Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis

Mark Thyer,[1] Benjamin Renard,[1] Dmitri Kavetski,[1] George Kuczera,[1] Stewart William Franks,[1] and Sri Srikanthan[2]

Received 11 January 2008; revised 2 October 2008; accepted 11 December 2008; published 1 April 2009.

[1] The lack of a robust framework for quantifying the parametric and predictive uncertainty of conceptual rainfall-runoff (CRR) models remains a key challenge in hydrology. The Bayesian total error analysis (BATEA) methodology provides a comprehensive framework to hypothesize, infer, and evaluate probability models describing input, output, and model structural error. This paper assesses the ability of BATEA and standard calibration approaches (standard least squares (SLS) and weighted least squares (WLS)) to address two key requirements of uncertainty assessment: (1) reliable quantification of predictive uncertainty and (2) reliable estimation of parameter uncertainty. The case study presents a challenging calibration of the lumped GR4J model to a catchment with ephemeral responses and large rainfall gradients. Postcalibration diagnostics, including checks of predictive distributions using quantile-quantile analysis, suggest that while still far from perfect, BATEA satisfied its assumed probability models better than SLS and WLS. In addition, WLS/SLS parameter estimates were highly dependent on the selected rain gauge and calibration period. This will obscure potential relationships between CRR parameters and catchment attributes and prevent the development of meaningful regional relationships. Conversely, BATEA provided consistent, albeit more uncertain, parameter estimates and thus overcomes one of the obstacles to parameter regionalization. However, significant departures from the calibration assumptions remained even in BATEA, e.g., systematic overestimation of predictive uncertainty, especially in validation. This is likely due to the inferred rainfall errors compensating for simplified treatment of model structural error.

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

1. Introduction

[2] Given the significance of water in terrestrial and aquatic ecosystems, hydrological models are an integral part of virtually all environmental models formulated at the catchment scale. This paper focuses on conceptual rainfall-runoff (CRR) models, which aim to capture the dominant catchment dynamics while remaining parsimonious and computationally efficient. However, their parameters are not directly measurable and must be inferred (''calibrated'') from the observed data.

[3] Characterizing the uncertainty in runoff predicted by a CRR model has attracted the attention of hydrologists over many years [*Beven and Binley*, 1992]. Yet recent reviews of CRR model calibration, for example, *Kuczera and Franks* [2002], *Kavetski et al.* [2002, 2006a, 2006b], *Vrugt et al.* [2005], and *Wagener and Gupta* [2005] note the lack of a robust framework that accounts for all sources of error (input, model structural and output error).

[4] The lack of a robust calibration framework raises three problems in CRR modeling: (1) quantifying the predictive uncertainty in runoff and other model outputs remains problematic, (2) the regionalization of CRR parameters continues to be confounded by biases in the calibrated parameters and unreliable assessment of parameter uncertainty; and (3) discriminating between competing CRR model hypotheses is difficult because the precise causes of poor model performance are unclear.

[5] In the quest for a more robust and comprehensive calibration and uncertainty estimation methodology, *Kavetski et al.* [2002, 2006a, 2006b] and *Kuczera et al.* [2006] developed the Bayesian total error analysis (BATEA) framework. Its core ideas are as follows: (1) specify explicit probability models for each source of uncertainty (input, output and model structural errors); (2) where necessary, use hierarchical techniques to implement these probability models within a Bayesian inference framework; (3) where available, include a priori information about the catchment behavior and data uncertainty; (4) jointly infer the parameters of the CRR model and any unknown parameters of the

[1] School of Engineering, University of Newcastle, Callaghan, New South Wales, Australia.
[2] Water Division, Bureau of Meteorology, Melbourne, Victoria, Australia.

0043-1397/09/2008WR006825

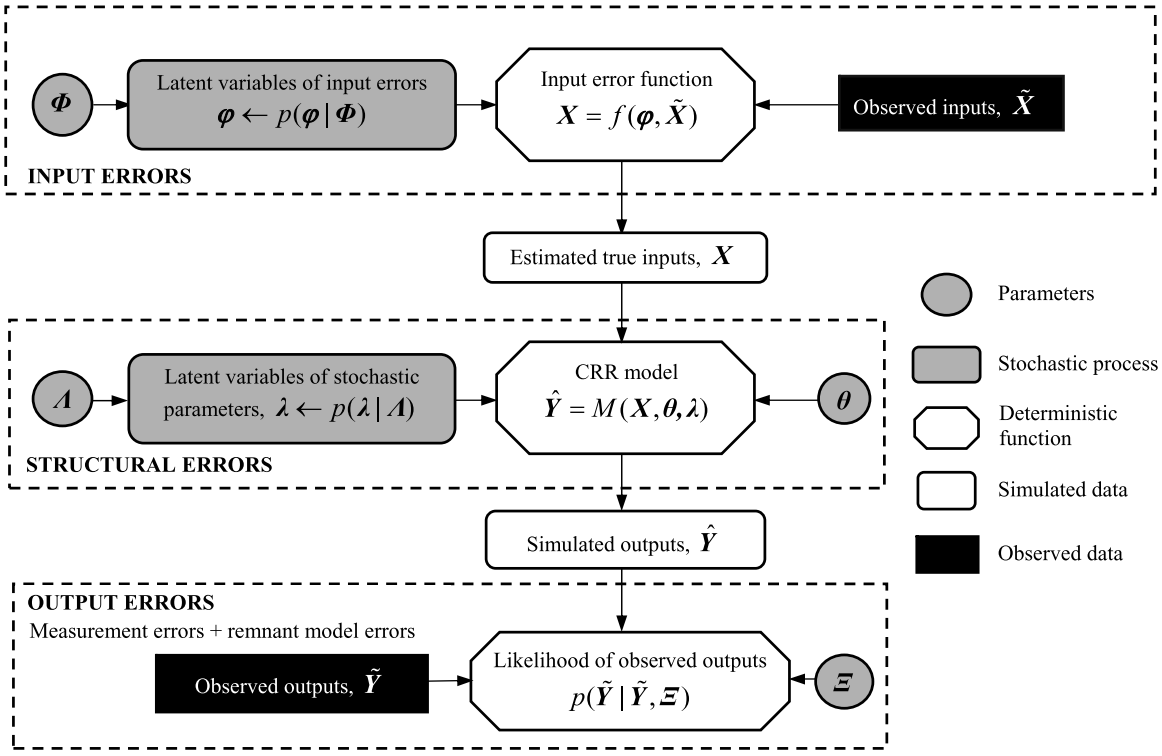


Figure 1. Schematic of BATEA in calibration mode.

error models; and (5) examine posterior diagnostics to check the assumptions made in step 1. BATEA allows, indeed, requires, modelers to explicitly hypothesize, infer and evaluate assumptions regarding each source of uncertainty, and generates model predictions accounting for all uncertainties included in the analysis.

[6] Earlier BATEA studies focused on the derivation of the posterior distribution given specific CRR and uncertainty models [*Kavetski et al.*, 2002, 2006a, 2006b]. Since the CRR model represents hypotheses describing hydrological dynamics and the uncertainty models represent hypotheses regarding the uncertainty in the calibration data, it is critical to evaluate these assumptions a posteriori and identify those that do not stand up to empirical scrutiny.

[7] The objective of this study is to compare and scrutinize the assumptions made in traditional least squares and BATEA calibrations. Specifically, the paper investigates the ability of these methods to provide (1) reliable quantification of predictive uncertainty and (2) consistent parameter estimation. The evaluation of competing CRR model hypotheses depends on successfully dealing with these two goals and will be undertaken in future work.

[8] The empirical assessment is based on a challenging case study of a catchment with markedly ephemeral hydrological dynamics and strong rainfall gradients. The quantification of predictive uncertainty is scrutinized by systematically assessing the credibility of the hypotheses underpinning four different calibration/prediction approaches, including two traditional least squares-based methods and two BATEA-based methods. The consistency of parameter estimates obtained by each calibration method is scrutinized by calibrating the same CRR model to different rainfall gauges and time periods.

[9] Of particular note is the application of a quantile-based diagnostic that directly evaluates whether the predictive distribution is consistent with the observed time series. This type of analysis, originally proposed in probabilistic forecasting [*Laio and Tamea*, 2007], is more comprehensive than traditional evaluation statistics such as the Nash-Sutcliffe index [*Nash and Sutcliffe*, 1970], which do not evaluate whether the predictive uncertainty is consistent with the observed data.

[10] This paper is structured as follows. Section 2 outlines the BATEA framework, including definitions of the error models. Section 3 describes the case study, including the catchment characteristics and the GR4J CRR model [*Perrin et al.*, 2003]. Section 4 outlines the calibration frameworks used in this paper: two traditional methods (standard least squares (SLS) and weighted least squares (WLS) schemes) and two BATEA methods (differing in the assumed error models). Section 5 applies posterior diagnostics to check the adequacy of the predictive distributions, while section 6 checks the consistency of the parameter inference. Section 7 discusses avenues for further improvements of the characterizations of predictive uncertainty, while section 8 discusses the potential of BATEA for model extrapolation and regionalization. Section 9 outlines future applications of BATEA to other types of hydrological models and catchments. The main conclusions are summarized in section 10.

2. BATEA Framework

[11] The BATEA framework conceptualizes the propagation of error in the CRR model using a hierarchical model. A schematic of this hierarchical model in calibration mode is depicted in Figure 1. Its components (the specific uncertainty models) represent hypotheses that will be scrutinized in the case study.

2.1. CRR Model Representation

[12] Let $\boldsymbol{X} = \{\boldsymbol{X}_t ; t = 1\ldots, T\}$ denote the true inputs of the CRR model (e.g., rainfall and potential evapotranspiration (PET)) and $\tilde{\boldsymbol{X}} = \{\tilde{\boldsymbol{X}}_t ; t = 1,\ldots, T\}$ be the observed values of these inputs. Similarly, let $\boldsymbol{Y} = \{\boldsymbol{Y}_t ; t = 1,\ldots, T\}$ denote the true outputs (e.g., runoff), $\tilde{\boldsymbol{Y}} = \{\tilde{\boldsymbol{Y}}_t ; t = 1,\ldots, T\}$ the observed outputs and $\hat{\boldsymbol{Y}} = \{\hat{\boldsymbol{Y}}_t ; t = 1,\ldots, T\}$ the outputs predicted by the model. Here, T is the total number of time steps. In this presentation, we assume equal length and resolution of inputs and outputs (this is not a necessary assumption; for example, *Kavetski et al.* [2006b] used hourly rainfalls and daily runoffs in a BATEA calibration).

[13] A CRR model $M()$ computes the simulated runoff value at time step t as follows

$$\hat{\boldsymbol{Y}}_t = M(\boldsymbol{X}_{1:t}, \boldsymbol{\theta}, \boldsymbol{S}_0) \tag{1}$$

where $\boldsymbol{X}_{1:t}$ is the history of inputs for time steps 1 to t, $\boldsymbol{\theta}$ are the CRR parameters and $\boldsymbol{S}_0$ denotes the initial conditions (which can be either inferred or handled using a warm-up).

2.2. Input Errors

[14] The observed input data of CRR models is often corrupted by sampling and measurement errors. In particular, areal rainfall estimates can have standard errors exceeding 30%, especially if the gauge network is sparse [*Linsley and Kohler*, 1988].

[15] The uncertainty in the inputs is described in BATEA using a probability model of the following general form [*Kavetski et al.*, 2006a]:

$$\boldsymbol{X} = f(\boldsymbol{\varphi}, \tilde{\boldsymbol{X}}) \tag{2}$$

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

where $\boldsymbol{\varphi}$ are variables that are used to estimate the true inputs $\boldsymbol{X}$ given the observations $\tilde{\boldsymbol{X}}$ and a hypothesized error function f(). Since the true inputs are not observable, $\boldsymbol{\varphi} = \{\varphi_{i(t)}; t = 1,\ldots, T\}$ are not observable and their values are inferred [*Renard et al.*, 2009]. In Bayesian hierarchical terminology, $\boldsymbol{\varphi}$ are referred to as "latent" variables, the distribution $p(\boldsymbol{\varphi}|\boldsymbol{\Phi})$ is referred to as the "hyperdistribution" and $\boldsymbol{\Phi}$ are the "hyperparameters."

[16] In this paper, we follow *Kavetski et al.* [2006a] and assume that observed rainfall is corrupted by multiplicative errors, hypothesizing the following relationship between observed and true rainfall:

$$\boldsymbol{X}_t = \varphi_{i(t)} \tilde{\boldsymbol{X}}_t \tag{4}$$

where $\varphi_{i(t)}$ is referred to as a rainfall multiplier.

[17] Following *Kuczera et al.* [2006], the rainfall multipliers are assumed to be statistically independent and follow a lognormal distribution with hyperparameters $\boldsymbol{\Phi} = (\mu, \sigma^2)$

$$\log \varphi_{i(t)} \sim N(\mu, \sigma^2) \tag{5}$$

If the data accuracy is unknown, the hyperparameters are also unknown, but can be included in the inference list.

[18] The "index function" $i(t)$ defines the hypothesized temporal structure of the input errors. Two competing hypotheses will be considered and scrutinized in this paper:

[19] 1. The first hypothesis is different rainfall multipliers for every day with significant rainfall. This yields the simple index function $i(t) = t$.

[20] 2. The second hypothesis is identical rainfall multipliers for time steps within the same storm event. This restricts the number of latent variables and is equivalent to assuming perfect autocorrelation of input errors within single storm events [*Kavetski et al.*, 2006a]. If the time series is partitioned into K epochs $\{(t_k, t_{k+1} - 1); k = 1,\ldots, K\}$, the index function is $i(t) = k$ for $t_k \leq t < t_{k+1} - 1$.

[21] Note that BATEA is not restricted to multiplicative errors, nor to the lognormal assumption. These merely represent specific initial hypotheses that should be tested and, if found unsupported by the empirical evidence, replaced by more adequate assumptions.

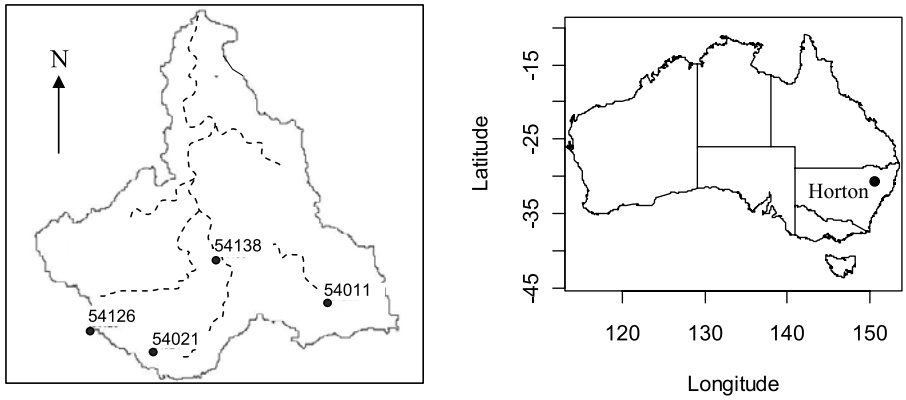
[22] Note that the input error model (4) is only applied to the rainfall, whereas the observed PET is assumed to be exact. More generally, BATEA can include probabilistic error models describing errors in both rainfall and PET. However, since rainfall is the primary driving input of CRR models and is more spatially and temporally variable than PET, this case study restricts its attention solely to uncertainties in rainfall. The implications of this are discussed in section 7.1.

