Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors

Benjamin Renard, D. Kavetski, M. Thyer, G. Kuczera, S. Franks

Water Resour. Res., doi:10.1029/2009WR008328
author-produced version of the final draft post-refeering

Understanding predictive uncertainty in hydrologic modeling: Le challenge of identifying input and structural errors

Benjamin Renard[(1)], Dmitri Kavetski[(2)],

George Kuczera[(2)], Mark Thyer[(2)] and Stewart W. Franks[(2)]

(1) Cemagref, UR HHLY, Hydrology-Hydraulics. 3 bis quai Chauveau - CP 220, F-69336 Lyon, France

(2) School of Engineering, University of Newcastle, Callaghan, NSW, 2308, Australia

Abstract

Meaningful quantification of data and structural uncertainties in conceptual rainfall-runoff modeling is a major scientific and engineering challenge. This paper focuses on the total predictive uncertainty and its decomposition into input and structural components under different inference scenarios. Several Bayesian inference schemes are investigated, differing in the treatment of rainfall and structural uncertainties, and in the precision of the priors describing rainfall uncertainty. Compared with traditional lumped additive-error approaches, the quantification of the total predictive uncertainty in the runoff is improved when rainfall and/or structural errors are characterized explicitly. However, the decomposition of the total uncertainty into individual sources is more challenging. In particular, poor identifiability may arise when the inference scheme represents rainfall and structural errors using separate probabilistic models. The inference becomes ill-posed unless sufficiently *precise* prior knowledge of data uncertainty is supplied; this ill-posedness can often be detected from the behavior of the Monte Carlo sampling algorithm. Moreover, the priors on the data quality must also be sufficiently *accurate* if the inference is to be reliable and support meaningful uncertainty decomposition. Our findings highlight the inherent limitations of inferring inaccurate hydrologic models using rainfall-runoff data with large unknown errors. Bayesian total error analysis (BATEA) can overcome these problems using independent prior information. The need for deriving independent descriptions of the uncertainties in the input and output data is clearly demonstrated.

Keywords: hydrologic calibration, identifiability, well-posedness, predictive uncertainty, uncertainty decomposition.

1. Introduction

1.1. Confronting uncertainty in hydrologic modeling

In any modeling endeavor, reducing the total predictive uncertainty requires a robust quantitative understanding of each of its sources. In hydrology, robust characterization of the uncertainties affecting rainfall-runoff models remains a major scientific and operational challenge. Generally speaking, hydrologic modeling is affected by four main sources of uncertainty: (i) input uncertainty, e.g., sampling and measurement errors in catchment rainfall estimates; (ii) output uncertainty, e.g., rating curve errors affecting runoff estimates; (iii) structural uncertainty (sometimes referred to as "model uncertainty"), arising from lumped and simplified representation of hydrological processes in hydrologic models; and (iv) parametric uncertainty, reflecting the inability to specify exact values of model parameters due to finite length and uncertainties in the calibration data, imperfect process understanding, model approximations, etc.

Numerous approaches for quantifying the uncertainty in hydrologic predictions have been proposed, including the Generalized Likelihood Uncertainty Estimation (GLUE) [*Beven and Binley*, 1992], frequentist approaches [*Montanari and Brath*, 2004], standard Bayesian approaches [*Feyen et al.*, 2007; *Krzysztofowicz*, 2002; *Kuczera and Parent*, 1998], Bayesian Recursive Estimation [*Thiemann et al.*, 2001], Bayesian hierarchical models [*Huard and Mailhot*, 2008; *Kavetski et al.*, 2006a; *Kuczera et al.*, 2006], instrumental-variable methods [*Young*, 1998], Bayesian model averaging [*Duan et al.*, 2007; *Marshall et al.*, 2007] and others.

The Bayesian Total Error Analysis (BATEA) framework [*Kavetski et al.*, 2002; *Kavetski et al.*, 2006a; *Kuczera et al.*, 2006] was developed to explicitly represent each source of uncertainty affecting calibration and prediction of hydrological models. Several studies have shown that,

especially in the presence of large rainfall errors, BATEA offers significant improvements over traditional approaches that lump all uncertainties into a single error term and yields: (i) reduced bias and more consistent parameter estimates; and (ii) more reliable estimates of predictive uncertainty [*Kavetski et al.*, 2006a; *Renard et al.*, 2009a; *Thyer et al.*, 2009].

Unlike data uncertainty, which can be estimated by analyzing sampling and measurement designs [*Refsgaard et al.*, 2006], structural error is much harder to characterize. Several approaches have been investigated in the context of conceptual rainfall-runoff (CRR) models, ranging from traditional additive Gaussian noise representation [e.g., *Huard and Mailhot*, 2008] to Kalman filters [e.g., *Moradkhani et al.*, 2005] and stochastic perturbations of model states [*Bras and Rodriguez-Iturbe*, 1984] and parameters [e.g., *Kuczera et al.*, 2006; *Young*, 1998]. None of the current approaches appears entirely satisfactory; the optimal methodology and implementation for handling structural errors remains to be established.

Recent work has aimed at quantifying the individual contributions of input, output and structural uncertainties to the total predictive uncertainty [*Huard and Mailhot*, 2008; *Kuczera et al.*, 2006; *Moradkhani et al.*, 2005]. This can be used for: (i) diagnosing the main causes of uncertainty, suggesting avenues for improving the predictive precision of CRR models; (ii) identifying CRR model deficiencies, indicating opportunities for model improvement; and (iii) comparing CRR models without obscuring the comparison by input/output data errors. However, significant challenges remain in the development of statistical techniques for achieving this decomposition, and in the adequate specification of error models and prior knowledge necessary for a meaningful and well-posed inference.

There is a broad recognition of the limitations of rainfall-runoff data in supporting a well-posed inference of complicated CRR models [e.g., *Beven*, 2006]. The inability to infer some or all

quantities of interest from the available data is often referred to as "non-identifiability" [e.g., *Wagener et al.*, 2001]; unless prior knowledge is available, non-identifiability leads to an "ill-posed" inference (more formal definitions are given in sections 3.1 and 3.3).

While this work focuses on lumped conceptual hydrological models, similar concerns hold for more complex physically-based distributed models. Indeed, since these models have increased data requirements to support the identification and resolution of additional catchment processes, the issue of data reliability and informativeness is likely even more critical.

This study presents a quantitative analysis of the identifiability of input and structural errors using a representative set of probabilistic calibration methods, several data-knowledge scenarios and two distinct treatments of structural error. It makes a step towards a deeper understanding of the different sources of uncertainty and their effect on model calibration, and opens avenues for improving the predictive capability of environmental models. The implications of our findings on the estimation of physically-based spatially-distributed models are also briefly discussed.

