan, Q., Di, Z., Quan, J., Wang, C., & Gong, W. (2017). Automatic model calibration: A new way to improve numerical weather forecasting. *Bulletin of the American Meteorological Society*, *98*(5), 959–970. https://doi.org/10.1175/bams-d-15-00104.1

Eckhardt, K. (2008). A comparison of baseflow indices, which were calculated with seven different baseflow separation methods. *Journal of Hydrology*, *352*(1–2), 168–173. https://doi.org/10.1016/j.jhydrol.2008.01.005

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*, 2350–2375. https://doi.org/10.1002/2013WR014185

Fenicia, F., Kavetski, D., & Savenije, H. H. G. (2011). Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. *Water Resources Research*, *47*, W11510. https://doi.org/10.1029/2010WR010174

Fenicia, F., McDonnell, J. J., & Savenije, H. H. G. (2008). Learning from model improvement: On the contribution of complementary data to process understanding. *Water Resources Research*, *44*, W06419. https://doi.org/10.1029/2007WR006386

Freer, J. E., McMillan, H., McDonnell, J. J., & Beven, K. J. (2004). Constraining dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures. *Journal of Hydrology*, *291*(3–4), 254–277. https://doi.org/10.1016/j.jhydrol.2003.12.037

Gassmann, M., Stamm, C., Olsson, O., Lange, J., Kummerer, K., & Weiler, M. (2013). Model-based estimation of pesticides and transformation products and their export pathways in a headwater catchment. *Hydrology and Earth System Sciences*, *17*(12), 5213–5228. https://doi.org/10.5194/hess-17-5213-2013

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed., xxv, 668. p.). Boca Raton, FL: Chapman & Hall/CRC.

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(2), 243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x

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

Gupta, H. V., Wagener, T., & Liu, Y. Q. (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

He, Z. H., Tian, F. Q., Gupta, H. V., Hu, H. C., & Hu, H. P. (2015). Diagnostic calibration of a hydrological model in a mountain area by hydrograph partitioning. *Hydrology and Earth System Sciences*, *19*(4), 1807–1826. https://doi.org/10.5194/hess-19-1807-2015

Hill, M. C., Kavetski, D., Clark, M., Ye, M., Arabi, M., Lu, D., et al. (2016). Practical use of computationally frugal model analysis methods. *Groundwater*, *54*(2), 159–170. https://doi.org/10.1111/gwat.12330

Hrachowitz, M., Savenije, H., Bogaard, T. A., Tetzlaff, D., & Soulsby, C. (2013). What can flux tracking teach us about water age distribution patterns and their temporal dynamics? *Hydrology and Earth System Sciences*, *17*(2), 533–564. https://doi.org/10.5194/hess-17-533-2013

Jennings, S. A., Lambert, M. F., & Kuczera, G. (2010). Generating synthetic high resolution rainfall time series at sites with only daily rainfall using a master–target scaling approach. *Journal of Hydrology*, *393*(3), 163–173. https://doi.org/10.1016/j.jhydrol.2010.08.013

Jepsen, S. M., Harmon, T. C., & Shi, Y. (2016). Watershed model calibration to the base flow recession curve with and without evapotranspiration effects. *Water Resources Research*, *52*, 2919–2933. https://doi.org/10.1002/2015WR017827

Kantelhardt, J. W., Koscielny-Bunde, E., Rybski, D., Braun, P., Bunde, A., & Havlin, S. (2006). Long-term persistence and multifractality of precipitation and river runoff records. *Journal of Geophysical Research*, *111*, D01106. https://doi.org/10.1029/2005JD005881

Kavetski, D., & Clark, M. P. (2010). Ancient numerical daemons of conceptual hydrological modeling: 2. Impact of time stepping schemes on model analysis and prediction. *Water Resources Research*, *46*, W10511. https://doi.org/10.1029/2009WR008896

Kavetski, D., & Fenicia, F. (2011). Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights. *Water Resources Research*, *47*, W11511. https://doi.org/10.1029/2011WR010748

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*. https://doi.org/10.1002/2017WR020258

Kavetski, D., & Kuczera, G. (2007). Model smoothing strategies to remove microscale discontinuities and spurious secondary optima in objective functions in hydrological calibration. *Water Resources Research*, *43*, W03411.https://doi.org/10.1029/2006WR005195

Liu, Y., Brown, J., Demargne, J., & Seo, D.-J. (2011). A wavelet-based approach to assessing timing errors in hydrologic predictions. *Journal of Hydrology*, *397*(3–4), 210–224. 10.1016/j.jhydrol.2010.11.040.