2.3. Structural Errors

[23] A CRR model is a simplified approximation of the catchment dynamics and therefore is unlikely to reproduce the true outputs even if the true inputs were known. We refer to these errors as structural errors of the hydrologic model. A major portion of these errors is likely to arise because of spatial and temporal averaging/lumping. For example, the response of a catchment to a rainfall event with a given total depth depends on the localization of the main mass of the rainfall field and its trajectory through the catchment. However, a lumped CRR model will not be able to reproduce the different runoff responses arising from rainfall fields with the same total depth but different spatial and temporal distributions.

[24] One possibility for characterizing structural errors using the hierarchical BATEA framework is to allow one or more CRR parameters to stochastically vary from storm to storm, leading to the "storm-dependent" parameter concept introduced by *Kuczera et al.* [2006]. These stochastic parameters can be modeled hierarchically using latent variables (analogously to using latent variables to describe input errors). In Figure 1, the latent variables for the stochastic CRR parameters are denoted $\boldsymbol{\lambda}$ and their hyperparameters are denoted $\boldsymbol{\Lambda}$. However, this paper does not use the hierarchical description of structural errors because investigations with synthetic data have indicated that joint inference of both input error and structural error with vague priors can become ill posed [*Renard et al.*, 2008]. Addressing this problem will require additional information about the input data corruption mechanisms and will be considered in future studies. See further discussion in section 7.1.

[25] It is also possible to account for model structural error using the output error model [*Huard and Mailhot*, 2008; *Kavetski et al.*, 2006a]. However, this approach is not



(a) Horton Catchment

(b) Location in Australia

Figure 2. Map of the Horton catchment, showing its stream network (dashed lines) and rain gauges.

used in this paper because the output error model is derived from rating curve analysis.

2.4. Output Errors

[26] Given a set of latent variables (multipliers), the true rainfall can be estimated using equation (4) and supplied to the CRR model to generate the simulated outputs $\hat{\boldsymbol{Y}}$. However, the simulated output $\hat{\boldsymbol{Y}}$ will not equal the observed output $\tilde{\boldsymbol{Y}}$ for several reasons: (1) observed outputs are affected by sampling and measurement error, e.g., runoff data are affected by rating curve errors (''output measurement errors'' in Figure 1); (2) a simple model such as equation (4) will not recover the true inputs exactly; and (3) structural errors are unlikely to be completely eliminated even if stochastic time-varying parameters are implemented. The errors associated with reasons 2 and 3 are labeled ''remnant model errors'' in Figure 1.

[27] It is therefore necessary to specify a distribution for residual errors $\boldsymbol{\varepsilon} = \tilde{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}$, or, equivalently, an output error model that describes the distribution of observed outputs given the simulated outputs,

$$\tilde{\boldsymbol{Y}}_t \sim p(\tilde{\boldsymbol{Y}}|\hat{\boldsymbol{Y}}_t, \boldsymbol{\Xi}) \tag{6}$$

where $\boldsymbol{\Xi}$ are the parameters of the output error model (these can be either inferred or fixed a priori).

[28] For example, if we assume that the output errors are independent and Gaussian, i.e., $\tilde{\boldsymbol{Y}} = \hat{\boldsymbol{Y}} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0};\sigma_\varepsilon^2\boldsymbol{I})$, it can be shown that $\tilde{\boldsymbol{Y}} \sim p(\tilde{\boldsymbol{Y}}|\hat{\boldsymbol{Y}}, \boldsymbol{\Xi}) = N(\hat{\boldsymbol{Y}}, \sigma_\varepsilon^2\boldsymbol{I})$. However, output errors are unlikely to have such simple form, and more complicated probability models allowing heteroscedasticity and autocorrelation might be necessary.

[29] In this paper, BATEA and WLS use a heteroscedastic output error model derived from rating curve analysis (section 4.2). Since this error model reflects solely output measurement errors, remnant errors are ignored in the case study.

2.5. Inference

[30] In general, BATEA can make an inference on all unknown quantities of the hierarchical structure, including (1) latent variables of input errors $\boldsymbol{\varphi}$, (2) hyperparameters of input errors $\boldsymbol{\Phi}$, (3) deterministic CRR parameters $\boldsymbol{\theta}$, (4) latent variables of stochastic CRR parameters $\boldsymbol{\lambda}$, (5) hyperparameters of stochastic CRR parameters $\boldsymbol{\Lambda}$, and (6) output errors parameters $\boldsymbol{\Xi}$.

[31] The BATEA posterior distribution is, up to a constant of proportionality, given by *Kavetski et al.* [2002, 2006a, 2006b] as follows:

$$p(\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\Lambda}, \boldsymbol{\varphi}, \boldsymbol{\Phi}, \boldsymbol{\Xi}|\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{X}}) \propto p(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\varphi}, \boldsymbol{\Xi}, \tilde{\boldsymbol{X}})p(\boldsymbol{\lambda}|\boldsymbol{\Lambda})p(\boldsymbol{\varphi}|\boldsymbol{\Phi})p(\boldsymbol{\theta})p(\boldsymbol{\Lambda})p(\boldsymbol{\Phi})p(\boldsymbol{\Xi}) \tag{7}$$

The quality of the inference using equation (7) is contingent on the strength of prior information on $\boldsymbol{\Xi}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Phi}$. For example, *Kavetski et al.* [2006a] show that in the absence of prior information on $\boldsymbol{\Xi}$ and $\boldsymbol{\Phi}$, the inference is ill posed even if there are no stochastic CRR parameters. To keep the inference well posed in this study, attention was restricted to input errors. Accordingly, no stochastic CRR parameters were used and the output errors parameters $\boldsymbol{\Xi}$ were specified a priori. Thus the posterior becomes

$$p(\boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\Phi}|\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{X}}, \boldsymbol{\Xi}) \propto p(\tilde{\boldsymbol{Y}}|\boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\Xi}, \tilde{\boldsymbol{X}})p(\boldsymbol{\varphi}|\boldsymbol{\Phi})p(\boldsymbol{\theta})p(\boldsymbol{\Phi}) \tag{8}$$

which comprises three parts: (1) the likelihood of observed outputs, (2) the population distribution of latent variables conditioned on hyperparameters, and (3) priors of deterministic parameters and hyperparameters.

[32] The use of latent variables for characterizing input and structural errors comes at the cost of the dimensionality of the posterior distributions (7) and (8): the number of quantities inferred by BATEA increases with the length of the calibration period and can include hundreds or more multipliers. As a result, maximizing and sampling the BATEA posterior distribution is computationally challenging, requiring efficient numerical methods and careful implementation. Strategies for implementation of Markov chain Monte Carlo (MCMC) methods that deal with these challenges are reported elsewhere [*Kuczera et al.*, 2007; *Renard et al.*, 2009].

3. Catchment and Data

[33] BATEA makes explicit the hypotheses used by the modeler to describe data errors and catchment dynamics (e.g., in this paper we assume multiplicative lognormal

Table 1. Rainfall Statistics for the Horton Catchment

Rain Gauge ID	Elevation (m)	Average Daily Rainfall (mm)
54011	567	1.83
54126	1465	3.40
54021	869	2.66
54138	392	2.15

rainfall errors and the GR4J CRR model). The posterior distribution (8) is conditioned on all these hypotheses. Therefore, merely reporting the posterior distribution of model parameters (and latent variables and hyperparameters) falls well short of an adequate analysis, since it fails to scrutinize the credibility of the underlying hypotheses. It is essential that these hypotheses be challenged with all available evidence and, if found wanting, revised. Naturally, the same scrutiny needs to be applied to any model identification methodology, but is seldom attempted in hydrology [see *Yang et al.*, 2007; *Feyen et al.*, 2007].

[34] The case study illustrates methods for assessing the credibility of the hypotheses made during hydrological calibrations using BATEA or any other probabilistic inference framework. To subject the calibration methods to a thorough evaluation, we selected a case study catchment with several challenging attributes:

[35] 1. The catchment should have a low runoff coefficient and ephemeral flow regime. This type of catchment is particularly difficult to model because sustained periods of little or no flow imply low information content of the runoff time series for parameter estimation [*Wooldridge et al.*, 2003].

[36] 2. The catchment should be subject to significant rainfall gradients and have multiple rain gauges, which is likely to produce significant input uncertainty.

[37] We selected the Horton catchment, located in northern inland New South Wales, Australia (Figure 2) [*Peel et al.*, 2000]. It has an average annual rainfall of 819 mm and an average annual runoff of 108 mm, yielding an annual runoff coefficient of 0.13. The catchment area is 1920 km^2 and it contains 4 daily rain gauges (Table 1 reports the elevation and average daily rainfall for each gauge). There is a strong rainfall gradient in the catchment, with higher rainfall in the southwestern areas of the catchment – indeed the average daily rainfall recorded at the wettest gauge (gauge 54126) is 86% higher than that of the driest gauge (gauge 54011).

[38] The ephemeral nature of the Horton catchment can be seen from the flow duration curve (Figure 3), where 97.5% of daily runoff is below 2 mm. The observed time series (Figure 6) indicate that the catchment experiences very few significant runoff generating events.

[39] Following the approach of *Peel et al.* [2000], the areal PET was taken as the mean monthly value based on estimates provided by the Australian Bureau of Meteorology [*Wang et al.*, 2001].

3.1. Calibration and Validation Periods

[40] Two calibration periods of different lengths and rain gauges were considered: a 2-year calibration period ranging from 21 April 1978 to 10 April 1980, and a 5-year period ranging from 1 January 1983 to 31 December 1987. In both cases, initial store variables were fixed using a 100-day warm-up prior to the calibration period. For the 5-year period only rain gauges 54138 and 54021 were used because the other gauges had too much missing data. Validation was undertaken for the 13-year period from 15 August 1990 to 21 December 2003 using rain gauges 54138 and 54021 because the other rain gauges were discontinued during this period.

3.2. CRR Model

[41] We used a lumped rather than distributed model because lumped models are predominant in hydrological practice because of their much lower data requirements and lower computational burden. The GR4J model was used because it has a parsimonious form with only four calibrated parameters and has been extensively tested over hundreds of catchments worldwide, with a range of climatic conditions from tropical to temperate and semiarid catchments [*Perrin et al.*, 2003]. Figure 4 shows a schematic of the GR4J model.

[42] GR4J has four parameters: the capacity of the production store x_1 (mm), the groundwater exchange coefficient x_2 (mm), the capacity of the nonlinear routing reservoir x_3 (mm) and the unit hydrograph time base x_4 (days).

[43] For computational convenience, we use transformed parameters $x_1 = \log(\tilde{x}_1)$, $x_2 = \tilde{x}_2$, $x_3 = \log(\tilde{x}_3)$, $x_4 = \log(\tilde{x}_4 - 0.5)$, where $\tilde{x}$ denotes the original parameter. This unconstrains the estimation problem with all transformed parameters being defined over $(-\infty, +\infty)$.

4. Calibration Frameworks

[44] The four calibration frameworks compared in the case study are summarized in Table 2 and are briefly described below.

4.1. Standard Least Squares

[45] SLS seeks to minimize the sum of the squares of the differences between observed and simulated responses. It is perhaps the most widely used calibration criterion in hydrology and is equivalent to maximizing the Nash-Sutcliffe statistic. SLS ignores input errors and assumes that the output errors are independently and normally distributed with zero mean and constant variance (homoscedastic errors). In this study, the output error variance is inferred as part of the calibration process. We also augment the SLS method with a single multiplier for the entire rainfall series

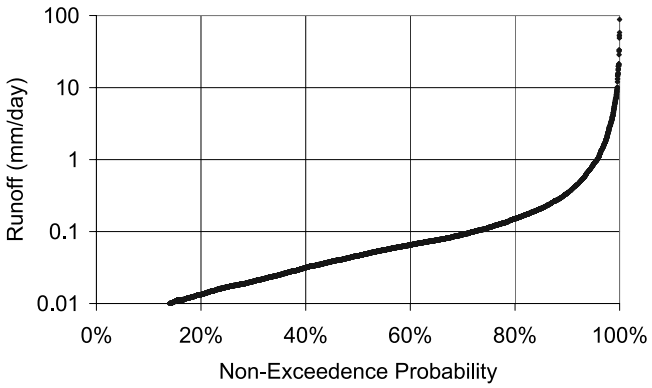


Figure 3. Flow duration curve of the Horton catchment.

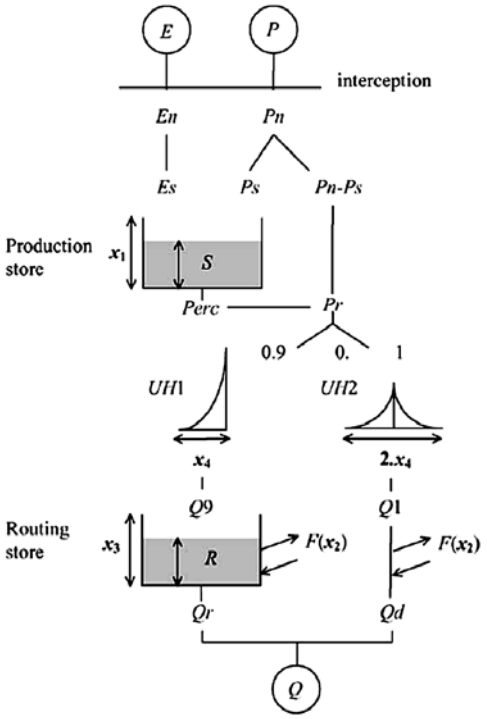


Figure 4. Diagram of the GR4J model (reprinted from *Perrin et al.* [2003], copyright 2003, with permission from Elsevier).

in an attempt to correct for systematic biases in the rainfall measurements.

4.2. Weighted Least Squares

[46] A major shortcoming of SLS in hydrological calibration is that output errors rarely, if ever, have a constant variance. To investigate this, the rating curve data for the Horton catchment was examined to estimate the runoff measurement errors due to errors in the rating curve.

[47] Figure 5 shows the runoff measurement errors (the difference between the runoff gaugings and the predicted runoff from the rating curve) as a function of the predicted rating curve runoff. Since practitioners are usually interested in high flows, we focused on runoff gaugings exceeding 0.5 mm. The data for defining the runoff measurement error model is quite sparse, with separate clusters of low/medium runoffs and a few high values. Nonetheless, there is a clear proportionality between runoff measurement error and the predicted runoff.

[48] A simple heteroscedastic model was used by assuming that the runoff measurement errors are normally distributed with zero mean and a standard deviation σ_ε linearly related to the predicted runoff,

$$\sigma_\varepsilon = a + b\tilde{y} \tag{9}$$

This relationship was fitted to the runoff measurement error data using WINBUGS [*Spieghalter et al.*, 2003], yielding the posterior distribution of a and b. The posterior uncertainty in a and b made little difference to the 90% probability limits of the runoff measurement error model. Consequently, we fixed these parameters to their expected values $a = 0.4$ and $b = 0.086$. The corresponding 90% probability limits of the runoff measurement error model are shown in Figure 5 and were judged to be satisfactory.

[49] Using the heteroscedastic runoff measurement error model as the output error model has two main effects on the calibration: (1) less weight is given to days with high runoff and (2) less weight is given to days with very low runoff (since the runoff measurement error equation has an intercept of 0.4).

[50] As with SLS, our WLS implementation also includes a single rainfall multiplier for all time steps.