1.2. Objectives

This study investigates the ability of statistical estimation, given uncertain rainfall-runoff data and an approximate hydrological model, to (i) infer reliable and precise predictive distributions of the runoff; and (ii) decompose the total predictive uncertainty, in particular, identify its input and structural components (and, moreover, identify individual input errors). Objective (i) is necessary to achieve objective (ii). We compare the ability of several distinct calibration schemes to achieve objectives (i) and (ii), and evaluate the impact of independent (prior) knowledge of the uncertainties in the calibration data.

It is stressed that this paper explores the properties of the predictive distributions of runoff and rainfall and does not attempt to investigate biases and identifiability issues in CRR models and their parameters. In particular, predictive distributions of runoff correspond to integrating over CRR parameter distributions and are the ultimate long-term objective of the majority of practical applications, especially given the growing emphasis on probabilistic risk analysis. Consequently, we limit the scope of this paper to predictive distributions and defer CRR parameter analysis to a separate study.

1.3. Outline of the presentation

The paper in organized as follows. Section 2 discusses data and structural uncertainties in further detail, while Section 3 defines and illustrates the key concepts of identifiability and well-posedness. Section 4 describes the data and CRR models, Section 5 details the Bayesian inference framework used for the analysis and Section 6 outlines the methodology. Three experiments are carried out next: Experiment A uses synthetic data and focuses solely on data errors (Section 7), Experiment B considers the effects of structural errors using synthetic data (Section 8), while Experiment C uses real-data to assess the relevance of the synthetic analysis (Section 9). The results are discussed in Section 10 and the conclusions are summarized in Section 11.

2. Data and structural uncertainties in hydrology

This section surveys distinctions between data and structural uncertainties and broadly classifies methods for treating structural errors.

2.1. The nature of data and structural uncertainties

There is a fundamental difference between the uncertainty in the data and the structural uncertainty in the CRR model itself:

(i) Data uncertainty stems from sampling, measurement and interpretation errors in the observed input/output data. Since these errors arise independently from the CRR model, their properties (e.g., means and variances of rainfall and runoff errors) can, at least in principle, be estimated prior to the calibration by analysis of the data-acquisition instruments and procedures. However, current practice seldom reports statistical measures of accuracy and precision of hydrological data (but see *Di Baldassarre and Montanari* [2009] and *Dottori et al.* [2009] for recent exceptions). This paper investigates the impact of this deficiency on the predictive capabilities of hydrological models and the decomposition of input and structural errors.

(ii) Structural uncertainty is an inherent feature of the CRR model: it is a consequence of the simplifying assumptions made in approximating the actual environmental system with a mathematical hypothesis. In general, the structural error of a CRR model depends on the model formulation (e.g. number and connectivity of stores, choice of constitutive functions, etc), on the specific catchment, and on the spatial and temporal scale of the analysis. Moreover it may vary from storm to storm, or on some other time scale. Since this uncertainty is poorly understood, specifying a meaningful prior for structural uncertainty, indeed, even formulating it mathematically, is problematic.

In practice, uncertainties in the calibration data and its finite length necessarily translate into uncertainties in the estimated CRR parameters and other inferred quantities (in a Bayesian context, "posterior parameter uncertainty"). This would occur even for an exact model, but can be

particularly pronounced when the model is approximate. In Bayesian (and frequentist) inferences, this "derived" parametric uncertainty declines as more data is included in the calibration. However, if the likelihood and/or priors are mis-specified (which, as discussed in this paper, can be detected using posterior diagnostics), the posterior will be in error (also see *Mantovan and Todini* [2006] and *Beven et al.* [2008]). Despite its asymptotic behavior, parametric uncertainty should not be ignored because it may contribute significantly to the total predictive uncertainty.

2.2. Characterizing structural uncertainty

This section outlines two broad classes of probabilistic approaches used in this paper for characterizing structural error. We also briefly survey alternative approaches.

Traditional approaches treat the CRR model as deterministic and represent structural error using an exogenous term, usually additive. Several options are possible:

A1. Lump output and structural errors into a single "residual" error term, defined as the difference between simulated and observed outputs, possibly after a transformation. This approach can be implemented both within schemes that ignore input errors (e.g., the standard least squares calibration), and within input-error sensitive methodologies [e.g., *Kavetski et al.*, 2006a].

A2. Represent output and structural errors using two separate terms, e.g., such that the difference between simulated and true outputs is structural error, while the difference between true and observed outputs is output error [e.g., *Huard and Mailhot*, 2008]. Though this allows using more specialized error models and priors, e.g., estimating streamflow uncertainty from independent gauge data, specifying a meaningful prior for structural errors remains problematic (see section 2.1).

More recent approaches abandon the notion that CRR models are deterministic. This is motivated by the stochastic nature of errors arising from spatial and temporal averaging of distributed and heterogeneous model inputs and internal fluxes, which are unavoidable in lumped models. Several related approaches have been proposed:

B1. Stochastic perturbations of the internal model states. This approach has been used in state-space approaches, such as the Ensemble Kalman Filter (EnKf) [e.g., *Moradkhani et al.*, 2005].

B2. Stochastic variation of one or more CRR parameters through time. This approach can be used with transfer function models estimated using instrumental variables [*Young*, 1998], or with general CRR models within BATEA [*Kuczera et al.*, 2006].

B3. Formulate the CRR model itself as a joint probability density function [*Bulygina and Gupta*, 2009].

In approaches A1-A2, the CRR model is deterministic in the sense that, given fixed inputs, parameters and initial conditions, it generates the same output. Conversely, in approaches B1-B3, the CRR model is viewed as stochastic: it generates a random output even for fixed inputs, parameters and initial conditions. More specifically, output randomness arises due to random variations of internal states (B1) or stochastic parameters (B2), or, more generally, due to probabilistic formulation of the model structure (B3).

As a result, in approaches A1-A2, as posterior CRR parameter uncertainty declines, the CRR model predictions quickly become deterministic and the total predictive uncertainty is dominated by the exogenous error term. Conversely, in approaches B1-B3, the CRR predictions are inherently stochastic even if the posterior uncertainty in its parameters is negligible.

Also note that approaches B1-B3 can be used to (implicitly or explicitly) reflect all sources of uncertainty, rather than just inadequacies of the model structure. Indeed, even when *intended* solely for structural errors, they may also capture at least some effects of data errors. This interaction is a key focus of our study.

