WATER RESOURCES RESEARCH, VOL. 45, W03603, doi:10.1029/2007WR006538, 2009

Comment on "An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction" by Newsha K. Ajami et al.

Benjamin Renard,[1,2] Dmitri Kavetski,[1] and George Kuczera[1]

Received 19 September 2007; revised 20 October 2008; accepted 3 November 2008; published 10 March 2009.

Citation: Renard, B., D. Kavetski, and G. Kuczera (2009), Comment on ''An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction'' by Newsha K. Ajami et al., *Water Resour. Res.*, *45*, W03603, doi:10.1029/2007WR006538.

1. Summary

[1] Uncertainty in the rainfall inputs, which constitute a primary forcing of hydrological systems, considerably affects the calibration and predictive use of hydrological models. In a recent paper, *Ajami et al.* [2007] proposed the Integrated Bayesian Uncertainty Estimator (IBUNE) to quantify input, parameter and model uncertainties. This comment analyzes two interpretations of the IBUNE method and compares them to the Bayesian Total Error Analysis (BATEA) method [*Kavetski et al.*, 2002, 2006a]. It is shown that BATEA and IBUNE are based on the same hierarchical conceptualization of the input uncertainty. However, in interpretation A of IBUNE, the likelihood function, and hence the posterior distribution, are random functions of the inferred variables, which violates a standard requirement for probability density functions (pdf). A synthetic study shows that IBUNE-A inferences are inconsistent with the correct parameter values and model predictions. In the second interpretation, IBUNE-B, it is shown that a specific implementation of IBUNE is equivalent to a special Metropolis-Hastings sampler for the full Bayesian posterior, directly including the rainfall multipliers as latent variables (but not necessarily storing their samples). Consequently, IBUNE-B does not reduce the dimensionality of the sampling problem. Moreover, the jump distribution for the latent variables embedded in IBUNE-B is computationally inefficient and leads to prohibitively slow convergence. Modifications of these jump rules can cause convergence to incorrect posterior distributions. The primary conclusion of this comment is that, unless the hydrological model and the structure of data uncertainty allow specialized treatment, Bayesian hierarchical models invariably lead to high-dimensional computational problems, whether working with the full posterior (high-dimensional sampling problem) or with the marginal posterior (high-dimensional integration problem each time the marginal posterior is evaluated).

2. Introduction

[2] Rigorous quantification of the uncertainties affecting the calibration of conceptual rainfall-runoff (CRR) models remains a challenging task in hydrological modeling. Several promising approaches have emerged to account for various sources of errors [*Kavetski et al.*, 2002; *Vrugt et al.*, 2005; *Ajami et al.*, 2006; *Kavetski et al.*, 2006a, 2006b; *Kuczera et al.*, 2006; *Vrugt and Robinson*, 2007]. Of significance is the recognition that rainfall uncertainty can be considerable and thus needs to be explicitly accounted for in the inference scheme. Recently, *Ajami et al.* [2007] (hereafter referred to as ADS2007) proposed the Integrated Bayesian Uncertainty Estimator (IBUNE). This inference framework characterizes input uncertainty using rainfall multipliers and addresses model structural uncertainty using Bayesian model averaging.

[3] This paper comments on the approach used by ADS2007 to incorporate rainfall uncertainty into the likelihood function, and compares IBUNE with the Bayesian total error analysis (BATEA) framework proposed by *Kavetski et al.* [2002, 2006a]. Although both schemes share a common conceptual framework, their formulation and implementation appear to be very different. In view of the significance of rainfall uncertainty in CRR modeling, it is important to critically evaluate and, where possible, reconcile these differences.

[4] The IBUNE approach represents rainfall errors using daily rainfall multipliers assumed to follow a Gaussian distribution with unknown mean and variance. The characterization of input errors using such multipliers (latent variables) corresponds to a hierarchical model of input uncertainty. This hierarchical model is the same as used in BATEA and leads to a high-dimensional full likelihood function because the number of latent variables is proportional to the length of calibration data. This makes its analysis computationally challenging compared to methods that ignore rainfall uncertainty.

[5] To reduce the computational cost of the BATEA approach, IBUNE was formulated to avoid inference of the latent variables (or rainfall multipliers). However, *Ajami et al.* [2007] do not state the complete IBUNE likelihood and posterior (and their precise dependence on the multipliers), making it difficult to ascertain IBUNE's theoretical

[1] School of Engineering, University of Newcastle, Callaghan, New South Wales, Australia.

[2] Cemagref, UR HHLY, Lyon, France.

0043-1397/09/2007WR006538$09.00

and practical properties and to implement the IBUNE equations using alternative analysis methods.

[6] This paper critically analyzes two interpretations of the IBUNE method and compares them to BATEA, which implements the full treatment of the Bayesian hierarchical model of input uncertainty and directly infers the latent variables. Interpretation A, referred to as IBUNE-A, is based strictly on the description in ADS2007 that the method does not infer the individual multipliers (only their mean and variance) and that a single set of rainfall multipliers is sampled from their Gaussian distribution at every evaluation of the likelihood function. It is shown that the IBUNE likelihood and posterior become random (nondeterministic) functions of their arguments, which violates a fundamental requirement for probability density functions. Alternatively, interpretation B (IBUNE-B) relates IBUNE to a Metropolis-Hastings algorithm for sampling from the full posterior of the Bayesian hierarchical model. This method is theoretically convergent, but operates in the full dimensional space including all the multipliers and does not resolve the ADS2007 aim of reducing the dimension of the problem.

[7] This paper is organized as follows. Section 3 details the Bayesian hierarchical framework for describing rainfall uncertainty. Section 4 describes the IBUNE-A interpretation and how it leads to a random likelihood function. Additional issues are raised related to optimization (section 4.3) and MCMC sampling (sections 4.4 and 4.5). Section 4.6 summarizes a synthetic case study that compares the inferences obtained using IBUNE-A and BATEA and assesses them given the known synthetic parameters and model simulations. (A full description of the case study can be found in Text S1[1].) Section 5 describes the IBUNE-B interpretation and examines its convergence and computational efficiency. The paper concludes with a discussion of the advantages and limitations of using Bayesian hierarchical methods to account for rainfall uncertainty in hydrological modeling.

3. Bayesian Hierarchical Treatment of Input Uncertainty

[8] In this paper, $\boldsymbol{X} = (\boldsymbol{x}_t)_{t\,=\,1...T}$ denotes the true model inputs (e.g., rainfall), $\boldsymbol{Y} = (\boldsymbol{y}_t)_{t\,=\,1...T}$ are the true outputs (e.g., runoff), $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$ are the corresponding observed values, and T is the total number of time steps. The hydrological model $\boldsymbol{M}$ approximates the mapping of the true inputs into the true outputs,

$$\boldsymbol{Y} = \boldsymbol{M}(\boldsymbol{\theta}, \boldsymbol{X}) \tag{1}$$

Traditional calibration methods such as standard least squares (SLS), which assume the observed inputs are error-free ($\boldsymbol{X} = \tilde{\boldsymbol{X}}$), can be shown to yield biased inferences of $\boldsymbol{\theta}$.

3.1. Hierarchical Model of Rainfall Errors

[9] One approach to describe input uncertainty is to assume that the observed rainfall depth at time step t is corrupted by a multiplicative error ϕ_t,

$$r_t = \phi_t \tilde{r}_t \tag{2}$$

The rainfall multipliers $\boldsymbol{\Phi} = (\phi_t)_{t\,=\,1...T}$ are then assumed to follow some distribution, i.e.,

$$\boldsymbol{\Phi} \sim p(\boldsymbol{\Phi}|\boldsymbol{\eta}) \tag{3}$$

If every multiplier ϕ_t is assumed to arise independently from an identical Gaussian distribution with parameters $\boldsymbol{\eta} = (m, \sigma)$, where m is the mean and σ is the standard deviation, then $p(\boldsymbol{\Phi}|\boldsymbol{\eta})$ is an uncorrelated Gaussian distribution,

$$p(\boldsymbol{\Phi}|\boldsymbol{\eta}) = N\left(\boldsymbol{\Phi}|m, \sigma^2\right) = \prod_{t=1}^{T} N\left(\phi_t|m, \sigma^2\right)$$

[10] This is the input error model adopted by ADS2007 (equation 7 in the ADS2007 paper). Since this conceptualization of rainfall errors corresponds to a hierarchical model, the rainfall multipliers $\boldsymbol{\Phi}$ are termed “latent variables” (whether or not they are explicitly estimated during the calibration) and $\boldsymbol{\eta}$ are the input error hyperparameters [*Kuczera et al.*, 2006]. Note that applying the multiplier model over storm event timescales [*Kavetski et al.*, 2002, 2006a, 2006b], rather than over fixed time steps [*Ajami et al.*, 2007; *Renard et al.*, 2007], does not alter the structure of the Bayesian inference, but corresponds to a specific assumption about the timescale of input errors.