4.3. BATEA With Daily Rainfall Multipliers

[51] BATEA with daily rainfall multipliers (BATEA_DAILY) uses a separate rainfall multiplier for each day with nonzero rainfall. Auxiliary investigations identified a significant number of insensitive rainfall multipliers, which exert little impact on the simulated runoff and are effectively redundant (they were associated with days with low rainfall which did not produce significant runoff). Such insensitive rainfall multipliers unnecessarily increase the computational cost of the inference. A preprocessing heuristic procedure was developed to identify and exclude the insensitive rainfall multipliers from the analysis (see Appendix A). Since this procedure makes a number of assumptions that could potentially result in excluding important multipliers, a conservative approach was adopted where only very insensitive rainfall multipliers were removed. Future work will refine this preprocessing heuristic using more robust techniques, but is tangential to this paper.

[52] Even excluding insensitive multipliers using the preprocessing heuristic, BATEA_DAILY was used only with a 2-year calibration period because of the computational burden associated with inferring many hundreds of latent variables.

[53] The output error model was the heteroscedastic runoff measurement error model also used in WLS. Since the parameters of this model were estimated independently

Table 2. Summary of Calibration Methods Used in Case Study

Statistical Method	Input Error Model	Input Temporal Structure	Output Error Model
SLS	$\boldsymbol{X}_t = \log(\mu)\tilde{\boldsymbol{X}}_t$	not applicable	$\tilde{\boldsymbol{Y}} \sim N(\hat{\boldsymbol{Y}}, \sigma_\varepsilon^2)$, σ_ε unknown
WLS	$\boldsymbol{X}_t = \log(\mu)\tilde{\boldsymbol{X}}_t$	not applicable	$\tilde{\boldsymbol{Y}} \sim N(\hat{\boldsymbol{Y}}, \sigma_\varepsilon^2)$, with $\sigma_\varepsilon = 0.4 + 0.086\tilde{y}$
BATEA_STORM	$\boldsymbol{X}_t = \phi_{i(t)}\tilde{\boldsymbol{X}}_t$ $\log(\phi_{i(t)}) \sim N(\mu, \sigma^2)$	storm epochs	$\tilde{\boldsymbol{Y}} \sim N(\hat{\boldsymbol{Y}}, \sigma_\varepsilon^2)$, with $\sigma_\varepsilon = 0.4 + 0.086\tilde{y}$
BATEA_DAILY	$\boldsymbol{X}_t = \phi_{i(t)}\tilde{\boldsymbol{X}}_t$ $\log(\phi_{i(t)}) \sim N(\mu, \sigma^2)$	daily epochs	$\tilde{\boldsymbol{Y}} \sim N(\hat{\boldsymbol{Y}}, \sigma_\varepsilon^2)$, with $\sigma_\varepsilon = 0.4 + 0.086\tilde{y}$

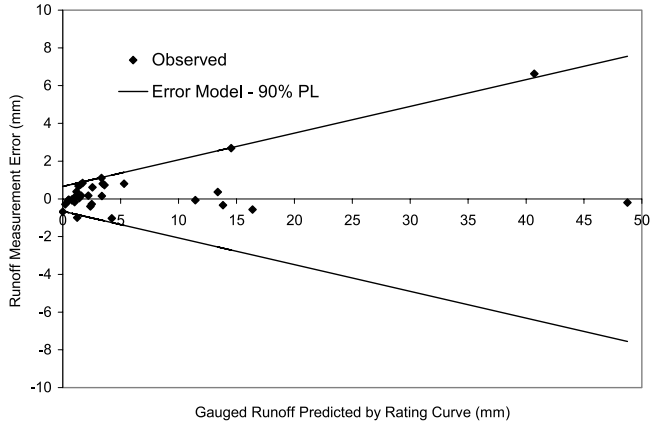


Figure 5. Probability model of runoff measurement errors estimated from rating curve analysis of the Horton catchment (PL, probability limits).

using rating curve data, they were not be inferred during the calibration.

4.4. BATEA With Storm Epoch Rainfall Multipliers

[54] BATEA with storm epoch rainfall multipliers (BATEA_STORM) applies individual rainfall multipliers to entire storm epochs rather than to individual days. This requires subdividing the rainfall series into storm epochs separated by interstorm dry spells of two or more days followed by a wet day with rainfall exceeding 0.5 mm. BATEA_STORM implements a coarser treatment of rainfall errors than BATEA_DAILY: it merely uses a rainfall multiplier to scale the storm depth while assuming the relative rainfall pattern of the storm is correct. This approach has two benefits: (1) a reduction in the number of inferred latent variables and (2) it is more likely that storm epoch multipliers are statistically independent from one epoch to the next because they operate over larger time scales. In contrast, consecutive daily multipliers may compensate for one another and thus their estimates can become negatively correlated.

5. Posterior Diagnostics

[55] An integral part of the Bayesian approach is a critical evaluation of its hypotheses given the available evidence using a range of posterior diagnostics.

[56] Throughout the rest of this paper, the notation "method_raingauge_period" is used to identify the calibration method, the rain gauge providing input to the GR4J model, and the calibration period with "2yr" referring to the period 21 April 1978 to 10 April 1980, and "5yr" to the period from 1 January 1983 to 31 December 1987.

5.1. Comparison of Predictive Distribution of Runoff to Observed Data

[57] Figure 6 shows the total predictive uncertainty of the simulated runoff time series for the 2-year calibration period using rain gauge 54138 and compares it to the observed runoff. The total predictive uncertainty for the simulated runoff series in calibration includes (1) the uncertainty due to input errors, (2) uncertainty due to output errors, and (3) the uncertainty in the inferred CRR parameters. The uncertainty arising from input errors was estimated by sampling from the posterior distribution of rainfall multipliers. For BATEA, the rainfall multipliers vary storm-by-storm or day-by-day, whereas for SLS and WLS a single rainfall multiplier is used for the entire time series.

[58] The poor fits obtained using the SLS and WLS methods are typical of practical applications of these methods, with significant runoff events missed (e.g., 50% errors during major flows). Note that WLS by construction will not fit the runoff peaks as well as SLS because it gives less weight to high runoffs (section 4.2). On the other hand, the errors of the SLS fit are strongly heteroscedastic, which violates the constant error variance assumption underpinning SLS. Although this is typical in practice, and is well known, SLS continues to be widely used as a fitting criterion.

[59] In contrast, BATEA_DAILY and BATEA_STORM produce much better fits to the observed data (for clarity, Figure 6 shows only BATEA_STORM results; the BATEA_DAILY results are very similar). However, since improved BATEA fits arise from estimating the rainfall errors during model calibration visual examination of observed and predicted responses are insufficient to conclusively determine whether the BATEA hypotheses are supported by the data. More probing diagnostics are required.

5.2. Predictive QQ Plot

[60] In the context of quantifying the uncertainty in the model predictions the outcome of the analysis takes the form of a predictive distribution. Regardless of the method used to derive this distribution two important points must be emphasized:

[61] 1. The predictive distribution is conditioned on the assumptions made during the inference. Consequently, unsupported assumptions may lead to inadequate predictive distributions. It follows that the estimated predictive distribution must be scrutinized ("validated"), as discussed by *Hall et al.* [2007, p. 987]: "Without validation, calibration is worthless, and so is uncertainty estimation."

[62] 2. The predictive uncertainty has to be validated using observations. From a methodological point of view, this requires a diagnostic approach that compares a time-varying distribution (the predictive distribution at all times t) to a time series of observations. This is a much more stringent test than validation methods currently used in hydrology, which simply compare two time series (observations and "optimal" simulations). Indeed, standard goodness-of-fit assessments (e.g., using the Nash-Sutcliff statistic) cannot check if the predictive distribution is consistent with the observed data.

[63] Consequently, the runoff time series shown in Figure 6 are insufficient to properly assess whether the predictive uncertainty is consistent with the observed data. For this task, we use the predictive QQ plot, adapted from the verification tools used for probabilistic forecasts of hydrological and meteorological variables [*Dawid*, 1984; *Gneiting et al.*, 2007; *Laio and Tamea*, 2007].

[64] The predictive QQ plot is constructed as follows: Let F_t be the cumulative distribution function (cdf) of the predictive distribution of runoff at time t, and $\tilde{y}_t$ the corresponding observed runoff. If the hypotheses in the calibration framework are consistent with the data, the observed value $\tilde{y}_t$ should be consistent with the distribution F_t. Hence, under the assumption that the observation $\tilde{y}_t$ is a realization from the predictive distribution, the p value

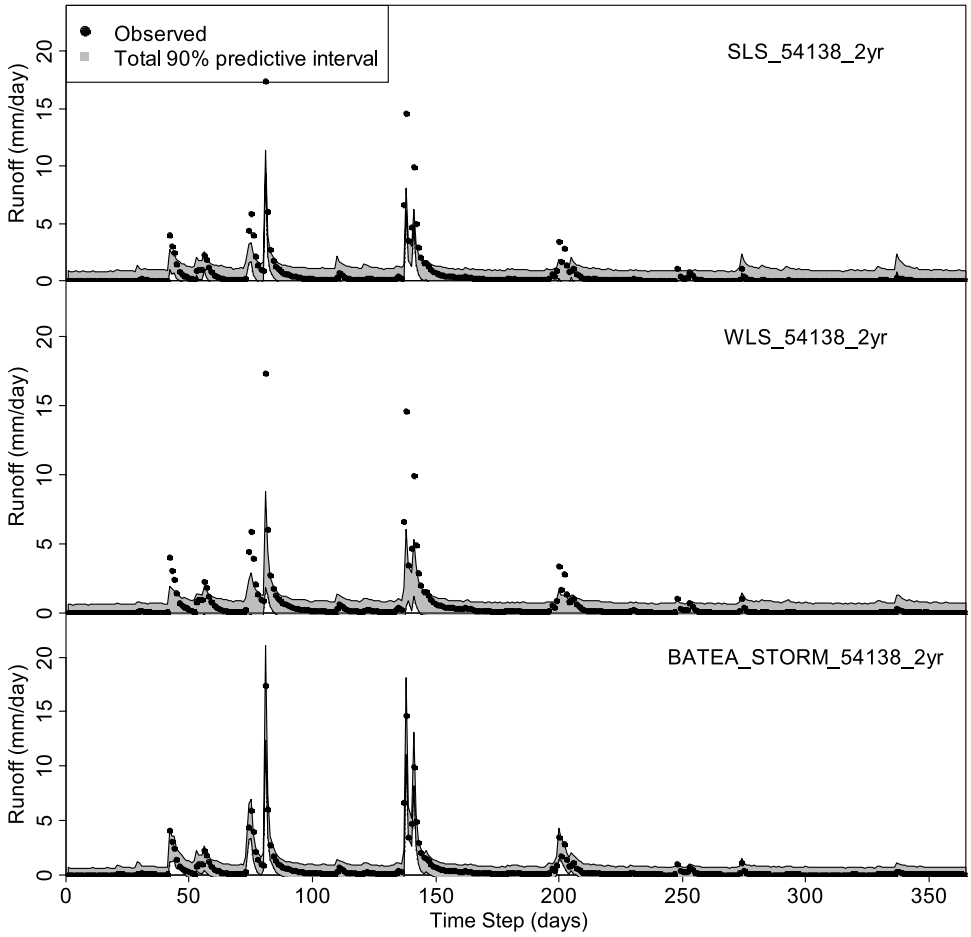


Figure 6. Observed versus simulated runoff for all calibration methods. (BATEA_DAILY_54138_2yr is similar to BATEA_STORM_54138_2yr and is omitted to avoid obscuring the plot.)

$F_t(\tilde{y}_t)$ is a realization from a uniform distribution on [0,1]. The predictive QQ plot compares the empirical cdf of the sample of p values $(F_t(\tilde{y}_t))_{t=1,\ldots,T}$ with the cdf of a uniform distribution to assess whether the hypotheses are consistent with the observations. The predictive QQ plot can be interpreted as follows (Figure 7): (1) If all points fall on the 1:1 line, the predicted distribution agrees perfectly with the observations. (2) If an observed p value is 1.0 or 0.0, the corresponding observed data lies outside the predicted range, implying that the predictive uncertainty is significantly underestimated. (3) If the observed p values cluster around the midrange (i.e., a low slope around theoretical quantile 0.4–0.6), the predictive uncertainty is overestimated. (4) If the observed p values cluster around the tails (i.e., a high slope around theoretical quantile 0.4–0.6), the predictive uncertainty is underestimated. (5) If the observed p values at the theoretical median are higher/lower than the theoretical quantiles, the modeled predictions systematically underpredict/overpredict the observed data.

[65] The predictive QQ plot provides a simple, intuitive and informative summary of the performance of probabilistic prediction frameworks. Very importantly, it is "distribution-assumption-free" in the sense of not making any additional assumptions beyond those used during the calibration. Indeed, it is a direct test of these assumptions.

[66] Figure 8 presents the predictive QQ plot for the 2-year calibration and 13-year validation period for SLS, WLS and BATEA_STORM and BATEA_DAILY for rain gauge 54138. We focus separately on the very low flows (runoff below 2 mm) and on the significant forcing events (runoff exceeding 2 mm).

[67] When BATEA is used in prediction mode the rainfall multipliers are not inferred (they can only be inferred in calibration mode). Since the rainfall errors cannot be in-

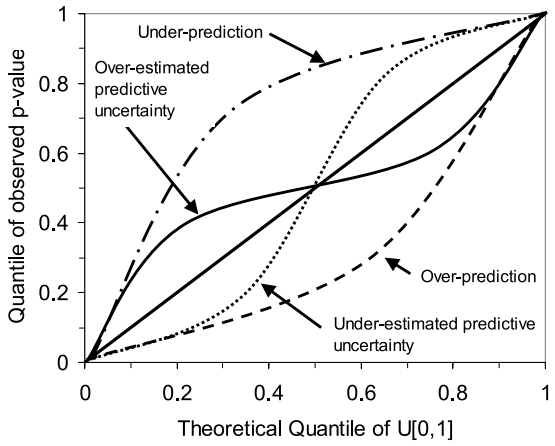


Figure 7. Interpretation of the predictive QQ plot (adapted from *Laio and Tamea* [2007]).