The list above is not exhaustive. Assuming that structural uncertainty is epistemic rather than strictly stochastic, some authors have abandoned the formal probabilistic framework, e.g., GLUE [*Beven and Binley*, 1992] and possibilistic methods [*Jacquin and Shamseldin*, 2007]. Yet even when structural errors are epistemic, i.e., arise as a consequence of lack of knowledge of catchment dynamics, they may still behave stochastically and be characterized using standard probability theory, in particular, Bayesian methods.

Alternatively, Bayesian Model Averaging (BMA) approaches [e.g., *Duan et al.*, 2007; *Marshall et al.*, 2007] attempt to quantify structural uncertainty by combining the predictions of multiple CRR models. However, BMA's key assumption that the supplied set of models is complete is difficult to achieve and scrutinize in practice; it is unclear what the posterior predictive uncertainty actually represents when this assumption is not met.

Consequently, the calibration methods investigated in this paper are based on the hypothesis that structural uncertainty, whatever its cause, can be described by an explicit probabilistic model that is then subjected to direct scrutiny.

2.3. Prior specification of data and structural uncertainties

A critical aspect of uncertainty quantification is the specification of the parameters of the data and structural error models (e.g., variances of rainfall and runoff errors, variance of structural errors).

Early applications of BATEA [*Kavetski et al.*, 2006a] used fixed rainfall-error parameters, while *Huard and Mailhot* [2008] used fixed input/output/structural-error parameters. In Bayesian theory, this corresponds to the strongest possible prior (parameters known exactly) and would be appropriate if the statistical properties of the errors were well understood. Since this remains a challenge in hydrology, a more general formulation of BATEA treats the error-model parameters as unknown quantities that are inferred along with CRR parameters and other quantities of interest [*Kuczera et al.*, 2006]. This corresponds to weaker (more vague) priors.

A major practical question considered in this paper is the accuracy and precision of prior information needed for (i) meaningful estimation of the total predictive uncertainty and (ii) accurate attribution of the predictive uncertainty to individual sources. The influence of the priors on the reliability of the inference is of critical practical significance because it motivates the development of accurate and precise independent prior knowledge, e.g., based on densely-gauged experimental basins, etc.

3. Identifiability and well-posedness

This section defines and contrasts the concepts of "identifiability" and "well-posedness". While these concepts are necessarily technical and must be defined and used very carefully, they are central to this study and for the broader topic of statistical model identification. A simple yet informative example is used for illustration.

3.1. Identifiability

The notion of identifiability in Bayesian inference can be formalized as follows. Let $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ denote the prior and posterior distributions of a parameter vector $\boldsymbol{\theta}$ given data $\boldsymbol{y}$. At least

one component of $\boldsymbol{\theta}$ is non-identifiable if there exists a one-to-one reparameterization from $\boldsymbol{\theta}$-space into $\boldsymbol{\psi}$-space such that

$$p(\boldsymbol{\psi}_2 \mid \boldsymbol{\psi}_1, \boldsymbol{y}) = p(\boldsymbol{\psi}_2 \mid \boldsymbol{\psi}_1) \tag{1}$$

for some partitioning of $\boldsymbol{\psi}$ into subsets $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$.

Equation (1) states that parameters $\boldsymbol{\psi}_2$ are non-identifiable if the data $\boldsymbol{y}$ do not provide any information on the conditional posterior distribution of $\boldsymbol{\psi}_2$ given $\boldsymbol{\psi}_1$ [see also *Gelfand and Sahu*, 1999].

Definition (1) is more intuitive when cast in terms of the likelihood function. Applying Bayes' theorem to the LHS of equation (1) yields

$$\begin{aligned} &p(\boldsymbol{y} \mid \boldsymbol{\psi}_2, \boldsymbol{\psi}_1)\, p(\boldsymbol{\psi}_2 \mid \boldsymbol{\psi}_1) / p(\boldsymbol{y} \mid \boldsymbol{\psi}_1) = p(\boldsymbol{\psi}_2 \mid \boldsymbol{\psi}_1) \Leftrightarrow \\ &p(\boldsymbol{y} \mid \boldsymbol{\psi}_1, \boldsymbol{\psi}_2) = p(\boldsymbol{y} \mid \boldsymbol{\psi}_1) \end{aligned} \tag{2}$$

Equation (2) states that $\boldsymbol{\psi}_2$ is non-identifiable when the likelihood does not depend on $\boldsymbol{\psi}_2$.

The simplest scenario for non-identifiability is when $\boldsymbol{\theta} = \boldsymbol{\psi}$ and equation (2) holds for at least one component of $\boldsymbol{\theta}$. This occurs when the model contains redundant parameters, or, more commonly, if a parameter $\boldsymbol{\theta}_2$ controls a specific model regime (e.g., extremely high flows) but the data does not force the model into this regime.

More generally, parameters $\boldsymbol{\theta}$ can be non-identifiable even if the likelihood function varies with respect to all inferred quantities in the original $\boldsymbol{\theta}$-parameterization. This occurs when parameters appear in groups that cannot be resolved into individual components (see example in section 3.2).

Non-identifiability has a strong connection to the properties of the parameter covariance matrix. For linear models, the covariance matrix of non-identifiable parameters is singular (i.e., has zero

eigenvalues), which can be detected using standard linear-algebraic methods. For nonlinear models, near-zero eigenvalues remain indicative (though not conclusively) of non-identifiability, but much more complex degeneracies can develop. *Kavetski et al.* [2006b] and *Tonkin et al.* [2007] further discuss the significance of the covariance/Hessian matrix and its eigenvalues for the estimation of nonlinear models.

In practice, the onset of non-identifiability is gradual. For example, likelihoods where

$$p(\boldsymbol{y} \mid \boldsymbol{\psi}_1, \boldsymbol{\psi}_2) \approx p(\boldsymbol{y} \mid \boldsymbol{\psi}_1) \tag{3}$$

do not strictly satisfy (2), but provide virtually no information about $\boldsymbol{\psi}_2$.

3.2. A simple illustration of non-identifiability

Consider the simple yet instructive example of non-identifiability [*Eberly and Carlin*, 2000]:

$$y_i \sim N\left(\theta_1 + \theta_2, 1^2\right), i = 1, ..., n \tag{4}$$

For illustrative purposes, θ_1 and θ_2 could be viewed as analogous to the parameters describing input and structural errors that we are trying to disaggregate in this study.