[11] Since the catchment responds to the true forcings, the response at a time step t is computed by forcing the model with the “corrected” rainfall $r_t = \phi_t\,\tilde{r}_t$. Equation (1) is then applied using $\tilde{\boldsymbol{X}}$ and $\boldsymbol{\Phi}$ to estimate the true inputs $\boldsymbol{X}$:

$$\boldsymbol{Y} = \boldsymbol{M}\left(\boldsymbol{\theta}, \{\boldsymbol{\Phi}, \tilde{\boldsymbol{X}}\}\right) \tag{4}$$

We now consider the Bayesian treatment of the hierarchical model (2)–(4).

3.2. Full Posterior of the Hierarchical Model

[12] Several unknown quantities appear in the hierarchical model: the CRR model parameters $\boldsymbol{\theta}$, the latent variables (here, the rainfall multipliers) $\boldsymbol{\Phi}$ and the hyperparameters $\boldsymbol{\eta}$. In a Bayesian framework, inference of these quantities is performed through their posterior distribution, which can be derived as follows:

$$\begin{aligned} p\left(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\Phi}|\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}\right) &\propto p\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\Phi}, \tilde{\boldsymbol{X}}\right) p\left(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\Phi}|\tilde{\boldsymbol{X}}\right) \\ &= p\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\Phi}, \tilde{\boldsymbol{X}}\right) p(\boldsymbol{\Phi}|\boldsymbol{\theta}, \boldsymbol{\eta}) p(\boldsymbol{\theta}, \boldsymbol{\eta}) \\ &= p\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}, \boldsymbol{\Phi}, \tilde{\boldsymbol{X}}\right) p(\boldsymbol{\Phi}|\boldsymbol{\eta}) p(\boldsymbol{\theta}, \boldsymbol{\eta}) \end{aligned} \tag{5}$$

The posterior distribution (5) is typical of error-in-variables problems treated using the Bayesian approach, in particular, the BATEA method [*Kavetski et al.*, 2006a]. The last line in equation (5) relies on three observations: (1) the prior $p(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\Phi}|\tilde{\boldsymbol{X}})$ does not depend on the observed inputs $\tilde{\boldsymbol{X}}$; (2) $p(\boldsymbol{\Phi}|\boldsymbol{\theta}, \boldsymbol{\eta}) = p(\boldsymbol{\Phi}|\boldsymbol{\eta})$ because the distribution of latent variables (here, multipliers) depends only on the hyperparameters $\boldsymbol{\eta}$ and not on the CRR parameters; and (3) $p(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\Phi}, \tilde{\boldsymbol{X}}) = p(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}, \boldsymbol{\Phi}, \tilde{\boldsymbol{X}})$ because the probability of observed outputs given the latent variables $\boldsymbol{\Phi}$ is independent of $\boldsymbol{\eta}$. (Equation (4) shows that $\boldsymbol{\eta}$ is not needed to estimate the true inputs and hence outputs.)

[13] The likelihood function can be specified after additional assumptions regarding the structure of residual un-

[1] Auxiliary materials are available in the HTML. doi:10.1029/2007WR006538.

certainty are made. For example, for a single-output model $\boldsymbol{M}(\cdot)$ where y_t is the runoff at time step t, the residuals $\boldsymbol{e} = (e_t)_{t=1\ldots T}$ can be defined as

$$e_t\left(\boldsymbol{\theta},\boldsymbol{\Phi},\tilde{\boldsymbol{X}}\right) = y_t\left(\boldsymbol{\theta},\{\boldsymbol{\Phi},\tilde{\boldsymbol{X}}\}\right) - \tilde{y}_t = M_t\left(\boldsymbol{\theta},\{\boldsymbol{\Phi},\tilde{\boldsymbol{X}}\}\right) - \tilde{y}_t \quad (6)$$

If the residuals are assumed to be independent realizations from a Gaussian distribution with zero mean and unknown variance σ_y^2 (consistently with ADS2007 assumption, p. 4), the following likelihood function is obtained

$$p\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta},\boldsymbol{\Phi},\tilde{\boldsymbol{X}},\sigma_y^2\right) = \prod_{t=1}^{T} N\left(e_t\left(\boldsymbol{\theta},\boldsymbol{\Phi},\tilde{\boldsymbol{X}}\right)|0,\sigma_y^2\right) \quad (7)$$

The unknown variance σ_y^2 in equation (7) can be integrated out

$$p\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta},\boldsymbol{\Phi},\tilde{\boldsymbol{X}}\right) = \int p\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta},\boldsymbol{\Phi},\tilde{\boldsymbol{X}},\sigma_y^2\right)p\left(\sigma_y^2\right)d\sigma_y^2 \quad (8)$$

When Jeffrey's prior $p(\sigma_y^2) \propto 1/\sigma_y^2$ is used, the likelihood becomes

$$p\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta},\boldsymbol{\Phi},\tilde{\boldsymbol{X}}\right) \propto \left[\sum_{t=1}^{T}\left(e_t\left(\boldsymbol{\theta},\boldsymbol{\Phi},\tilde{\boldsymbol{X}}\right)\right)^2\right]^{-\frac{T}{2}} \quad (9)$$

Note that more general distributions of the residuals $\boldsymbol{e}$ can be used, e.g., to allow for heteroscedasticity in runoff errors, but this issue is tangential to this paper. A critical observation in the context of inferring input errors is that both the full posterior (5) and the likelihood (9) are functions of the latent variables (rainfall multipliers) $\boldsymbol{\Phi}$.

3.3. Expected Likelihood Method

[14] If explicit estimates of the latent variables $\boldsymbol{\Phi}$ are not requested, Bayes theorem can be used to obtain the marginal posterior distribution of the model parameters $\boldsymbol{\theta}$ and input error hyperparameters $\boldsymbol{\eta}$ as follows:

$$p\left(\boldsymbol{\theta},\boldsymbol{\eta}|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}\right) \propto p\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta},\boldsymbol{\eta},\tilde{\boldsymbol{X}}\right)p\left(\boldsymbol{\theta},\boldsymbol{\eta}\right) \quad (10)$$

By definition, the marginal posterior (10) corresponds to integrating the latent variables out of the full posterior (5) using the total probability integral

$$\begin{aligned} p\left(\boldsymbol{\theta},\boldsymbol{\eta}|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}\right) &= \int p\left(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\Phi}|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}\right)d\boldsymbol{\Phi} \\ &\propto p\left(\boldsymbol{\theta},\boldsymbol{\eta}\right)\int p\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta},\boldsymbol{\Phi},\tilde{\boldsymbol{X}}\right)p\left(\boldsymbol{\Phi}|\boldsymbol{\eta}\right)d\boldsymbol{\Phi} \quad (11) \end{aligned}$$

It is then possible to define the "expected" likelihood as follows:

$$p\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta},\boldsymbol{\eta},\tilde{\boldsymbol{X}}\right) = \int p\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta},\boldsymbol{\Phi},\tilde{\boldsymbol{X}}\right)p\left(\boldsymbol{\Phi}|\boldsymbol{\eta}\right)d\boldsymbol{\Phi} \quad (12)$$

We stress the distinction between the likelihood $p(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}, \boldsymbol{\Phi}, \tilde{\boldsymbol{X}})$ in the full posterior (5) and the expected likelihood $p(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}, \boldsymbol{\eta}, \tilde{\boldsymbol{X}})$ in the marginal posterior (10). The latter requires integration over the T-dimensional distribution of latent variables. In general the T-dimensional integral (12) is analytically intractable and has to be approximated numerically (see *Huard and Mailhot* [2006] for an example of analytical treatment for a simple hydrological model).

