AGU100 ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1002/2017WR020528

This article is a companion to Fenicia et al. [2018], https://doi.org/10.1002/2017WR021616.

Key Points:
- Signature-domain calibration can be implemented using Approximate Bayesian Computation (ABC)
- ABC is a class of sampling techniques that does not require evaluation of the likelihood function (e.g., if unavailable in closed form)
- Applications that omit randomness in the hydrological model and/or use a coarse ABC tolerance do not achieve the full potential of ABC

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

Citation:
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*, 4059–4083. https://doi.org/10.1002/2017WR020528

Received 3 FEB 2017
Accepted 19 MAR 2018
Accepted article online 6 APR 2018
Published online 30 JUN 2018

Signature-Domain Calibration of Hydrological Models Using Approximate Bayesian Computation: Theory and Comparison to Existing Applications

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

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

Abstract This study considers Bayesian calibration of hydrological models using streamflow signatures and its implementation using Approximate Bayesian Computation (ABC). If the modeling objective is to predict streamflow time series and associated uncertainty, a probabilistic model of streamflow must be specified but the inference equations must be developed in the signature domain. However, even starting from simple probabilistic models of streamflow time series, working in the signature domain makes the likelihood function difficult or impractical to evaluate (in particular, as it is unavailable in closed form). This challenge can be tackled using ABC, a general class of numerical algorithms for sampling from conditional distributions, such as (but not limited to) Bayesian posteriors given any calibration data. Using ABC does not avoid the requirement of Bayesian inference to specify a probability model of the data, but rather exchanges the requirement to *evaluate* the pdf of this model (needed to evaluate the likelihood function) by the requirement to *sample* model output realizations. For this reason ABC is attractive for inference in the signature domain. We clarify poorly understood aspects of ABC in the hydrological literature, including similarities and differences between ABC and GLUE, and comment on previous applications of ABC in hydrology. An error analysis of ABC approximation errors and their dependence on the tolerance is presented. An empirical case study is used to illustrate the impact of omitting the specification of a probabilistic model (and instead using a deterministic model within the ABC algorithm), and the impact of a coarse ABC tolerance.

1. Introduction

Hydrological models describing the relationship between rainfall and runoff are widely used in environmental sciences and applications. Hydrological model estimation and predictive use must be robust and efficient in the face of multiple sources of uncertainty. For example, both rainfall and streamflow can be subject to major observational errors (e.g., McMillan et al., 2011a) and many catchments worldwide are still poorly gauged or ungauged (e.g., Sivapalan et al., 2003). In addition, model deficiencies inevitably arise from the simplified nature of hydrological models. Model deficiencies not only degrade model predictions (e.g., by introducing systematic predictive biases), but often lead to the need to balance tradeoffs between the ability of the model to meet multiple objectives (Gupta et al., 1998; Reichert & Schuwirth, 2012, and others).

The calibration of hydrological models has received tremendous attention in the literature. Model calibration has the goal of finding parameter values and uncertainty ranges that lead to a close match of simulated and observed responses (typically streamflow). In hydrological modeling, this aim has been pursued using various strategies, ranging from single-objective optimization (e.g., Duan et al., 1992; Tolson & Shoemaker, 2007) and multi-objective optimization (e.g., Gupta et al., 1998; Madsen, 2000; Vrugt et al., 2003), to Bayesian methods (e.g., Evin et al., 2014; Kuczera & Parent, 1998; Reichert & Mieleitner, 2009; Renard et al., 2011; Smith et al., 2015; Vrugt et al., 2008), "informal" techniques (e.g., Beven & Binley, 1992; Freer et al., 2004), and others. The representation of data and model uncertainties, and the quantification of the resulting uncertainty in the calibrated model parameters and predictions, has been pursued using paradigms such as probability theory (Bayesian and frequentist) (e.g., Ang & Tang, 2007; Box & Tiao, 1973, see references above for hydrological applications), fuzzy set theory and possibilistic approaches (e.g., Franks et al., 1998; Wang et al., 2016), and informal techniques (e.g., Beven & Binley, 1992; Smith et al., 2008).

This study considers "signature-domain" calibration – an alternative to traditional "time-domain" calibration where the modeler seeks to match not the streamflow time series themselves, but specific features, or "signatures," of these time series. Typical signatures used in hydrological calibration include the Flow Duration Curve (FDC) and master recession curves, baseflow indices, and other streamflow characteristics (e.g., Castellarin et al., 2013; Lamb & Beven, 1997; Vogel & Fennessey, 1995; Westerberg & McMillan, 2015; Yadav et al., 2007). In this paper, we define signature-domain calibration as using signatures directly for parameter estimation (e.g., Shafii & Tolson, 2015; Vrugt & Sadegh, 2013; Westerberg et al., 2011; Yilmaz et al., 2008), rather than solely for posteriori diagnostics (e.g., Kavetski et al., 2011; McMillan et al., 2011b).

Signature-domain calibration offers appealing benefits for hydrological applications, yet it is not immune to many of the same challenges that affect traditional time series calibration. Uncertainty quantification, in the estimated parameters and predicted streamflow, is a major focus of this work.

The simplest approach to signature-domain calibration in hydrology is to optimize an objective function that quantifies the mismatch between one or more signatures of observed and simulated streamflow (Castiglioni et al., 2010; Shafii & Tolson, 2015; Yu & Yang, 2000; Zhang et al., 2008). The optimization approach can provide point estimates of model parameters and predictions, but does not provide estimates of uncertainty in these quantities. Given the demand for uncertainty quantification in environmental modeling and decision-making (e.g., Reichert et al., 2015), this is an important limitation. In addition, uncertainty quantification is important when assessing and comparing the effectiveness of signatures in capturing the information content of hydrological time series.

Current approaches for uncertainty quantification in signature-domain hydrological calibration can be classified into the following broad categories:

a. Applications based on the Generalized Likelihood Uncertainty Estimation (GLUE) approach of Beven and Binley (1992). For example, the study of Winsemius et al. (2009) on prediction in ungauged basins defined a pseudo-likelihood function based on a combination of hydrograph characteristics and monthly water balance estimates. In another study, Westerberg et al. (2011) defined a pseudo-likelihood function in terms of quantiles of the FDC and incorporated the effects of rating curve uncertainty. The limitations of GLUE approaches in the context of probabilistic uncertainty quantification have been debated in the hydrological literature (e.g., Beven et al., 2008; Mantovan & Todini, 2006; Stedinger et al., 2008);
b. Multi-objective calibration "extended" to provide uncertainty estimates. For example, Yilmaz et al. (2008) assessed the sensitivity of parameter estimates to the choice of calibration signatures, including the runoff coefficient, streamflow lag times and FDC characteristics. However they found only a single parameter set that produced hydrographs with signatures within a specified tolerance of observed signatures. In an ungauged basin study, Yadav et al. (2007) used a similar approach to construct prediction envelopes from hydrographs corresponding to parameter sets that satisfied a signature-domain tolerance (defined using regionalization). These studies are conceptually similar to the GLUE applications listed above and share the same limitations;
c. Simple Bayesian inference with the likelihood function formulated in the signature domain. For example, Ye et al. (2012) calibrated a conceptual rainfall-runoff model to a "regime" curve constructed using daily streamflow averages over multiple calendar years, with a likelihood function that assumed independent constant-variance Gaussian errors in the regime curve. This approach can be used to quantify the uncertainty in predicted streamflow signatures, but not in predicted streamflow itself (section 2.2);
d. Applications of "Approximate Bayesian Computation" (ABC), where parameter samples are accepted or rejected based on the similarity of corresponding probabilistic predictions to the observed data (with similarity quantified using a numerical tolerance). ABC methods originated in the field of statistics (e.g., Blum et al., 2013; Diggle & Gratton, 1984; Marjoram et al., 2003), in particular motivated by applications in genetics (e.g., Csilléry et al., 2010; Pritchard et al., 1999; Tavare et al., 1997). The study by Nott et al. (2012) was amongst the first in hydrology to consider ABC, and to comment on its apparent semblance to GLUE. Subsequently, Vrugt and Sadegh (2013) proposed ABC as a tractable way to address uncertainties in signature-domain calibration, and as a vehicle for diagnostic model evaluation.

From the point of view of probabilistic uncertainty quantification, ABC algorithms are of clear interest. However, although signature-domain inference and ABC algorithms are well rooted in Bayesian statistics and probability theory (e.g., Csilléry et al., 2010; see Turner & Van Zandt, 2012 for a tutorial), their applications in

hydrology have been generally limited and incomplete. We identify three common concerns: the usage of deterministic models as if they were probabilistic models within ABC applications (Nott et al., 2012; Sadegh & Vrugt, 2013), the misconception that ABC approaches are statistical methods for "cases when an explicit likelihood (objective) function cannot be justified" (Vrugt & Sadegh, 2013), and the oversight of important conceptual and algorithmic differences between ABC and GLUE (Nott et al., 2012; Sadegh & Vrugt, 2013). Given the potential of signature-domain techniques and ABC to address problems of hydrological interest, we consider it important that these techniques be judiciously translated into hydrological context.

This paper pursues the following aims:

1. Present the theory of Bayesian signature-domain calibration of hydrological models, with emphasis on the relation between probability models in the time domain and likelihood functions in the signature domain;
2. Articulate the Approximate Bayesian Computation (ABC) approach as a general numerical sampling strategy, describe its suitability for signature-domain inference and interpret its error properties;
3. Critically appraise similarities and differences between ABC and GLUE, and clarify misunderstandings that have arisen in previous applications of ABC in the hydrological literature.

These aims are investigated using theoretical and empirical analyses. Aims 1 and 2 are pursued using theoretical analysis to highlight the conceptual and mathematical connections between different inference approaches. Aim 3 is pursued through an extension of the empirical case study in Fenicia et al. (2018), using rainfall-runoff models representative of current hydrological modeling research and operation.

The paper is structured as follows. Section 2 describes the Bayesian framework for time-domain and signature-domain inference, and contrasts two approaches for sampling from Bayesian posteriors. Approximate Bayesian Computation (ABC) is introduced as a sampling algorithm attractive for inference problems where sampling from the assumed probability model is much easier than evaluating its probability density function (e.g., if the latter is not known in closed form). Section 3 discusses similarities and differences between ABC and GLUE. Section 4 describes the empirical case study used to illustrate signature-domain inference using ABC and highlight the effect of changes such as omitting the random terms from the model, or using a coarse ABC tolerance. Section 5 discusses the case study results and interprets them from a theoretical perspective. Section 6 summarizes the key conclusions of the study and outlines directions for future research.

2. Theory

2.1. Hydrological Model

Consider a probabilistic hydrological model that represents streamflow time series as a random vector $\boldsymbol{Q}(\theta, \boldsymbol{x})$, and is assumed to characterize the observed streamflow time series $\tilde{\boldsymbol{q}}$,

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

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

In many hydrological applications, $\boldsymbol{Q}(\theta, \boldsymbol{x})$ is constructed from a deterministic hydrological model $\boldsymbol{h}$ such as HyMod (Boyle, 2001) or GR4J (Perrin et al., 2003), by adding a random residual error term $\boldsymbol{\mathcal{E}}$ intended to represent the combined effect of all data and model uncertainties,

$$\boldsymbol{Q}(\theta, \boldsymbol{x}) = \boldsymbol{h}(\theta_h, \boldsymbol{x}) + \boldsymbol{\mathcal{E}}(\theta_{\mathcal{E}}) \tag{2}$$

where $\theta = (\theta_h, \theta_{\mathcal{E}})$ comprises hydrological model parameters θ_h and error model parameters $\theta_{\mathcal{E}}$. The simplest residual error model is an independent Gaussian, $\boldsymbol{\mathcal{E}} \sim \mathcal{N}(0, \sigma_{\mathcal{E}})$ with standard deviation $\sigma_{\mathcal{E}}$, but it can be easily extended to reflect error heteroscedasticity, persistence and other properties (e.g., McInerney et al., 2017; Schoups & Vrugt, 2010; Smith et al., 2015).

More generally, $\boldsymbol{Q}(\theta, \boldsymbol{x})$ can represent a genuine stochastic model where the representation of predictive uncertainty is internal to the model structure rather than represented by external error terms (e.g., Albert et al., 2016; Lockart et al., 2015; Reichert & Mieleitner, 2009; Renard et al., 2011).

The probability density function (pdf) describing $\boldsymbol{Q}(\theta,\boldsymbol{x})$ will be denoted as $p(\boldsymbol{q}|\theta,\boldsymbol{x})$. In many cases, $p(\boldsymbol{q}|\theta,\boldsymbol{x})$ is straightforward to evaluate. For example, for common choices of residual error models in equation (2), $p(\boldsymbol{q}|\theta,\boldsymbol{x})$ has a convenient closed-form expression (usually related to the Gaussian pdf) and is computationally cheap once the modeler furnishes the value of $\boldsymbol{h}(\theta_h,\boldsymbol{x})$ (e.g., McInerney et al., 2017; see equation (19) in Fenicia et al., 2018). In other cases, particularly if $\boldsymbol{Q}(\theta,\boldsymbol{x})$ represents a non-trivial stochastic model, $p(\boldsymbol{q}|\theta,\boldsymbol{x})$ may be infeasible to evaluate directly, especially if it does not have a convenient closed-form expression. Note that our usage of "closed-form" to describe computational quantities is necessarily quite loose and context-specific: it is used to denote any 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). For example, although in principle $p(\boldsymbol{q}|\theta,\boldsymbol{x})$ can be approximated by brute force from multiple realizations drawn from $\boldsymbol{Q}(\theta,\boldsymbol{x})$, e.g., using Gaussian or kernel approximations (e.g., Jennings et al., 2010; Lockart et al., 2015), in practice this approach would entail a forbidding undertaking whenever $\boldsymbol{q}$ and/or θ are high-dimensional, especially if $\boldsymbol{Q}(\theta,\boldsymbol{x})$ has a complicated time-dependence structure and/or is computationally expensive.

2.2. Bayesian Inference Framework

Consider the calibration of the probabilistic model $\boldsymbol{Q}(\theta,\boldsymbol{x})$ to a general set of observed data $\tilde{\boldsymbol{y}}$, which might comprise streamflow time series, hydrological signatures, and/or other observable quantities. Let us introduce the notation $\boldsymbol{Y}(\theta,\boldsymbol{x})$ for a more general probability model, with pdf $p(\boldsymbol{y}|\theta,\boldsymbol{x})$, in order to accommodate two scenarios of interest, namely time-domain inference and signature-domain inference.