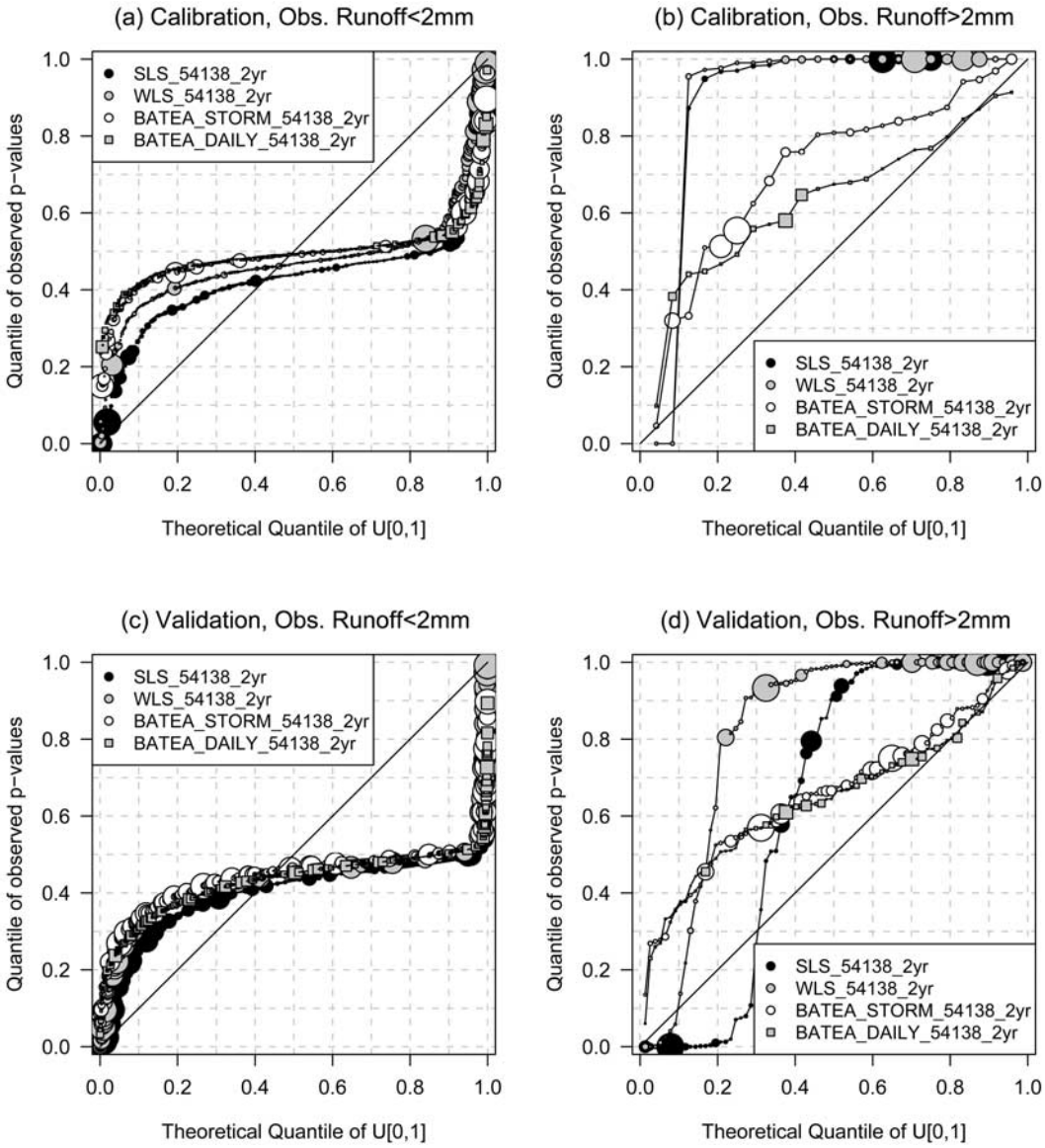


Figure 8. Predictive QQ plot (symbol size is scaled proportionally to the magnitude of the observed runoff).

ferred in validation they are sampled from the hyperdistribution inferred during the calibration.

[68] All QQ plots are far from ideal. For runoffs below 2 mm (Figures 8a and 8c), there is significant overestimation of the uncertainty. In SLS this occurs because the output error variance is assumed homoscedastic, while for WLS, BATEA_STORM and BATEA_DAILY this occurs because the assumed output error model uses a standard deviation of 0.4 mm for near-zero runoffs, which appears too high.

[69] The differences in the calibration methods for runoffs exceeding 2 mm are more distinct (Figures 8b and 8d). Both SLS and WLS underpredict the high flows, with numerous observations lying outside the predicted range. The QQ plot for BATEA_STORM/BATEA_DAILY is much closer to the 1:1 line than both WLS and SLS, but there remains systematic underprediction of the high flows. The reason for this underprediction is currently unclear.

[70] Overall, while still far from perfect, BATEA yields a noticeable improvement on SLS and WLS in terms of the adequacy of predictive uncertainty in calibration and validation.

5.3. Residual Error Diagnostics

[71] The predictive QQ plot provides an overall assessment of whether the total predictive uncertainty is consistent with the observations. More specific diagnostics are required to verify the assumptions of the individual error models. In particular, the residual errors should conform to the output error model. Following *Carlin and Louis* [2000], the residuals were computed as the difference between the observed runoff and the expected value from the predictive distribution. Note that collapsing the posterior (whether to modal, expected, or median statistics) in this way can result in a substantial loss of information [*Bernardo and Smith*, 2000].

[72] In order to simplify the comparison between calibration methods that use different output error models, all residuals are standardized by the standard deviation estimated using the output error model.

[73] In BATEA, the residual error diagnostics are meaningful only in calibration mode, because the rainfall multipliers are unknown in validation and are sampled from the

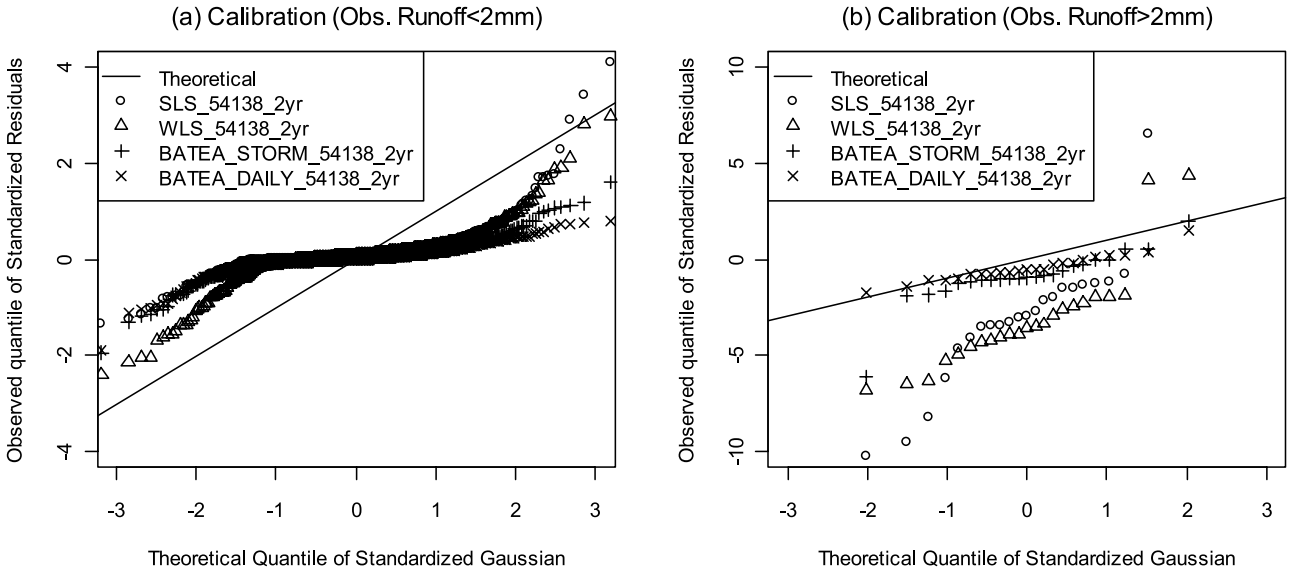


Figure 9. QQ plot of standardized residuals. To show detail, the y axis limits are fixed at $[-4, 4]$ for runoffs below 2 mm and to $[-10, 10]$ for runoffs exceeding 10 mm.

(posterior) hyperdistribution. In that case the predictive distribution of runoff includes a significant input error contribution and cannot be expected to be consistent with the output error model (even if remanent errors were included in the analysis). This does not imply that BATEA cannot be scrutinized in validation mode: the runoff predictive distribution should be consistent with the observed data and this evaluation has already been undertaken using the predictive QQ plot (section 5.2).

5.3.1. Distributional Properties

[74] Figure 9 presents a quantile-quantile (QQ) plot for the residual errors for the calibration periods on days with high (>2 mm) and low (<2 mm) runoff. If the output error hypothesis (Table 2) is adequate, the residuals should follow the straight line labeled "theoretical. "None of the QQ plots for low runoffs are ideal. All plots exhibit fat-tail behavior characteristic of outliers, with SLS and WLS notably worse than the BATEA.

[75] For all runs, the slope at the center of the distribution is less than the assumed slope, implying that the residual variance is less than expected from the hypothesized output error model. In the case of BATEA, this discrepancy arises because on most days the observed and simulated runoffs are virtually zero, whereas it is assumed that the standard deviation of output errors is 0.4 mm when the simulated runoff is zero.

[76] Adequate treatment of errors in near-zero runoffs remains problematic. Allowing the standard deviation of output error to approach zero as simulated runoff goes to zero can introduce major statistical artifacts: the likelihood of near-zero observed outputs then exerts enormous (and usually undue) leverage on the objective function.

[77] For high runoff, BATEA outperforms both WLS and SLS providing a much better, yet still imperfect match to the theoretical distribution. On the one hand, both SLS and WLS exhibit fat tails and a systematic bias: most of the standardized residuals are negative, which highlights a systematic underestimation of high flows. On the other hand, BATEA_STORM and BATEA_DAILY yield residuals in far better agreement with the output error model assumptions.

5.3.2. Autocorrelation

[78] Another important assumption of the output error model is that the residuals are statistically independent. To test this assumption, Figure 10 presents partial autocorrelation functions (PACF) of the residuals. In calibration, the independence assumption is clearly violated by SLS and WLS. In the case of BATEA, the lag 1 correlation, though statistically significant, is relatively low: 0.3 for BATEA_STORM and 0.4 for BATEA_DAILY. Note that the autocorrelation at lag 2 is statistically significant for BATEA_DAILY but not for BATEA_STORM.

5.4. Diagnostics for Latent Variables

[79] Hierarchical methods using latent variables need a posteriori checks of the adequacy of the hyperdistribution. Since in this BATEA application we hypothesized that the rainfall multipliers (both for daily and storm epochs) are independent and lognormal, these assumptions need posterior checks.

[80] Examination of the posteriors of the individual rainfall multipliers revealed that the majority of multipliers had a relatively large posterior uncertainty. Figure 11 shows

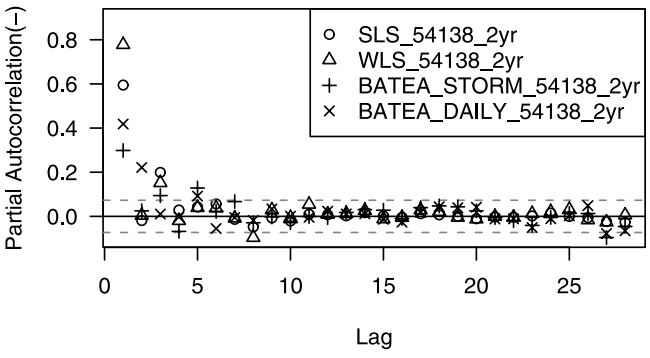


Figure 10. Partial autocorrelation of residuals during calibration. Dashed line shows the 95% probability limits provided by the R statistical package [*R Development Core Team*, 2008].

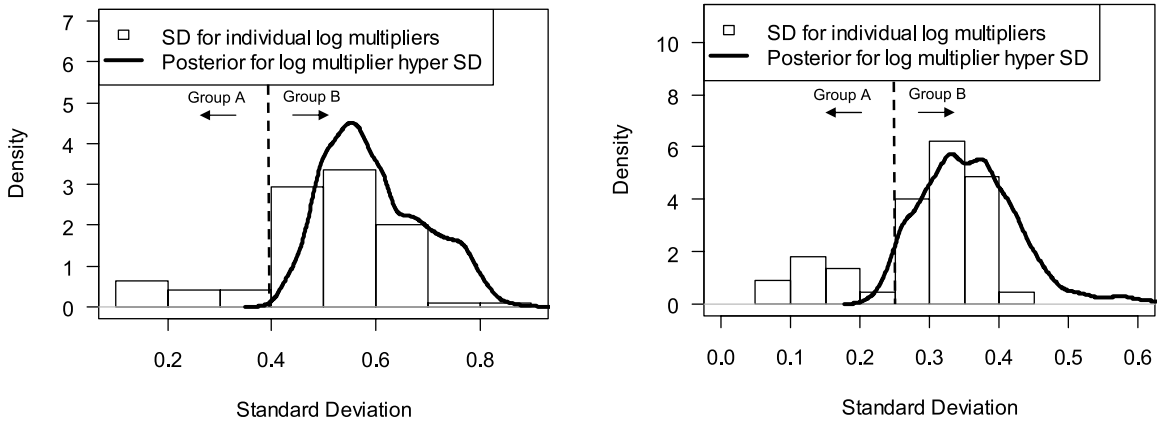


Figure 11. Histogram of the posterior standard deviation (SD) of individual rainfall multipliers. Group A contains multipliers with low SDs, and group B contains multipliers with high SDs. The solid line denotes the posterior distribution of the standard deviation of the multipliers (the "hyper–standard deviation" σ in equation (5), abbreviated "hyper SD"). The analysis is presented in log space.

a histogram of the posterior standard deviations of all rainfall multipliers in the BATEA_DAILY_54138_2yr and BATEA_STORM_54138_2yr calibrations. In both cases, there is evidence of a mixture of two different types of multipliers; group A with a low posterior standard deviation and group B with a high posterior standard deviation.

[81] Figure 11 also shows the posterior distribution of the standard deviation for the rainfall multipliers. In both cases, the standard deviation of the individual multipliers in group B (high standard deviation) corresponds to the posterior standard deviation of the multipliers. This indicates that these rainfall multipliers remain uncertain and are not informed by the rainfall/runoff data (i.e., they are insensitive multipliers that were missed by the heuristic procedure outlined in section 4.3).

[82] The posterior of insensitive multipliers becomes near-identical to the posterior hyperdistribution (as seen in Figure 12). This can by demonstrated by considering the marginal posterior of the insensitive rainfall multipliers, $\boldsymbol{\varphi}_b$ (using equation (8)):

$$p(\boldsymbol{\varphi}_b|\tilde{\boldsymbol{Y}},\tilde{\boldsymbol{X}}) = \int p(\boldsymbol{\varphi}_b|\boldsymbol{\Phi},\boldsymbol{\theta},\tilde{\boldsymbol{Y}},\tilde{\boldsymbol{X}})p(\boldsymbol{\Phi},\boldsymbol{\theta}|\tilde{\boldsymbol{Y}},\tilde{\boldsymbol{X}})d\boldsymbol{\Phi}\, d\boldsymbol{\theta} \quad (10)$$

Given the insensitivity of the multipliers to the data, this further simplifies to

$$p(\boldsymbol{\varphi}_b|\tilde{\boldsymbol{Y}},\tilde{\boldsymbol{X}}) = \int p(\boldsymbol{\varphi}_b|\boldsymbol{\Phi})p(\boldsymbol{\Phi}|\tilde{\boldsymbol{Y}},\tilde{\boldsymbol{X}})d\boldsymbol{\Phi} = \underset{p[\boldsymbol{\Phi}|\tilde{\boldsymbol{Y}},\tilde{\boldsymbol{X}}]}{\mathrm{E}}[p(\boldsymbol{\varphi}_b|\boldsymbol{\Phi})] \quad (11)$$

where the notation $\underset{p[\boldsymbol{\Phi}|\tilde{\boldsymbol{Y}},\tilde{\boldsymbol{X}}]}{\mathrm{E}}[p(\boldsymbol{\varphi}_b|\boldsymbol{\Phi})]$ denotes the "expected posterior hyperdistribution." If the posterior uncertainty in the hyperparameters $\boldsymbol{\Phi}$ is low, the expected posterior hyperdistribution is almost identical to the hyperdistribution evaluated with the expected posterior hyperparameters.

[83] This derivation explains why the marginal posterior of the insensitive rainfall multipliers becomes near-identical to the expected posterior hyperdistribution (Figure 12). In particular, their expectations are near-identical to the posterior expectation of μ and their standard deviation becomes very similar to the posterior expectation of σ.

[84] Moreover, Figure 11 shows that these insensitive rainfall multipliers comprise 85% of the multipliers for BATEA_DAILY_54138_2yr and 80% for BATEA_STORM_54138_2yr. The likely cause for this is a combination of the ephemeral nature of the catchment and the lower bound of 0.4 mm on the standard deviation of output errors (this lower bound appears too high).