[15] A simple numerical approximation of the expected likelihood using Monte Carlo integration is [*Kavetski et al.*, 2002]

$$p\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta},\boldsymbol{\eta},\tilde{\boldsymbol{X}}\right) \approx \frac{1}{N}\sum_{j=1}^{N} p\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta},\boldsymbol{\Phi}^{(j)},\tilde{\boldsymbol{X}}\right); \boldsymbol{\Phi}^{(j)} \leftarrow p\left(\boldsymbol{\Phi}|\boldsymbol{\eta}\right) \quad (13)$$

where $\boldsymbol{\Phi}^{(j)}$ is a multiplier set sampled from the hyperdistribution $p(\boldsymbol{\Phi}|\boldsymbol{\eta})$. As $N \to \infty$, the approximation converges to the integral (12).

[16] Although the full posterior (5) has T more dimensions than the marginal posterior (10), numerical analyses of both (5) and (10) require T-dimensional computation. Indeed, using the marginal posterior is generally more computationally expensive than using the full posterior, because the T-dimensional integration in the expected likelihood (12) has to be performed every time the marginal posterior distribution is evaluated.

[17] Furthermore, it is methodologically preferable to infer the multipliers through the full posterior distribution (as done in the BATEA approach) because it allows a much more thorough posterior analysis of the assumed input error model, including assessing the assumed hyperdistribution of the multipliers, the independence of multipliers, etc. [*Kuczera et al.*, 2006]. These checks are impossible if the multipliers are integrated out before sampling the posterior distribution.

[18] We now analyze two interpretations of the IBUNE method, focusing on its treatment of the marginal posterior (10) and on its relationship to MCMC methods for sampling from the full posterior (5).

4. IBUNE Interpretation A: Random Likelihood Function

[19] Interpretation A of the IBUNE method is motivated by the marginal posterior $p(\boldsymbol{\theta}, \boldsymbol{\eta}|\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$ stated in equation (9) of ADS2007 (identical to equation (10) in this paper). Under this interpretation, IBUNE-A does not infer the rainfall multipliers and instead samples "a random multiplier to each time step, drawn from the same normal distribution with unknown mean and variance" [ADS2007, p. 10].

[20] Assuming the residuals $\boldsymbol{e}$ are independent and follow a Gaussian distribution with zero mean and variance σ_y^2 (which is integrated out identically to equations (8)–(9)), the following IBUNE-A likelihood can be constructed on the basis of the description in ADS2007 (pp. 9–10)

$$p_{\Phi}\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta},m,\sigma^2,\tilde{\boldsymbol{X}}\right) \propto \left[\sum_{t=1}^{T}\left(e_t\left(\boldsymbol{\theta},\boldsymbol{\Phi},\tilde{\boldsymbol{X}}\right)\right)^2\right]^{-\frac{T}{2}}; \boldsymbol{\Phi} \leftarrow N\left(\boldsymbol{\Phi}|m,\sigma^2\right) \quad (14)$$

where the single set of multipliers $\boldsymbol{\Phi} = (\phi_1,\ldots,\phi_T)$ is sampled from the Gaussian hyperdistribution $N(\boldsymbol{\Phi}|m, \sigma^2)$.

[21] More generally, the relationship between the IBUNE-A likelihood and the likelihood $p(\tilde{\boldsymbol{Y}}|\ \boldsymbol{\theta},\ \boldsymbol{\Phi},\ \tilde{\boldsymbol{X}})$ in the full posterior (5) is

$$p_{\boldsymbol{\Phi}}\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta},\boldsymbol{\eta},\tilde{\boldsymbol{X}}\right) = p\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta},\boldsymbol{\Phi},\tilde{\boldsymbol{X}}\right); \boldsymbol{\Phi} \leftarrow p(\boldsymbol{\Phi}|\boldsymbol{\eta}) \qquad (15)$$

The IBUNE-A posterior is then

$$p_{\boldsymbol{\Phi}}\left(\boldsymbol{\theta},\boldsymbol{\eta},|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}\right) = p_{\boldsymbol{\Phi}}\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta},\boldsymbol{\eta},\tilde{\boldsymbol{X}}\right)p(\boldsymbol{\theta},\boldsymbol{\eta}) \qquad (16)$$

which closely resembles the marginal posterior (10).

[22] Equations (14)–(16) require further comment. Neither the IBUNE-A likelihood $p_{\boldsymbol{\Phi}}(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta},\ \boldsymbol{\eta},\ \tilde{\boldsymbol{X}})$, nor the IBUNE-A posterior $p_{\boldsymbol{\Phi}}(\boldsymbol{\theta},\ \boldsymbol{\eta}\ |\tilde{\boldsymbol{X}},\ \tilde{\boldsymbol{Y}})$, include the multipliers $\boldsymbol{\Phi}$ in their list of arguments (this is consistent with equation (9) in ADS2007 and the ensuing discussion, pp. 9–10). However, the right-hand side in equation (15) shows that a multiplier set $\boldsymbol{\Phi}$ is associated with the evaluation of the likelihood function because it is used to correct the observed inputs. Consequently, the implicit dependence of the IBUNE-A likelihood and posterior on the sampled multipliers is recorded using a subscript $\boldsymbol{\Phi}$.

4.1. IBUNE-A Likelihood and Posterior are Random Functions

[23] As indicated in equation (15), a new set of multipliers is randomly sampled every time the IBUNE-A likelihood is evaluated. However, the use of randomly sampled multipliers makes the IBUNE-A likelihood (15) a random function. In this paper, a function is termed "random" (nondeterministic) if it returns a random variable (i.e., the function value is not uniquely determined by its arguments). Conversely, a deterministic function returns a fixed value for a given set of arguments.

[24] In statistics, the likelihood function returns the value of a probability density (here, of $\tilde{\boldsymbol{Y}}$ given $\{\boldsymbol{\theta},\ \boldsymbol{\eta},\ \tilde{\boldsymbol{X}}\}$) and should evaluate to a constant for specific values of its arguments and the data (here, $\{\boldsymbol{\theta},\ \boldsymbol{\eta},\ \tilde{\boldsymbol{X}}\}$ and $\tilde{\boldsymbol{Y}}$). However, in the IBUNE-A formulation (15), the likelihood is a random function, since, given $\{\boldsymbol{\theta},\ \boldsymbol{\eta},\ \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}\}$, its value depends on the randomly sampled multipliers $\boldsymbol{\Phi}$. In turn, the IBUNE-A posterior (16) also becomes a random function. This violates a fundamental requirement for probability density functions.

4.2. Comparison With the Expected Likelihood Approach

[25] It is stressed that it is not the form of the posterior distribution (10) that gives rise to a random likelihood function. Indeed, as stated in section 3.3, posterior (10) is the marginal distribution of the full posterior (5), with the term $p(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta},\ \boldsymbol{\eta},\ \tilde{\boldsymbol{X}})$ being the expected likelihood (12). It is the evaluation of this expected likelihood term that is critical for the marginal posterior to be meaningfully (deterministically) defined.

[26] Consider the Monte Carlo approximation (13) of the expected likelihood. Comparison with the likelihood (15) shows that IBUNE-A is equivalent to a Monte Carlo approximation of the expected likelihood (12) using a single random sample ($N = 1$). However, such "single-sample" T-dimensional integration is unreliable and inaccurate, especially since T is proportional to the length of calibration data (e.g., 5 years of calibration data with daily multipliers yields $T \approx 1800$). The approximation error of the Monte Carlo integration is a random variable (dependent on the sampled $\boldsymbol{\Phi}$), which yields an alternative interpretation of the randomness of the IBUNE-A posterior.

4.3. Optimization of the IBUNE-A Posterior

[27] The nondeterministic nature of the posterior (which is the objective function in calibration) complicates parameter inference using IBUNE-A in several ways and reduces the range of tools available for analyzing its parameter distributions.

[28] The most likely model parameters, which are often used to generate operational model predictions and/or to initialize MCMC sampling, can be obtained by maximizing the objective function. However, a random function cannot be meaningfully maximized: (1) the maximum of a random function is itself a random variable (thus making irrelevant the concept of a point-estimate); and (2) numerical optimization methods will behave poorly if the objective function is noisy, especially if it returns different values when called with the same arguments. These problems are exacerbated as the variance of the multipliers σ^2 increases (and hence the sampled multipliers vary significantly from call to call). Consequently, it is difficult, if not impossible, to meaningfully maximize the IBUNE-A posterior using optimization methods, precluding its exploration using a basic analysis tool.

4.4. MCMC Analysis of the IBUNE-A Posterior

[29] Applying an MCMC sampling algorithm to the random IBUNE-A posterior (16) is also problematic. Indeed, the notion of sampling from a "random" pdf is mathematically undefined and leads to ambiguity when implemented using standard sampling algorithms.