In this study, time-domain calibration is defined by $\boldsymbol{Y}(\theta,\boldsymbol{x})$ being a model of streamflow time series,

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

and signature-domain calibration is defined by $\boldsymbol{Y}(\theta,\boldsymbol{x})$ being a model of a set of signatures,

$$\tilde{\boldsymbol{y}}=\boldsymbol{g}(\tilde{\boldsymbol{q}}),\quad \boldsymbol{Y}(\theta,\boldsymbol{x})=\boldsymbol{g}(\boldsymbol{Q}(\theta,\boldsymbol{x})),\quad \mathcal{Y}(\theta_h,\boldsymbol{x})=\boldsymbol{g}(\boldsymbol{h}(\theta_h,\boldsymbol{x})) \tag{4}$$

where $\boldsymbol{g}(\boldsymbol{q})$ is a vector of signatures computed from streamflow time series $\boldsymbol{q}$. For convenience, we define $\mathcal{Y}(\theta,\boldsymbol{x})$ to represent the deterministic model of observations $\tilde{\boldsymbol{y}}$ (in either the time or signature domains).

Given the probability model $\boldsymbol{Y}(\theta,\boldsymbol{x})$, a prior distribution of model parameters $p(\theta)$, and a set of observed data $\tilde{\boldsymbol{y}}$, the posterior distribution of model parameters $p(\theta|\tilde{\boldsymbol{y}},\boldsymbol{x})$ is given by Bayes' theorem,

$$p(\theta|\tilde{\boldsymbol{y}},\boldsymbol{x})=\frac{p(\tilde{\boldsymbol{y}}|\theta,\boldsymbol{x})p(\theta)}{p(\tilde{\boldsymbol{y}}|\boldsymbol{x})}\propto p(\tilde{\boldsymbol{y}}|\theta,\boldsymbol{x})p(\theta) \tag{5}$$

where $p(\tilde{\boldsymbol{y}}|\theta,\boldsymbol{x})$ is the likelihood function, defined as the pdf of $\boldsymbol{Y}(\theta,\boldsymbol{x})$ evaluated at the observed data $\tilde{\boldsymbol{y}}$ and viewed as a function of the parameters θ; for this reason it is often written as $\mathcal{L}(\theta;\tilde{\boldsymbol{y}})$.

The "signature-domain posterior" $p(\theta|\boldsymbol{g}(\tilde{\boldsymbol{q}}),\boldsymbol{x})$ will resemble the "time-domain posterior" $p(\theta|\tilde{\boldsymbol{q}},\boldsymbol{x})$ to the extent that the parameter-related information content in the signatures resembles the parameter-related information content in the time series. The signature-domain posterior can be expected to be "consistent with, but wider than" the time-domain posterior, because most signatures throw away at least some information from the time series they are computed from. However, the signature-domain posterior might deviate substantially from the time-domain posterior if the probability model $\boldsymbol{Q}$ is substantially mis-specified, as shown in the empirical case studies of Fenicia et al. (2018).

An important exception to the behavior above arises when *sufficient statistics* are used as signatures. Sufficient statistics are defined as functions of the data that encapsulate its entire information content relevant to (the parameters of) the assumed model (e.g., the mean and variance of a data set are sufficient statistics under the Gaussian model assumption; see section 5.3.1 for a discussion). For a given model, if a complete set of sufficient statistics is used as signatures, the signature-domain posterior will coincide with the time-domain posterior. In practice, sufficient statistics are difficult or impossible to derive for general probability models (see section 5.3.1), and furthermore there are circumstances where the modeler may deliberately use non-sufficient statistics (e.g., see Experiment 1.3 in Fenicia et al., 2018).

In terms of practical application, the immediate challenge in using equation (4) is the specification of the likelihood function $p(\boldsymbol{g}(\tilde{\boldsymbol{q}})|\theta,\boldsymbol{x})$. Two distinct approaches can be contemplated:

1. Formulate the predictive model in streamflow space, as in equation (1). But, even if the pdf $p(\boldsymbol{q}|\theta,\boldsymbol{x})$ is known in closed form, the closed form of the pdf $p(\boldsymbol{g}(\boldsymbol{q})|\theta,\boldsymbol{x})$ is often difficult or impossible to derive, especially when the signature $\boldsymbol{g}()$ is defined by multiple "non-analytical" algorithmic operations such as sorting, etc. For example, if the streamflow errors are assumed to be Gaussian, what is the distribution of errors in the FDC, or in the flashiness index (section 4.2.2)?
2. Formulate the predictive model directly in the signature space. For example, start with a deterministic model such as $\boldsymbol{g}(\boldsymbol{h}(\theta,\boldsymbol{x}))$, assume an error model for the signatures, such as Gaussian errors in the regime curve (Ye et al., 2012), carry out the inference, check the signature error model assumptions, and refine as appropriate. This approach readily yields parameter posteriors, but is restrictive when making probabilistic predictions of streamflow – for example, if errors in the regime curve are assumed to be Gaussian, what is the corresponding distribution of errors in the predicted streamflow time series?

These challenges are related, respectively, to the difficulty in analytically propagating streamflow uncertainty into the signatures, and the non-uniqueness of inverting (most) signatures to yield the streamflow time series.

We hence consider numerical techniques for uncertainty propagation. Note that, given the probabilistic model $\boldsymbol{Q}(\theta,\boldsymbol{x})$, sampling from the (derived) probability distribution of signatures $\boldsymbol{g}(\boldsymbol{Q})$ is straightforward: just sample a streamflow prediction $\boldsymbol{q}^{(i)} \leftarrow \boldsymbol{Q}(\theta,\boldsymbol{x})$ and compute its signature $\boldsymbol{g}^{(i)}=\boldsymbol{g}(\boldsymbol{q}^{(i)})$. It is the underlying pdf $p(\boldsymbol{g}(\boldsymbol{q})|\theta,\boldsymbol{x})$, needed to evaluate the likelihood function $p(\boldsymbol{g}(\tilde{\boldsymbol{q}})|\theta,\boldsymbol{x})$, that is difficult or impractical to compute because it does not have a closed-form expression for a general choice of $\boldsymbol{Q}$ and $\boldsymbol{g}$. Once again, an approximation to $p(\boldsymbol{g}(\boldsymbol{q})|\theta,\boldsymbol{x})$ could be constructed by brute force from realizations drawn from $\boldsymbol{g}(\boldsymbol{Q}(\theta,\boldsymbol{x}))$; in practice this approach is cumbersome and often infeasible to implement.

The situation where the assumed probability model of the data is much easier to sample from than to evaluate its probability density function motivates the class of numerical methods known as Approximate Bayesian Computation (ABC) (Pritchard et al., 1999; Tavare et al., 1997).

2.3. Sampling From Bayesian Posteriors

Two distinct approximation strategies are available for sampling from a Bayesian posterior $p(\theta|\tilde{\boldsymbol{y}},\boldsymbol{x})$:

Strategy A. Approaches that evaluate the right-hand-side of equation (5). For example, sampling algorithms such as importance and Markov Chain Monte Carlo (MCMC) methods can be applied directly to the (potentially un-normalized) posterior distribution $p(\theta|\tilde{\boldsymbol{y}},\boldsymbol{x})$ (e.g., Kuczera & Parent, 1998), which in turn requires evaluating the likelihood function $p(\tilde{\boldsymbol{y}}|\theta,\boldsymbol{x})$ for many different values of θ. These sampling approaches are standard in the calibration of time series models in hydrology (and elsewhere in science and engineering), because the likelihood functions are typically available in closed form and are (relatively) easy to evaluate (e.g., Evin et al., 2014; Kuczera & Parent, 1998; Reichert & Schuwirth, 2012; Renard et al., 2011; Smith et al., 2015; Vrugt et al., 2008). Strategy A is well suited for implementing time-domain inference of standard hydrological models, especially when the decomposition in equation (2) holds (see section 4.2.2).

Strategy B. Approaches that first sample from the joint probability distribution of $\boldsymbol{y}$ and θ,

$$p(\boldsymbol{y},\theta|\boldsymbol{x})=p(\boldsymbol{y}|\theta,\boldsymbol{x})p(\theta) \tag{6}$$

and then condition on the observations $\tilde{\boldsymbol{y}}$ using an approximation of the identity

$$p(\theta|\tilde{\boldsymbol{y}},\boldsymbol{x})=\frac{1}{p(\tilde{\boldsymbol{y}}|\boldsymbol{x})}p(\tilde{\boldsymbol{y}},\theta|\boldsymbol{x})=C\times\int p(\boldsymbol{y},\theta|\boldsymbol{x})\ \mathcal{D}(\boldsymbol{y}-\tilde{\boldsymbol{y}})\,\mathrm{d}\boldsymbol{y} \tag{7}$$

where $C=1/p(\tilde{\boldsymbol{y}}|\boldsymbol{x})$ is a constant independent of θ, and $\mathcal{D}(\boldsymbol{y}-\tilde{\boldsymbol{y}})$ is the Dirac delta function at $\boldsymbol{y}=\tilde{\boldsymbol{y}}$ (e.g., Albert et al., 2015; Csilléry et al., 2010; Diggle & Gratton, 1984).

The distinctive feature of Strategy B is that the likelihood function $p(\tilde{\boldsymbol{y}}|\theta,\boldsymbol{x})$ is never *evaluated*. Instead we *sample* from the probability model $\boldsymbol{Y}(\theta,\boldsymbol{x})$. This feature is well suited to the challenges arising when attempting to calibrate $\boldsymbol{Q}(\theta,\boldsymbol{x})$ to streamflow signatures, as noted at the end of section 2.2.

2.4. Approximate Bayesian Computation (ABC)

Sampling algorithms that follow Strategy B are known collectively as "Approximate Bayesian Computation," or ABC (Csilléry et al., 2010; Turner & Van Zandt, 2012). Although this name is indicative of numerical approximations made when applying equation (7), it is also somewhat misleading because many sampling algorithms routinely used to implement Strategy A, such as MCMC, are themselves approximate.

The archetypical ABC algorithm is an acceptance-rejection algorithm based on equations (6)–(7) as follows:

1. Draw a sample $\boldsymbol{\theta}^{(i)}$ from the prior $p(\boldsymbol{\theta})$;
2. Draw a sample from the joint distribution $p(\boldsymbol{y}, \boldsymbol{\theta}|\boldsymbol{x})$, by sampling the probabilistic model $\boldsymbol{Y}(\boldsymbol{\theta}, \boldsymbol{x})$ given the parameters $\boldsymbol{\theta}^{(i)}$ from step 1, i.e., $\boldsymbol{y}^{(i)} \leftarrow \boldsymbol{Y}(\boldsymbol{\theta}^{(i)}, \boldsymbol{x})$. For example, when undertaking signature-domain inference using the hydrological model in equation (2), this step requires running the deterministic model to compute $\boldsymbol{h}^{(i)} = \boldsymbol{h}(\boldsymbol{\theta}_h^{(i)}, \boldsymbol{x})$, drawing a sample from the error distribution, i.e., $\varepsilon^{(l)} \leftarrow \boldsymbol{\mathcal{E}}(\boldsymbol{\theta}_{\mathcal{E}}^{(i)})$, computing the streamflow time series $\boldsymbol{q}^{(i)} = \boldsymbol{h}^{(i)} + \varepsilon^{(l)}$, and, finally, computing the signatures $\boldsymbol{y}^{(i)} = \boldsymbol{g}(\boldsymbol{q}^{(i)})$;
3. Accept $\boldsymbol{\theta}^{(i)}$ if $\tilde{\boldsymbol{y}} \approx \boldsymbol{y}^{(i)}$, which can be formalized as $\rho(\tilde{\boldsymbol{y}}, \boldsymbol{y}^{(i)}) \leq \tau_\rho$, where ρ is a distance metric such as the RMSE or similar and τ_ρ is a small tolerance. In view of equation (7), this step represents an *approximate* conditioning on $\tilde{\boldsymbol{y}}$, because exact conditioning, i.e., accepting $\boldsymbol{\theta}^{(i)}$ only if $\tilde{\boldsymbol{y}} = \boldsymbol{y}^{(i)}$, is usually unachievable;
4. Repeat Steps 1–3 for $i = 1, 2, \ldots, N_{sam}$, where N_{sam} is the number of samples required.

The practical performance of an ABC algorithm depends on a number of factors relevant to our presentation.

First, the distance metric needs to be selected such that $\rho(\tilde{\boldsymbol{y}}, \boldsymbol{y}^{(i)}) = 0$ if and only if (iff) $\tilde{\boldsymbol{y}} = \boldsymbol{y}^{(i)}$. Any such distance metric will have a vanishing impact on the ABC samples as the tolerance τ_ρ is tightened (to see this, consider that in the limit as $\tau_\rho \to 0$ any distance metric that satisfies the condition above will only accept $\boldsymbol{\theta}^{(i)}$ iff $\tilde{\boldsymbol{y}} = \boldsymbol{y}^{(i)}$, and hence the choice of metric becomes immaterial). It follows that, when $\tau_\rho \approx 0$, the ABC samples will converge to $p(\boldsymbol{\theta}|\tilde{\boldsymbol{y}}, \boldsymbol{x})$ irrespective of whether the RMSE, or any other vector norm, are used as the distance metric. In other words, *provided* τ_ρ *is sufficiently tight*, the distance metric ρ can be specified "subjectively," in the sense of not requiring a statistical interpretation. Conversely, *when* τ_ρ *is loose*, the choice of ρ and value of τ_ρ will have an increasing impact on the ABC results (see section 4.3).

Second, to reduce the computational cost associated with satisfying the acceptance test, especially when $\tilde{\boldsymbol{y}}$ is high-dimensional, many ABC applications replace the criterion $\tilde{\boldsymbol{y}} \approx \boldsymbol{y}^{(i)}$ with an even more lenient criterion of the form $\boldsymbol{\omega}(\tilde{\boldsymbol{y}}) \approx \boldsymbol{\omega}(\boldsymbol{y}^{(i)})$, where $\boldsymbol{\omega}(\boldsymbol{y})$ is a vector of "summary statistics" computable from any $\boldsymbol{y}$. Mathematically, the use of summary statistics $\boldsymbol{\omega}(\boldsymbol{y})$ has the same type of effect on posteriors as the use of data signatures $\boldsymbol{g}(\boldsymbol{q})$, and in this respect data signatures can be viewed as summary statistics, and vice versa. However, conceptually, we prefer to view the choice of signatures $\boldsymbol{g}(\boldsymbol{q})$ as a modelling choice made by the hydrologist at the stage of developing their inference model (much like the choice of hydrological and error models, and calibration data sets), and the choice of summary statistics $\boldsymbol{\omega}(\boldsymbol{y})$ as a numerical choice made as part of practical ABC computation (much like the choice of the numerical tolerance τ_ρ). Ideally, summary statistics used within ABC should be as close to "sufficiency" as possible (section 5.3.1), to reduce approximation errors resulting from their use. In this work, we make no use of $\boldsymbol{\omega}(\boldsymbol{y})$.

Third, additional algorithmic enhancements are often made, such as adaptively tightening an initially loose tolerance (Albert et al., 2015; Lenormand et al., 2013), weighting the ABC samples according to the distance metric (Beaumont et al., 2002; Blum & François, 2010), using "MCMC-within-ABC" (Albert et al., 2015; Marjoram et al., 2003), using differential evolution and continuous acceptance kernels (Sadegh & Vrugt, 2014), and others, in order to reduce computational costs while achieving high numerical accuracy in approximating $p(\boldsymbol{\theta}|\tilde{\boldsymbol{y}}, \boldsymbol{x})$. These algorithmic aspects are clearly important for practical computation, but are tangential to the focus of this study on the conceptual aspects of signature-domain inference and ABC.