Assuming the y_i's are independent, the likelihood of observing the data $\boldsymbol{y}$ is:

$$p(\boldsymbol{y} \mid \theta_1, \theta_2) = \prod_{i=1}^{n} N\left(y_i \mid \theta_1 + \theta_2, 1^2\right) \tag{5}$$

Although this likelihood depends on both θ_1 and θ_2, there is no information in the data to discriminate between (θ_1, θ_2) pairs that add up to the same value.

More formally, the one-to-one reparameterization from $\left(\theta_1, \theta_2\right)$ to $\left(\psi_1, \psi_2\right)^{(1)} = \left(\eta, \theta_2\right)$, where $\eta = \theta_1 + \theta_2$, yields

$$p(\boldsymbol{y} \mid \eta, \theta_2) = \prod_{i=1}^{n} N\left(y_i \mid \eta, 1^2\right) \tag{6}$$

Since the re-parameterized likelihood (6) is independent from θ_2, it satisfies the definition (2) and therefore θ_2 is not identifiable. Similarly, reparameterization from (θ_1, θ_2) to $(\psi_1, \psi_2)^{(2)} = (\eta, \theta_1)$ shows that θ_1 is not identifiable either. On the other hand, the group η is identifiable – even though its individual components θ_1 and θ_2 are not!

3.3. Well-posedness

It is stressed that, given definitions (1) and (2), non-identifiability is a property solely of the likelihood function, and is completely independent of the prior distribution.

While the concept of identifiability is sufficient in maximum-likelihood estimation, Bayesian inference requires an analogous measure of informativeness of the posterior distribution. For this purpose, we adapt the distinction between "well-posed" and "ill-posed" problems, which is central in mathematics and physics [*Hadamard*, 1902].

We term a Bayesian inference "well-posed" if the associated posterior has the following properties: (a) it integrates to unity; (b) it is "informative"; and (c) it depends reasonably continuously on the inference data. These characteristics mimic Hadamard's criteria, originally developed in the context of mathematical models of physical phenomena [see also *Tarantola*, 2005, for a discussion in the context of inverse problems].

Criterion (b) can be formulated in direct analogy to condition (2): a posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ is non-informative with respect to at least one element of $\boldsymbol{\theta}$ if it can be re-parameterized such that

$$p(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2 \mid \boldsymbol{y}) = p(\boldsymbol{\psi}_1 \mid \boldsymbol{y}) \tag{7}$$

Equation (7) effectively defines an ill-posed posterior as the product of a non-identifiable likelihood with a non-informative prior.

An ill-posed posterior does not yield a useful inference of $\boldsymbol{\psi}_2$. In many cases, especially in the absence of prior bounds, a posterior that satisfies (7) does not integrate to a constant.

Finally, in practice it is common to see posteriors where

$$p(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2 \mid \boldsymbol{y}) \approx p(\boldsymbol{\psi}_1 \mid \boldsymbol{y}) \tag{8}$$

These are effectively ill-posed and yield very little useful inference. The sensitivity of the posterior to $\boldsymbol{\psi}_2$ before the inference is judged ill-posed is problem- and context- dependent.

3.4. Use of prior information

Since Bayesian analysis incorporates additional (prior) information into the analysis, it can obtain well-posed inferences from the posterior even if the likelihood function alone does not. *Indeed, the ability to bring in such information is a key strength of the Bayesian paradigm*. Yet this does not imply that a Bayesian modeler can disregard whether it is the prior or the likelihood that controls the well-posedness of a specific inference application.

In hydrology, independent (prior) information about data uncertainty can be obtained, e.g., from geostatistical analysis of spatial rainfall data [*Kuczera and Williams*, 1992] and rating curve analysis [*Thyer et al.*, 2009]. On the other hand, since meaningful characterization of structural errors remains a major challenge, it is unclear how to develop informative priors for structural errors (see section 2.1).

This section illustrates how prior knowledge can be used to produce a well-posed posterior inference. We simulate n=100 data from model (4), with true parameter values $\breve{\theta}_1 = -1$ and $\breve{\theta}_2 = 1$.

θ_1 and θ_2 are then inferred using standard Bayesian analysis. Two distinct prior knowledge scenarios are investigated:

1) The prior π_1 represents some prior knowledge of θ_1 and no prior knowledge of θ_2:

$$\pi_1(\theta_1, \theta_2) = p(\theta_1)p(\theta_2), \text{ with} \quad (9)$$
$$p(\theta_1) = N(-1, 0.1^2) \text{ and } p(\theta_2) \propto 1$$

2) The prior π_2 corresponds to no prior knowledge of θ_1 and θ_2:

$$\pi_2(\theta_1, \theta_2) = p(\theta_1)p(\theta_2), \text{ with} \quad (10)$$
$$p(\theta_1) \propto 1 \text{ and } p(\theta_2) \propto 1$$

Inference using the (informative) prior π_1 (Figure 1a) yields a posterior that is approximately Gaussian. The non-identifiability of θ_1 and θ_2 does not induce statistical problems; we refer to this situation as a "well-posed inference".

In contrast, the inference using the (non-informative) prior π_2 is "ill-posed" (Figure 1b). In particular, the posterior is constant along infinite-size sub-spaces $\theta_1 + \theta_2 = \eta$. This posterior does not yield any useful information on (θ_1, θ_2). However, the inference on $\eta = \theta_1 + \theta_2$ is well-posed (Figure 1c).

It is critical to note that, as discussed in section 3.2, (θ_1, θ_2) are non-identifiable from the data regardless of the prior distribution (identifiability as defined in equation (2) is strictly a property of the likelihood function). However, η is identifiable (and its inference well-posed) for both priors.

3.5. Practical diagnosis of well-posedness and identifiability

The instructive example (4) shows that parameter identifiability cannot be assessed by simply checking that the likelihood is sensitive to a change in individual parameter values. Furthermore,

the parameter grouping fulfilling condition (2) was obvious in the preceding example, but might be very difficult to uncover for more complicated hydrological models. Consequently, in practice non-identifiability and ill-posedness are more likely to be detected through their empirical symptoms, rather than through formal mathematical analysis.

In general, the posterior distributions of nonlinear hydrological models are too complicated to be described analytically and therefore are usually explored using Markov Chain Monte Carlo (MCMC) methods [e.g., *Kuczera and Parent*, 1998]. Since well-posedness is a key characteristic of the posterior, it controls the convergence of MCMC methods. Consequently, the behavior of the latter, in conjunction with an evaluation of prior knowledge, can be used to indirectly detect non-identifiability.