[30] Consider a Metropolis sampler applied to the marginal posterior (10). At iteration i, a candidate sample $\{\boldsymbol{\theta}^{*(i)},\ \boldsymbol{\eta}^{*(i)}\}$ is generated from a symmetric jump distribution. The acceptance/rejection step is carried out next: the probability of accepting the candidate, i.e., setting $\{\boldsymbol{\theta}^{(i)},\ \boldsymbol{\eta}^{(i)}\} = \{\boldsymbol{\theta}^{*(i)},\ \boldsymbol{\eta}^{*(i)}\}$, is given by the acceptance ratio r_i, defined as

$$r_i = \frac{p\left(\boldsymbol{\theta}^{*(i)},\boldsymbol{\eta}^{*(i)}|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}\right)}{p\left(\boldsymbol{\theta}^{(i-1)},\boldsymbol{\eta}^{(i-1)}|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}\right)} \qquad (17)$$

If the candidate is rejected, the chain does not move and $\{\boldsymbol{\theta}^{(i)},\ \boldsymbol{\eta}^{(i)}\} = \{\boldsymbol{\theta}^{(i-1)},\ \boldsymbol{\eta}^{(i-1)}\}$.

[31] If the Metropolis method is applied to the random posterior (16), the acceptance ratio becomes

$$r_i = \frac{p_{\Phi_i\left(\eta^{*(i)}\right)}\left(\boldsymbol{\theta}^{*(i)},\boldsymbol{\eta}^{*(i)}|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}\right)}{p_{\Phi_i\left(\eta^{(i-1)}\right)}\left(\boldsymbol{\theta}^{(i-1)},\boldsymbol{\eta}^{(i-1)}|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}\right)} \qquad (18)$$

where $\boldsymbol{\Phi}_k(\boldsymbol{\eta})$ is the multiplier set sampled during the evaluation of $p_{\boldsymbol{\Phi}}(\boldsymbol{\theta},\ \boldsymbol{\eta}|\tilde{\boldsymbol{X}},\ \tilde{\boldsymbol{Y}})$ at Metropolis iteration k.

[32] Unlike the standard Metropolis ratio (17), it is not clear how to evaluate the denominator of the "random" Metropolis ratio (18). Indeed, the precise meaning of $\boldsymbol{\Phi}_i\ (\boldsymbol{\eta}^{(i-1)})$ in the subscript of the denominator is not evident. At least two strategies are possible:

[33] 1. Freshly reevaluate the random function in the denominator of (18) at each Metropolis iteration. In this case, $\boldsymbol{\Phi}_i(\boldsymbol{\eta}^{(i-1)}) \neq \boldsymbol{\Phi}_{i-1}(\boldsymbol{\eta}^{(i-1)})$; that is, the multipliers are resampled in both the numerator and the denominator terms of ratio (18) at every Metropolis iteration. Although perhaps counterintuitively, it could be argued that this treatment is consistent with the assumption in IBUNE-A that the multipliers are not part of the inference, and therefore are not accepted/rejected during iterations, but are sampled randomly at every occurrence of the likelihood function.

[34] 2. Save the density of the $(i - 1)$th sample after the acceptance/rejection test is carried out and use it in the denominator of the acceptance ratio at iteration i. This treatment is consistent with practical MCMC computer codes, which save the densities between their iterations because (for a deterministic target distribution) this saves function calls. Here, it implies that $\boldsymbol{\Phi}_i\,(\boldsymbol{\eta}^{(i-1)}) = \boldsymbol{\Phi}_{i-1}\,(\boldsymbol{\eta}^{(i-1)})$; that is, the random function in the denominator of (18) is not reevaluated at every iteration. Conceptually, this corresponds to the multiplier set $\boldsymbol{\Phi}^{*(i)} = \boldsymbol{\Phi}_i(\boldsymbol{\eta}^{*(i)})$ associated with the ith candidate $\{\boldsymbol{\theta}^{*(i)}, \boldsymbol{\eta}^{*(i)}\}$ being accepted/rejected along with the candidate sample itself.

[35] Since sampling from a random pdf is mathematically undefined, there are no theoretical grounds for favoring either of the two strategies outlined above. However, section 5 will show that strategy (2) is identical to a particular Metropolis-Hastings method for sampling from the full posterior $p(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\Phi}|\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$. Since this posterior is a standard (deterministic) pdf and is thus fundamentally distinct from the random IBUNE-A posterior (16), we defer analysis of strategy (2) to section 5, where it is treated as interpretation B of IBUNE. Consequently, empirical assessment of IBUNE-A (Text S1) uses strategy (1) in the MCMC sampler.

4.5. Convergence of IBUNE-A MCMC Samples

[36] This section compares the acceptance ratio corresponding to Metropolis sampling from the marginal posterior $p(\boldsymbol{\theta}, \boldsymbol{\eta}|\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$ to the acceptance ratio arising in IBUNE-A. Using the definition of the marginal posterior stated in equation (11) yields the ratio

$$
\begin{aligned}
r_i &= \frac{\int p\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}^{*(i)}, \boldsymbol{\Phi}, \tilde{\boldsymbol{X}}\right) p\left(\boldsymbol{\Phi}|\boldsymbol{\eta}^{*(i)}\right) d\boldsymbol{\Phi}}{\int p\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\Phi}, \tilde{\boldsymbol{X}}\right) p(\boldsymbol{\Phi}|\boldsymbol{\eta}^{(i-1)}) d\boldsymbol{\Phi}} \frac{p\left(\boldsymbol{\theta}^{*(i)}, \boldsymbol{\eta}^{*(i)}\right)}{p\left(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\eta}^{(i-1)}\right)} \\
&= \lambda^{(i)} \frac{p\left(\boldsymbol{\theta}^{*(i)}, \boldsymbol{\eta}^{*(i)}\right)}{p\left(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\eta}^{(i-1)}\right)}
\end{aligned}
\tag{19}
$$

Conversely, using the IBUNE-A posterior (16) yields

$$
\begin{aligned}
r_i &= \frac{p_{\Phi_i\left(\eta^{*(i)}\right)}\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}^{*(i)}, \boldsymbol{\eta}^{*(i)}, \tilde{\boldsymbol{X}}\right)}{p_{\Phi_i(\eta^{(i)})}\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\eta}^{(i-1)}, \tilde{\boldsymbol{X}}\right)} \frac{p\left(\boldsymbol{\theta}^{*(i)}, \boldsymbol{\eta}^{*(i)}\right)}{p\left(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\eta}^{(i-1)}\right)} \\
&= \lambda_{\Phi}^{(i)} \frac{p\left(\boldsymbol{\theta}^{*(i)}, \boldsymbol{\eta}^{*(i)}\right)}{p\left(\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\eta}^{(i-1)}\right)}
\end{aligned}
\tag{20}
$$

Comparison of equations (19) and (20) shows that IBUNE-A leads to the correct acceptance ratio for sampling from the marginal posterior (10) only if the ratio of (random) IBUNE-A likelihoods $\lambda_{\Phi}^{(i)}$ is close to the ratio of the (deterministic) expected likelihoods $\lambda^{(i)}$, which in general is not true. It is hence unclear whether MCMC samplers applied to IBUNE-A converge to a stationary distribution, and whether this stationary distribution, if it exists, is the marginal posterior (10). Indeed the case study shows (see section 4.1 of Text S1) the susceptibility of IBUNE-A to convergence difficulties.

4.6. Empirical Analysis of IBUNE-A and BATEA

[37] The ability of IBUNE-A to operate under input uncertainty was empirically examined using a synthetic problem with no model error and specified "true" inputs and "true" parameters. By excluding model error, we focused on the primary issue of this study, namely the treatment of input uncertainty. True inputs were corrupted using the multiplier model (2)–(3). True outputs were generated by propagating the true inputs through the conceptual rainfall-runoff model LogSPM [*Kuczera et al.*, 2006] and were then corrupted using Gaussian noise.

[38] The IBUNE-A and BATEA methods were then used to estimate the CRR parameters and the predictive uncertainty in the runoff using the same corrupted input/output data and the same input/output uncertainty models (the rainfall multiplier model (2)–(3) and Gaussian output noise). It is stressed that the same statistical models were used and the only difference was the treatment of latent variables in the likelihood function (with BATEA working directly with the full posterior (5) and explicitly estimating the latent variables). The same MCMC sampling strategy (multiblock Metropolis scheme with Gaussian jump distribution) was used for both the IBUNE-A and BATEA inferences. A full description of the case study and results can be found in Text S1.