From a computational perspective, the art of using ABC is to choose the distance metric ρ, tolerance τ_ρ and (if necessary) summary statistics that are sufficiently lenient to make the ABC acceptance test feasible to satisfy (otherwise the algorithm converges too slowly) – but not so lax that they nullify the conditioning power of the likelihood function (otherwise the ABC estimate of the posterior becomes identical to the prior). In the present paper, we employ an ABC algorithm, named SABC, which uses Metropolis sampling from a jump distribution instead of sampling from the prior and, in the spirit of Simulated Annealing algorithms, adaptively tightens the tolerance during the course of the algorithm (Albert et al., 2015).

We caution the reader to not confuse the concept of the likelihood function $p(\tilde{\boldsymbol{y}}|\theta,\boldsymbol{x})=\mathcal{L}(\theta;\tilde{\boldsymbol{y}})$, which in Bayesian inference is determined by the probability model $\boldsymbol{Y}(\theta,\boldsymbol{x})$ of observed data $\tilde{\boldsymbol{y}}$, versus the concept of the distance metric $\rho(\tilde{\boldsymbol{y}},\boldsymbol{y})$, which is a numerical device used in conjunction with a numerical tolerance τ_ρ by ABC algorithms to approximate a given posterior $p(\theta|\tilde{\boldsymbol{y}},\boldsymbol{x})$. The specification of distance metric and numerical tolerance in ABC is no different to the specification of numerical accuracy in any other numerical approximation technique – as tight as feasible for the given computational budget, and, ideally, *well below* the likely magnitude of data and structural errors associated with the model $\boldsymbol{Y}(\theta,\boldsymbol{x})$. If the tolerance is made small enough, ABC results become independent from the choice of distance metric and the tolerance value – a behavior that cannot be achieved for (most) likelihood functions. These subtle concepts are illustrated and discussed in more detail in sections 5.1.3, 5.2.2 and 5.4 (see also section 4.1.3 of Fenicia et al., 2018).

Conceptually, ABC algorithms rely on the interesting premise that a probability model (a random variable) can be defined either by the equation of its probability density function (or probability mass function if the variable is discrete), or by its sampling algorithm. In theory, these attributes are interchangeable, because in principle once we know the pdf (or pmf) we can always design a sampling algorithm, and, conversely, if we can sample values of a random variable we can always construct an approximator of its pdf or pmf. For many probability models, including the ubiquitous choice of a deterministic model plus random noise, we can usually derive the pdfs of their outputs *and* design efficient sampling algorithms. However, it can also happen that the pdf of a stochastic model is not readily accessible—e.g., if it has no convenient closed-form expression and the modeler is unable or unwilling to invest in developing an efficient numerical approximation—but sampling from this stochastic model is (relatively) straightforward. This is precisely the case when formulating a Bayesian inference of a hydrological model in the signature domain, as described in section 2.2.

Our presentation of ABC emphasizes its generality as a class of numerical algorithms for sampling from a conditional distribution: it can be used to sample from Bayesian posteriors conditioned on any type of data $\tilde{\boldsymbol{y}}$, including streamflow time series and/or signatures. However, ABC without summary statistics $\omega(\boldsymbol{y})$ would be a computationally prohibitive choice for implementing time-domain inference because the acceptance criterion in Step 2 is near-impossible to satisfy when $\tilde{\boldsymbol{y}}$ is high-dimensional; indeed ABC is expensive even when calibrating to signatures (section 5.3 of Fenicia et al., 2018). The choice to use ABC is purely numerical – in applications where the pdf $p(\boldsymbol{g}(\boldsymbol{q})|\theta,\boldsymbol{x})$ is (relatively) easy to compute, standard MCMC sampling approaches can be used as usual. For example, if $\boldsymbol{g}(\boldsymbol{q})$ represents a log transformation, the pdfs $p(\boldsymbol{g}(\boldsymbol{q})|\theta,\boldsymbol{x})$ and $p(\boldsymbol{q}|\theta,\boldsymbol{x})$ are readily derived and evaluated, and do not require ABC to be used for inference or prediction (e.g., section 5.3.4).

3. Comments on Previous Applications of ABC in Hydrology

The richness of concepts underlying ABC, signature-domain calibration, and their relationship to general Bayesian estimation, have already attracted the attention of the hydrological community. An objective of this section is to identify and clarify some related misconceptions, to help hydrologists take better advantage of ABC algorithms and put signature-domain applications in hydrology on a more solid theoretical and computational basis.

3.1. Relationship Between ABC and GLUE

In GLUE applications, the similarity between observations and predictions of a deterministic model $\boldsymbol{y}(\theta_h,\boldsymbol{x})$ is quantified using a subjectively specified "pseudo-likelihood" function Υ and an acceptability threshold α_Υ:

$$\Upsilon\left(\tilde{\boldsymbol{y}},\mathscr{y}(\theta_h,\boldsymbol{x})\right)=\mathcal{L}_\Upsilon(\theta_h;\tilde{\boldsymbol{y}})\geq\alpha_\Upsilon \tag{8}$$

Equation (8) can be applied in both the time and signature domains.

The pseudo-likelihood function is typically used for two purposes: (i) weigh parameter samples, usually drawn from the prior, e.g., within an importance sampling algorithm; and (ii) reject parameter samples for which the pseudo-likelihood is below a user-specified acceptability threshold.

Equation (8) in the GLUE framework is at first sight analogous to Step 3 in the ABC algorithm (section 2.4): both ρ and Υ are subjective measures of similarity of observations and simulations, and both are applied with user-specified thresholds, τ_ρ and α_Υ respectively, to accept or reject parameter sets sampled from the prior. This similarity led several publications in the hydrological literature to suggest a theoretical relationship between ABC and GLUE (e.g., Nott et al., 2012; Sadegh & Vrugt, 2013).

However, ABC algorithms differ from GLUE applications in several important aspects:

1. In ABC, the distance metric in the acceptance criterion is applied to realizations from $\boldsymbol{Y}(\theta, \boldsymbol{x})$, i.e., to realizations from a stochastic model including all represented sources of uncertainty. Conversely, in GLUE, the pseudo-likelihood function is computed using the deterministic model $\mathscr{y}(\boldsymbol{\theta}_h, \boldsymbol{x})$. This is a fundamental difference because $\mathscr{y}(\boldsymbol{\theta}_h, \boldsymbol{x})$ has no stochastic terms;
2. In ABC, the tolerance τ_ρ is intended to be as tight as possible, to accurately approximate equation (7) and to ensure the ABC results are controlled by the probability model $\boldsymbol{Y}(\boldsymbol{\theta}, \boldsymbol{x})$ rather than by the (subjective) choice of ρ (see section 2.4). Conversely, in GLUE, the threshold value α_Υ is typically set to large values, in accordance with the modeler's belief in the quality of the data and model; e.g., model simulations with Nash-Sutcliffe Efficiency (NSE) above 0.6 are accepted in Freer et al. (2004). In this case, the choice of pseudo-likelihood function Υ and threshold α_Υ will determine the shape and width of parameter posteriors and streamflow predictive distributions;
3. In ABC (and Bayesian applications in general), the total predictive distribution of streamflow is given by the probability model $\boldsymbol{Y}(\boldsymbol{\theta}, \boldsymbol{x})$ in combination with the propagation of posterior uncertainty in $\boldsymbol{\theta}$. In the simplest case of equation (2), total predictive uncertainty will generally be dominated by the residual error term $\mathcal{E}$. Conversely, in GLUE, predictive distributions are (virtually) always generated using parametric uncertainty alone (Freer et al., 1996; Westerberg et al., 2011); indeed, given a typical pseudo-likelihood function such as the NSE, it is not even clear what the corresponding distribution of residual errors $\mathcal{E}$ is, let alone how to scrutinize its assumptions or sample from it to obtain total predictive limits.

More generally, we emphasize that ABC is a family of numerical algorithms for sampling from conditional distributions such as Bayesian posteriors (section 2.4), whereas GLUE is a distinct inference/prediction framework, with some similarities to the Bayesian paradigm and some major differences, as detailed (and debated) elsewhere (e.g., Beven, 2006; Beven et al., 2008; De Finetti, 1972; Mantovan & Todini, 2006; Reichert et al., 2015; Stedinger et al., 2008). *Implementations* of Bayesian methods and GLUE – whether using MCMC, ABC, importance sampling, or any other sampling technique – are *algorithmically* similar because they both seek to draw parameter samples from an (often un-normalized) product of functions of the form $\mathcal{L}(\boldsymbol{\theta}; \tilde{\boldsymbol{y}})p(\boldsymbol{\theta})$, where $p(\boldsymbol{\theta})$ is a prior and $\mathcal{L}(\boldsymbol{\theta}; \tilde{\boldsymbol{y}})$ is a function that quantifies differences between observations and predictions (the GLUE threshold effectively truncates the "tails" of the likelihood function). The role of the pseudo-likelihood function in GLUE is closer to the role of the likelihood function in Bayesian inference, not of the distance metric in an ABC algorithm. The crucial difference between Bayesian methods and GLUE is that the latter does not require $\mathcal{L}(\boldsymbol{\theta}; \tilde{\boldsymbol{y}})$ to be formulated using probability theory, in which case the inference and prediction cannot be expected to provide a probabilistic description of uncertainty. For example, a GLUE modeler may choose to formulate $\mathcal{L}(\boldsymbol{\theta}; \tilde{\boldsymbol{y}})$ using fuzzy set theory (Freer et al., 2004), in which case its inference and prediction can be expected to provide a fuzzy set description of uncertainty.

A discussion of merits and limitations of different paradigms for uncertainty quantification is important and fascinating in its own right, but is well beyond our scope here. We refer the reader to debates such as Mantovan and Todini (2006), Beven et al. (2008), Stedinger et al. (2008), as well as to general references such as *De* Finetti (1972), Ang and Tang (2007), Reichert et al. (2015), and others. Our intention in this work is limited to highlighting the conceptual and algorithmic differences between ABC and GLUE, in order to reduce the potential for confusion in the hydrological modeling community.

3.2. A Comment on Previous Applications of ABC in Hydrology

In this section we argue that previous attempts to use ABC in hydrology have not exploited the full potential of the ABC approach, and in some cases have created confusion regarding what ABC is, and what it can achieve.

First, most previous applications of ABC in hydrology have not specified probabilistic hydrological models $\boldsymbol{Q}(\boldsymbol{\theta}, \boldsymbol{x})$ and instead used deterministic models $\boldsymbol{h}(\boldsymbol{\theta}_h, \boldsymbol{x})$ in the acceptance test of ABC algorithms (e.g.,

Nott et al., 2012; Sadegh & Vrugt, 2013; Vrugt & Sadegh, 2013). The resulting algorithm is much closer to GLUE than to ABC, unless the modeler is explicitly working on the assumption that $\boldsymbol{Q}(\theta,\boldsymbol{x})=\boldsymbol{h}(\theta_h,\boldsymbol{x})$ and hence $\boldsymbol{Y}(\theta,\boldsymbol{x})=_{\mathscr{Y}}(\theta_h,\boldsymbol{x})$, i.e., that the observed data originates from a deterministic model and there are no data or model errors. The consequences of using a deterministic model in the ABC acceptance test will be illustrated and interpreted in the empirical case study (sections 4 and 5.1).

Second, previous ABC applications in hydrology often used quite coarse tolerances τ_ρ. For example, Sadegh and Vrugt (2013) define the tolerance using a "limits of acceptability" approach, on the basis of the assumed "effective" observation error in the streamflow data (defined as the combined effect of input/output data and model structural errors). This specification mimics the way GLUE is used in "limits of acceptability" applications (e.g., Westerberg et al., 2011), but contradicts the fundamental premise of ABC, which is to provide a numerical approximation to a given posterior using a numerical tolerance τ_ρ that is as tight as possible (e.g., Toni et al., 2009). As the tolerance is coarsened, ABC results will be increasingly controlled by the selection of the distance metric and the value of the tolerance, rather than by the specification of the probabilistic model $\boldsymbol{Q}$. The consequences of using a coarse tolerance within ABC sampling is demonstrated through the empirical case study in section 4; the usage of the ABC tolerance is discussed further in section 5.2.2.

More generally, the practice of using $\boldsymbol{h}$ instead of $\boldsymbol{Q}$, which appears to sidestep the need to specify a residual error model let alone a more complex stochastic model of the data, led to the misconception that ABC approaches are statistical methods for "cases when an explicit likelihood (objective) function cannot be justified" (Vrugt & Sadegh, 2013). This mis-understanding may have originated because, as noted in section 5.4, ABC methods are often referred to as "likelihood-free methods" in the statistical literature (e.g., Marjoram et al., 2003). However, ABC is "likelihood-free" only in the sense that it spares the modeler from having to *evaluate (compute)* the likelihood function, i.e., from having to derive and compute the pdf of $\boldsymbol{Q}$. Instead, ABC requires the modeler to *sample from* $\boldsymbol{Q}$, i.e., the modeler still has to specify – and justify! - the probabilistic model $\boldsymbol{Q}$, which by definition has a pdf and hence a likelihood function. We re-iterate that a modeler seeking to make probabilistic inference and prediction cannot escape the need to develop a probability model of the system, and to check its assumptions against available evidence.

4. Empirical Case Study

4.1. Case Study Objectives

The empirical case study in this paper has the following objectives:

1. Demonstrate the general viability of signature-domain inference of a probabilistic model $\boldsymbol{Q}(\theta,\boldsymbol{x})$ using an ABC algorithm, and its ability to generate predictive distributions of streamflow time series;
2. Demonstrate the impact of using a deterministic model $\boldsymbol{h}(\theta_h,\boldsymbol{x})$ in lieu of a probabilistic model $\boldsymbol{Q}(\theta,\boldsymbol{x})$ in the ABC acceptance test, while aiming for a tight tolerance τ_ρ as expected in standard ABC applications;
3. Demonstrate the impact of using $\boldsymbol{h}$ in lieu of $\boldsymbol{Q}$, while setting a coarse ABC tolerance τ_ρ. In this case the inference closely resembles an application of GLUE.

These objectives serve the general aims of our paper to illustrate Bayesian signature-domain inference using ABC, and demonstrate the impact of using ABC algorithms as reported in the hydrological literature (see section 3.2). The case study represents an extension of the real data case study reported in (Fenicia et al., 2018), where the properties of Bayesian signature-domain inference are thoroughly investigated and compared to Bayesian time-domain inference under a variety of scenarios.