Consider MCMC sampling from the posteriors in Figure 1. Figure 2 shows the evolution of two parallel Metropolis chains for parameters θ_1, θ_2 and $\eta = \theta_1 + \theta_2$. The top three panels refer to the posterior obtained with prior π_1: the two chains mix and converge quickly for all inferred quantities. However, the behavior in the case of the prior π_2 (Figure 2, bottom panels) is totally different: the chains for θ_1 and θ_2 diverge (note the wide scale of the y-axis). Moreover, the posterior correlation between θ_1 and θ_2 is almost –1, suggesting complete interaction between these parameters. Yet convergence is almost immediate for parameter η: despite its individual components θ_1 and θ_2 being non-inferable, the inference of η is perfectly well-posed.

The poor convergence and near-perfect cross-correlation of MCMC samples from the ill-posed posterior is emphasized, since a qualitatively similar behavior will be observed in the case studies using conceptual hydrological models (sections 8 and 9).

3.6. Non-identifiability, ill-posedness and predictive ability

While non-identifiability is generally undesirable, its practical consequences depend on the objective of the analysis. If parameter estimation is the chief objective, non-identifiability is a serious impediment, especially with weak prior knowledge. Yet in some cases, non-identifiability does not prevent reliable predictions. For example, prediction of y using the model (4) is straightforward because the (sufficient) parameter $\eta = \theta_1 + \theta_2$ is perfectly identifiable. However, if θ_1 and/or θ_2 are used to predict quantities other than y, using the ill-posed inference can result in very poor predictions. In hydrology, this corresponds to using the model to predict environmental variables that the model has not been calibrated to. Similar problems develop when attempting to extrapolate ill-inferred models beyond the range of calibration data.

4. Experimental setup

4.1. Validity of synthetic experiments

Recent literature debates the value of synthetic experiments [e.g. *Beven*, 2006; *Montanari*, 2007]. Our view is that synthetic tests are a necessary step to ensure the internal consistency of a statistical method and identify its strengths and weaknesses. However, synthetic tests using exact models say little about the robustness of the method in the common case when the CRR model is inaccurate.

The strategy used in this study to partially overcome the latter limitation is to generate the "true" data using model *M0* and calibrate another model, *M1*, to this data, possibly corrupting the latter with synthetic "observation" errors. The advantages of this approach are: (i) all quantities are known, so that exact and estimated values can be compared, and (ii) by using different models *M0*

and *M1*, the calibration scheme can be tested in cases where the notion of "true parameter values" is not applicable (since in general there is no *M1*-parameter set leading to the *M0*-generated data, even if the true input/output is used).

Since it remains to be seen whether the discrepancies between two hydrological models are representative of the discrepancies between a hydrological model and actual physical processes, a real-data study is used to check whether qualitatively similar results are obtained as in the synthetic analysis. Agreement in this respect suggests, though does not conclusively prove, that the same conclusions hold.

4.2. Calibration data and models

This paper uses two synthetic and one real dataset. The synthetic set *D0* is generated using the logSPM model (with parameters summarized in Table 1 and model equations detailed in Appendix A) and is corrupted with input/output errors. This dataset is used for basic analysis in the absence of structural error (Experiment A). The synthetic set *D1* is generated using the GR4J model [*Perrin et al.*, 2003] and is also corrupted with data errors. Calibrating logSPM to *D1* (Experiment B) tests the ability of the calibration methodology to handle structural errors (see Section 4.1).

Five years of daily rainfall and potential evapotranspiration (PET) from the Abercrombie catchment (2770 km^2, New South Wales, Australia) are treated as the true inputs (***r*** and ***pet***) and used to generate synthetic runoffs.

The observed rainfall ($\tilde{\boldsymbol{r}}$) is generated by corrupting the true rainfall as follows:

$$\begin{aligned} &\tilde{r}_t = r_t / \exp(m_t) \quad \text{(a)} \\ &m_t \sim N(-0.2, 0.2^2) \quad \text{(b)} \end{aligned} \tag{11}$$

The lognormal distribution used to generate rainfall errors in equation (11)-b leads to a systematic over-prediction of about 20% and a standard error of about 20%.

Since the sensitivity of CRR models to PET errors is low [e.g., *Oudin et al.*, 2006], we assume the PET data is error-free, i.e., $\tilde{\boldsymbol{pet}} = \boldsymbol{pet}$.

The "true" outputs $\boldsymbol{q}$ are generated using logSPM (dataset *D0*) and GR4J (dataset *D1*) and are corrupted to produce observed outputs $\tilde{\boldsymbol{q}}$:

$$\begin{aligned} \tilde{q}_t &= q_t + e_t && \text{(a)} \\ e_t &\sim N\left(0, (0.1 q_t)^2\right) && \text{(b)} \end{aligned} \qquad (12)$$

The real-data study (Experiment C) uses the observed rainfall, PET and runoff for the calibration and validation periods.

In all three experiments, the calibration period includes days 529-1083 (1.5 years) and is preceded by a warm-up period of 100 days. Days 1084-1827 (2 years) are used for validation.

5. Bayesian inference framework

The calibration schemes investigated in this study differ in their treatment of each source of uncertainty. They can be obtained from the general Bayesian Total Error Analysis (BATEA) framework by supplying specific error models and priors. Following an outline of the overall framework in sections 5.1-5.8, the calibration schemes are summarized in section 5.9.

5.1. Basic notation

Let $\boldsymbol{R} = (r_t)_{t=1,\ldots,T}$ denote the true areal rainfall at day t and $\tilde{\boldsymbol{R}} = (\tilde{r}_t)_{t=1,\ldots,T}$ be the corresponding observed rainfall. Similarly, let $\boldsymbol{Q} = (q_t)_{t=1,\ldots,T}$ and $\tilde{\boldsymbol{Q}} = (\tilde{q}_t)_{t=1,\ldots,T}$ denote the true and observed runoffs.

In general, a CRR model $M()$ predicts the runoff $\hat{\boldsymbol{Q}} = (\hat{q}_t)_{t=1,\ldots,T}$ given rainfall, PET, parameters and initial conditions:

$$\hat{q}_t = M(\boldsymbol{R}_{1:t}, \boldsymbol{PET}_{1:t}, \boldsymbol{\theta}, \boldsymbol{S}_0) \tag{13}$$

where $\boldsymbol{R}_{1:t}$ and $\boldsymbol{PET}_{1:t}$ are the inputs for time indices 1 to t, $\boldsymbol{\theta}$ are the deterministic CRR parameters and $\boldsymbol{S}_0$ is the vector of initial store values.

The initial conditions $\boldsymbol{S}_0$ are not inferred because their influence is minimized using a warm-up.