[39] Figure S4 shows the CRR parameters estimated using BATEA and IBUNE-A from 1000 days of data corrupted by rainfall multipliers ϕ_0, with $\log\phi_0 \sim N(m_0, \sigma_0^2)$, with mean $m_0 = -0.4$ and a range of standard deviations σ_0 (increasing from 0.002 up to 0.8). It was found that IBUNE-A produced CRR parameter estimates with rapidly increasing posterior variances. Similar results were found for the estimated hypermean of rainfall log multipliers (Figure S3). Figure S3 also shows that IBUNE-A underestimated the standard deviation of the rainfall multipliers for $\sigma_0 \geq 0.02$. In contrast, BATEA identified the CRR parameters and rainfall error hyperparameters accurately and precisely regardless of the magnitude of random corruption σ_0. Indeed, Figure S3 shows that the BATEA estimates agree well with the true values and have a low posterior uncertainty.

[40] An important characteristic of an inference scheme is correct attribution of predictive uncertainty. Figure S4 shows a decomposition of predictive uncertainty in the case of calibrating 4 CRR parameters to 1000 days of data corrupted by lognormal rainfall multipliers with $m_0 = -0.4$ and $\sigma_0 = 0.2$. The first row of Figure S4 compares the full predictive interval with a partial predictive interval that ignores posterior CRR parameter uncertainty (obtained by fixing all CRR parameters to their modal values). In the BATEA results, the two intervals are almost identical, implying that in this case study CRR parameter uncertainty contributes very little to predictive uncertainty. This is not surprising given the long calibration period and demon-

strates the increased precision of Bayesian posterior pdf's as the data length is increased (in this synthetic study, we can ascertain that the parameter estimates are accurate). A very different result is observed for IBUNE-A: during low-flow periods, about half of the uncertainty is attributed to CRR parameter uncertainty.

[41] In the second row of Figure S4, a more restrictive partial interval is shown, which also excludes predictive uncertainty due to the random errors in the observed rainfalls (it was obtained by fixing the CRR parameters at their modal values and setting $\sigma = 0$). In the BATEA case, the narrowness of this predictive interval during high flows shows that the predictive uncertainty is dominated by rainfall uncertainty. However, in the IBUNE-A case, random rainfall errors contribute minimally to the overall predictive uncertainty during both recessions and high flows. This is because IBUNE-A significantly underestimates the standard deviation of rainfall errors (Figure S3).

[42] It is noted that despite misidentifying the sources of errors contributing to the overall predictive uncertainty, IBUNE-A produced 90% predictive uncertainty bounds that captured about 90% of the observed runoffs. However, the previous analysis shows that it compensated for a major underestimation of random rainfall errors by increasing the uncertainty in the CRR parameters and in the mean of the rainfall multipliers *m* (this was verified because the study used known synthetic data). This shows that enveloping the correct fraction of observations in the prediction limits is not a sufficient condition for these limits to be statistically meaningful and insightful.

5. IBUNE Interpretation B: Using the Hyperdistribution as a Jump Distribution

[43] This section analyzes an interpretation of the IBUNE method that avoids most of the problems of interpretation A, but operates in the full inference space including the latent variables (and hence does not offer dimensionality-reduction benefits). Interpretation B is based on a particular MCMC scheme for sampling from the full posterior distribution (5).

5.1. IBUNE-B: Markov-Independent MCMC Sampler

[44] Consider an MCMC scheme for sampling from the full posterior (5) using the composite jump distribution

$$\begin{aligned}Q&\left(\boldsymbol{\theta}^{*(i)},\boldsymbol{\eta}^{*(i)},\boldsymbol{\Phi}^{*(i)}|\boldsymbol{\theta}^{(i-1)},\boldsymbol{\eta}^{(i-1)},\boldsymbol{\Phi}^{(i-1)}\right)\\&=q\left(\boldsymbol{\theta}^{*(i)},\boldsymbol{\eta}^{*(i)}|\boldsymbol{\theta}^{(i-1)},\boldsymbol{\eta}^{(i-1)}\right)p\left(\boldsymbol{\Phi}^{*(i)}|\boldsymbol{\eta}^{*(i)}\right)\end{aligned}\tag{21}$$

to propose a candidate sample $\{\boldsymbol{\theta}^{*(i)}, \boldsymbol{\eta}^{*(i)}, \boldsymbol{\Phi}^{*(i)}\}$ at iteration *i*.

[45] Drawing from the jump distribution (21) consists of two steps: (1) Markov step: draw a candidate sample $\{\boldsymbol{\theta}^{*(i)}, \boldsymbol{\eta}^{*(i)}\}$ of CRR parameters and hyperparameters from the jump distribution $q(\boldsymbol{\theta}^{*(i)}, \boldsymbol{\eta}^{*(i)}|\boldsymbol{\theta}^{(i-1)}, \boldsymbol{\eta}^{(i-1)})$, which depends on the previous sample (the SCEM algorithm [*Vrugt et al.*, 2003] can be used in this step); and (2) Independent step: draw a set of rainfall multipliers $\boldsymbol{\Phi}^{*(i)}$ from the hyperdistribution $p(\boldsymbol{\Phi}|\boldsymbol{\eta}^{*(i)})$, which does not directly depend on the previous sample. Note that the acceptance/rejection step is performed on the entire proposed sample $\{\boldsymbol{\theta}^{*(i)}, \boldsymbol{\eta}^{*(i)}, \boldsymbol{\Phi}^{*(i)}\}$, rather than separately for the Markov and Independent steps.

[46] The Metropolis-Hastings acceptance probability of a candidate sample from the jump distribution (21) is

$$\begin{aligned}r_i=&\frac{p\left(\boldsymbol{\theta}^{*(i)},\boldsymbol{\eta}^{*(i)},\boldsymbol{\Phi}^{*(i)}|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}\right)}{p\left(\boldsymbol{\theta}^{(i-1)},\boldsymbol{\eta}^{(i-1)},\boldsymbol{\Phi}^{(i-1)}|\tilde{\boldsymbol{X}},\tilde{\boldsymbol{Y}}\right)}\\&\times\frac{Q\left(\boldsymbol{\theta}^{(i-1)},\boldsymbol{\eta}^{(i-1)},\boldsymbol{\Phi}^{(i-1)}|\boldsymbol{\theta}^{*(i)},\boldsymbol{\eta}^{*(i)},\boldsymbol{\Phi}^{*(i)}\right)}{Q\left(\boldsymbol{\theta}^{*(i)},\boldsymbol{\eta}^{*(i)},\boldsymbol{\Phi}^{*(i)}|\boldsymbol{\theta}^{(i-1)},\boldsymbol{\eta}^{(i-1)},\boldsymbol{\Phi}^{(i-1)}\right)}\end{aligned}\tag{22}$$

Substituting equations (5) and (21) into (22) yields

$$\begin{aligned}r_i=&\frac{p\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}^{*(i)},\boldsymbol{\Phi}^{*(i)},\tilde{\boldsymbol{X}}\right)p\left(\boldsymbol{\Phi}^{*(i)}|\boldsymbol{\eta}^{*(i)}\right)p\left(\boldsymbol{\theta}^{*(i)},\boldsymbol{\eta}^{*(i)}\right)}{p\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}^{(i-1)},\boldsymbol{\Phi}^{(i-1)},\tilde{\boldsymbol{X}}\right)p\left(\boldsymbol{\Phi}^{(i-1)}|\boldsymbol{\eta}^{(i-1)}\right)p\left(\boldsymbol{\theta}^{(i-1)},\boldsymbol{\eta}^{(i-1)}\right)}\\&\times\frac{q\left(\boldsymbol{\theta}^{(i-1)},\boldsymbol{\eta}^{(i-1)}|\boldsymbol{\theta}^{*(i)},\boldsymbol{\eta}^{*(i)}\right)p\left(\boldsymbol{\Phi}^{(i-1)}|\boldsymbol{\eta}^{(i-1)}\right)}{q\left(\boldsymbol{\theta}^{*(i)},\boldsymbol{\eta}^{*(i)}|\boldsymbol{\theta}^{(i-1)},\boldsymbol{\eta}^{(i-1)}\right)p\left(\boldsymbol{\Phi}^{*(i)}|\boldsymbol{\eta}^{*(i)}\right)}\end{aligned}\tag{23}$$

Cancelling the terms $p(\boldsymbol{\Phi}^{*(i)}|\boldsymbol{\eta}^{*(i)})$ and $p(\boldsymbol{\Phi}^{(i-1)}|\boldsymbol{\eta}^{(i-1)})$, and assuming the jump distribution in the Markov step is symmetric yields

$$r_i=\frac{p\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}^{*(i)},\boldsymbol{\Phi}^{*(i)},\tilde{\boldsymbol{X}}\right)p\left(\boldsymbol{\theta}^{*(i)},\boldsymbol{\eta}^{*(i)}\right)}{p\left(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}^{(i-1)},\boldsymbol{\Phi}^{(i-1)},\tilde{\boldsymbol{X}}\right)p\left(\boldsymbol{\theta}^{(i-1)},\boldsymbol{\eta}^{(i-1)}\right)}=\frac{\gamma^{*(i)}}{\gamma^{(i-1)}}\tag{24}$$

This Markov-Independent (MI) algorithm is a Metropolis-Hastings method [*Gelman et al.*, 1995] that asymptotically produces samples from the full posterior (5).