4.2. Case Study Material and Methods

4.2.1. Study Area and Hydrological Model

The study area and models used for the case study are described in Fenicia et al. (2018); here we provide a succinct description of essential details. The study area is the Lacmalac catchment in South-East Australia (Australian Bureau of Meteorology, http://www.bom.gov.au/water/hrs/, Gauge 410057). The hydrological model HyMod (*Boyle*, 2001) is used to generate the deterministic streamflow predictions $\boldsymbol{h}$ in equation (2). The period of 1 March 1996 to 1 March 2000 is used for calibration, and the period of 1 March 2002 to 1 March 2006 is used for validation; each period is preceded by a 1 year warmup period. Note that this selection of analysis periods excludes periods where inter-annual variability in streamflow characteristics appears too strong to be captured using the HyMod hydrological model (see section 3.1 in Fenicia et al., 2018).

The probabilistic hydrological model $\boldsymbol{Q}(\theta,\boldsymbol{x})$ is defined as additive in transformed space,

$$\boldsymbol{Q}(\theta,\boldsymbol{x})=z^{-1}[z[\boldsymbol{h}(\theta_h,\boldsymbol{x});\lambda]+\boldsymbol{\mathcal{E}}(\theta_{\mathcal{E}});\lambda] \quad (9)$$

where $z[q;\lambda]=(q^{\lambda}-1)/\lambda$ is the Box-Cox transformation (here, used with fixed $\lambda=0.4$).

The random error term $\boldsymbol{\mathcal{E}}(\theta_{\mathcal{E}})$, which represents the residual errors of the model $\boldsymbol{h}(\theta_h,\boldsymbol{x})$, is described by an AR(1) process, $\mathcal{E}_t=\phi\varepsilon_{t-1}+W_t$, with truncated Gaussian innovations, $W_t\sim\mathcal{TN}(0,\sigma_W,L_{W,t})$. The lower bound $L_{W,t}$ is defined such that $Q_t>0$, the autoregressive parameter ϕ is fixed to 0.8, and the standard deviation of innovations σ_W is treated as unknown. These settings are based on the findings of McInerney et al. (2017).

4.2.2. Calibration Data, Signatures and Inference Methods

Signature-domain inference is set to calibrate jointly to the following observed streamflow signatures $\tilde{\boldsymbol{y}}=\boldsymbol{g}(\tilde{\boldsymbol{q}})$, selected based on the explorations in Fenicia et al. (2018):

i. Flow Duration Curve, represented by the streamflow values corresponding to the 10%, 50%, 75% and 95% percentiles of the cumulative distribution of streamflow;
ii. Flashiness index, $g_f[\boldsymbol{q}_{1:N_T}]=\sum_{t=2}^{N_T}|q_t-q_{t-1}|/\sum_{t=2}^{N_T}q_t$, which provides a measure of the "responsiveness" of a catchment (Baker et al., 2004).

For this choice of calibration data $\tilde{\boldsymbol{y}}$, the data model is given by $\boldsymbol{Y}(\theta,\boldsymbol{x})=\boldsymbol{g}(\boldsymbol{Q}(\theta,\boldsymbol{x}))$. In this case, the pdf of $\boldsymbol{Y}(\theta,\boldsymbol{x})$ is difficult or impossible to derive in closed form (see end of section 2.2), and hence the likelihood function and posterior are not available in closed form. For this reason, we employ an ABC approach to sample from the posterior $p(\theta|\boldsymbol{g}(\tilde{\boldsymbol{q}}),\boldsymbol{x})$, by first sampling $\boldsymbol{q}^{(i)}\leftarrow\boldsymbol{Q}(\theta,\boldsymbol{x})$ using equation (9) and a draw from $\mathcal{E}(\theta_{\mathcal{E}})$, and then computing $\boldsymbol{y}^{(i)}=\boldsymbol{g}(\boldsymbol{q}^{(i)})$. The Simulated Annealing ABC (SABC) algorithm of Albert et al. (2015) is used, with the distance metric being specified as the largest (by magnitude) relative error in the individual FDC quantiles and the flashiness index (section 2.4.2 of Fenicia et al., 2018).

As a basic check of the signature-domain inference, we also report the results of time-domain inference using observed streamflow time series, $\tilde{\boldsymbol{y}}=\tilde{\boldsymbol{q}}$. In this case, the pdf of the probability model $\boldsymbol{Y}(\theta,\boldsymbol{x})=\boldsymbol{Q}(\theta,\boldsymbol{x})$, and hence the likelihood function $\mathcal{L}(\theta;\tilde{\boldsymbol{y}})$, have convenient closed-form expressions (Bates & Campbell, 2001; McInerney et al., 2017). The posterior $p(\theta|\tilde{\boldsymbol{q}},\boldsymbol{x})$ is then also known in closed form, and can be sampled directly using a standard MCMC algorithm (see section 2.5 of Fenicia et al., 2018).

The set of signatures above is not expected to be sufficient for HyMod (or any similar rainfall-runoff model) and the selected residual error model. Therefore we expect at least some loss of information when switching from time- to signature- domain inference (see Fenicia et al., 2018, for a detailed analysis).

4.2.3. Inference Setups

The following inference setups are compared:

1. Signature-domain calibration of $\boldsymbol{Q}(\theta,\boldsymbol{x})$ using SABC with $\tau_\rho\rightarrow 0$. This setup represents our "theoretically" recommended usage of ABC, as per the original statistical literature (Albert et al., 2015; Toni et al., 2009). The notation $\tau_\rho\rightarrow 0$ is intended to represent the SABC algorithm driving τ_ρ as tight as possible. SABC convergence was assessed by testing the sensitivity of results to the choice of ABC distance metric: we found that replacing the max function with the arithmetic average had virtually no impact on the final estimates of predictive distributions of streamflow (see section 2.4.2 in Fenicia et al., 2018);
2. Signature-domain calibration of $\boldsymbol{h}(\theta_h,\boldsymbol{x})$ using SABC with $\tau_\rho\rightarrow 0$. This setup is included to examine the effect of omitting random terms from the hydrological model while still attempting to make the tolerance as tight as possible (as expected in an ABC application);
3. Signature-domain calibration of $\boldsymbol{h}(\theta_h,\boldsymbol{x})$ using ABC with $\tau_\rho=0.15$. This setup is included to examine the effect of omitting random terms from the hydrological model and using a (relatively) coarse tolerance. It is similar to a GLUE application, but with equal weighting of parameter samples that satisfy the acceptability threshold (roughly speaking, corresponding to a rectangular pseudo-likelihood function);
4. Signature-domain calibration of $\boldsymbol{h}(\theta_h,\boldsymbol{x})$ using ABC with $\tau_\rho=0.2$. Same as above, but coarser tolerance;
5. Signature-domain calibration of $\boldsymbol{h}(\theta_h,\boldsymbol{x})$ using ABC with $\tau_\rho=0.4$. Same as above, even coarser tolerance;
6. Time-domain calibration of $\boldsymbol{Q}(\theta,\boldsymbol{x})$ using MCMC applied to the closed-form likelihood function. This setup is included as a representation of "traditional" inference, and provides a basic check of the setups above.

All inference setups calibrate the hydrological parameters θ_h. In addition, when calibrating $\boldsymbol{Q}(\theta,\boldsymbol{x})$, in both the signature and time domains, we also infer the residual error variance σ_W^2 within equation (9). In setups

1–2, 5,000 parameter samples are collected after 2×10^6 SABC iterations; in setups 3–5, basic ABC sampling proceeds until 5,000 parameter samples are collected.

The predictive distributions are constructed as follows. When predicting using $\boldsymbol{Q}(\theta, \boldsymbol{x})$, as in inference setups 1 and 6, the predictive distribution is constructed by propagating parametric and residual uncertainty. When predicting using $\boldsymbol{h}(\theta_h, \boldsymbol{x})$, as in inference setups 2–5, the predictive distribution is constructed solely by propagating parametric uncertainty, as $\boldsymbol{h}$ itself has no random terms.

The inference setups are compared as follows: (i) plots of predicted versus observed streamflow time series in the validation period; (ii) plots of predicted versus observed signatures (FDC and flashiness index) in the validation period, and (iii) box-plots of the distributions of ABC distance metric values of the sampled parameter sets (reported for the ABC-implemented inference setups 1–5 only).

Note that our demonstration is intended to illustrate the fundamental difference between the use of probabilistic versus deterministic models within an ABC implementation of signature-domain inference. In the context of this aim, we utilize a standard "off-the-shelf" hydrological and residual error model. The case study is not intended to achieve the best predictions possible for this particular catchment, which might require tailoring the hydrological model structure (e.g., Fenicia et al., 2016), in-depth analysis of data and structural errors (e.g., McInerney et al., 2017; Renard et al., 2011), and diagnosis/mitigation of potential hydrological non-stationarity (e.g., Westra et al., 2014).

4.3. Case Study Results

4.3.1. Streamflow Predictions

Figure 1 shows the streamflow predictive distributions estimated using the inference setups detailed in section 4.2.3. The inset zooms in on a storm event within the subset of the period.

Signature-domain inference of $\boldsymbol{Q}(\theta, \boldsymbol{x})$, shown in row 1, generally captures the observed streamflow, with the 95% prediction limits enveloping the majority of the observations. The predictive distribution is reasonably centered on the observed data in the first half of the simulation period, though there appears to be a mild pattern of over-prediction in the second half of the period (observations close to lower bound).

The use of the deterministic model $\boldsymbol{h}(\theta_h, \boldsymbol{x})$ while seeking a tight ABC tolerance τ_ρ, shown in row 2, produces clearly unacceptable results in terms of uncertainty quantification. The deterministic prediction is generally accurate in the sense of tracking the observed hydrographs with an NSE of 0.75, but the predictive distributions are extremely narrow, essentially collapsing to a single line.

The use of the deterministic model $\boldsymbol{h}(\theta_h, \boldsymbol{x})$ while relaxing the ABC tolerance τ_ρ produces a markedly different behaviour: the predictive limits "inflate" around the deterministic predictions depending on the value of τ_ρ, as shown in rows 3–5. When $\tau_\rho = 0.15$, the predictions are still much too tight to capture the observed streamflow. In contrast, when $\tau_\rho = 0.4$, the predictions appear too wide, especially for the storm peaks (e.g., for the largest storm event, the upper 95% prediction limit is almost 3 times higher than the observed flow). The setting $\tau_\rho = 0.2$ appears to provide a reasonable compromise, generally capturing the shape of recessions without excessively compromising precision during peak flows, as seen in row 4. However, this predictive distribution is clearly different from the predictive distribution obtained using $\boldsymbol{Q}(\theta, \boldsymbol{x})$ in row 1: there is a tradeoff between predictive precision during recession periods versus during storm events, and the over-prediction appears more pronounced in row 4 than in row 1.

Finally, it can be seen that the signature-domain and time-domain inferences of $\boldsymbol{Q}(\theta, \boldsymbol{x})$ are generally consistent with each other. The shape of the predictive distributions is qualitatively similar, though the signature-based predictions are somewhat less precise (wider prediction limits) and the dynamics are smoother (predicted hydrographs less peaky than the observed ones). Here, we refer the reader to Fenicia et al. (2018) for an in-depth comparison of signature- versus time- domain inferences.

4.3.2. Signature Predictions

Figure 2 shows the predictive distributions of the FDC and flashiness index signatures corresponding to the streamflow time series shown in Figure 1 above. The signature-domain calibration of $\boldsymbol{Q}(\theta, \boldsymbol{x})$ achieves a generally good capture of the signatures, with the predictive distributions centered on the observed values. The signature-domain calibration of $\boldsymbol{h}(\theta_h, \boldsymbol{x})$ using SABC ($\tau_\rho \to 0$) is clearly over-confident in predicting the

signatures. When the signature-domain calibration of $\boldsymbol{h}(\boldsymbol{\theta}_h, \boldsymbol{x})$ is implemented using ABC with a coarse τ_ρ, the width of the predictive limits once again becomes controlled by the value of τ_ρ. For example, when τ_ρ=0.1, the predictions resemble those obtained from signature-domain inference of $\boldsymbol{Q}(\boldsymbol{\theta}, \boldsymbol{x})$ using SABC, whereas when τ_ρ=0.4 the predictive uncertainty is very wide in both the FDC and the flashiness index. Finally, in comparison to the time-domain calibration, signature-domain calibration of $\boldsymbol{Q}(\boldsymbol{\theta}, \boldsymbol{x})$ is slightly less precise, though in terms of the FDC predictions the predictive limits appear better centered.

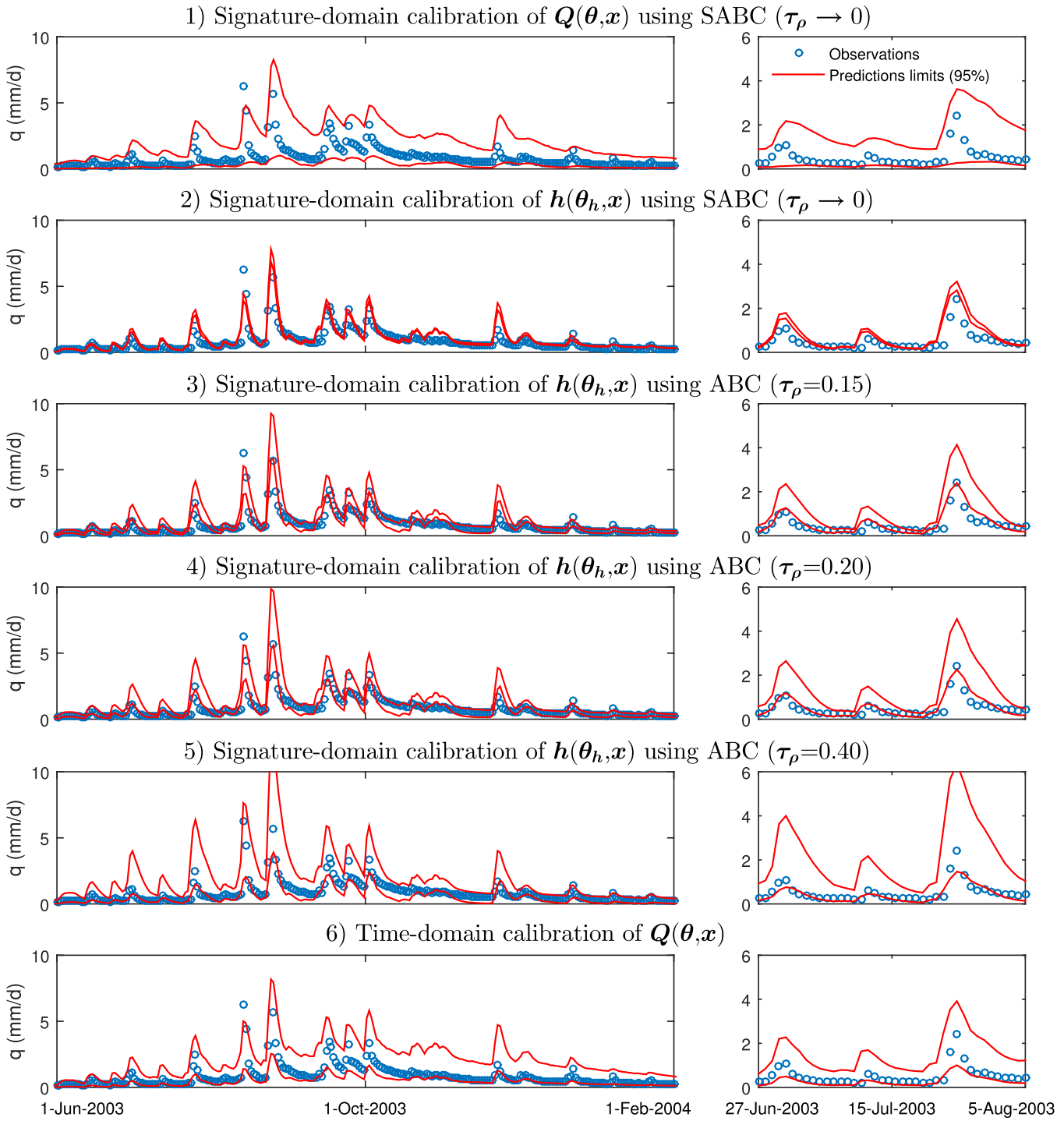

Figure 1. Predictions of streamflow time series over the validation period, obtained using signature-domain inference (multiple setups) and time-domain inference.