Lockart, N., Kavetski, D., & Franks, S. W. (2015). A new stochastic model for simulating daily solar radiation from sunshine hours. *International Journal of Climatology*, *35*(6), 1090–1106. https://doi.org/10.1002/joc.4041

Lyne, V., & Hollick, M. (1979). *Stochastic time-variable rainfall runoff modelling*. Paper presented at Proceedings of the Hydrology and Water Resources Symposium, Perth, 10–12 September (*79*/10, pp. 89–92).

Marjoram, P., Molitor, J., Plagnol, V., & Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences United States of America*, *100*(26), 15324–15328. https://doi.org/10.1073/pnas.0306899100

McGuire, K. J., McDonnell, J. J., Weiler, M., Kendall, C., McGlynn, B. L., Welker, J. M., et al. (2005). The role of topography on catchment-scale water residence time. *Water Resources Research*, *41*, W05002. https://doi.org/10.1029/2004WR003657

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*, 2199–2239. https://doi.org/10.1002/2016WR019168

Montanari, A., & Toth, E. (2007). Calibration of hydrological models in the spectral domain: An opportunity for scarcely gauged basins? *Water Resources Research*, *43*, W05434. https://doi.org/10.1029/2006WR005184

Montanari, A., Young, A., Savenije, H. H. G., Hughes, D., Wagener, T., Ren, L. L. (2013). Panta Rhei-everything flows'': Change in hydrology and society-The IAHS Scientific Decade 2013–2022. *Hydrological Sciences Journal*, *58*(6), 1256–1275. https://doi.org/10.1080/02626667.2013.809088

Nott, D. J., Marshall, L., & Brown, J. (2012). Generalized likelihood uncertainty estimation (GLUE) and approximate Bayesian computation: What's the connection? *Water Resources Research*, *48*, W12602. https://doi.org/10.1029/2011wr011128

Paillex, A., Reichert, P., Lorenz, A. W., & Schuwirth, N. (2017). Mechanistic modelling for predicting the effects of restoration, invasion and pollution on benthic macroinvertebrate communities in rivers. *Freshwater Biology*, *62*(6), 1083–1093. https://doi.org/10.1111/fwb.12927

Press, W. H. (1992). *Numerical recipes in Fortran the art of scientific computing* (2nd ed., 963. pp.). Cambridge, UK: University Press.

Priestley, M. B. (1981). *Spectral analysis and time series*. New York, NY: Academic Press.

Razavi, S., & Tolson, B. A. (2013). An efficient framework for hydrologic model calibration on long data periods. *Water Resources Research*, *49*, 8418–8431. https://doi.org/10.1002/2012WR013442

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

Sadegh, M., & Vrugt, J. A. (2013). Bridging the gap between GLUE and formal statistical approaches: Approximate Bayesian computation. *Hydrology and Earth System Sciences*, *17*(12), 4831–4850. https://doi.org/10.5194/hess-17-4831-2013

Sadegh, M., & Vrugt, J. A. (2014). Approximate Bayesian computation using Markov Chain Monte Carlo simulation: DREAM((ABC)). *Water Resources Research*, *50*, 6767–6787. https://doi.org/10.1002/2014WR015386

Sadegh, M., Vrugt, J. A., Gupta, H. V., & Xu, C. (2016). The soil water characteristic as new class of closed-form parametric expressions for the flow duration curve. *Journal of Hydrology*, *535*(Suppl. C), 438–456. https://doi.org/10.1016/j.jhydrol.2016.01.027

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

Seibert, J., & McDonnell, J. J. (2002). On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration. *Water Resources Research*, *38*(11), 1241. https://doi.org/10.1029/2001WR000978

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

Sivapalan, M. (2006). Pattern, process and function: Elements of a unified theory of hydrology at the catchment scale. In M. G. Anderson (Ed.), *Encyclopedia of hydrological sciences*. Hoboken, NJ: John Wiley.