[85] Figure 12 shows the QQ plot of the rainfall multipliers for BATEA_STORM_54138_2yr and includes the posterior 95% probability limits and a comparison to the theoretical distribution (equation (5) with posterior expected value of the hyperparameters). Again, the slope is less than expected around the midrange of the distribution, suggesting that the lognormal multiplier assumption is not sup-

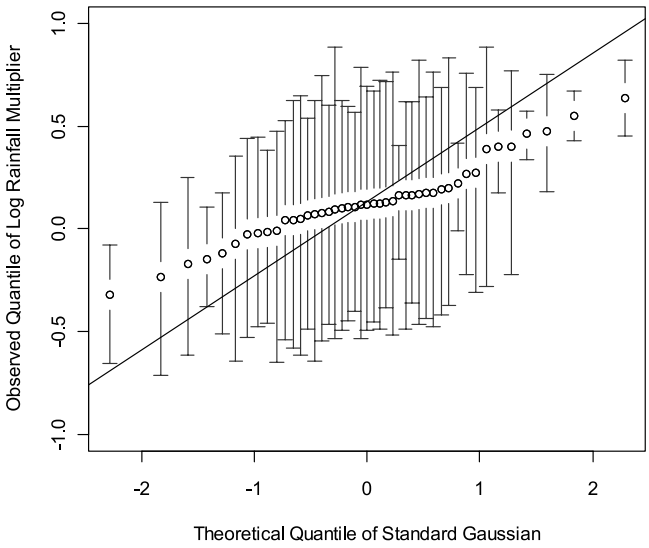


Figure 12. QQ plot of rainfall multipliers for BATEA_STORM_54138_2yr. Circles show the posterior expected value of each multiplier, and the bars provide the posterior 95% probability limits. The solid line denotes the theoretical distribution based on the posterior expected values of the hyperparameters.

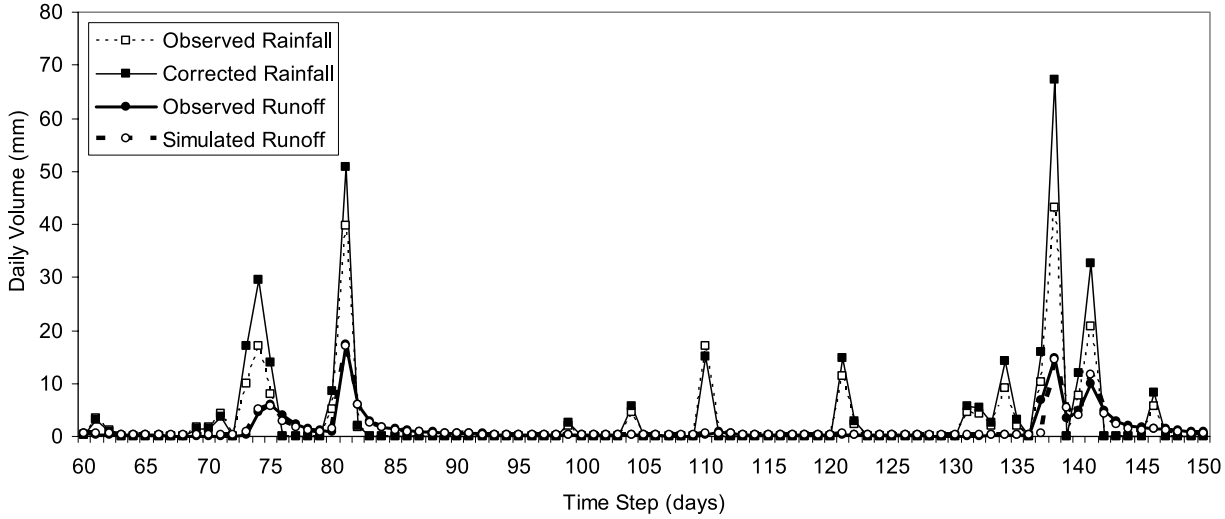


Figure 13. Comparison of observed and "corrected" rainfall and observed and simulated runoff for BATEA_STORM_54138_2yr.

ported and a distribution with fatter tails would be more appropriate. Similar findings have been reported by *Kuczera et al.* [2006].

[86] However, it is clear that the insensitive multipliers cluster in this midrange because they remain near-identical to the posterior expectation of μ. Therefore the lower slope in the midrange of the QQ plot is caused by insensitive multipliers that are poorly identified from the data. Hence, it remains unclear if the lognormal assumption is violated. In addition, because of the large number of insensitive multipliers, it is difficult to ascertain the autocorrelation properties of these multipliers. Further work is needed to refine the selection and analysis of the rainfall multipliers that can be informed by the rainfall-runoff data.

5.5. Analysis of "Optimal" Parameters and Simulations

[87] Analysis of the "optimal" parameter set, defined here as the parameter set that maximizes the posterior probability distribution (hence referred to as "modal parameter set") is of hydrological interest because it provides a continuous model run using the most likely CRR model parameters and, for BATEA, rainfall multipliers. However, its significance should not be overestimated: focusing solely on the modal predictions in lieu of the entire predictive distribution can cause a substantial loss of information from the full posterior and corresponds to a 0–1 loss function [*Bernardo and Smith*, 2000].

5.5.1. Comparison of Observed and Simulated Rainfall/Runoff Series for BATEA

[88] Applying the rainfall multipliers to the observed rainfall time series provides an estimate of the true rainfall, hereafter referred to as the "corrected" rainfall. Figure 13 compares the corrected rainfall to the observed rainfall/runoff and the simulated streamflow for BATEA_STORM_54138_2yr (similar results were found for BATEA_DAILY_54138_2yr). Figure 13 shows that (1) the rainfall correction is relatively moderate and is consistent with other estimates of rain gauge measurement errors (see section 7.3) and (2) runoff estimates provided by BATEA modal parameter estimates are a close match to the observed runoff (Nash Sutcliffe Statistic = 0.93).

[89] However, as discussed by *Huard and Mailhot* [2008], the comparison of observed and optimal simulated outputs can be misleading if the calibration uses extra degrees of freedom to account for input errors. Indeed, since the rainfall multipliers are inferred (and hence optimized) along with the CRR parameters they can potentially compensate for structural errors of the model. This can yield near-perfect matches of the simulated runoff to the observed runoff even for a severely flawed CRR model. Consequently, calibration under conditions of uncertain inputs requires more probing diagnostics than calibrating to error-free inputs. The predictive QQ plot in validation (see section 5.2) is very useful in this respect. It is also stressed that the potential interaction of the multipliers and the structural errors of the CRR model does not imply that input errors should be ignored (see section 7.1 for further discussion).

5.5.2. Analysis of State Variables

[90] No model assessment is complete without an analysis of the internal state variables, even if these do not have a direct physical interpretation. In GR4J the state variables are the production and routing stores. The state variables corresponding to the modal parameter values were compared for each of the calibration runs (Figures B1–B4 in Appendix B).

[91] In general, we found no obvious anomalies in the state variables. Empirical analysis suggests that the state variables estimated using SLS and, to a lesser extent, WLS, depended strongly on the rain gauge used in the calibration. Conversely, the state variables estimated using BATEA were more consistent. These findings held for both calibration periods and both stores (differences between the calibration methods were more pronounced for the production store). This is not surprising because BATEA yields more consistent parameters and predictions with respect to the rain gauges than SLS/WLS. Since the simulated runoff is a function of the state variables, consistent runoff predictions imply consistent state variable behavior.

6. Consistency of Parameter Estimates

[92] A fundamental assumption made by most calibration frameworks is that the CRR parameters are stationary over

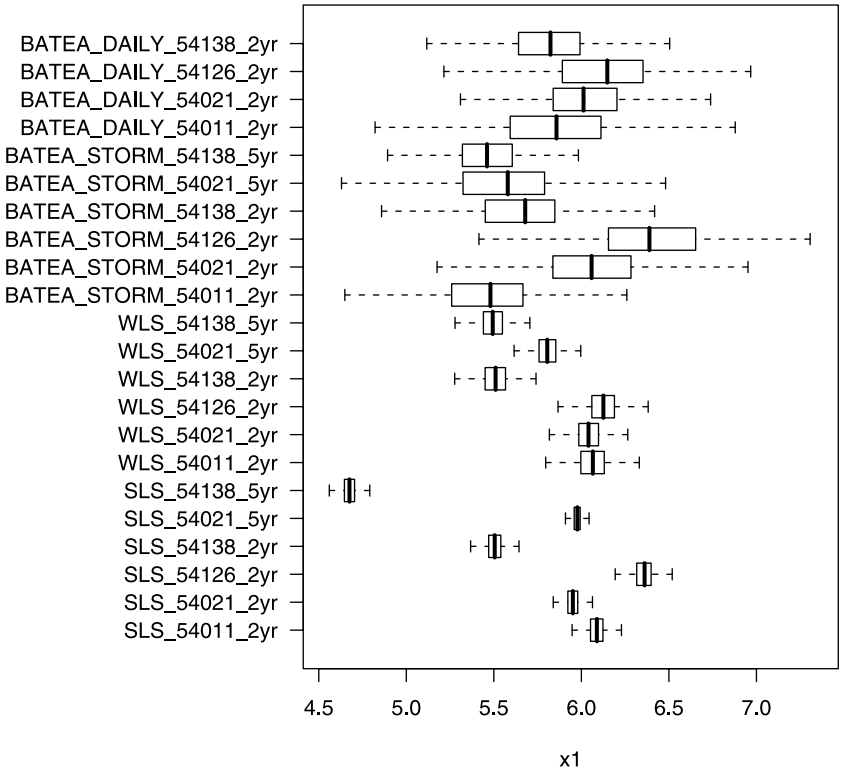


Figure 14. Box plots of marginal posterior distributions of GR4J parameter. The marginal posterior is described by a box plot where the ends of the box represent the 25% and 75% quantiles, the whiskers are the 5% and 95% quantiles, and the vertical bar denotes the median.

time. Moreover, one would like the parameters to be consistent regardless of the choice of rain gauge. To test this assumption, the marginal posterior distributions of the GR4J parameters and rainfall multiplier hyperparameters are compared for different combinations of calibration periods and rain gauges.

6.1. Visual Assessment of Parameter Consistency

[93] Figure 14 shows the marginal posterior distributions of the GR4J parameter x_1 (similar behavior was observed for the other GR4J parameters). Analysis of the marginal distributions suggests the following:

[94] 1. SLS parameter estimates are highly inconsistent between different calibration periods. Likewise the SLS distributions are inconsistent for different rain gauges, even if a rainfall multiplier (constant over the calibration period) is calibrated to compensate for rainfall gradient effects.

[95] 2. WLS parameter estimates display a more consistent behavior, although some of the distributions do not overlap.

[96] 3. The posterior spread of SLS estimates is typically much tighter than that of WLS and BATEA distributions. The parameter uncertainty reported by BATEA is larger because of its recognition of input uncertainty. In general, SLS underestimates the parameter uncertainty [e.g., *Beven and Binley*, 1992; *Kavetski et al.*, 2002].

[97] 4. The parameter distributions inferred using BATEA_STORM and BATEA_DAILY are significantly more consistent than WLS and SLS for all calibration periods and rain gauges. This is an important finding, since a necessary (though not sufficient) condition for successful regionalization of CRR parameters is that the parameter estimates be robust with respect to choice of rain gauge and calibration period.

6.2. Quantitative Measures of Consistency

[98] A quantitative measure of the parameter consistency was developed to compare marginal posterior distributions across different data sets. This measure was based on decomposing the total variance of estimated parameters into within- and between-group variances, where the groups are defined according to the rain gauge used for calibration, or alternatively, according to the calibration period. The measure closely resembles analogous criteria used in other statistical applications, including (1) in cluster analysis to optimize the cluster groupings [*Mirkin*, 2005], (2) convergence assessment of MCMC chains [*Gelman et al.*, 2004], and (3) in standard ANOVA methods.

[99] Let $\boldsymbol{\theta}_{i,j}$ be a collection of samples of parameter $\boldsymbol{\theta}$, where $i = 1,\ldots,N_{sim}$ indexes the posterior samples of the parameters, and $j = 1,\ldots,K$ indexes the calibration data sets (e.g., for different rain gauges or calibration periods). The estimated overall variance of parameter $\boldsymbol{\theta}$ can be decomposed into the sum of between-group and within-group variances, var_B and var_W respectively, as follows:

$$\begin{aligned}\mathrm{var}_T[\boldsymbol{\theta}] &= \frac{1}{KN_{sim}}\sum_{j=1}^{K}\sum_{i=1}^{N_{sim}}(\boldsymbol{\theta}_{i,j}-\boldsymbol{\theta}_{\cdot,\cdot})^2 \\ &= \frac{1}{K}\sum_{j=1}^{K}(\boldsymbol{\theta}_{\cdot,j}-\boldsymbol{\theta}_{\cdot,\cdot})^2 + \frac{1}{K}\sum_{j=1}^{K}\frac{1}{N_{sim}}\sum_{i=1}^{N_{sim}}(\boldsymbol{\theta}_{i,j}-\boldsymbol{\theta}_{\cdot,j})^2 \\ &= \mathrm{var}_B[\boldsymbol{\theta}] + \mathrm{var}_W[\boldsymbol{\theta}] \end{aligned} \quad (12)$$

Table 3. Parameter Consistency Across Different Rain Gauges for the 2-Year Calibration Period

Method	Parameter					
	x_1	x_2	x_3	x_4	μ	σ
SLS	0.03	0.15	0.06	0.29	0.02	na
WLS	0.12	0.18	0.19	0.77	0.04	na
BATEA_STORM	0.42	0.86	0.44	0.87	0.12	0.80
BATEA_DAILY	0.88	0.91	0.93	0.89	0.20	0.57

where $\boldsymbol{\theta}_{.j} = \frac{1}{N_{sim}} \sum_{i=1}^{N_{sim}} \boldsymbol{\theta}_{i,j}$ is the mean value of $\boldsymbol{\theta}$ inferred from data set j, and $\boldsymbol{\theta}_{...} = \frac{1}{K} \sum_{j=1}^{K} \boldsymbol{\theta}_{.j}$ is the overall mean.

[100] If the parameter estimates are consistent across different data sets, the between-group variance var_B (differences between the mean parameters inferred from each data set) should be close to zero. Conversely, inconsistent parameter estimates will result in the total variance being dominated by var_B. Consequently, the statistic $R = \text{var}_W/(\text{var}_W + \text{var}_B)$ can be used to quantify the consistency of parameters, with $R \approx 0$ implying inconsistent parameters and $R \approx 1$ implying consistent parameters.

[101] Table 3 reports the parameter consistency of the four calibration methods across the four rain gauges for the 2-year calibration period. Table 3 confirms the visual assessment that SLS and, to a lesser extent, WLS produce inconsistent CRR parameter estimates. In contrast, the BATEA-inferred parameters are generally more consistent. Lower consistency is observed for the estimated means and standard deviations of rainfall errors, which is not surprising because rainfall errors are unlikely to have the same distribution at different rain gauges.

[102] Note the distinctly better performance of BATEA_DAILY over BATEA_STORM. This is likely due to BATEA_DAILY being more flexible than BATEA_STORM: the latter imposes much more structure on the rainfall errors, with only a single multiplier for an entire storm epoch that could last several days. Whenever this is inappropriate (e.g., significant variability of errors within a single storm epoch) it may degrade the parameter consistency.

[103] Table 4 reports the parameter consistency with respect to the two different calibration periods (2 and 5 years). The results are similar: significant inconsistencies appear in SLS and WLS, whereas, with a few exceptions, BATEA_STORM yields consistent parameter estimates.