[47] Equation (24) shows that the MI sampler does not require evaluating the full posterior (5). Indeed, iteration *i* of the MI method consists of the following steps:

[48] 1. Draw a candidate sample of CRR parameters and input error hyperparameters $\{\boldsymbol{\theta}^{*(i)}, \boldsymbol{\eta}^{*(i)}\}$ from a symmetric jump distribution.

[49] 2. Evaluate the quantity $\gamma^{*(i)} = p(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}^{*(i)}, \boldsymbol{\Phi}^{*(i)}, \tilde{\boldsymbol{X}})\, p(\boldsymbol{\theta}^{*(i)}, \boldsymbol{\eta}^{*(i)})$ as follows: (1) Draw a candidate set of multipliers $\boldsymbol{\Phi}^{*(i)}$ from the hyperdistribution $p(\boldsymbol{\Phi}|\boldsymbol{\eta}^{*(i)})$; (2) Run the CRR model with parameters $\boldsymbol{\theta}^{*(i)}$, forcing it with rainfalls corrected using $\boldsymbol{\Phi}^{*(i)}$, and evaluate the residuals $\boldsymbol{e}^{(i)}$; (3) Evaluate the likelihood (9) using residuals $\boldsymbol{e}^{(i)}$ and calculate $\gamma^{*(i)}$.

[50] 3. Compute the acceptance ratio (24) using $\gamma^{*(i)}$ and carry out the acceptance/rejection test. The denominator is evaluated using the quantity $\gamma^{(i-1)}$ corresponding to the $(i-1)$th sample.

[51] Step 2 is identical to the evaluation of the IBUNE posterior described in ADS2007 (indeed, $\gamma(\boldsymbol{\theta}, \boldsymbol{\eta})$ corresponds to $p_{\boldsymbol{\Phi}}(\boldsymbol{\theta}, \boldsymbol{\eta}\,|\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$). Therefore, the MI algorithm (1)–(3) is equivalent to a Metropolis sampler (such as SCEM) applied directly to the IBUNE posterior (16). This makes the MI sampler a plausible interpretation of the IBUNE description in ADS2007.

5.2. Dimensional Complexity of IBUNE-B

[52] Although IBUNE-B does not require evaluating the full posterior $p(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\Phi}|\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$ it is stressed that it operates in the full dimensional $\{\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\Phi}\}$ space including the latent variables $\boldsymbol{\Phi}$, rather than in the much lower-dimensional

$\{\boldsymbol{\theta}, \boldsymbol{\eta}\}$ space of $p(\boldsymbol{\theta}, \boldsymbol{\eta} | \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$. Consequently, the sampling process must converge not only in the $\{\boldsymbol{\theta}, \boldsymbol{\eta}\}$ space of the Markovian jump, but also in the T-dimensional $\boldsymbol{\Phi}$ space of the Independent step.

[53] It follows that IBUNE-B offers no dimensionality-reduction benefits compared to BATEA-type methods applied directly to the full posterior (5). Indeed, IBUNE-B differs from the MCMC samplers currently used in BATEA solely in its choice of jump distributions. To date, BATEA has been applied with single-block Gaussian jump distributions [*Kavetski et al.*, 2006a, 2006b] and with a multiblock Gibbs sampler [*Kuczera et al.*, 2007], while IBUNE-B utilizes the Markov-Independent jump rule.

[54] Also note that not storing the multipliers $\{\boldsymbol{\Phi}^{(i)}, i = 1 \ldots N\}$ sampled during N Metropolis iterations does not affect the computational speed/results of the algorithm, but saves $4NT$ bytes of computer memory (e.g., 60 MB for 10,000 samples when calibrating to 5 years with daily multipliers in 4-byte (single) precision). This adjustment can be implemented for any MCMC sampler, but makes it difficult to apply posterior diagnostics. Importantly, it can also undermine MCMC convergence checks (since convergence in $\boldsymbol{\Phi}$ space cannot be ascertained).

5.3. Relationship with IBUNE-A

[55] The IBUNE-B acceptance ratio (24) is similar to the acceptance ratio (20) derived by applying a Metropolis algorithm to the IBUNE-A posterior (16). More precisely, the equality of these ratios depends on the implementation of the IBUNE-A Metropolis sampler, with IBUNE-B corresponding to strategy 2 outlined in section 4.4.

[56] Therefore, the computation and samples produced using IBUNE-B are identical to those obtained with a Metropolis sampler (implementing strategy 2, which is standard in MCMC computer codes) applied to the random posterior (16). This interesting result shows that applying a Metropolis algorithm to the "random marginal posterior" (16) is equivalent to sampling from the full posterior (5) using the Markov-Independent method. However, the sampling unavoidably operates in the full space including the latent variables (multipliers), and therefore using a random pdf does not reduce the dimensionality of the sampling space.

5.4. Computational Efficiency of IBUNE-B

[57] In general, the efficiency of MCMC methods depends on how closely the jump distribution resembles the target distribution (here, the full posterior (5)) [*Gelman et al.*, 1995]. This requires either special insight into the target distribution, or some tuning of the jump distribution (e.g., using adaptive schemes such as SCEM).

[58] The MI method (i.e., IBUNE-B) uses the hyperdistribution $p(\boldsymbol{\Phi}|\boldsymbol{\eta}^{*(i)})$ as an importance function to sample from the target distribution of the Independent step, $p(\boldsymbol{\Phi}|\boldsymbol{\theta}^{*(i)}, \boldsymbol{\eta}^{*(i)}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$. The hyperdistribution is a poor choice for sampling candidate multipliers $\boldsymbol{\Phi}^{*(i)}$ for several reasons:

[59] 1. The hyperdistribution is off-centered relative to, and is usually wider than, the target distribution. Indeed, (1) the posterior mean of a multiplier ϕ_t can be very different from its hypermean m (e.g., if the rainfall error at day t is large); and (2) $p(\boldsymbol{\Phi}|\boldsymbol{\theta}, \boldsymbol{\eta}, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$ is generally peakier than $p(\boldsymbol{\Phi}|\boldsymbol{\eta})$ because it is conditioned on the data and the CRR parameters. Figure S13 illustrates the difference between the individual multipliers posterior distributions and the hyperdistribution: some multipliers have a posterior distribution which is significantly off-centered and much peakier than the hyperdistribution. This occurs for multipliers that significantly affect the model predictions and can be accurately identified from the data.

[60] 2. The hyperdistribution does not depend on the current location of the MCMC chain (in this respect, it is closely related to standard importance sampling). Therefore, unlike "Markovian" jump distributions (e.g., Gaussians centered on the current sample), the hyperdistribution will not favor the chains to remain and move within high posterior probability regions in the $\boldsymbol{\Phi}$ space (once these are reached).

[61] 3. Once the Markov chains reach higher-density regions of the $\{\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\Phi}\}$ space, a random T-dimensional multiplier set proposed from the exceedingly wide and off-centered hyperdistribution is likely to significantly degrade the current posterior density and hence has a very low acceptance probability. This leads to exceedingly low jump rates.

[62] 4. The hyperdistribution is fixed for a given $\boldsymbol{\eta}^{*(i)}$ and does not contain tunable parameters. Therefore the Independent step cannot be tuned to its target distribution.