5.2. Input errors

Traditional calibration methods, e.g., standard least squares (SLS), assume all observed inputs are error-free, in particular, $\boldsymbol{R} = \tilde{\boldsymbol{R}}$. With this assumption, the only quantities requiring inference in equation (13) are the CRR parameters. However, ignoring input uncertainty can significantly degrade the inference [*Kavetski et al.*, 2002]. One possibility, used in BATEA, is to treat input uncertainty using a hierarchical formalism, where each rainfall error is represented using a latent variable. The full posterior then yields a joint inference of the true inputs and the CRR parameters given the model and the observed input/output data [*Kavetski et al.*, 2006a].

In this study, rainfall errors at each wet day are represented using rainfall multipliers sampled from an uncorrelated lognormal distribution. More formally, we assume Gaussian log-multipliers $\boldsymbol{\Phi} = (\phi_\tau)_{\tau=1,\ldots,N_{wet}}$ as follows:

$$\begin{aligned} r_t &= \tilde{r}_t \exp(\phi_{\tau(t)}) && \text{(a)} \\ \phi_{\tau(t)} &\sim N(\mu_r, \sigma_r^2) && \text{(b)} \\ \mu_r &\sim N(-0.2, 1/\nu^2) && \text{(c)} \\ \sigma_r^2 &\sim Inv\chi^2(\nu, 0.2) && \text{(d)} \end{aligned} \qquad (14)$$

where $\tau(t)$ is the index of the log-multiplier affecting time step t, $N(a,b^2)$ is the Gaussian distribution with mean a and variance b^2 and $inv\chi^2(a,b)$ is the inverse-χ^2 distribution with degrees of freedom a and scale b.

Equation (14)-b is the hyper-distribution of latent variables (Gaussian distribution), with hyper-parameters μ_r ("hyper-mean") and σ_r ("hyper-standard-deviation", "hyper-SD" hereafter).

Equation (14)-c represents the prior distribution of the hyper-mean. The prior mean is set to -0.2, which, given equation (14)-a, centers the prior on the actual mean of the rainfall errors. The precision parameter v controls the sharpness of the prior distribution. Three values of v are investigated:

(1) $v=10^3$: high prior precision: hyper-mean can be considered as virtually known;

(2) $v=10^2$: medium prior precision: appreciable prior information;

(3) $v=10$: low prior precision: little prior knowledge.

Similarly, equation (14)-d represents the prior on the hyper-SD. The scale parameter is set to 0.2, so that the prior encompasses the true value of σ_r^2 and becomes progressively more concentrated

around it as the prior precision v increases. The three values of v described above are also used when specifying the precision of this prior.

5.3. Structural errors via stochastic CRR parameters

Structural uncertainty can be represented hierarchically using stochastic variations of some CRR parameters (section 2.2). Following *Kuczera et al*. [2006], the parameter k_S of logSPM is allowed to vary across storm epochs delimited by rainfall events exceeding 2 mm/day. Since $k_S > 0$, we assumed a lognormal hyper-distribution at each epoch ω:

$$\begin{aligned} \log(k_{S\omega}) &= \lambda_\omega && \text{(a)} \\ \lambda_\omega &\sim N(\mu_{k_S}, \sigma^2_{k_S}) && \text{(b)} \\ \mu_{k_S} &\sim N(-2, 4^2) && \text{(c)} \\ \sigma^2_{k_S} &\sim Inv\chi^2(1, 0.5) && \text{(d)} \end{aligned} \qquad (15)$$

Similarly to rainfall log-multipliers $\boldsymbol{\Phi}$, the values $(\lambda_\omega)_{\omega=1,\ldots,N_{epochs}} = \boldsymbol{\Lambda}$ are unknown and are therefore treated as latent variables. Since specifying meaningful informative priors for the hyper-parameters of structural errors is problematic, the priors in equations (15)-c and (15)-d correspond to vague knowledge of the stochastic parameter.

5.4. Output errors

The uncertainty in the observed runoff is due mainly to rating curve errors. Previous studies suggested that these errors are heteroscedastic [*Huard and Mailhot*, 2008; *Thyer et al.*, 2009], e.g.,

$$\begin{aligned} \tilde{q}_t &= q_t + \gamma_t && \text{(a)} \\ \gamma_t &\sim N\left(0, (\zeta \tilde{q}_t)^2\right) && \text{(b)} \end{aligned} \qquad (16)$$

Here we assume a relative standard error ζ =0.1, though in general it should be determined from rating curve analysis [*Thyer et al.*, 2009] or added to the inference itself. However, since this study focuses on input and structural uncertainties, the output error model (16) is fully specified prior to calibration. Note a minor inconsistency between equation (16) above and equation (12): the synthetic data was corrupted using observation errors proportional to the true flows, whereas in BATEA we assumed observation errors proportional to the observed flows. Empirical checks suggested the effect of this inconsistency is minor. Importantly, Experiment A (see section 7) suggests that it does not introduce any bias into the analysis.

Note that while operational interest is usually in the *actual* runoff, both calibration and validation are necessarily limited to comparison to *observed* values. This requires a meaningful consideration of the uncertainty in observed streamflows, e.g., as described in equation (16). In addition, the predictive uncertainty communicated to decision-makers must clearly state whether it includes output observation uncertainty.

5.5. Remnant errors

The output error model (16) links the observed runoff with the true runoff. Since the latter is unknown, an additional model linking the true runoff with the simulated runoff must be specified. Here, we use an additive Gaussian error model with unknown variance σ^2,

$$\begin{aligned} q_t &= \hat{q}_t + \varepsilon_t && \text{(a)} \\ \varepsilon_t &\sim N(0,\sigma^2) && \text{(b)} \\ \sigma^2 &\sim Inv\chi^2(1,0.2) && \text{(c)} \end{aligned} \qquad (17)$$

In this paper, errors ε_t are termed "remnant" because their interpretation depends on the error sources remaining due to omission of sources of uncertainty in the calibration scheme or due to

imperfect representation of these sources (see Section 5.9 for further discussion). This makes them subtly different from the notions of "model inadequacy" and "discrepancy" introduced elsewhere when discussing model structural errors [*Goldstein and Rougier*, 2009; *Kennedy and O'Hagan*, 2001]. Note that the remnant error variance σ^2 is expected to decrease as improved input/output/structural error models are specified (see Section 10.2.3).