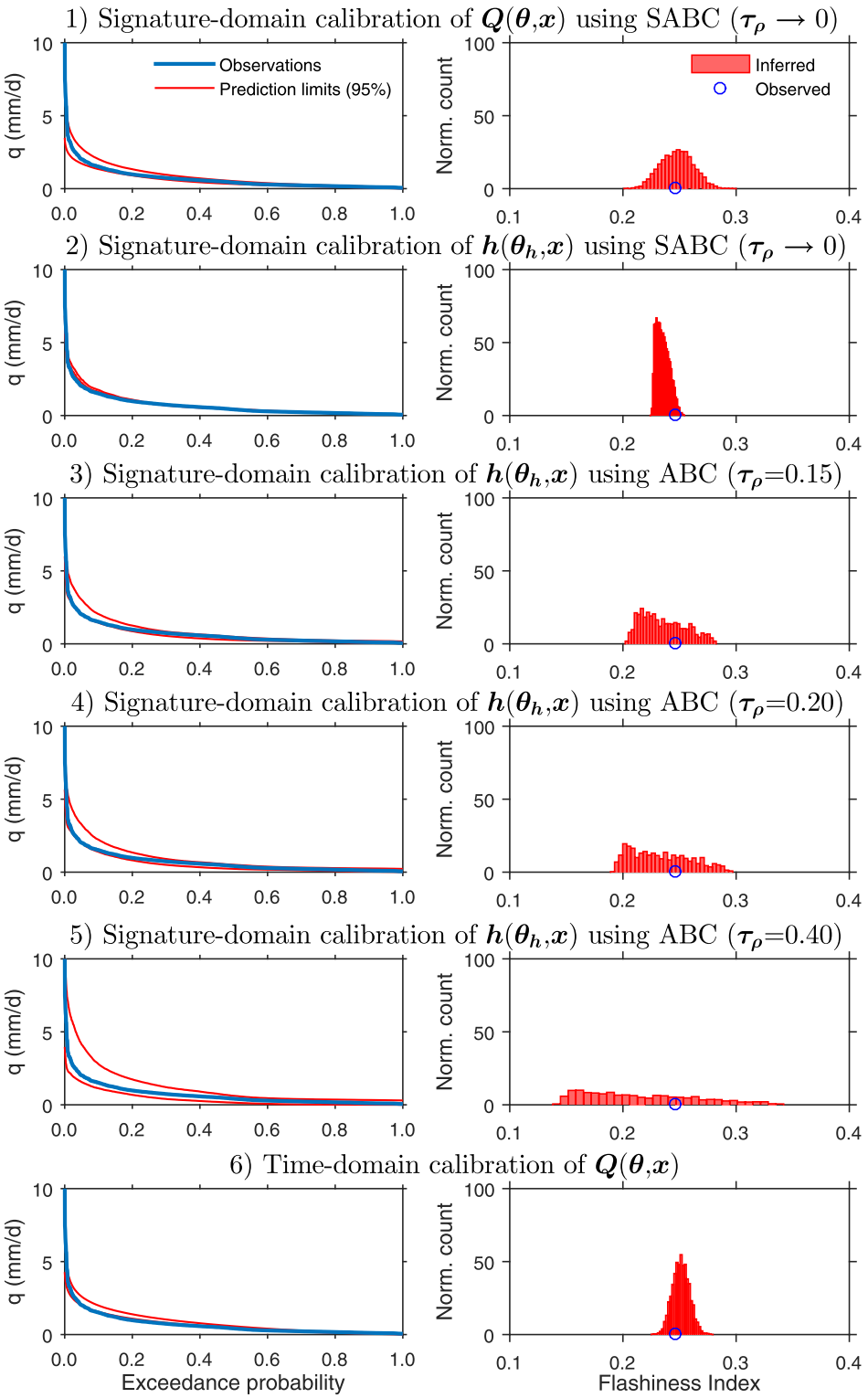

Figure 2. Predictions of streamflow signatures (Flow Duration curve and flashiness index) over the validation period, obtained using signature-domain inference (multiple setups) and time-domain inference.

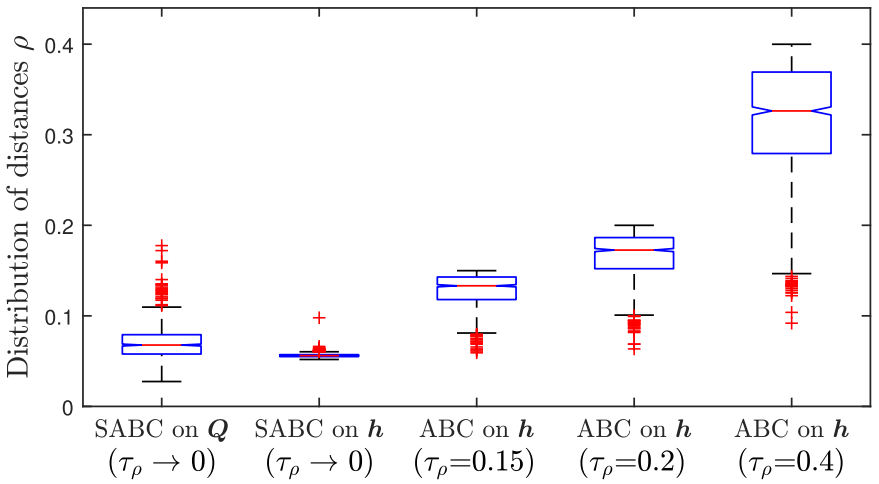

Figure 3. Distribution of distances ρ associated with the ABC parameter samples generated during the signature-domain inference of the probabilistic model $\boldsymbol{Q}(\theta, \boldsymbol{x})$ using SABC, and the signature-domain inference of the deterministic model $\boldsymbol{h}(\theta_h, \boldsymbol{x})$ using SABC and ABC with a range of tolerances τ_ρ.

4.3.3. Distribution of Distance Metrics of ABC Samples

Figure 3 reports the distribution of distances ρ associated with the ABC samples from inference setups 1–5. Recall that, for a given ABC sample of parameters and streamflow time series, the value of ρ reflects the largest relative error across all signatures. Two findings are of relevance.

First, the SABC algorithm applied to the probabilistic model $\boldsymbol{Q}(\theta, \boldsymbol{x})$ ("SABC-on-*Q*") achieves appreciably lower values of the distance metric (i.e., closer match of simulated and observed signatures) than the ABC algorithm with tolerances τ_ρ=0.1−0.4 applied to the deterministic model $\boldsymbol{h}(\theta_h, \boldsymbol{x})$ ("ABC-on-*h*"). The distance distribution of the SABC-on-*Q* samples has a median of about 0.07 (7%), whereas the distances of the ABC-on-*h* samples have medians of approximately 0.13, 0.17 and 0.33 (just below the tolerances of 0.15, 0.2 and 0.4 respectively). In addition, the distance distributions have opposite skewness patterns: the SABC-on-*Q* distribution has a thick *lower* tail (lower distances more frequent), whereas the ABC-on-*h* distributions have thick *upper* tails (larger distances more frequent).

Second, SABC applied to the deterministic model ("SABC-on-*h*") achieves a median distance of around 0.06, which is slightly lower than the median distance of 0.07 achieved by SABC-on-*Q*. However, the spread of the distributions is different: SABC-on-*h* distances fall in the narrow range from 0.07 to 0.05 (excluding the outlier at 0.1), whereas the SABC-on-*Q* distances range from as high as 0.18 to as low as 0.02.

5. Discussion

5.1. Interpretation of Case Study Results

5.1.1. General Viability of Bayesian Signature-Domain Inference Using ABC

The case study results in Figure 1 demonstrate the viability of Bayesian signature-domain calibration of a probabilistic model $\boldsymbol{Q}(\theta, \boldsymbol{x})$. Even though the calibration is carried out entirely in the signature domain, the predictive distributions of streamflow time series are generally consistent with the observed data, even in an independent validation period. Figure 1 also demonstrates some loss of precision when going from time-domain to signature-domain calibration, as expected because the FDC and flashiness index are rather unlikely to represent the complete set of sufficient signatures for the HyMod model. The correspondence between time-domain and signature-domain inference of $\boldsymbol{Q}(\theta, \boldsymbol{x})$, including the effects of differences in the information content of the streamflow signatures versus the streamflow time series, is analyzed in Fenicia et al. (2018), which provides detailed insights into the influence of the number and type of signatures, the impact of substantial deficiencies in the model $\boldsymbol{Q}(\theta, \boldsymbol{x})$, and so forth.

Figure 2 row 1 demonstrates the general ability of signature-domain calibration to reproduce the signatures of the observed data (once again, including in a validation period). This is not surprising because the ABC

algorithm is specifically constructed to only accept parameter samples that can generate random realizations close to the observed signatures. For example, Figure 3 shows that, in this study, the SABC algorithm applied to $\boldsymbol{Q}(\boldsymbol{\theta}, \boldsymbol{x})$ (labeled "SABC-on-*Q*") is able to drive the distances to a median of 0.07, with some distances as low as 0.02. As per section 4.2.2, these distance values correspond to the largest relative errors in the signatures across all SABC samples having a median of 7% and a minimum of 2%, respectively. As noted in section 4.2.3, these distance values are small enough to make the SABC inference and prediction virtually insensitive to the choice of distance metric (see also section 4.1.4 in Fenicia et al., 2018).

That said, Figure 3 also demonstrates the difficulty encountered by any ABC algorithm (including SABC) in achieving tolerances (distances values) that are small in an absolute sense, with some SABC-on-*Q* distances as high as 0.18 (i.e., some samples incurred a relative error of 18% in matching at least one of the signatures). A separate analysis revealed that in this study the distance values are generally dominated by mismatches in the low quantiles of FDC curves, where small absolute errors can translate into large relative errors; the errors in the other signatures are often substantially lower. Two additional points can be made:

a. The ABC distance metric is a *random* function of $\boldsymbol{\theta}$: even if the same value of $\boldsymbol{\theta}$ is drawn, different realizations from $\boldsymbol{Q}$ (in our case, different realizations of residual errors) will alter the value of the distance metric. Finding near-optimal regions of a random function is algorithmically challenging – this is the likely reason for outliers with large distance values in the SABC-on-*Q* results in Figure 3. Censoring these samples is an option, though not attempted here because SABC results appear already converged (see above).
b. In principle, $\boldsymbol{Q}$ can match the signatures (and the streamflow time series themselves) closer than $\boldsymbol{h}$ on its own – all we need is a realization of residual errors that cancels a portion of the data and model errors. In fact, under the Gaussian residual error model, for any parameter set $\boldsymbol{\theta}$, there is always a realization that yields a perfect match of model and observations. This ability of $\boldsymbol{Q}(\boldsymbol{\theta}, \boldsymbol{x})$ is no magic coincidence – the very role of the residual error model is to explain mismatches between simulated and observed data! Finding these "perfect" realizations through random sampling is unlikely (mildly put), yet finding realizations that are closer and closer to this ideal is, in theory, merely a matter of running SABC long enough. We expect that, if SABC-on-*Q* were run *indefinitely*, the distances would continue decreasing until they hit zero (but this would take a very long time, far beyond any practical computational time frame).

The performance of signature-domain inference is subject to the same limitations as time-domain inference. For example, the ability to predict streamflow and signatures in periods other than calibration is contingent on the probability model (hydrological and error model) being able to represent any potential nonstationarities, such as changed catchment conditions, changes in data quality, and so forth. Dealing with these effects is beyond the scope of this demonstration, but clearly presents a major challenge for hydrological and broader environmental sciences (e.g., Montanari et al., 2013; Sadegh et al., 2015; Westra et al., 2014).

The analysis above just scratches the surface of signature-domain inference of probabilistic models using ABC techniques. For a detailed analysis of the properties of signature-domain inference, including its potential ability to mitigate against deficiencies in the probability model of the data, we refer the reader to Fenicia et al. (2018). Our focus now shifts to interpreting the impact of departures from standard usage of ABC.

5.1.2. Omission of Random Terms From the Hydrological Model

Figures 1 row 2 vividly demonstrate that using the deterministic model $\boldsymbol{h}(\boldsymbol{\theta}_h, \boldsymbol{x})$ when computing the ABC distance metric, while using SABC to reach a tight tolerance τ_ρ, fails to capture any uncertainty at all, with the predictive limits essentially collapsing to a line. Although the streamflow predictions might be acceptable according to deterministic performance metrics (here, an NSE of 0.75 is achieved), the lack of uncertainty quantification is a major limitation from the perspective of this work.

The collapse of this inference setup to point predictions is hardly surprising because $\boldsymbol{h}(\boldsymbol{\theta}_h, \boldsymbol{x})$ is purely deterministic. Consider that the starting point of the ABC derivation is the posterior in equation (5), which itself is based on the assumption in equation (1) that the model $\boldsymbol{Y}(\boldsymbol{\theta}, \boldsymbol{x})$ provides a suitable probabilistic description of observed data $\tilde{\boldsymbol{y}}$. In other words, $\boldsymbol{Y}(\boldsymbol{\theta}, \boldsymbol{x})$ is a model that could have generated $\tilde{\boldsymbol{y}}$. Using the deterministic streamflow model $\boldsymbol{h}(\boldsymbol{\theta}, \boldsymbol{x})$ in the distance metric of an ABC algorithm corresponds to the assumption that $\boldsymbol{h}(\boldsymbol{\theta}, \boldsymbol{x})$ provides an exact representation of observed streamflow data $\tilde{\boldsymbol{q}}$, i.e., that there is no data or model uncertainty at all! This correspondence is noted by Nott et al. (2012), and it hardly represents a useful

assumption given the typical uncertainties affecting hydrological modeling; in fact it represents a step backward compared to even the simplest additive Gaussian error model in equation (2).