[63] The MI sampler is therefore likely to be inefficient, with a continuously declining jump rate. Adaption of the Markov step, e.g., using SCEM as done in ADS2007, cannot remedy this inefficiency because the Markov step operates in the $\{\boldsymbol{\theta}, \boldsymbol{\eta}\}$ space, whereas the inefficiency occurs in the $\boldsymbol{\Phi}$ space. Moreover, even if the Markov step is able to identify and sample near-optimal values of $\boldsymbol{\eta}$, it still cannot resolve the inefficiency of the Independent step. Indeed, even if the true value $\boldsymbol{\eta}_0$ was known, the hyperdistribution $p(\boldsymbol{\Phi}|\boldsymbol{\eta}_0)$ would generally be a poor approximation of $p(\boldsymbol{\Phi}|\boldsymbol{\theta}, \boldsymbol{\eta}_0, \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$.

[64] The case study (section 5.1 in Text S1) shows that the jump rate of the MI sampler quickly declines to almost zero. For example, during a 100,000-iteration simulation, the jump rate declined from 7% during the first 100 samples down to 0.005% over the last 90,000 iterations. In addition, the sampled multipliers $\boldsymbol{\Phi}$ are rarely close to their modal (most likely) values $\hat{\boldsymbol{\Phi}}$. Indeed, this would require randomly sampling a specific T-dimensional Gaussian deviate ($\hat{\boldsymbol{\Phi}}$ within some tolerance), which becomes exceedingly improbable if T is large. For example, if the multiplier hypermean m is 1, the probability of randomly sampling a vector $\log\boldsymbol{\Phi}^{(i)}$ with all components having the same signs as $\log\hat{\boldsymbol{\Phi}}$ is $(1/2)^T$. Consequently, while the MI approach is asymptotically convergent in theory, its convergence is exceedingly slow for practical computation, unless, as discussed in section 6, σ is constrained near zero.

5.5. Fixed-Pool MI Method

[65] Consider an implementation of the MI method that pregenerates a fixed finite pool of standardized multiplier sets. This corresponds to the practical implementation used by ADS2007 (Ajami, personal communication, 2007). More precisely, a fixed pool of M standard Gaussian deviates $(\boldsymbol{\Omega}^{(k)})_{k=1\ldots M} = (\omega_{\mathrm{t}}^{(k)})_{k=1\ldots M;\, t=1\ldots T}$ is generated prior to MCMC sampling. At MCMC iteration *i,* the multipliers are obtained as $\boldsymbol{\Phi}^{*(i)} = m^{*(i)} + \boldsymbol{\Omega}^{(k)}\boldsymbol{\sigma}^{*(i)}$, where $\boldsymbol{\eta}^{*(i)} = (m^{*(i)}, \sigma^{*(i)})$ are the hyperparameters proposed in the

Markov step and $k = i \bmod M$ (i.e., k is the remainder of i/M). The number of MCMC samples N exceeds M, so the algorithm cycles through the same pregenerated standardized multipliers $\mathbf{\Omega}$ every M iterations (in the same sequential order). If $M \geq N$, the fixed-pool MI method reduces to the standard MI method.

[66] The empirical study (section 5.2 in Text S1) suggests that using a fixed pool, especially with $M \ll N$, improves the jump rate. While the precise reasons for this are unclear, it was observed that (1) the fixed-pool algorithm will always eventually resample the same multiplier sets (but with potentially improved candidate $\{\boldsymbol{\theta}^{*(i)}, \boldsymbol{\eta}^{*(i)}\}$): indeed, in the case study, 98% of the jumps were obtained with the same member $\mathbf{\Omega}^{(k_0)}$ of the fixed pool; and (2) the posterior density is never close to its near-optimal values because the probability that a random M pool of multipliers contains near-optimal multipliers is very low. Consequently, the jump probability r increases as M decreases, and the fixed-pool method is "jumpier" than the standard MI method (which is equivalent to $M \rightarrow \infty$).

[67] However, the fixed-pool strategy loses the asymptotic convergence of the standard MI method because a $\mathbf{\Phi}$ pool with finite $M < N$ cannot provide the complete coverage of the $\mathbf{\Phi}$ space (especially when T is large) required for asymptotic convergence as $N \rightarrow \infty$. It is equivalent to sampling the multipliers using a random number generator with a smaller period (M) than the number of Monte Carlo samples (N).

[68] Note that adequate jump rates and numerical convergence of the fixed-pool MI method to a stationary distribution do not imply that the samples correspond to the target full posterior (5) (see *Gelman et al.* [1995] for the distinction between the stationary and target distributions of an MCMC method). Indeed, Figure S12 shows that the fixed-pool MI method with $M < N$ does not accurately approximate the correct posterior and the true parameter values. It shows that the marginal pdf estimated using the fixed-pool MI method differs significantly from the correct marginal posterior and does not encompass the known true value of sK (see also section 5.2 in Text S1).

6. Using IBUNE With the Assumption of Small Random Errors in the Inputs

[69] The prior distributions used by ADS2007 correspond to relatively small rainfall uncertainty. Indeed, the priors $m \sim U\,[0.9, 1.1]$ and $\sigma^2 \sim U\,[10^{-5}, 10^{-3}]$ (ADS2007, p. 9) correspond to a systematic bias of up to 10%, but almost negligible random errors (with a standard error not exceeding $\sigma_{\max} = \surd\,(10^{-3}) \approx 3\%$). Note the difference between inputs biased by large systematic errors (m deviates significantly from 1, but σ is small) and unbiased inputs with large random errors ($m \approx 1$, but σ is large). A large standard error implies that the observed rainfalls are corrupted by large random errors, whereas a large mean error implies that the observations contain large systematic errors (i.e., the observations are systematically biased).

[70] Assuming input uncertainty with very small random errors (i.e., constraining $\sigma \approx 0$) improves the computational behavior of both interpretations of IBUNE. First, the random likelihood in IBUNE-A becomes near-deterministic because the sampled multipliers are then near-identical at each evaluation of the likelihood with a given set of arguments. Second, IBUNE-B becomes more efficient because both the hyperdistribution (used in the jump distribution) and the marginal posterior density of the multipliers converge to the same Dirac function as $\sigma_{\max} \rightarrow 0$. More generally, IBUNE behavior improves as σ decreases, but does not depend on m.

[71] However, the assumption that random errors in the observed inputs are small is very restrictive in hydrological calibration, especially given the spatial variability of rainfall fields. For example, *Linsley and Kohler* [1958] report that standard errors of storm-depth estimates may exceed 30%, especially for large sparsely gauged catchments.

[72] If the actual standard deviation σ_0 of rainfall multipliers exceeds the upper bound $\sigma_{\max}$ assumed a priori to the inference, the prior becomes incompatible with the data and excludes regions of the posterior that otherwise would be highly probable. Any Bayesian method may then appear to behave well computationally, but will converge to poor posterior estimates (because input uncertainty is underestimated). Empirical analysis suggests two problems when σ_0 is large but σ is constrained near zero: (1) parameters are estimated with significant biases; and (2) the uncertainty in the predicted runoffs is underestimated.

[73] In the ADS2007 study, IBUNE inferred the standard deviation of rainfall multipliers to be in the range [0.003, 0.01] (Figure 10 in ADS2007). This corresponds to a standard error in the observed daily catchment rainfall of about 1%, which, as discussed above, appears unduly optimistic.

7. Discussion

[74] A major motivation for IBUNE was overcoming two drawbacks of the BATEA approach reported by ADS2007. This section considers these criticisms in more detail to clarify the key aspects of Bayesian hierarchical inference.

[75] The first reported drawback was that "it is impossible to assess the input error model likelihood $L(\tilde{\boldsymbol{X}}|\boldsymbol{X})$ because the true inputs are unknown in practice." Here, $L(\tilde{\boldsymbol{X}}|\boldsymbol{X})$ refers to the notation used by *Kavetski et al.* [2002] to denote the probability of observing rainfall $\tilde{\boldsymbol{X}}$ if the true rainfall is $\boldsymbol{X}$. Using $L(\tilde{\boldsymbol{X}}|\boldsymbol{X})$ in the BATEA inference equations does not imply that the true inputs must be known (indeed, they are estimated as part of the BATEA inference). Rather, $L(\tilde{\boldsymbol{X}}|\boldsymbol{X})$ is used as a function of $\boldsymbol{X}$ and can be constructed given an assumed input error model, e.g., the Gaussian multipliers (2)–(3). Moreover, the input error model represented by $L(\tilde{\boldsymbol{X}}|\mathbf{X})$ can and should be scrutinized a posteriori, e.g., by examining whether the calibrated multipliers are consistent with their hyperdistribution [*Kavetski et al.*, 2006b; *Kuczera et al.*, 2006]. Finally, while *Kavetski et al.* [2002] assumed the hyperparameters $\boldsymbol{\eta}$ (here, mean and variance of the multipliers) are known, they can be added to the BATEA inference and themselves estimated from the data [*Kavetski et al.*, 2006a; *Kuczera et al.*, 2006].