[104] The hyperparameters of rainfall errors also appear to be consistent across different periods, supporting the hypothesis that the rainfall errors are stationary. Note that BATEA could also be used with nonstationary errors, but this would require additional information (knowledge of the trends, etc.) that is currently unavailable.

[105] Figure 15 shows posterior marginal distributions for the standard deviation of the rainfall multipliers for the two BATEA methods. It shows that BATEA_STORM yields more precise estimates than BATEA_DAILY. This is not surprising because, for a multiday storm, BATEA_DAILY would have daily rainfall multipliers for each day within the storm, whereas BATEA_STORM would have only a single multiplier for the same storm. We would hence expect larger variability in the multipliers estimated using BATEA_DAILY.

6.3. Parameter Precision

[106] Sections 6.1 and 6.2 illustrate that BATEA provides significantly more consistent parameter estimates than WLS and SLS. However, while BATEA-based parameters are much more consistent (suggesting, though not proving, higher accuracy and robustness), they are also more uncertain (i.e., are less precise). This is not surprising because BATEA recognizes the additional data uncertainty (in the rainfall data), which is ignored in SLS and WLS. In turn, recognizing additional uncertainties in the data generally yields larger uncertainty in the inferred parameters. *Vrugt et al.* [2005] report similar results when comparing the Simultaneous Optimization and Data Assimiliation (SODA) framework, which incorporates an additive combined structural/input error and output measurement error, to the Bayesian SLS approach.

[107] This finding raises the question of whether the total predictive uncertainty becomes dominated by CRR parameter uncertainty. One of the advantages of the BATEA methodology is that it allows an evaluation of the contribution of each source of uncertainty to the total predictive uncertainty. As shown in Figure 16, in this case study the parameter uncertainty is not the major contributor to the total predictive uncertainty (which is dominated by the uncertainty in the rainfall). Therefore, our preference is to seek more reliable and consistent parameter estimates even if they come at the expense of reduced precision.

7. Improvements Required for Robust Estimates of Predictive Uncertainty of CRR Models

7.1. Is Input Error Compensating for Other Sources of Error?

[108] This study used BATEA to infer input errors, given a rainfall-runoff model and an output error model derived from the rating curve. While BATEA can explicitly incorporate structural error using storm-dependent parameters [*Kuczera et al.*, 2006], this was not attempted because synthetic studies suggest that simultaneous inference of input and structural errors can become ill posed if vague priors are used on both sources of error [*Renard et al.*, 2008; see also *Kavetski et al.*, 2006a, section 3.3]. A possible alternative is to use the output error model to absorb structural errors (*Kavetski et al.* [2006a], see also *Huard and Mailhot* [2008] for a similar approach).

Table 4. Parameter Consistency Across the 2- and 5-Year Calibration Periods

Method	Rain Gauge 54021						Rain Gauge 54138					
	x_1	x_2	x_3	x_4	μ	σ	x_1	x_2	x_3	x_4	μ	σ
SLS	0.88	0.00	0.01	0.28	0.01	na	0.01	0.18	0.16	0.51	0.04	na
WLS	0.30	0.09	0.22	0.70	0.99	na	0.99	0.06	0.14	0.82	0.06	na
BATEA_STORM	0.61	0.41	0.89	0.90	0.65	0.86	0.86	0.95	0.95	1.00	0.81	0.69

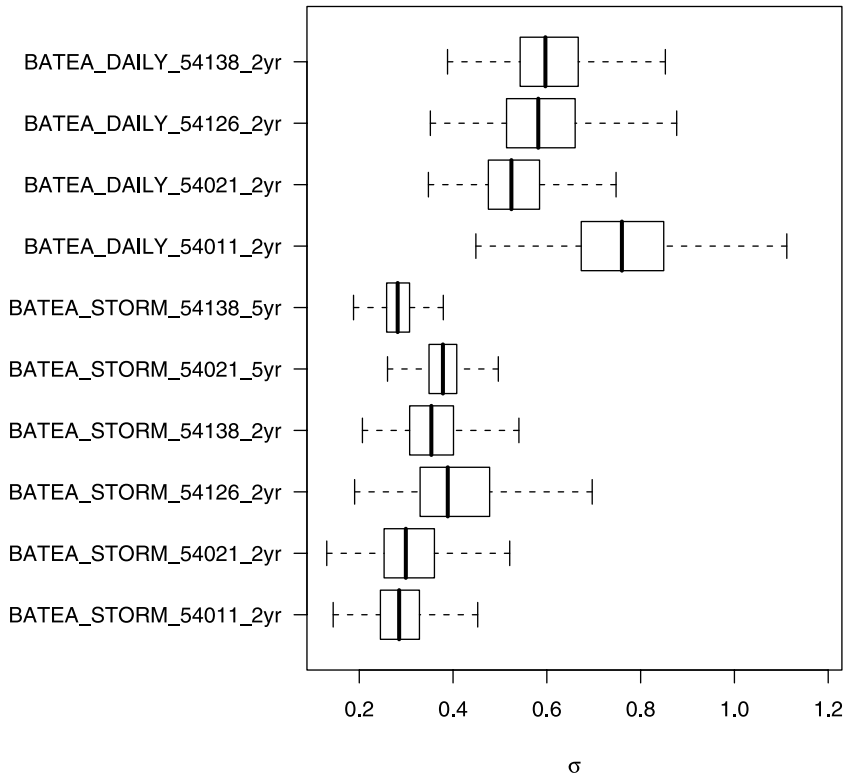


Figure 15. Box plots of marginal posterior distributions of the standard deviation of the rainfall multipliers, inferred for different rain gauges and calibration periods.

[109] Given that structural error was not incorporated explicitly, it is possible that the input error latent variables, intended to represent input errors, are also compensating for structural error. For example, in the GR4J model (as in many other models), the net rainfall is computed by subtracting the potential evapotranspiration from the rainfall [*Perrin et al.*, 2003]. Hence, random and/or systematic errors in the potential evapotranspiration, or in the conceptualization of the rainfall interception store, would have effects indistinguishable from rainfall input errors. In addition, GR4J has a groundwater exchange term (parameterized by x_2), which represents the catchment water gains/losses via groundwater. This process also has the potential to interact with rainfall errors.

[110] This issue was investigated by determining the posterior correlation between the mean rainfall multiplier and the groundwater exchange coefficient x_2 for the different calibration methods. For SLS/WLS, the mean rainfall multiplier, i.e., the single value scaling the entire rainfall series (section 4.1), had an expected posterior correlation of

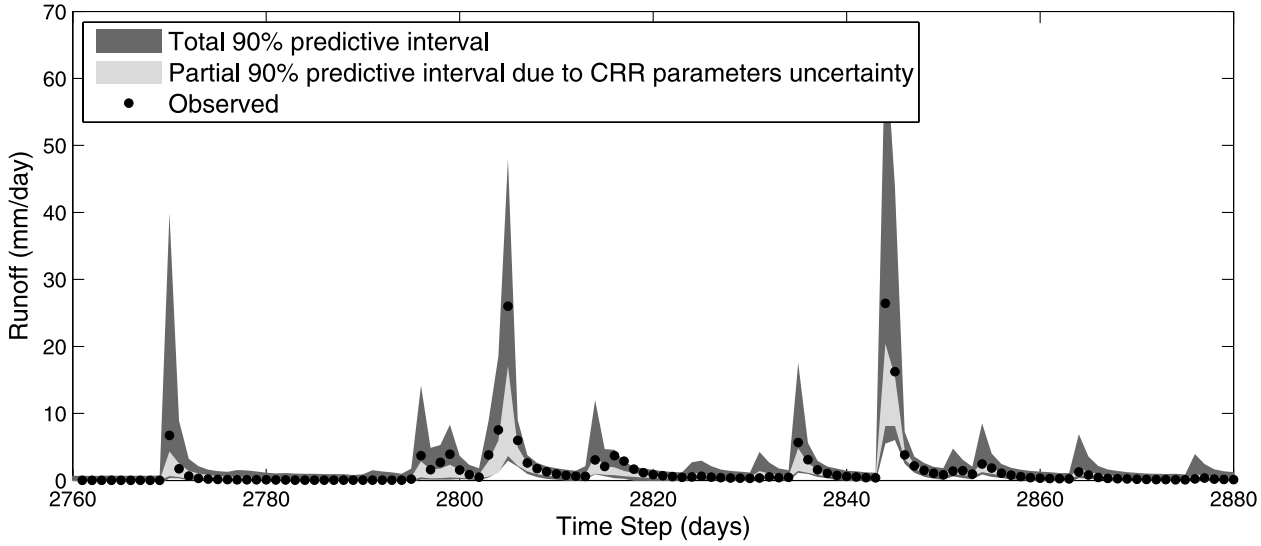


Figure 16. Total and partial predictive uncertainty in the validation period obtained using BATEA_STORM_54138_2yr.

−0.80 with x_2, while for WLS the correlation was −0.88. This indicates that the SLS/WLS multiplier and the x_2 parameter were indeed interacting with each other. For the BATEA methods, the expected posterior correlation between the hypermean μ of the rainfall multipliers and x_2 was somewhat lower, −0.58 for BATEA_STORM and −0.38 for BATEA_DAILY.

[111] It follows that reducing the time scale at which the rainfall multipliers operate (from the entire time series for SLS/WLS, to storm epochs for BATEA_STORM, to daily for BATEA_DAILY) reduces the correlation between x_2 and the rainfall multipliers. This is likely due to the rainfall errors and the groundwater structural errors operating at different time scales and impacting on different parts of the hydrograph.

[112] Synthetic studies have shown that joint inference of both input error and structural error with vague priors can become ill posed [*Renard et al.*, 2008]. Hence it is preferable to use independent information to set appropriate priors on the rainfall errors and/or other model parameters (e.g., groundwater exchange coefficient). Since the hyper-distribution must reflect the modelers' understanding of the corruption mechanisms affecting observed inputs, this also calls for a more careful data collection in hydrology: each observation must be accompanied by its uncertainty estimate. We expect that accurate estimates of input uncertainty would significantly reduce compensatory effects between input and structural errors and yield a more precise inference. In the absence of such information, reliable separation of input and structural errors remains highly problematic.

7.2. Comparison of Daily and Storm Epoch Approaches for Characterizing Input Errors

[113] The storm epoch approach [*Kavetski et al.*, 2002, 2006b; *Kuczera et al.*, 2006] reduces the computational cost of the inference by reducing the number of latent variables to be inferred. Statistically, using storm epochs corresponds to an assumption of perfect correlation between daily multiplicative rainfall errors within a single storm epoch.

[114] Comparison of BATEA_STORM and BATEA_DAILY results revealed no major differences in terms of the posterior diagnostics. The predictive QQ plot is slightly better for BATEA_DAILY than BATEA-STORM in calibration, but both methods perform similarly in validation. The residual QQ plot also shows little difference, while the residual PACF shows BATEA_STORM has a slightly lower autocorrelation than BATEA_DAILY in calibration.

[115] Examination of the correlation between individual rainfall multipliers for both BATEA_DAILY_54138_2yr and BATEA_STORM_54138_2yr revealed no significant correlations between inferred multipliers.

[116] In terms of parameter consistency, it appears that BATEA_DAILY yields more consistent parameter estimates than BATEA_STORM and is less likely to be compensating for groundwater exchange errors. Finally, the most significant difference between the daily and the storm epoch approaches is that BATEA_DAILY produced estimates of the standard deviation of rainfall multipliers that were double that of BATEA_STORM (see Figure 15). Overall, current results suggest that BATEA_DAILY may be a more appropriate temporal resolution for the input errors, but further analysis using dense rain gauge networks and radar data is needed to more conclusively identify the most appropriate temporal resolution for the rainfall errors.

[117] It would also be very useful to develop relationships between storm types and rainfall multipliers. This would reduce the total predictive uncertainty in the validation period, with rainfall multipliers sampled conditionally on the storm type. This analysis could be carried out at both the daily and storm time scale and will be attempted in future work.

7.3. Comparison to Rainfall Error Estimates Reported in the Literature

[118] *Linsley and Kohler* [1988] report a rainfall error analysis of the 2000 km^2 Muskingum Basin, Ohio using a dense gauge network. They report a standard error of 35% for single-gauge rainfall estimates. This estimate is identical to the posterior median of the standard deviation of the multipliers inferred using BATEA_STORM (note that the standard deviation of multipliers corresponds to the standard error in the rainfalls). However, for BATEA_DAILY the posterior median is 0.6–0.7, which is twice higher than Linsley and Kohler's estimate.

[119] The Muskingum basin has a humid continental climate, while the Horton catchment is humid subtropical (according to the Koppen Classification System [*Peel et al.*, 2007]). Both catchments have reasonably similar rainfall patterns, with annual average rainfall of 800–900 mm and summer-dominated rainfall patterns (which can be extremely variable). The elevation range of the Muskingum Basin is considerably less than in the Horton, 200–300 m versus 1000 m respectively. Given the similar climates, but the more varied topography of the Horton catchment, the standard error from *Linsley and Kohler* [1988] could be viewed as a conservative estimate of the rainfall errors expected in the Horton catchment. We do warn that analyses and interpretations of multipliers inferred using a BATEA application that ignores structural errors should be viewed with caution because of the potential for interaction (see section 7.1). In addition, many of these multipliers are insensitive (see section 5.4).

7.4. Limitations of the Multiplicative Input Error Model

[120] To date, all BATEA applications have used rainfall multipliers to represent rainfall errors. While this accounts for the likely heteroscedasticity of rainfall errors, it cannot handle situations where a rainfall event is not recorded by a rain gauge. This type of input error is particularly challenging. While inspection of the time series for the Horton catchment revealed little evidence of missed rainfall (all significant runoff events were associated with rainfall events), in larger catchments, especially with low rain gauge density and convective type rainfall events, the rainfall gauges may miss significant rainfall events. More sophisticated error models that allow for errors in near-zero recorded rain measurements will be presented in future work.

7.5. Improving the Output Error Model

[121] As outlined in section 2.4, the output error model represents two sources of uncertainty: (1) sampling and measurement errors in the observed runoff and (2) "remanent errors" that were not accounted by the input and structural error models. In this study, the output error model is based solely on rating curve analysis and thus represents

solely sampling and measurement error only. Since no other treatment of model structural error was implemented, it is possible that the latent variables, here intended strictly for input errors, can also be compensating for structural errors.

[122] A more conceptually appealing approach would be to infer the component of the output errors representing remanent errors unaccounted by the input error model and/or storm-dependent parameters. However, this can result in poor identifiability unless accurate and precise prior information on the input uncertainty is available [*Renard et al.*, 2008].

[123] While BATEA_STORM and BATEA_DAILY appear to satisfy the assumptions of the output error model better than SLS and WLS, there remains a need for a better characterization of output errors, particularly for near-zero runoffs, where misspecification of the error distribution can exert undue leverage on the likelihood function.

8. Implications for Model Extrapolation and Regionalization

[124] The ability of BATEA to infer parameters that are not biased by input error is an important advance for two key practical challenges of catchment modeling: model extrapolation and regionalization of CRR models.

[125] Model extrapolation can be as simple as investigating the impact of including an additional rain gauge in the catchment on the runoff estimated using a CRR model, or as sophisticated as assessing the impact of climate change on runoff given rainfalls modified using a climate change model. In both cases, the rainfall (and therefore the input errors) used in the extrapolation is different than in the calibration period. This study indicates that neither type of extrapolation can be done reliably using SLS-calibrated parameter estimates because they are biased by rainfall errors. BATEA has the potential to overcome these limitations because its parameter estimates are less dependent on the specific realization of input errors in the selected calibration period (Figure 14).