If runoff measurement errors γ_t and remnant errors ε_t are independent, the distribution of observed runoff conditioned on simulated runoff is

$$\begin{aligned} &\tilde{q}_t = q_t + \gamma_t = \hat{q}_t + \varepsilon_t + \gamma_t = \hat{q}_t + \eta_t && \text{(a)} \\ &\eta_t \sim N\left(0, (\zeta \tilde{q}_t)^2 + \sigma^2\right) && \text{(b)} \\ &\tilde{q}_t \sim N\left(\hat{q}_t, (\zeta \tilde{q}_t)^2 + \sigma^2\right) && \text{(c)} \end{aligned} \qquad (18)$$

This equation is used to evaluate the likelihood of observed runoff.

5.6. Improving error models: an open frontier

The BATEA framework described in sections 5.1-5.5 integrates probabilistic error models describing individual sources of uncertainty. Its reliability evidently depends on the adequacy of these error models. While this study focuses on fundamental aspects of identifiability and therefore uses synthetic data, significant further work is needed to derive and evaluate realistic models of uncertainties in hydrological data. In particular, the following limitations need to be addressed:

1. The multiplicative treatment of input errors in equation (14) does not handle the situation where a rainfall event or time step is missed by the raingauge network.

2. The characterization of structural errors using stochastic variations of CRR parameter (equation (15)) is a hypothesis that needs empirical scrutiny. This assessment requires the

disaggregation of input and structural errors; the feasibility of this disaggregation is precisely the aim of this paper.

3. Improved treatment of rating curve errors (equation (16)) is needed. Recent literature [e.g., *Di Baldassarre and Montanari*, 2009; *Dottori et al.*, 2009; *Moyeed and Clarke*, 2005; *Neppel et al.*, 2009; *Reitan and Petersen-Overleir*, 2009] suggests promising avenues, including treatment of stochastic uncertainty (e.g., in the height-discharge measurements used to establish the rating curve) and systematic errors (e.g. in the extrapolation necessary when measuring high and low flows).

4. The treatment of remnant errors (equation (17)) is arguably the most challenging tusk, because their interpretation depends on the treatment of other error sources (input, output, structural). Moreover, their dependence on the catchment dynamics and on the temporal and spatial resolution of the analysis is poorly understood. The remnant error model (17) used in this paper is quite simple, in particular, it does not account for autocorrelation. An interesting approach that represents remnant errors as (discrete) realizations from a continuous-time stochastic process [e.g., *Reichert and Mieleitner*, 2009; *Yang et al.*, 2007] will be evaluated in future work.

As shown in this paper, the adequacy of the entire likelihood function, as well as its individual components representing remnant errors, input errors, etc., can and should be directly scrutinized using stringent diagnostics such as QQ plots, autocorrelation measures, etc. While disappointingly rare in most hydrological applications to-date, such posterior scrutiny is an essential part of Bayesian analysis and aids model improvement [see *Thyer et al.*, 2009, for a recent illustration].

5.7. Posterior distribution

The posterior distribution of all inferred quantities is given by Bayes' theorem as follows [see *Kavetski et al.*, 2006a; *Kuczera et al.*, 2006; *Thyer et al.*, 2009 for details]:

$$p(\boldsymbol{\theta}, \boldsymbol{\Phi}, \mu_r, \sigma_r, \boldsymbol{\Lambda}, \mu_{kS}, \sigma_{kS}, \sigma \mid \tilde{\boldsymbol{Q}}, \tilde{\boldsymbol{R}}) \propto \\ p(\tilde{\boldsymbol{Q}} \mid \boldsymbol{\theta}, \boldsymbol{\Phi}, \boldsymbol{\Lambda}, \sigma, \tilde{\boldsymbol{R}}) p(\boldsymbol{\Phi} \mid \mu_r, \sigma_r) p(\boldsymbol{\Lambda} \mid \mu_{kS}, \sigma_{kS}) p(\boldsymbol{\theta}, \mu_r, \sigma_r, \mu_{kS}, \sigma_{kS}, \sigma) \tag{19}$$

The full posterior (19) comprises the following three parts:

(i) The likelihood of observed runoffs, derived from (18) as

$$p(\tilde{\boldsymbol{Q}} \mid \boldsymbol{\theta}, \boldsymbol{\Phi}, \boldsymbol{\Lambda}, \sigma, \tilde{\boldsymbol{R}}) = \prod_{t=1}^{T} N\left(\tilde{q}_t \mid \hat{q}_t, (0.1\tilde{q}_t)^2 + \sigma^2\right) \\ = \prod_{t=1}^{T} N\left(\tilde{q}_t \mid M\left(\left\{\tilde{\boldsymbol{R}}_{1:t}, \boldsymbol{\Phi}_{1:\tau(t)}\right\}, \left\{\boldsymbol{\theta}, \boldsymbol{\Lambda}_{1:\omega(t)}\right\}\right), (0.1\tilde{q}_t)^2 + \sigma^2\right) \tag{20}$$

(ii) The prior distribution of deterministic parameters and hyper-parameters $p(\boldsymbol{\theta}, \mu_r, \sigma_r, \mu_{kS}, \sigma_{kS}, \sigma)$. In this study, independent priors are used.

(iii) The terms $p(\boldsymbol{\Phi} \mid \mu_r, \sigma_r)$ and $p(\boldsymbol{\Lambda} \mid \mu_{kS}, \sigma_{kS})$ represent the hierarchical parts of the model and are derived from (14) and (15),

$$p(\boldsymbol{\Phi} \mid \mu_r, \sigma_r) = \prod_{\tau=1}^{N_{wet}} N(\phi_\tau \mid \mu_r, \sigma_r) \tag{21}$$

$$p(\boldsymbol{\Lambda} \mid \mu_{kS}, \sigma_{kS}) = \prod_{\omega=1}^{N_{epochs}} N(\lambda_\omega \mid \mu_{kS}, \sigma_{kS}) \tag{22}$$

5.8. Distinction between posterior distributions of latent variables and their hyper-distribution

A subtle but important aspect of hierarchical models such as (19) is the distinction between the posterior distributions of individual latent variables and their prior/posterior hyper-distributions. This distinction is highly germane to the analyses carried out in this paper.

In the case of rainfall errors, the prior hyper-distribution describes the prior knowledge of rainfall uncertainty. The calibration data supports the inference of individual rainfall multipliers, yielding the posterior distributions of individual latent variables (i.e. of individual rainfall errors). The Bayesian formulation jointly uses these distributions to refine the prior hyper-distribution, yielding the posterior hyper-distribution. The posterior hyper-distribution of rainfall multipliers represents a refined description of rainfall uncertainty given the observed data and the CRR model.

The same mechanism applies to the latent variables describing structural errors.