Interestingly, Figure 2 row 2 shows that this setup can reproduce observed signatures to a high precision. Indeed Figure 3 shows that SABC applied to $\boldsymbol{h}(\boldsymbol{\theta}_h, \boldsymbol{x})$ (labeled "SABC-on-*h*") can find parameters for which the (deterministic) model matches the signatures to within 5–10% relative accuracy, and from Figure 2 row 2 we see that this precision generally holds in the selected validation period. Yet Figure 1 row 2 shows that this precision proves illusory when predicting streamflow itself, and leads to much overconfident predictions. It is intuitive that a good model fit in the signature domain is not a reliable indicator of the predictive ability of the model in the streamflow time series domain, due to the non-uniqueness of the signature inverse (the mapping from a signature or set of signatures to streamflow time series). In addition, as noted in section 5.1.1, predictive performance under independent conditions may be challenging in the presence of non-stationarity, especially when the model predictions are so tight.

Figure 3 raises an interesting question: why does SABC appear able to achieve tighter distances when applied to a deterministic model rather than to a probabilistic model? Doesn't this finding contradict our preference for tight values of τ_ρ expressed earlier? Consider the following numerical and statistical insights. First, when ABC is applied to $\boldsymbol{h}(\boldsymbol{\theta}_h, \boldsymbol{x})$, the distance metric ρ becomes a *deterministic* function of the model parameters $\boldsymbol{\theta}_h$ – making it much easier to reduce than the random distance metric arising when ABC is applied to $\boldsymbol{Q}(\boldsymbol{\theta}, \boldsymbol{x})$ (section 5.1.1). This is the likely reason why, in Figure 3, the SABC-on-*h* distance values have less spread (fewer outliers) and, on average, are slightly closer to zero than the SABC-on-*Q* distances. Yet this is only a part of the picture! Second, since $\boldsymbol{h}$ is a deterministic model, it has no random terms that could bridge the gap between the simulated and observed responses. Hence, unlike the SABC-on-*Q* setup, there may be a hard limit on how close can SABC-on-*h* get to the observed data (here, the signatures). This is the likely reason why none of the SABC-on-*h* samples reach distance values as low as the lower tails of the SABC-on-*Q* distance distribution. In fact, direct numerical optimization of ρ with respect to $\boldsymbol{\theta}_h$ would find the minimal distance and corresponding parameter values much faster than SABC. Third, and most importantly from a statistical perspective, the benefits of tight signatures distances achieved by SABC-on-*h* are illusory because of grossly under-estimated uncertainty in the streamflow space (Figure 1). This behavior accords with probability theory: a deterministic model on its own has no terms to describe predictive uncertainty (cf section 3.2).

The analyses above and in the earlier section 5.1.1 help understand the theory and empirical behavior of the ABC approach, and help illustrate the concepts presented in sections 2 and 3.2.

5.1.3. Omission of Random Terms From the Hydrological Model and Use of Coarse ABC Tolerance

The marked over-confidence that arises when the deterministic model $\boldsymbol{h}(\boldsymbol{\theta}_h, \boldsymbol{x})$ is used in the ABC distance metric can be counteracted by relaxing the ABC tolerance τ_ρ. As seen from Figure 2 rows 4–5, a modeler willing to experiment with different values of τ_ρ, while monitoring the prediction limits in the *streamflow* space, might be able to "calibrate" τ_ρ to achieve a desired coverage of observed data (provided the model parameterization is sufficiently flexible). This is possible because coarsening the ABC tolerance widens the posterior parameter ranges, which in turn inflates predictive uncertainty. The "calibrated ABC tolerance" strategy is very similar in spirit to GLUE approaches where the width of predictive limits was "calibrated" through an exponent within a sum-of-squared-errors pseudo-likelihood function (e.g., Franks et al., 1998).

In the context of predictive uncertainty estimation, the calibration of τ_ρ is broadly analogous to inferring σ_W^2 (or similar error model parameters). However, there are important practical differences:

1. The inference of σ_W^2 requires a single pass though the calibration process (e.g., ABC sampling when working in the signature domain). In contrast, the "calibration" of τ_ρ entails multiple consecutive calibrations, as well as the definition of a separate "prediction limit fitting" criterion;
2. The inference of σ_W^2 can proceed entirely in the *signature* domain (section 5.1.1), with no reference to observed streamflow time series. This capability is essential, e.g., for probabilistic prediction in ungauged basins where signature regionalization is possible (e.g., Castellarin et al., 2013). In contrast, the calibration of τ_ρ as described above is only possible if observed *streamflow time series* are available for inspection. If observed hydrographs are not available, a reasonable value of τ_ρ might be difficult or impossible to

identify: section 5.1.2 vividly illustrates that the performance of prediction limits in the signature domain can be very misleading of their performance in the time domain.

Even after a calibration of τ_ρ, the predictions obtained using the deterministic model will generally differ from those obtained from a probabilistic model: in the former case predictive uncertainty is due solely to parametric uncertainty, whereas in the latter case it is dominated by the residual error term (see section 3.1). In Figure 1, both approaches produce streamflow predictions that are quite vague – reflecting large modelling uncertainties – but the types of deficiencies are different. For example, omitting the error model not only changes the width and shape of the predictive limits, but also their general location (e.g., see the over-prediction of streamflow in rows 3–5). In Figure 2, the preditive limits on the signatures vary substantially depending on the model setup, and become increasingly wide when τ_ρ is coarse. Figure 3 confirms that, when working with the deterministic model, the distances required to produce streamflow predictions with reasonable coverage of observed time series are clearly larger than when the probabilistic model is used.

Overall, when ABC is applied to a deterministic model using a coarse tolerance, it bears little resemblance to its intended use: the distance metric is used in a markedly different way, the rationale behind the selection of tolerance value is different, and the prediction limits are generated differently. It is arguably misleading to consider such computation to be "ABC" – instead it essentially represents a GLUE application (see section 3.1). In the context of this study, this statement is not intended as a criticism of GLUE, but as a statement of fundamental conceptual and algorithmic differences between GLUE as an inference framework and ABC as a numerical sampling technique (see sections 3.1 and 5.4).

5.2. Errors Analysis of ABC Algorithms

5.2.1. How Do Approximation Errors of ABC Algorithms Manifest Themselves?

To gain further insights into the behavior of approximation errors of ABC algorithms, consider the SABC algorithm (Albert et al., 2015), which generates samples from the approximate posterior

$$p^{\mathrm{ABC}}(\boldsymbol{\theta}|\tilde{\boldsymbol{y}},\boldsymbol{x}) \propto \int p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{x})\exp\left(-\rho(\tilde{\boldsymbol{y}},\boldsymbol{y})/\tau_\rho\right)\mathrm{d}\boldsymbol{y} \tag{10}$$

where $\exp\left(-\rho(\tilde{\boldsymbol{y}},\boldsymbol{y})/\tau_\rho\right)$ is the approximation used in the SABC algorithm to replace the exact conditioning by the Dirac delta $\mathcal{D}(\boldsymbol{y}-\tilde{\boldsymbol{y}})$ in equation (7); note that $\lim_{\tau_\rho\to 0}\exp\left(-\rho(\tilde{\boldsymbol{y}},\boldsymbol{y})/\tau_\rho\right) \propto \mathcal{D}(\boldsymbol{y}-\tilde{\boldsymbol{y}})$.

We begin with the case where the model $\boldsymbol{Y}$ is specified as purely deterministic, i.e., $\boldsymbol{Y}=\mathscr{y}$. In this case, only $\boldsymbol{\theta}_h$ can be inferred, and equation (10) yields

$$p^{\mathrm{ABC}}(\boldsymbol{\theta}_h|\tilde{\boldsymbol{y}},\boldsymbol{x}) \propto p(\boldsymbol{\theta}_h)\exp\left(-\rho(\tilde{\boldsymbol{y}},\mathscr{y}(\boldsymbol{\theta}_h,\boldsymbol{x}))/\tau_\rho\right) \tag{11}$$

which can be interpreted as a Bayesian posterior corresponding to the likelihood function

$$p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}_h,\boldsymbol{x})=\mathcal{L}(\boldsymbol{\theta}_h;\tilde{\boldsymbol{y}}) \propto \exp\left(-\rho(\tilde{\boldsymbol{y}},\mathscr{y}(\boldsymbol{\theta}_h,\boldsymbol{x}))/\tau_\rho\right) \tag{12}$$

In turn, equation (12) can be interpreted as an (un-normalized) exponential density with kernel $\rho(\tilde{\boldsymbol{y}},\boldsymbol{y})$. For example, if we set the distance metric $\rho(\tilde{\boldsymbol{y}},\boldsymbol{y})=\sum_i(\tilde{y}_i-y_i)^2$, equation (12) corresponds to the common Gaussian likelihood with mean $\mathscr{y}(\boldsymbol{\theta}_h,\boldsymbol{x})$ and variance $\tau_\rho/2$, i.e., to a homoscedastic Gaussian residual error model with zero mean and fixed variance (assumed known). The specified value of $\tau_\rho/2$ then corresponds to an assumed variance of total (data and structural) errors; larger values of τ_ρ increase posterior parametric uncertainty which, once propagated through the model, generates wider prediction limits in the streamflow (and signature) space. Similar analysis holds when a different distance metric is used: the corresponding "$\rho-$related error model" may not be Gaussian but its variance (or, more generally, dispersion) will still depend on τ_ρ. In other words, in the context of parameter estimation, the specification of an ABC distance metric corresponds to the specification of a residual error model – and quite a simple one because the distance metric is generally a simple function.

Next, consider the case of a probabilistic streamflow model constructed from a deterministic model with additive errors. To simplify the analysis, assume the errors to be iid Gaussian with fixed variance σ_ε^2 (e.g., as in the study of Sadegh & Vrugt, 2013). In this case, it can be shown by extension of equations (10)–(12) that ABC approximation errors take the form of an *additional* term in the residual error model. The ABC samples are then describing the posterior corresponding to an "augmented" error model,

$$p^{ABC}_{(\tau_\rho)}(\boldsymbol{\theta}|\cdot) = p_{(\text{aug})}(\boldsymbol{\theta}|\tilde{\boldsymbol{y}}, \boldsymbol{x}, \sigma^2_{\mathcal{E}}) : \text{"augmented" residual error model } \boldsymbol{\mathcal{E}} \sim \mathcal{N}(0, \sigma^2_{\mathcal{E}}) + O(\tau_\rho) \quad (13)$$

where the notation $O(\tau_\rho)$ denotes the additional error term associated with the ABC tolerance τ_ρ. While the precise dependence of this additional error term on τ_ρ will generally be difficult to derive in closed form, equation (13) provides important clues into the way ABC approximation errors manifest themselves.

First, equation (13) shows that when τ_ρ is small, ABC preserves the residual error model specified by the modeler. In contrast, if τ_ρ is too coarse in relation to $\sigma^2_{\mathcal{E}}$, it modifies and eventually swamps the residual error variance. Relationships analogous to equation (13) hold for more complex probabilistic models – the ABC tolerance acts to represent additional model/data uncertainty and increase posterior parametric uncertainty.

Next, we can interpret the behavior of the hydrological model calibrated under different inference setups:

1. When a *probabilistic* model $\boldsymbol{Q}(\boldsymbol{\theta}, \boldsymbol{x})$ is specified and the ABC tolerance τ_ρ is tight, the description of uncertainty via $\boldsymbol{Q}(\boldsymbol{\theta}, \boldsymbol{x})$ dominates the contribution of the "additional" ρ-related error model. Ideally, the tolerance should be set well below $\sigma^2_{\mathcal{E}}$ to prevent numerical ABC errors from obscuring the behavior of the specified probabilistic model. Similar arguments hold if $\sigma^2_{\mathcal{E}}$ is inferred rather than fixed;
2. When a *deterministic* model $\boldsymbol{h}(\boldsymbol{\theta}_h, \boldsymbol{x})$ is specified and the tolerance τ_ρ is tight, the inference lacks any description of uncertainty and essentially collapses to point estimation;
3. When calibrating a deterministic model $\boldsymbol{h}(\boldsymbol{\theta}_h, \boldsymbol{x})$ using a coarse tolerance τ_ρ, the parameter inference corresponds to calibrating $\boldsymbol{h}(\boldsymbol{\theta}_h, \boldsymbol{x})$ with a residual error model with variance (or dispersion) given by some function of τ_ρ; the form of this function might be worked out through an analysis similar to the one following equation (12) above. The modeler can then use τ_ρ to control – to an extent – the predictive limits (see section 5.1.3); as such, it is the only way to characterize predictive uncertainty in the absence of an explicit error model parameter such as the residual error variance σ^2_W. Note that the predictions would be generated solely by propagation of posterior uncertainty in $\boldsymbol{\theta}_h$ through $\boldsymbol{h}(\boldsymbol{\theta}_h, \boldsymbol{x})$, and would largely forego opportunities to refine the characterization of uncertainty afforded by probabilistic models, e.g., through residual error terms (e.g., Evin et al., 2014; McInerney et al., 2017; Pianosi & Raso, 2012; Wang et al., 2016; and others) and/or internal stochastic terms (e.g., Bulygina & Gupta, 2011; Lockart et al., 2015; Reichert & Mieleitner, 2009; Renard et al., 2011; and others).

These considerations prompt us to elaborate on the role of the tolerance within ABC algorithms.

5.2.2. Role of Tolerance Within ABC—Can Not the Modeler Tweak it to Get "Better" Results?

The tolerance τ_ρ in an ABC algorithm has exactly the same role as the tolerance in any other numerical technique, and should be set according to accuracy requirements and computational budgets. Ideally, τ_ρ should be sufficiently tight that it has little if any impact on the results (e.g., Toni et al., 2009); that is, the intent of ABC is to approximate the posterior, not alter its shape. This is the behavior we generally expect from a numerical approximation algorithm, and is achievable using efficient ABC algorithms such as SABC (section 4.2.3). It makes good practical sense to run ABC with different values of τ_ρ, including potentially coarse ones, to check if the results are numerically stable for the particular application purposes. But, when using ABC as a numerical approximation, it makes little sense to coarsen the tolerance to seek a "better" inference. For example, when comparing signature-domain inference using ABC to time-domain inference using MCMC (Fenicia et al., 2018), a coarse ABC tolerance will introduce approximation errors that manifest in wider parameter posteriors (section 2.4) and could be confused with the effect of information loss due to the use of (non-sufficient) signatures. An analogy with numerical optimization is apt here – a modeler may loosen the optimization convergence criteria (metrics and tolerance) to reduce computational costs, but not to yield "better" optima. And indeed at some point an excessively coarse tolerance will lead to results that cannot be considered a meaningful solution of the original optimization problem.