[76] The second reported drawback was the high dimensionality of the BATEA posterior (5) when input errors are characterized using a large number of latent variables. This occurs for long calibration time periods, or for high temporal resolution of rainfall uncertainty (e.g., if BATEA is applied with daily, rather than storm event multipliers). The high dimensionality of the Bayesian hierarchical model

(2)–(5) certainly poses a significant computational challenge. However, it is unavoidable. Although working with the marginal posterior (10) reduces the apparent dimension of the objective function, section 3.3 shows that this merely exchanges a single *T*-dimensional sampling problem for a series of *T*-dimensional integration problems. Yet numerical exploration of the full posterior is by no means prohibitive. For example, *Kavetski et al.* [2006b] sampled the full posterior using a standard Metropolis scheme in a real-data calibration involving 3 years of data (~70 storm-based multipliers), while *Kuczera et al.* [2006] estimated the posterior parameter mode using a quasi-Newton method in a real-data calibration involving 2 years of daily data (~150 storm-based multipliers).

[77] Note that a single evaluation of the full posterior (5) is relatively cheap, only slightly more expensive than a single evaluation of the common least squares objective function. The primary challenges of sampling from high-dimensional pdf's are (1) a large number of samples may be needed to explore a high-dimensional sampling space; and (2) the difficulties encountered by MCMC sampling of pdf's with complicated shapes are exacerbated when the dimension of the sampling space increases. It is stressed that the intrinsic *T*-dimensional structure of the Bayesian hierarchical model does not rule out the existence of more efficient (but *T*-dimensional!) numerical sampling algorithms, nor the use of fast analytical or numerical solution methods whenever possible.

[78] Finally this study highlights that IBUNE is based on the same Bayesian hierarchical conceptualization of the input uncertainty as BATEA and therefore is subject to the same issues regarding the statistical and computational complexity of the inference problem, sensitivity to assumed models of data uncertainty, assumed timescale of input errors, prior specifications and posterior diagnostics.

8. Conclusions

[79] The IBUNE method proposed by ADS2007 accounts for input uncertainties in hydrological calibration using a Bayesian hierarchical model, with input errors represented using latent variables (rainfall multipliers). Unlike BATEA, IBUNE does not explicitly infer the multipliers and instead samples them from their assumed hyperdistribution at each evaluation of the likelihood function. Though intuitive, this creates several theoretical and practical difficulties. This paper considered two different interpretations of the IBUNE method, analyzed their theoretical and practical properties, and compared them with BATEA.

[80] Interpretation A is based on the ADS2007 description that the approach does not estimate the rainfall multipliers and uses a single multiplier sample in the evaluation of the likelihood function. In this case the IBUNE likelihood, and hence the IBUNE posterior and objective functions, become random functions of the inferred variables, which violates a general requirement for probability density functions. IBUNE-A can also be viewed as an approximation to the marginal posterior distribution of the CRR parameters and input error hyperparameters, which requires integrating the full likelihood over the hyperdistribution of latent variables. However, it is shown that IBUNE-A is equivalent to a single-sample Monte Carlo integration and is numerically inaccurate. It is empirically shown to cause a significant underestimation of the rainfall multiplier variance and a significant misidentification of the sources contributing to the uncertainty in the predicted runoffs.

[81] In Interpretation B, the IBUNE method is related to a special two-step Metropolis-Hastings scheme for sampling from the full posterior including the rainfall multipliers. IBUNE-B is an asymptotically convergent MCMC method, but does not reduce the dimensionality of the problem compared to BATEA-type methods. Indeed, IBUNE-B differs from MCMC methods currently used in BATEA solely in its jump distribution: it uses the hyperdistribution as a jump distribution for the latent variables. Yet this jump distribution is inefficient because it is a poor approximation of the target distribution and leads to exceedingly low jump rates. Modifications based on drawing from a pregenerated pool of multiplier samples can raise jump rates, but the method then no longer converges to the correct posterior distribution.

[82] The behavior of IBUNE (both A and B) improves if the variance of the multipliers is a priori constrained to be small, which is appropriate if the input errors are known to be largely systematic (rather than random). In this case, the random variability in the IBUNE-A posterior decreases, while the hyperdistribution becomes more adequate as a jump distribution in IBUNE-B. However, the evidence suggests that input errors can be considerable.

[83] A primary conclusion of this study is that, unless the hydrological model and the structure of data uncertainty allow specialized treatment, Bayesian hierarchical models of input uncertainty invariably lead to *T*-dimensional computational problems, whether working with the full posterior (*T*-dimensional sampling problem) or with the marginal posterior (*T*-dimensional integration problem each time the marginal posterior is evaluated).

[84] Finally retaining the latent variables in the posterior distribution (rather than integrating them out) provides valuable posterior diagnostic information about the hypotheses used to construct the hyperdistribution (e.g., independence and distribution of input errors). This is not possible for methods that integrate the multipliers out of the posterior distribution. In this context, the high dimensionality of BATEA-type objective functions is not a theoretical deficiency, but rather a computational challenge.

[85] **Acknowledgments.** The authors thank Newsha Ajami, Qingyun Duan, and Soroosh Sorooshian for helpful discussions and for providing the code implementing the IBUNE approach. The valuable comments by Paul Smith, Peter Reichert, and an anonymous reviewer significantly improved this paper and are gratefully acknowledged.

References

Ajami, N. K., Q. Y. Duan, X. G. Gao, and S. Sorooshian (2006), Multimodel combination techniques for analysis of hydrological simulations: Application to Distributed Model Intercomparison Project results, *J. Hydrometeorol.*, *7*(4), 755–768, doi:10.1175/JHM519.1.

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

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995), *Bayesian Data Analysis*, 526 pp. pp., CRC Press, Boca Raton, Fla.

Huard, D., and A. Mailhot (2006), A Bayesian perspective on input uncertainty in model calibration: Application to hydrological model "abc", *Water Resour. Res.*, *42*, W07416, doi:10.1029/2005WR004661.

Kavetski, D., S. Franks, and G. Kuczera (2002), Confronting input uncertainty in environmental modelling in calibration of watershed models, in

Water Sci. Appl. Ser., vol. 6, edited by Q. Y. Duan et al., pp. 49–68, AGU, Washington, D.C.

Kavetski, D., G. Kuczera, and S. W. Franks (2006a), Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, *Water Resour. Res.*, *42*, W03407, doi:10.1029/2005WR004368.

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

Kuczera, G., D. Kavetski, S. Franks, and M. Thyer (2006), Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters, *J. Hydrol.*, *331*(1–2), 161–177, doi:10.1016/j.jhydrol.2006.05.010.

Kuczera, G., D. Kavetski, B. Renard, and M. Thyer (2007), Bayesian total error analysis for hydrologic models: Markov Chain Monte Carlo methods to evaluate the posterior distribution, in *MODSIM 2007 International Congress on Modelling and Simulation*, edited by L. Oxley and D. Kulasiri, pp. 2466–2472, Modell. amd Simulation Soc. of Australia and N. Z., Christchurch, N. Z.

Linsley, R. K., and M. A. Kohler (1958), *Hydrology for Engineers*, 340 pp., McGraw Hill, New York.

Renard, B., M. Thyer, G. Kuczera, and D. Kavetski (2007), Bayesian total error analysis for hydrologic models: Sensitivity to error models, in *MODSIM 2007 International Congress on Modelling and Simulation*, edited by L. Oxley and D. Kulasiri, pp. 2473–2479, Modell. amd Simulation Soc. of Australia and N. Z., Christchurch, N. Z.

Vrugt, J. A., and B. A. Robinson (2007), Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging, *Water Resour. Res.*, *43*, W01411, doi:10.1029/2005WR004838.

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

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

D. Kavetski and G. Kuczera, School of Engineering, University of Newcastle, Callaghan, NSW 2308, Australia.

B. Renard, Cemagref, UR HHLY, 3 bis quai Chauveau, CP 220, F-69336 Lyon, France. (benjamin.renard@cemagref.fr)