[126] Regionalization refers to the determination of CRR parameters without recourse to calibration and is a key challenge in hydrology because the majority of catchments are ungauged and have little or no streamflow observations. Regionalization is a type of spatial model extrapolation that requires the development of ''regional relationships,'' e.g., relating CRR parameters to catchment characteristics. However, since input errors are likely to vary from catchment to catchment depending on rain gauge density and location, as well as on catchment climate and topography, it is unlikely that input error biased SLS parameter estimates could be meaningfully regionalized. Indeed, previous studies using SLS/WLS-type methods to calibrate the CRR parameters prior to regionalizing them have shown poor predictive power. For example, *Chiew and Siriwardena* [2005, p. 2889] conclude ''The modelled monthly runoffs [. . .] are reasonable in about three quarters of the catchments, where the Nash-Sutcliffe model efficiency is greater than 0.6 and the total modelled runoff is within 30% of the total recorded runoff.'' This is not an encouraging conclusion.

[127] This study suggests that the poor predictive power of SLS-based regionalization is explained at least in part by parameter biases arising from ignoring rainfall uncertainty (which results in troublesome parameter sensitivity to the calibration period and rain gauge). This bias arises because SLS forces the CRR parameters to compensate for the specific realization of rainfall errors, which varies unpredictably between data periods and rain gauges. In contrast, BATEA-based parameter estimates were consistent for all rain gauges and calibration periods, eliminating one of the obstacles to CRR parameter regionalization.

[128] The regionalization of BATEA-based CRR parameters to an ungauged site does not require regionalizing the input/output error models, precisely because BATEA estimates are (relatively) independent of input/output data errors. However, if this regionalized CRR model was used for streamflow predictions, it would provide the uncertainty in streamflow due to the regionalized CRR model parameters only. More reliable estimates of the total predictive uncertainty in the streamflow predictions at an ungauged catchment would require the development of an input error model for this catchment. This could be accomplished by analysis of the rain gauge network and may further benefit from regionalizing the input error model. Further research is needed to determine how this could be achieved for different catchments.

[129] It is stressed that the reliance of BATEA on explicit input error models is a strength, rather than weakness of the approach vis-à-vis methods that do not explicitly use such models. While superficially, SLS or WLS calibration do not ''use'' input error models, they actually correspond to a special case of BATEA with all multipliers fixed at 1.0 (the Dirac hyperdistribution, see also *Kavetski et al.* [2002]). Consequently, they correspond to using an error model that is known to be highly incorrect.

[130] It also follows that regionalizing SLS parameter estimates and using them for prediction simply corresponds to assuming a regionalized input error model that ignores rainfall uncertainty. It is stressed that BATEA does not necessarily make any more assumptions than SLS or WLS, it merely makes its assumptions transparent and explicit, and offers a systematic procedure for checking these assumptions against empirical evidence.

[131] An alternative approach for regionalization is based on calibrating a CRR model to estimates of runoff statistics. This has had encouraging results for European catchments [*Bardossy*, 2007], but average to poor results for Australian catchments [*Boughton and Chiew*, 2007]. Combinations of these two approaches, i.e., regionalizing both parameters and runoff statistics, could be necessary for meaningful regionalization and these will be investigated in future research.

9. Future BATEA Applications

[132] The performance of BATEA in other modeling contexts, including using semidistributed and distributed hydrological models, wetter catchments, etc, is of interest. In general, most calibration methods and models perform better in wetter climates because (1) the catchment dynamics are less threshold driven (and hence less nonlinear), (2) more runoff information is available to infer CRR parameters, and (3) the development of reliable output error model for near-zero runoffs is less critical.

[133] On the other hand, the application of BATEA to semidistributed and distributed models requires care to avoid prohibitive computational costs. If a separate input

error model is specified for each modeling unit, with only a single runoff series available for the entire network, the problem would likely become ill posed and computationally intractable. Instead, rainfall and topographic information should be used as prior information to develop more precise data uncertainty and CRR models. Kriging of rainfall fields, radar data, etc, could be used for this purpose [e.g., *Kuczera and Williams*, 1992].

10. Conclusions

[134] Three calibration frameworks, including the widely used SLS and WLS methods and the more recent BATEA methodology, were used to calibrate the rainfall-runoff model GR4J to a difficult-to-model ephemeral catchment. The key assumptions of each method were scrutinized, focusing on (1) evaluating predictive uncertainty and (2) parameter consistency. The Horton catchment (New South Wales, Australia) was used because of its challenging ephemeral hydrological dynamics and large rainfall gradients. These types of catchments are notoriously difficult to calibrate.

[135] Assessment of requirement 1 using standard diagnostics (tests of probability model assumptions) showed that BATEA provided a significant improvement over SLS and WLS. Furthermore, a diagnostic of the total predictive uncertainty in validation was presented. This simple quantile-based plot provides an excellent summary of the performance of probabilistic prediction methods. Here, it showed that all calibration methods performed poorly during low-flow periods, while BATEA provided more reliable estimates of predictive uncertainty during higher flows than both WLS and SLS.

[136] Requirement 2 was evaluated by examining the parameter consistency for each of the calibration methods when calibrating the same CRR model to the same catchment runoff data using different rain gauges/time periods. The results showed that BATEA provided much more consistent parameter estimates than both SLS and WLS, with the latter yielding results highly dependent on the calibration period and rain gauge. These results suggest that regionalization of SLS/WLS-based estimates of model parameters is likely to be unreliable because of input-error-induced biases. BATEA offers a way to overcome these problems. Moreover, its Bayesian foundation offers opportunities to incorporate additional knowledge in the calibration and in regionalization, including relationships between rainfall errors and storm types, etc. This information cannot be utilized by standard methods such as SLS and WLS.

[137] The fundamental difference in the modeling philosophy between the three calibration frameworks considered in this work is that BATEA provides a systematic methodology to hypothesize, infer and evaluate models for input error, model structural error and output error. Conversely, neither SLS nor WLS can account for input uncertainty, and they both assume that model structural error and output error are simple additive random noise. Moreover, the capacity of WLS to use a more sophisticated output error model (heteroscedastic response uncertainty) is insufficient to produce reliable parameter estimates and predictions.

[138] The implementation of BATEA used in this case study incorporated input and output uncertainties, but did not explicitly consider model structural error. The treatment of structural errors using the hierarchical BATEA framework while avoiding identifiability problems remains a research challenge and will be tackled in future work.

Appendix A: Preprocessing Heuristic to Identify "Insensitive" Rainfall Multipliers

[139] The rationale behind the heuristic for identifying "insensitive" rainfall multipliers is to evaluate if a perturbation of rainfall at time t with a rainfall multiplier leads to a significant difference in simulated runoff $\hat{y}_t$. If not, rainfall at this time step is classified as "insensitive" and a rainfall multiplier is not inferred for this time step.

[140] For this heuristic the runoff is simulated using the CRR model (GR4J in this case study) with some prior estimate of its parameters (e.g., from SLS or WLS calibration).

[141] First, the range of possible rainfall perturbations needs to be specified. Prior distributions of the hyperparameters can be used for this purpose. However, exceedingly diffuse priors allowing unrealistically large values for rainfall multiplier variance (e.g., 1000% 's of errors) can cause numerical overflows in the CRR model computation. In this case study, the log-rainfall multipliers were sampled from a Gaussian distribution, $\log \varphi_t \sim N(0, 0.5^2)$.

[142] The evaluation of whether a perturbation of rainfall at time step t produces a significant different in runoff at time t proceeds as follows:

[143] 1. A series of rainfall multipliers for time steps 1 to t is first sampled. These rainfall multipliers are applied to the observed rainfall series to produce a perturbed rainfall series. Runoffs for time steps between 1 and t are simulated from the CRR model, using the perturbed rainfall series as input. This provides the first runoff value $\hat{y}_t^{(i)}$.

[144] 2. Another runoff value $\hat{y}_t^{*(i)}$ is then simulated with the same perturbed rainfall inputs except that the rainfall at time t is not perturbed by a rainfall multiplier: the observed rainfall at time step t is used.

[145] 3. The difference $\varepsilon_t^{(i)} = \hat{y}_t^{*(i)} - \hat{y}_t^{(i)}$ quantifies the difference in the runoff due to rainfall perturbation at time t taking into account the perturbations of preceding rainfalls from 1 to $t - 1$.

[146] Steps 1–3 are replicated N_{rep} times thus providing a distribution of runoff differences $\{\varepsilon_t^{(i)} : i = 1, \ldots, N_{rep}\}$ that will be used to decide if a rainfall multiplier at time t is insensitive or not.

[147] Since the impact of rainfall errors decreases with time, a memory threshold τ is used to reduce computational requirements. Rainfall perturbations prior to time step $t - \tau$ are assumed to have negligible impact on simulated runoff at time step t. Hence only simulated runoffs between $t - \tau$ and t are computed in each of the N_{rep} replicates, where τ is chosen prior to the analysis. For this case study $\tau = 10$ was used because of the fast catchment response and low base flow. For catchments with a slower response, a longer memory threshold would be more suitable.

[148] The last step is to define a criterion for deciding what is meant by a significant difference in the simulated runoff. This requires defining a range of runoff errors $\boldsymbol{I}_y$ based on the runoff measurement error model. In this case study, the 90% probability interval from the runoff measurement error model is used. If perturbing a rainfall value leads to a runoff

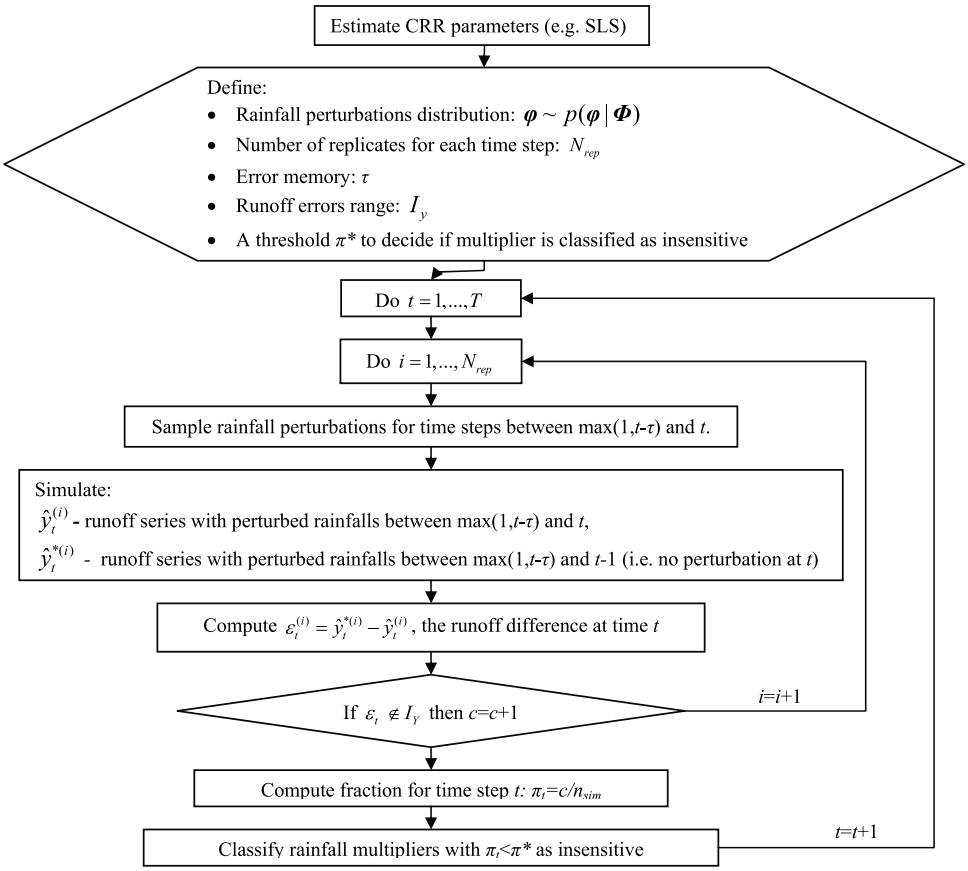


Figure A1. Flowchart of preprocessing heuristic for identifying "insensitive" rainfall multipliers.

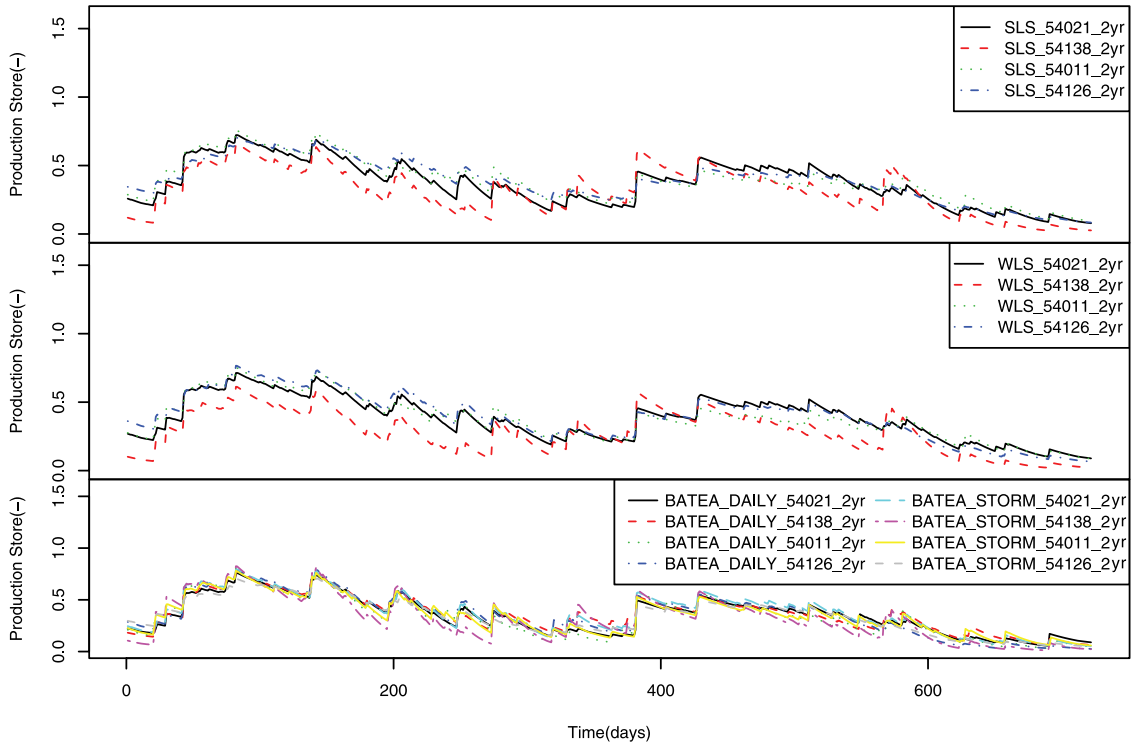


Figure B1. Production store state variable for the 2-year calibration period.

difference lying inside this runoff errors range, the corresponding rainfall multiplier is deemed "insensitive." This is applied by determining π_t, the fraction of $\{\varepsilon_t^{(i)} : i = 1,\ldots, N_{rep}\}$ lying outside the interval, $\boldsymbol{I}_y$. If this fraction π_t is below a prespecified threshold π^* ($\pi^* = 0.1$ was used in this study), then the rainfall multiplier at time step t is classified as insensitive and not inferred.

[149] A flowchart of this preprocessing heuristic is provided in Figure A1.