Smith, T., Sharma, A., Marshall, L., Mehrotra, R., & Sisson, S. (2010). Development of a formal likelihood function for improved Bayesian inference of ephemeral catchments. *Water Resources Research*, *46*, W12551. https://doi.org/10.1029/2010WR009514

Son, K., & Sivapalan, M. (2007). Improving model structure and reducing parameter uncertainty in conceptual water balance models through the use of auxiliary data. *Water Resources Research*, *43*, W01415. https://doi.org/10.1029/2006WR005032

Su, C.-H., Peterson, T. J., Costelloe, J. F., & Western, A. W. (2016). A synthetic study to evaluate the utility of hydrological signatures for calibrating a base flow separation filter. *Water Resources Research*, *52*, 6526–6540. https://doi.org/10.1002/2015WR018177

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W., & Srikanthan, S. (2009). Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis. *Water Resources Research*, *45*, W00B14. https://doi.org/10.1029/2008WR006825

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., & Stumpf, M. P. H. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, *6*(31), 187–202. https://doi.org/10.1098/rsif.2008.0172

Vaze, J., Post, D. A., Chiew, F. H. S., Perraud, J. M., Viney, N. R., & Teng, J. (2010). Climate non-stationarity: Validity of calibrated rainfall–runoff models for use in climate change studies. *Journal of Hydrology*, *394*(3), 447–457. https://doi.org/10.1016/j.jhydrol.2010.09.018

Vrugt, J. A., Diks, C. G. H., Gupta, H. V., Bouten, W., & Verstraten, J. M. (2005). Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation. *Water Resources Research*, *41*, W01017. https://doi.org/10.1029/2004WR003059

Vrugt, J. A., Gupta, H. V., Bouten, W., & Sorooshian, S. (2003). A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resources Research*, *39*(8), 1201. https://doi.org/10.1029/2002WR001642

Vrugt, J. A., & Sadegh, M. (2013). Toward diagnostic model calibration and evaluation: Approximate Bayesian computation. *Water Resources Research*, *49*, 4335–4345. https://doi.org/10.1002/wrcr.20354

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., & Sorooshian, S. (2001). A framework for development and application of hydrological models. *Hydrology and Earth System Sciences*, *5*(1), 13–26.

Weiss, G., & von Haeseler, A. (1998). Inference of population history using a likelihood approach. *Genetics*, *149*(3), 1539–1546.

Westerberg, I. K., Guerrero, J. L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., et al. (2011). Calibration of hydrological models using flow-duration curves. *Hydrology and Earth System Sciences*, *15*(7), 2205–2227. https://doi.org/10.5194/hess-15-2205-2011

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

Westra, S., Thyer, M., Leonard, M., Kavetski, D., & Lambert, M. (2014). A strategy for diagnosing and interpreting hydrological model nonstationarity. *Water Resources Research*, *50*, 5090–5113. https://doi.org/10.1002/2013WR014719

Winsemius, H. C., Savenije, H. H. G., van de Giesen, N. C., van den Hurk, B. J. J. M., Zapreeva, E. A., & Klees, R. (2006). Assessment of gravity recovery and climate experiment (GRACE) temporal signature over the upper Zambezi. *Water Resources Research*, *42*, W12201. https://doi.org/10.1029/2006WR005192

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, *466*, 1102–1104.

Yadav, M., Wagener, T., & Gupta, H. (2007). Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. *Advances in Water Resources*, *30*(8), 1756–1774. https://doi.org/10.1016/j.advwatres.2007.01.005

Ye, W., Bates, B. C., Viney, N. R., Sivapalan, M., & Jakeman, A. J. (1997). Performance of conceptual rainfall-runoff models in low-yielding ephemeral catchments. *Water Resources Research*, *33*(1), 153–166. https://doi.org/10.1029/96WR02840

Yilmaz, K. K., Gupta, H. V., & Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resources Research*, *44*, W09417. https://doi.org/10.1029/2007WR006716

Yu, P. S., & Yang, T. C. (2000). Using synthetic flow duration curves for rainfall-runoff model calibration at ungauged sites. *Hydrological Processes*, *14*(1), 117–133. https://doi.org/10.1002/(SICI)1099-1085(200001)14:1<117::AID-HYP914>3.0.CO;2-Q

Zhu, A., Guo, J., Ni, B.-J., Wang, S., Yang, Q., & Peng, Y. (2015). A novel protocol for model calibration in biological wastewater treatment, *Science Reports*, *5*, 8493, https://doi.org/10.1038/srep08493, https://www.nature.com/articles/srep08493#supplementary-information