This role of the ABC tolerance τ_ρ is fundamentally distinct from the role of the threshold α_Υ applied to the pseudo-likelihood function in GLUE applications, where it is intended to reflect the modeler's understanding of likely errors in the data. ABC applications where the distance metric and tolerance are used in the latter way (e.g., Sadegh & Vrugt, 2013, our usage in section 5.1.3, and others) are much closer to GLUE than to ABC. As such, the usage of the distance metric and tolerance to describe predictive uncertainty corresponds to the reasoning behind the statistical modeling choice of likelihood (and pseudo-likelihood) functions, rather than to the reasoning behind the numerical specification of an ABC algorithmic setting.

5.3. General Comments

5.3.1. Sufficient Statistics: What Are They Good for?

This section elaborates on the connection of sufficient statistics to statistical inference using likelihood functions (e.g., Box & Tiao, 1973; Edwards, 1992), and contrasts the role of sufficient statistics in signature-domain inference versus applications of ABC algorithms.

A statistic is said to be "sufficient," with respect to an assumed model, if it encapsulates the entire information content of the data relevant to the model parameters. More formally, given the likelihood function $\mathcal{L}(\boldsymbol{\theta};\tilde{\boldsymbol{y}})$ associated with the assumed probability model $\boldsymbol{Y}(\boldsymbol{\theta},\boldsymbol{x})$, a sufficient statistic $\boldsymbol{\varpi}()$ is a function such that the full dependence of $\mathcal{L}(\boldsymbol{\theta};\tilde{\boldsymbol{y}})$ on $\boldsymbol{\theta}$ can be reconstructed (up to a multiplicative constant) using $\boldsymbol{\varpi}(\tilde{\boldsymbol{y}})$ on its own, even in the absence of $\tilde{\boldsymbol{y}}$ itself. This definition implies that a function $v(\boldsymbol{\varpi},\boldsymbol{\theta})$ exists such that

$$\mathcal{L}(\boldsymbol{\theta};\tilde{\boldsymbol{y}}) = c(\tilde{\boldsymbol{y}})\, v(\boldsymbol{\varpi}(\tilde{\boldsymbol{y}}),\boldsymbol{\theta}) \tag{14}$$

where $c(\tilde{\boldsymbol{y}})$ is independent of the model parameters (e.g., see Edwards, 1992, p. 15). Under these conditions, once $\boldsymbol{\varpi}(\tilde{\boldsymbol{y}})$ is known, knowledge of the entire original data $\tilde{\boldsymbol{y}}$ does not alter the likelihood (nor the posterior), and the inference of $\boldsymbol{\theta}$ remains unchanged.

Practical interest is in sufficient statistics of (much) lower dimension than the data vector itself (otherwise there is the trivial solution $\boldsymbol{\varpi}(\boldsymbol{y}) := \boldsymbol{y}$). For example, in the simple case of a Gaussian probability model, a set of sufficient statistics is given by the empirical (sample) mean and variance of the data. Unfortunately, for a general probability model, finding a $\boldsymbol{\varpi}(\boldsymbol{y})$, say of dimension comparable to the number of model parameters, can be difficult or impossible; indeed we are not aware of a *general* procedure for deriving a low-dimensional $\boldsymbol{\varpi}(\boldsymbol{y})$ from a likelihood function $\mathcal{L}(\boldsymbol{\theta};\tilde{\boldsymbol{y}})$ – even if the latter is available in closed form.

Armed with this understanding, we can now contrast the potential role and utility of sufficient statistics in signature-domain inference versus applications of ABC algorithms.

Role of sufficient statistics ("sufficient signatures") in signature-domain inference. When using signatures to overcome lack of time series data, the ability to derive sufficient signatures (or as sufficient as possible) would be very helpful if the values of these signatures could then be estimated in locations where full time series are not available. In contrast, when using signatures to reduce or eliminate the impact of certain data characteristics on the estimated model parameters (e.g., Experiment 1.3 of Fenicia et al., 2018), the non-sufficiency of a signature is a desired – and designed – property!

Role of sufficient statistics in applications of ABC algorithms. As such, sufficient statistics (or any summary statistics) are not required by ABC as a sampling strategy (see section 2.4). For example, our application of ABC to signature-domain inference here and in Fenicia et al. (2018) does not use any summary statistics, because the set of signatures is already low-dimensional. However, summary statistics are essential as a dimension-reduction tool when using ABC with large observational data sets, e.g., when using ABC to implement an inference directly in the time domain (e.g., Fearnhead & Prangle, 2012). To the extent that such summary statistics are sufficient – and the ABC tolerance is tight! – ABC approximation errors will be minimized.

Note that equation (14) and the definitions given earlier highlight that the sufficiency of any given statistic depends on the adopted probability model. Hence, even if a statistic is sufficient in "theory," its practical utility depends on the validity of this assumed model in the specific modelling application.

For these reasons, practical interest is not necessarily in formally deriving the complete set of strictly sufficient statistics for a given model, but rather in obtaining a set of summary statistics that are "approximately sufficient," or even just "descriptive" ("informative") in a looser sense. Such statistics can be obtained from a variety of sources, including: (i) hydrological insights, i.e., hydrological signatures relevant to a particular catchment, as explored in this work; (ii) by deriving sufficient statistics for a simple model and assuming it is (approximately) sufficient for a larger set of models; (iii) surrogate modelling, e.g., where the mapping between the parameters and outputs of the probability model is approximated using regression, and the summary statistics are then given by the parameter estimators obtained by formulating the surrogate model as a function of observed data (Fearnhead & Prangle, 2012). Hence, even if the sufficiency of a set of statistics (signatures) cannot be proven, it may still be of tremendous practical value to a modeler.

5.3.2. Do "Aggregate" Signatures Obviate the Need to Include Random Terms in the Model?

Signatures that summarize aggregate (average) system behavior tend to exhibit an "uncertainty filtering" effect due to averaging-out of data fluctuations applications (e.g., Westerberg et al., 2011). In light of this, provided the noise in the data and model is not biased, it may appear that there should be little difference between signature-domain calibration of $\boldsymbol{Q}(\boldsymbol{\theta}, \boldsymbol{x})$ or $\boldsymbol{h}(\boldsymbol{\theta}, \boldsymbol{x})$. For example, Figure 15 of Sadegh and Vrugt (2013) suggests that ABC inference of $\boldsymbol{\theta}_h$ (in our notation) using signatures is insensitive to the a priori specification of the error variance $\sigma_{\mathcal{E}}^2$ of an iid Gaussian residual error model (derived from a limits of acceptability approach). However, there are important subtleties.

First, not all metrics that capture aggregate behavior are insensitive to data noise. For example, a question raised in Fenicia et al. (2018) is whether it is possible to estimate the streamflow error (noise) parameter σ_W^2 from signatures alone – and this is shown to be feasible for a certain type of signatures. In particular, the flashiness index (Baker et al., 2004), which is defined by an integral (average) over time but a derivative in streamflow space (section 4.2.2), is sensitive to errors in the data; indeed, its form, based on differences between consecutive data points, is similar to the measurement error estimator used by Vrugt et al. (2005). Despite its "aggregate" nature, the flashiness index can support the estimation of σ_W^2.

Second, the results in Figure 15 of Sadegh and Vrugt (2013) were obtained using a coarse ABC tolerance. In this case, ABC approximation errors can swamp genuine differences between posteriors obtained under different residual error model assumptions. Suppose we are comparing ABC samples of posteriors conditioned on two different values of $\sigma_{\mathcal{E}}^2$, say $\sigma_{\mathcal{E}1}^2$ and $\sigma_{\mathcal{E}2}^2$. Equation (13) shows that if τ_ρ is too coarse in relation to $\sigma_{\mathcal{E}1}^2$ and $\sigma_{\mathcal{E}2}^2$, it will obscure genuine differences between these posteriors.

Finally, even in a hypothetical case where the calibration gives the same estimates of $\boldsymbol{\theta}_h$ despite omitting the error model, this might not meet modelling objectives. In many operational contexts, including streamflow forecasting at a range of lead times (Demargne et al., 2014), error model parameters are needed to make probabilistic predictions, e.g., using residual error models to characterize predictive uncertainty. If the modeler is not interested in inferring unknown error parameters, the appropriate approach is to integrate over them in the posterior (Box & Tiao, 1973).

5.3.3. What If the Inference Cannot Match the Signatures?

It is sometimes impossible to find a model parameter set $\boldsymbol{\theta}^{(i)}$ that satisfies the distance metric to the specified tolerance. For example, in our empirical study, the SABC algorithm could not find parameter sets for which the deterministic model in setup 2 from section 4.2.3 had less than 5% relative errors in the signatures. Similar behavior was reported by Yilmaz et al. (2008) for multi-objective calibration in signature-space: only a single parameter set met the tolerance.

In general, if a parameter set that satisfies all metrics (to a sufficiently tight tolerance) cannot be found even after a large number of trials, it follows that the assumed model is unable to characterize the magnitude of *total* uncertainty in the model structure and/or data. This is clearly relevant to any setup where a purely deterministic model is used without any allowance for predictive uncertainty (here understood to correspond to data and model errors), but could also occur when trying to match signatures such as low FDC quantiles to a high relative accuracy (e.g., see section 5.1.1 here and section 4.1.4 in Fenicia et al., 2018).

If the modeler has specified a residual error model to reflect solely the observational uncertainty in the data, failure to find an acceptable parameter could point to model structural error (as suggested by Sadegh & Vrugt, 2013). However, unless the modeler trusts their a priori estimates of observation error (in both input and output data) including their effect on any signatures used, failure to meet the ABC tolerance could just as likely be due to unaccounted observation errors, or to a *mix* of unaccounted data and structural errors. Uncertainty decomposition to reliably distinguish between data and structural errors requires detailed analysis similar to Renard et al. (2011). ABC on its own cannot provide such information – it is just a numerical sampling algorithm and cannot replace statistical analysis of data and structural errors.

If the modeler cannot furnish reliable observation error models, it is preferable to use an aggregate residual error model and infer its variance from the calibration data. Whether such error parameter inference is feasible in the signature domain depends on the selection of signatures. For example, the empirical studies in section 4 and in Fenicia et al. (2018) suggest that the flashiness index captures information that can support the inference of a residual error model formulated in the streamflow time domain.

5.3.4. Is Working With Signatures Similar to Working With Data Transformations?

Readers may notice the mathematical similarity between the formulation of Bayes equation in terms of streamflow signatures, and the formulation of Bayes equation in terms of data transformations, such as the log and Box-Cox transformations. Indeed, both cases represent Bayesian inference in a transformed domain. However, most hydrological signatures of interest are not invertible, whereas data transformations such as log and Box-Cox have well-defined inverse transformations. For example, the Flow Duration Curve transformation cannot be uniquely inverted to obtain the original streamflow time series, in contrast it is trivial to go back-and-forth between the streamflow time series in raw versus log-transformed space. As a result, an error model in the transformed space can be uniquely mapped to an error model in the raw space (and vice versa), and the likelihood function for the inference using log and Box-Cox transformations can be formulated in closed form (e.g., see section 2.5.1 in Fenicia et al., 2018).

Conceptually, we could say that most signatures "throw away" some information in order to isolate the characteristic of interest, whereas data transformations do not. Nor can numerical values of log- or Box-Cox-transformed streamflow be "regionalized" in the same way as streamflow signatures. As a result, data transformations can be used to represent features such as heteroscedasticity and skewness (e.g., McInerney et al., 2017), but cannot fulfil the promises offered by signature-domain inference.

5.4. ABC: What It Is, and What It Is Not

Given the number and variations of inference frameworks and associated sampling techniques, and the potential for confusion given relatively subtle algorithmic aspects, this section revisits some earlier questions on the relationship between signature-domain inference, ABC algorithms and general inference.

Our presentation of ABC emphasizes its generality as a class of numerical algorithms for sampling from a conditional distribution: it can be used to sample from Bayesian posteriors conditioned on any type of data, including streamflow time series and/or signatures. The hallmark feature of ABC is that it exchanges the requirement to *evaluate* the likelihood function with the requirement to *sample* from the probability model of the data. This feature makes ABC an attractive choice for implementing signature-domain inference: for many signatures of hydrological interest, the corresponding likelihood functions may not have convenient closed-form expressions even if the original time-domain probabilistic models do (e.g., as in section 4).

ABC is not limited to implementing signature-domain inference – it can also be used to calibrate time-domain models, offering advantages for any stochastic model for which it is much easier to draw a sample than to evaluate its probability distribution (pdf or pmf). This class of models is of major interest in hydrology and broader environmental sciences, with examples including, as noted by Nott et al. (2012), explicit representations of input uncertainty (e.g., extending studies such as Kavetski et al., 2006; Renard et al., 2011; Wright et al., 2017), state-dependent representation of model structural errors (e.g., extending studies of Albert et al., 2016; Reichert & Mieleitner, 2009; Renard et al., 2011), stochastic disaggregation of coarse input data (e.g., the sunshine model of Lockart et al., 2015), and so forth. Given the broad utility of ABC, hydrologists should not (mistakenly) assume that ABC is a synonym for signature-domain calibration.

ABC algorithms are often described as "likelihood-free" techniques in the statistical literature (e.g., Marjoram et al., 2003). This wording refers to the likelihood function never being *computed* within an ABC algorithm. It should not be mis-interpreted to imply that ABC algorithms somehow spare the modeler from having to develop a probability model of the system of interest, including making explicit assumptions about the statistical properties of data and structural errors, and from inquisitively testing these assumptions against available evidence. Rather, ABC algorithms liberate the modeler from having to derive closed-form expressions for the pdfs of their assumed models (or constructing direct numerical approximations), and instead allow them to specify these models in the form of sampling algorithms (see Step 2 of the ABC algorithm in section 2.4 for a simple example). The likelihood function is of course "still there" by the virtue that every probability model and stochastic algorithm have an associated pdf (or pmf) – which inevitably appears in the Bayes inference equation (5) whether it is evaluated explicitly or not. How does the likelihood function exert its influence on the posterior if never computed? It acts through the sampling distribution generated by the probability model $\boldsymbol{Y}(\boldsymbol{\theta}, \boldsymbol{x})$: values of $\boldsymbol{\theta}$ that lead frequently to model realizations that match the observations closely will appear more frequently in the set of ABC samples.

Neither does the subjectivity in the choice of the ABC distance metric imply that a Bayesian inference implemented using an ABC algorithm is more "subjective" than, e.g., a standard setup where an MCMC algorithm is applied to the posterior distribution given a closed-form likelihood function. The ABC distance metric ρ and tolerance τ_ρ are numerical approximation tools – in the same sense as, e.g., the metrics and tolerances used when assessing the convergence of MCMC samples. These numerical choices are "subjective" and can impact on practical computation. For example, as noted in section 2.4, ABC with a coarse tolerance introduces approximation errors that manifest as wider posteriors ("information loss"). To prevent such numerical artifacts from obscuring model analysis, it is a good modelling practice to set numerical accuracy settings as tight as practical, to reduce numerical errors to small or negligible values (e.g., as achieved by the SABC-on-*Q* inference in this work; see Figure 8 in Fenicia et al. (2018) for a direct illustration).