Appendix B: Analysis of State Variables

[150] Figures B1–B4 show the time series of the production and routine store state variable for the 2-year and 5-year

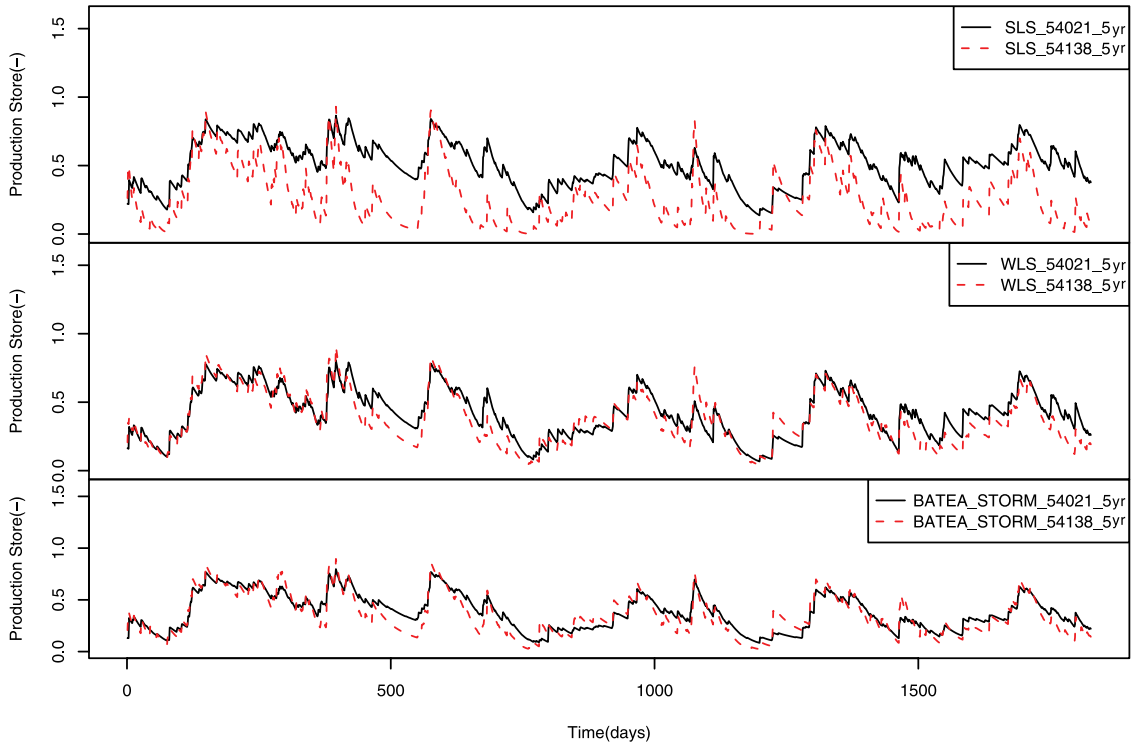


Figure B2. Production store state variable for the 5-year calibration period.

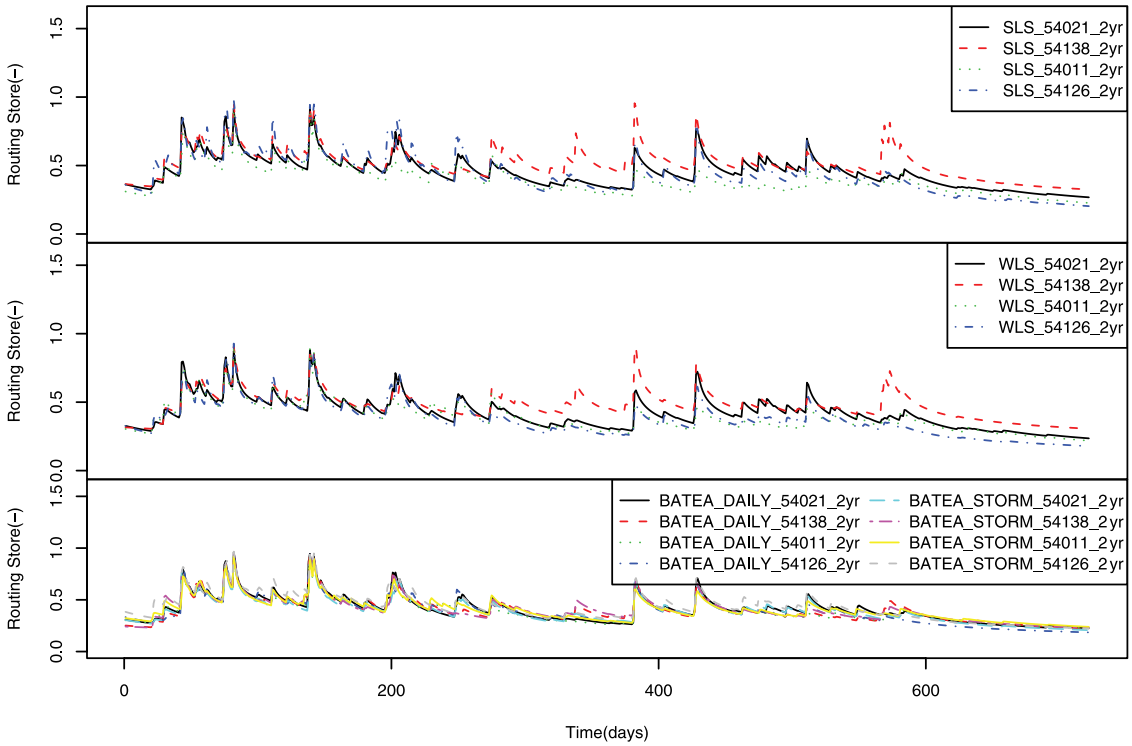


Figure B3. Routing store state variable for the 2-year calibration period.

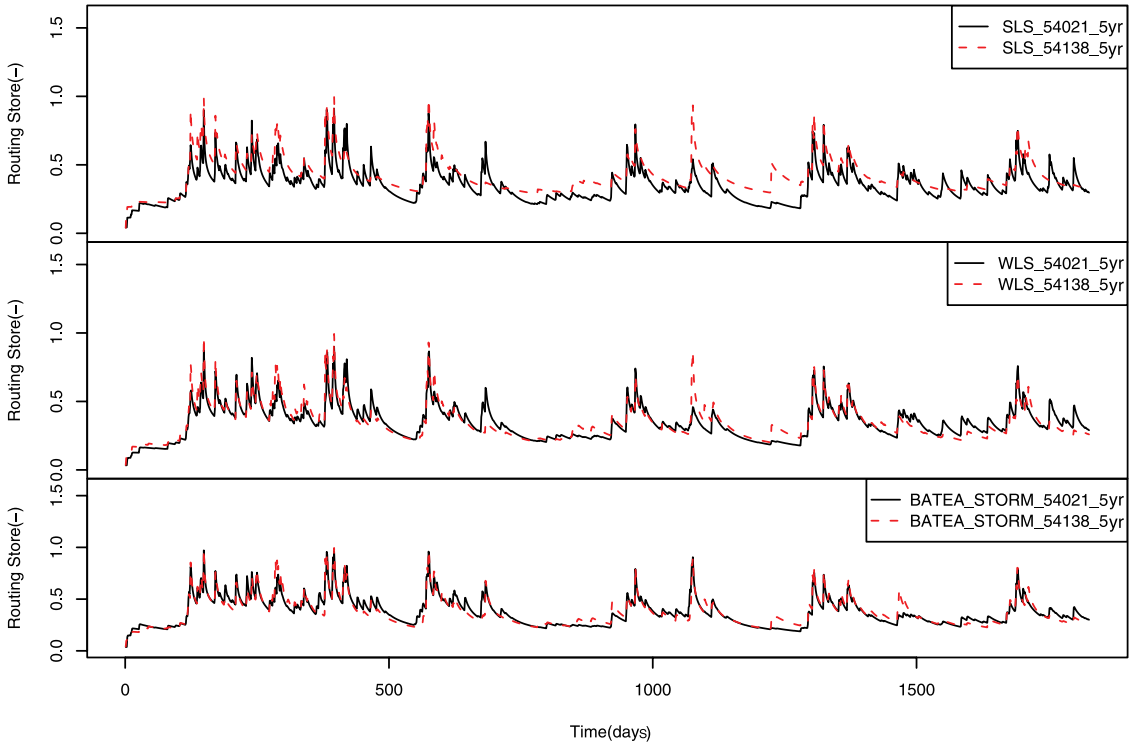


Figure B4. Routing store state variable for the 5-year calibration period.

calibration periods. The general trend is that the state variables inferred using SLS and WLS are much more dependent on the rain gauge and calibration period compared to BATEA.

[151] **Acknowledgments.** This work was partly supported by a grant from the Australia Research Council. The authors would like to thank Bettina Schaefli and two anonymous reviewers whose thoughtful comments greatly improved this paper.

References

Bardossy, A. (2007), Calibration of hydrological model parameters for ungauged catchments, *Hydrol. Earth Syst. Sci.*, *11*(2), 703–710.

Bernardo, J. M., and A. F. M. Smith (2000), *Bayesian Theory*, 2nd ed., John Wiley, Chichester, U. K.

Beven, K. J., and A. M. Binley (1992), The future of distributed models: Model calibration and uncertainty prediction, *Hydrol. Processes*, *6*, 279–298, doi:10.1002/hyp.3360060305.

Boughton, W., and F. Chiew (2007), Estimating runoff in ungauged catchments from rainfall, PET and the AWBM model, *Environ. Modell. Softw.*, *22*(4), 476–487, doi:10.1016/j.envsoft.2006.01.009.

Carlin, B. P., and T. A. Louis (2000), *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed., Chapman and Hall, Boca Raton, Fla.

Chiew, F. H., and L. Siriwardena (2005), Estimation of SIMHYD parameter values for application in ungauged catchments, paper presented at MODSIM 2005 International Congress on Modelling and Simulation, Modell. and Simul. Soc. of Aust. and N. Z., Melbourne, Vic., Australia.

Dawid, A. P. (1984), Statistical theory: The prequential approach (with discussion), *J. R. Stat. Soc., Ser. A*, *147*, 278–292, doi:10.2307/2981683.

Feyen, L., J. A. Vrugt, B. Ó. Nualláin, J. van der Knijff, and A. De Roo (2007), Parameter optimisation and uncertainty assessment for large-scale streamflow simulation with the LISFLOOD model, *J. Hydrol.*, *332*(3–4), 276–289, doi:10.1016/j.jhydrol.2006.07.004.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004), *Bayesian Data Analysis*, 2nd ed., Chapman and Hall, New York.

Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007), Probabilistic forecasts, calibration and sharpness, *J. R. Stat. Soc., Ser. B*, *69*, 243–268, doi:10.1111/j.1467-9868.2007.00587.x.

Hall, J., E. O'Connell, and J. Ewen (2007), On not undermining the science: Coherence, validation and expertise. Discussion of invited commentary by Keith Beven Hydrological Processes, 20, 3141–3146 (2006), *Hydrol. Processes*, *21*(7), 985–988, doi:10.1002/hyp.6639.

Huard, D., and A. Mailhot (2008), Calibration of hydrological model GR2M using Bayesian uncertainty analysis, *Water Resour. Res.*, *44*, W02424, doi:10.1029/2007WR005949.

Kavetski, D., S. W. Franks, and G. Kuczera (2002), Confronting input uncertainty in environmental modelling, in *Calibration of Watershed Models*, *Water Sci. and Appl.*, vol. 6, edited by Q. 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), Calibration of conceptual hydrological models revisited: 2. Improving optimisation and analysis, *J. Hydrol.*, *320*(1–2), 187–201.

Kuczera, G. A., and S. W. Franks (2002), Testing hydrologic models: Fortification or falsification?, in *Mathematical Modelling of Large Watershed Hydrology*, edited by V. P. Singh and D. K. Frevert, pp. 141–185, Water Resour. Publ., Littleton, Colo.

Kuczera, G., and B. J. Williams (1992), Effect of rainfall errors on accuracy of design flood estimates, *Water Resour. Res.*, *28*(4), 1145–1154.

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, paper presented at MODSIM 2007 International Congress on Modelling and Simulation, Modell. and Simul. Soc. of Aust. and N. Z., Christchurch, N. Z., Dec.

Laio, F., and S. Tamea (2007), Verification tools for probabilistic forecasts of continuous hydrological variables, *Hydrol. Earth Syst. Sci.*, *11*(4), 1267–1277.

Linsley, R. K., and M. A. Kohler (1988), *Hydrology for Engineers*, McGraw-Hill, London.

Mirkin, B. (2005), *Clustering for Data Mining: A Data Recovery Approach*, Chapman and Hall, Boca Raton, Fla.

Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models part I—A discussion of principles, *J. Hydrol.*, *10*(3), 282–290, doi:10.1016/0022-1694(70)90255-6.

Peel, M., F. Chew, A. Weston, and T. McMahon (2000), Extension of unimpaired monthly streamflow data and regionalisation of parameter values to estimate streamflow in ungauged catchments, 37 pp., Natl. Land and Water Resour. Audit, Canberra, A. C. T. (Available at http://www.anra.gov.au/topics/water/pubs/national/streamflow/streamflow.pdf)

Peel, M. C., B. L. Finlayson, and T. A. McMahon (2007), Updated world map of the Koppen-Geiger climate classification, *Hydrol. Earth Syst. Sci.*, *11*(5), 1633–1644.

Perrin, C., C. Michel, and V. Andreassian (2003), Improvement of a parsimonious model for streamflow simulation, *J. Hydrol.*, *279*, 275–289, doi:10.1016/S0022-1694(03)00225-7.

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Found. for Stat. Comput., Vienna, Austria.

Renard, B., G. Kuczera, D. Kavetski, M. Thyer, and S. Franks (2008), Bayesian total error analysis for hydrologic models: Quantifying uncertainties arising from input, output and structural errors, paper presented at 31st Hydrology and Water Resources Symposium, Eng. Aust., Adelaide, South Aust., Australia.

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.

Spieghalter, D. J., A. Thomas, and N. Best (2003), WinBugs, version 1.4, user manual, Biostat. Unit. Med. Res. Counc., Cambridge, U. K.

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.

Wagener, T., and H. V. Gupta (2005), Model identification for hydrological forecasting under uncertainty, *Stochastic Environ. Res. Risk Assess.*, *19*(6), 378–387, doi:10.1007/s00477-005-0006-5.

Wang, Q. J., F. L. N. McConachy, F. H. S. Chiew, R. James, G. C. deHoedt, and W. J. Wright (2001), Climatic atlas of Australia: Maps of evapotranspiration, 39 pp., Aust. Bur. of Meteorol., Canberra, A. C. T., Australia.

Wooldridge, S. A., J. D. Kalma, and J. P. Walker (2003), Importance of soil moisture measurements for inferring parameters in hydrologic models of low-yielding ephemeral catchments, *Environ. Modell. Softw.*, *18*(1), 35–48, doi:10.1016/S1364-8152(02)00038-5.

Yang, J., P. Reichert, and K. C. Abbaspour (2007), Bayesian uncertainty analysis in distributed hydrologic modeling: A case study in the Thur River basin (Switzerland), *Water Resour. Res.*, *43*, W10401, doi:10.1029/2006WR005497.

S. W. Franks, D. Kavetski, G. Kuczera, B. Renard, and M. Thyer, School of Engineering, University of Newcastle, Callaghan, NSW 2308, Australia. (mark.thyer@newcastle.edu.au)

S. Srikanthan, Water Division, Bureau of Meteorology, GPO Box 1289, Melbourne, Vic 3001, Australia.