Once seen from this perspective, it should be clear that ABC is not an "alternative" to time-domain inference: time-domain inference refers to a particular data choice, whereas ABC refers to a numerical strategy choice. ABC is not even an alternative to MCMC, because MCMC techniques can be used within ABC algorithms (Albert et al., 2015; Marjoram et al., 2003). As such, the sequence of modeling choices facing a Bayesian modeler could be described as follows: (i) choose the probability model (e.g., hydrological model plus residual errors) and calibration data (e.g., streamflow time series or streamflow signatures), (ii) choose between Strategy A and Strategy B for sampling from the posterior (section 2.3), (iii) choose a numerical sampling algorithm suitable for the strategy selected in the previous step (e.g., SABC, MCMC, etc.), and (iv) iterate until the modelling assumptions and objectives are met to the required standard.

Finally, an application of ABC in itself is neither "formal" nor "informal." Neither does the use of signatures or time series as observed data determine whether a probabilistic inference is "formal" or "informal." Rather it is the nature of the model $\boldsymbol{Y}(\boldsymbol{\theta}, \boldsymbol{x})$ assumed to describe the observed data $\tilde{\boldsymbol{y}}$ that is important: if this model is "formally" constructed using probability theory, the inference and prediction can be said to be "formally" probabilistic – irrespective of whether it is formulated in the time domain or in the signature domain, and irrespective of whether the modeler works directly with the likelihood function $p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}, \boldsymbol{x})$ or uses an ABC algorithm in conjunction with a sampler from $\boldsymbol{Y}(\boldsymbol{\theta}, \boldsymbol{x})$. And then it is a separate question whether the resulting setup meets the required probabilistic performance criteria. Conversely, if the model is not constructed using probability theory, the inference and prediction using such a model cannot be meaningfully described as probabilistic. This would not preclude the modeler from specifying their own set of performance criteria, and checking if these criteria are met with their particular modelling choices.

6. Conclusions

This study provides a Bayesian perspective on signature-domain inference of hydrological model parameters, its implementation using Approximate Bayesian Computation (ABC), and its connection to traditional Bayesian inference based on streamflow time series. A major focus is on quantification of uncertainty in the calibrated parameters and streamflow predictions, even when calibrating to signatures alone.

The following major conclusions are obtained:

1. Theoretical analysis shows that:
   a. Signature-domain calibration can be formulated within a Bayesian inference framework, which provides a mechanism for using signatures within uncertainty estimation and prediction applications;
   b. When starting from a probabilistic model of streamflow time series, even a relatively simple one such as a deterministic model with additive Gaussian errors, the likelihood function in the signature-domain is difficult or impossible to derive in closed-form, making it impractical to evaluate (unless some kind of efficient approximation is developed). Traditional sampling techniques, which require the pdf to be evaluated, are hence not readily applicable. Instead, Approximate Bayesian Computation (ABC) algorithms provide a general approach for sampling from conditional distributions without evaluating their pdf, but instead sampling from the related joint distribution and applying an approximate conditioning step. ABC algorithms avoid the need to evaluate the likelihood function and are hence particularly attractive for signature-domain inference;
   c. Previous attempts to use ABC in hydrology have important limitations. Most notably, the use of the deterministic model in lieu of the probabilistic model within the acceptance-rejection test corresponds to ignoring all data and model errors and is hardly reasonable. In addition, the use of a coarse

ABC tolerance contradicts the intention of ABC to provide an accurate numerical approximation to a given posterior, and corresponds to an additional assumed error model.

d. Despite some general resemblance, ABC algorithms differ from GLUE applications in several fundamental aspects: most notably, in their starting point being a probability model of the data, as well as in the expectation of driving the acceptance tolerance tight enough that the estimated posteriors and predictive distributions become insensitive to the choice of distance metric. With this in mind, GLUE is neither more nor less similar to ABC than to any other Bayesian inference scheme.

2. Empirical analyses using data from the Lacmalac catchment in Australia suggest that:

   a. Bayesian signature-domain inference implemented using ABC with a sufficiently tight tolerance produces predictive distributions that generally capture the observed streamflow time series and their signatures. This confirms the general viability of the probabilistic estimation framework even when the inferred hydrological model is articulated in a different domain than the calibration data.
   b. Using the deterministic model alone (omitting random terms) while still using a tight ABC tolerance might produce reasonable deterministic predictions, but completely overlook predictive uncertainty. This behavior is readily explained by theoretical considerations;
   c. Using the deterministic model alone (omitting random terms) while using a coarse acceptance (ABC) tolerance cannot be reasonably described as ABC and instead is much closer to a GLUE application. Under these circumstances, the inference is controlled by the distance metric and tolerance (rather than by the formulation of the probability model, as expected in Bayesian modeling). In turn, the modeler may be able to control the width of predictive limits by changing the tolerance;

The ability of 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 study by Fenicia et al. (2018) provides an in-depth analysis of the properties of signature-domain inference, including its comparison to the time-domain inference under a diverse range of scenarios, including cases where the predictive model is substantially mis-specified. Furthermore, the utility of ABC goes well beyond signature-domain inference and includes the inference of stochastic models for which sampling is more efficient than evaluating the probability distribution; this research direction will be pursued in future investigations.

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 first 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 second author (FF). The SABC algorithm, including parallelization facilities, is available in a Python library from https://github.com/eth-cscs/abcpy.

References

Albert, C., Künsch, H., & Scheidegger, A. (2015). 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.

Ang, A. H.-S., & Tang, W. H. (2007). *Probability concepts in engineering emphasis on applications to civil & environmental engineering* (2nd ed., 406 pp.). Hoboken, NJ: Wiley.

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.

Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, *162*(4), 2025–2035.

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*(1–2), 18–36.

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

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

Blum, M. G. B., & François, O. (2010). Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*, *20*(1), 63–73. https://doi.org/10.1007/s11222-009-9116-0

Blum, M. G. B., Nunes, M. A., Prangle, D., & Sisson, S. A. (2013). A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science, 28*(2), 189–208. https://doi.org/10.1214/12-STS406

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

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

Bulygina, N., & Gupta, H. (2011). Correcting the mathematical structure of a hydrological model via Bayesian data assimilation. *Water Resources Research*, *47*, W05514. https://doi.org/10.1029/2010WR009614

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* (pp. 135–162). New York, NY: Cambridge University Press.

Castiglioni, S., Lombardi, L., Toth, E., Castellarin, A., & Montanari, A. (2010). Calibration of rainfall-runoff models in ungauged basins: A regional maximum likelihood approach. *Advances in Water Resources*, *33*(10), 1235–1242. https://doi.org/10.1016/j.advwatres.2010.04.009

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, *25*(7), 410–418. https://doi.org/10.1016/j.tree.2010.04.001

De Finetti, B. (1972). *Probability, induction and statistics: The art of guessing* (XXIV, 266 pp.). London, UK: John Wiley.

Demargne, J., Wu, L., Regonda, S. K., Brown, J. D., Lee, H., He, M., et al. (2014). The science of NOAA's operational hydrologic ensemble forecast service. *Bulletin of the American Meteorological Society*, *95*(1), 79–98. https://doi.org/10.1175/BAMS-D-12-00081.1

Diggle, P. J., & Gratton, R. J. (1984). Monte Carlo Methods of Inference for Implicit Statistical Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, *46*(2), 193–227.

Duan, Q. Y., Sorooshian, S., & Gupta, V. (1992). Effective and Efficient Global Optimization for Conceptual Rainfall-Runoff Models. *Water Resources Research*, *28*(4), 1015–1031.

Edwards, A. W. F. (1992). *Likelihood* (expanded ed., xix, 275 pp.). Baltimore, MD: Johns Hopkins University Press.

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

Fearnhead, P., & Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, *74*, 419–474. https://doi.org/10.1111/j.1467-9868.2011.01010.x

Fenicia, F., Kavetski, D., Reichert, P., & Albert, C. (2018), Understanding signature-domain calibration of hydrological models using Approximate Bayesian Computation: Empirical analysis of fundamental properties. *Water Resources Research*, *54*. https://doi.org/10.1002/2017WR021616

Fenicia, F., Kavetski, D., Savenije, H. H. G., & Pfister, L. (2016). From spatially variable streamflow to distributed hydrological models: Analysis of key modeling decisions. *Water Resources Research*, *52*, 954–989. https://doi.org/10.1002/2015WR017398

Franks, S. W., Gineste, P., Beven, K. J., & Merot, P. (1998). On constraining the predictions of a distributed moder: The incorporation of fuzzy estimates of saturated areas into the calibration process. *Water Resources Research*, *34*(4), 787–797.

Freer, J., Beven, K., & Ambroise, B. (1996). Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach. *Water Resources Research*, *32*(7), 2161–2173.

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

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.

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/https://doi.org/10.1016/j.jhydrol.2010.08.013

Kavetski, D., Fenicia, F., & Clark, M. P. (2011). Impact of temporal data resolution on parameter inference and model identification in conceptual hydrological modeling: Insights from an experimental catchment. *Water Resources Research*, *47*, W05501. https://doi.org/10.1029/2010WR009525

Kavetski, D., Kuczera, G., & Franks, S. W. (2006). Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resources Research*, *42*, W03407. https://doi.org/10.1029/2005WR004368

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

Lamb, R., & Beven, K. (1997). Using interactive recession curve analysis to specify a general catchment storage model. *Hydrology and Earth System Sciences*, *1*(1), 101–113. https://doi.org/10.5194/hess-1-101-1997

Lenormand, M., Jabot, F., & Deffuant, G. (2013). Adaptive approximate Bayesian computation for complex models. *Computational Statistics*, *28*(6), 2777–2796. https://doi.org/10.1007/s00180-013-0428-3

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

Madsen, H. (2000). Automatic calibration of a conceptual rainfall-runoff model using multiple objectives. *Journal of Hydrology*, *235*(3–4), 276–288.

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

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

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

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

McMillan, H. K., Clark, M. P., Bowden, W. B., Duncan, M., & Woods, R. A. (2011b). Hydrological field data from a modeller's perspective: Part 1. Diagnostic tests for model structure. *Hydrological Processes*, *25*(4), 511–522. https://doi.org/10.1002/hyp.7841

Montanari, A., Young, G., Savenije, H. H. G., Hughes, D., Wagener, T., Ren, L. L., et al. (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

Perrin, C., Michel, C., & Andreassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology*, *279*(1–4), 275–289. https://doi.org/10.1016/S0022-1694(03)00225-7

Pianosi, F., & Raso, L. (2012). Dynamic modeling of predictive uncertainty by regression on absolute errors. *Water Resources Research*, *48*, W03516. https://doi.org/10.1029/2011WR010603

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. (1999). Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molecular Biology and Evolution*, *16*(12), 1791–1798.

Reichert, P., Langhans, S. D., Lienert, J., & Schuwirth, N. (2015). The conceptual foundation of environmental decision support. *Journal of Environmental Management*, *154*, 316–332. 10.1016/j.jenvman.2015.01.053.

Reichert, P., & Mieleitner, J. (2009). Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters. *Water Resources Research*, *45*, W10402. https://doi.org/10.1029/2009WR007814
Reichert, P., & Schuwirth, N. (2012). Linking statistical bias description to multiobjective model calibration. *Water Resources Research*, *48*, W09543. https://doi.org/10.1029/2011WR011391
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., 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
Schoups, G., & Vrugt, J. A. (2010). A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resources Research*, *46*, W10531. https://doi.org/10.1029/2009WR008933
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., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., et al. (2003). IAHS decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. *Hydrological SciencesJournal*, *48*(6), 857–880. https://doi.org/10.1623/hysj.48.6.857.51421
Smith, P., Beven, K. J., & Tawn, J. A. (2008). Informal likelihood measures in model assessment: Theoretic development and investigation. *Advances in Water Resources*, *31*(8), 1087–1100. https://doi.org/10.1016/j.advwatres.2008.04.012
Smith, T., Marshall, L., & Sharma, A. (2015). Modeling residual hydrologic errors with Bayesian inference. *Journal of Hydrology*, *528*, 29–37. https://doi.org/10.1016/j.jhydrol.2015.05.051
Stedinger, J. R., Vogel, R. M., Lee, S. U., & Batchelder, R. (2008). Appraisal of the generalized likelihood uncertainty estimation (GLUE) method. *Water Resources Research*, *44*, W00B06. https://doi.org/10.1029/2008WR006822
Tavare, S., Balding, D. J., Griffiths, R. C., & Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, *145*(2), 505–518.
Tolson, B. A., & Shoemaker, C. A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, *43*, W01413. https://doi.org/10.1029/2005WR004723
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
Turner, B. M., & Van Zandt, T. (2012). A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology*, *56*(2), 69–85. https://doi.org/10.1016/j.jmp.2012.02.005
Vogel, R. M., & Fennessey, N. M. (1995). Flow duration curves II: A review of applications in water resources planning 1. *Water Resourxes Bulletin*, 31(6), 1029–1039. https://doi.org/10.1111/j.1752-1688.1995.tb03419.x
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., Bastidas, L. A., Bouten, W., & Sorooshian, S. (2003). Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resources Research*, *39*(8), 1214. https://doi.org/10.1029/2002WR001746
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
Vrugt, J. A., ter Braak, C. J. F., Clark, M. P., Hyman, J. M., & Robinson, B. A. (2008). Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resources Research*, *44*, W00B09. https://doi.org/10.1029/2007WR006720
Wang, S., Huang, G. H., Baetz, B. W., & Huang, W. (2016). Probabilistic Inference Coupled with Possibilistic Reasoning for Robust Estimation of Hydrologic Parameters and Piecewise Characterization of Interactive Uncertainties. *Journal of Hydrometeorology*, *17*(4), 1243–1260. https://doi.org/10.1175/jhm-d-15-0131.1
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., Schaefli, B., Montanari, A., & Savenije, H. H. G. (2009). On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information. *Water Resources Research*, *45*, W12422. https://doi.org/10.1029/2009WR007706
Wright, A. J., Walker, J. P., & Pauwels, V. R. N. (2017). Estimating rainfall time series and model parameter distributions using model data reduction and inversion techniques. *Water Resources Research*, *53*, 6407–6424. https://doi.org/10.1002/2017WR020442
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, S., Yaeger, M., Coopersmith, E., Cheng, L., & Sivapalan, M. (2012). Exploring the physical controls of regional patterns of flow duration curves – Part 2: Role of seasonality, the regime curve, and associated process controls. *Hydrology and Earth System Sciences*, *16*(11), 4447–4465. https://doi.org/10.5194/hess-16-4447-2012
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
Zhang, Z. X., Wagener, T., Reed, P., & Bhushan, R. (2008). Reducing uncertainty in predictions in ungauged basins by combining hydrologic indices regionalization and multiobjective optimization. *Water Resources Research*, *44*, W00B04. https://doi.org/10.1029/2008WR006833