5.9. Summary of Calibration Schemes

Table 2 summarizes the nine calibration schemes used in this paper. They correspond to special cases of the Bayesian framework described in sections 5.2-5.7 and differ in their representation of each source of uncertainty.

SLS refers to standard least squares regression (equivalent to maximizing the Nash-Sutcliffe statistic). In the application of SLS in this paper, the residual standard deviation σ in equation (17)-b is inferred rather than specified a priori. It lumps the effects of input, output and structural errors affecting the CRR model in the remnant ("residual") error model. This can be obtained by setting $\mu_r = 0$ and $\sigma_r = 0$ (so that $r_t = \tilde{r}_t$) in equation (14) and $\zeta = 0$ in (18)-c (so that $q_t = \tilde{q}_t$).

Scheme O is similar to SLS, except that output uncertainty is represented directly ($\zeta = 0.1$ in (18)-c). This can be viewed as a special case of the weighted least squares (WLS) method, where σ in equation (18) is inferred. In the formulation (18), the remnant error term ε lumps the effects of input and structural errors, as well as imperfections of the output error model (16).

The SLS and O schemes treat the CRR model as deterministic and use an additive error term to represent all other sources of error (see section 2.2). They are used in this paper as baseline methods representing common practice.

Scheme OP, in addition to representing output uncertainty, describes structural errors using a single stochastic CRR parameter. Consequently, the remnant error term lumps the effect of input errors and imperfections of the output and structural error models.

Schemes OI represent the case where input and output errors are included (sections 5.2 and 5.4 respectively). The suffixes 1, 2 and 3 represent the specified precision of prior information on input errors, with 1 denoting the highest precision and 3 the lowest. In the OI scheme, the remnant error term lumps structural errors and imperfections of the input/output error models.

Schemes OIP represent the case where input/output errors are included and structural errors are represented using a single stochastic CRR parameter (with suffixes 1, 2 and 3 denoting the specified prior precision of input errors). In this case, remnant errors solely represent imperfections of the input/output/structural error models.

5.10. Dimensionality of the inference and MCMC strategy

Introducing and inferring latent variables representing input and/or structural errors in the CRR model comes at the cost of increased dimensionality of the inference. This can be seen in Table 2 (column 8), where schemes accounting for input errors and/or allowing parameter stochasticity

require the inference of a large number of latent variables. For example, the calibration data in experiment B yields 251 rainfall log-multipliers (one for each wet day) and 157 latent variables for the stochastic parameter s_K (one for each epoch).

Sampling from high-dimensional posteriors is computationally challenging but not insurmountable. In particular, the evaluation of the BATEA posterior distribution for a given set of CRR parameters is only marginally more expensive than that of SLS or WLS (the extra cost of evaluating (21)-(22) is trivial). The increased cost of the BATEA inference comes almost exclusively from a larger number of samples needed to adequately characterize high-dimensional distributions. In particular, the adaptation of high-dimensional MCMC jump distributions can be very challenging, with few theoretical guidelines [e.g., *Haario et al.*, 2005].

In this study, the BATEA posterior (19) is explored using a two-stage MCMC strategy [*Kuczera et al.*, 2007; *Thyer et al.*, 2009]. The sampler evolves four parallel chains until the Gelman-Rubin criteria [*Gelman et al.*, 1995] are below 1.2 for all inferred quantities. The number of MCMC iterations and the total CPU times needed to satisfy the Gelman-Rubin criterion are reported in Table 2. The longest run did not exceed 5 hrs on a standard desktop computer (2.2 GHz CPU, 4 GB RAM, Windows XP).

The increase in dimensionality and its implications for inference are further discussed in Section 10.1.3.

6. Experimental methodology

6.1. Evaluation strategy

Several analyses are necessary to achieve the objectives of this study:

1. Examine the well-posedness of the inference. This is done by inspecting convergence diagnostics and the correlation structure of the MCMC samples (Section 3.5).

2. Evaluate the predictive distribution (PD) of the observed runoff during the validation period (see *Thyer et al.* [2009] for details). This establishes the adequacy of the total predictive uncertainty.

3. [Synthetic studies only] Evaluate the PD of the true rainfall. This establishes whether the sources of uncertainty have been accurately and precisely identified. This check can only be carried out for the synthetic datasets *D0* and *D1*, where the true rainfall is known.

6.2. Evaluating time-varying predictive distributions

In time series analysis, evaluating a predictive distribution (PD) requires comparing a time-varying random variable X_t (with cdf F_t) to a time series of realizations x_t. For the rainfall PD, x_t represents the true rainfall, while for the runoff PD, x_t represents the observed runoff. However, model performance measures currently predominant in hydrology, such as the Nash-Sutcliffe statistic, are unsuitable for analyzing PD's, because they merely compare two time series of values and disregard their associated uncertainties. Instead, following the terminology used in meteorological ensemble predictions [*Atger*, 1999], this paper considers two criteria: "reliability" to quantify the statistical consistency between the time series of x_t and its PD, and "resolution" to quantify the sharpness of the PD.

6.3. Reliability

If the PD is reliably quantified, the observations correspond to realizations from the PD. This can be examined using the predictive QQ-plot [*Laio and Tamea*, 2007; *Thyer et al.*, 2009]. If the realizations x_t are consistent with F_t, the p-values $F_t(x_t) = p(X_t \le x_t)$ will follow a uniform

distribution on the interval [0,1]. This can be checked graphically: deviations from the bisector (the 1:1 line) denote interpretable deficiencies (see Figure 3). To simplify the comparison of QQ-plots, they are summarized using two indexes that quantify the reliability of the PD:

$$\alpha_x = 1 - 2\alpha'_x \qquad \text{(a)} \tag{23}$$
$$\alpha'_x = \sum_{i=1}^{N_x} \left| p_{x(i)} - p_{x(i)}^{(th)} \right| / N_x \qquad \text{(b)}$$

$$\xi_x = 1 - \xi'_x \qquad \text{(a)} \tag{24}$$
$$\xi'_x = \sum_{i=1}^{N_x} (1_{\{0,1\}}(p_{x(i)})) / N_x \qquad \text{(b)}$$
$$1_{\{0,1\}}(z) = \begin{cases} 1 \text{ if } z = 0 \text{ or } z = 1 \\ 0 \text{ otherwise} \end{cases} \qquad \text{(c)}$$

where $p_{x(i)}$ and $p_{x(i)}^{(th)}$ are the *i*th observed and theoretical *p*-values of x_t, N_x is the number of x_t values and $1_A(x)$ is the indicator function of the set *A*.

The index α is related to the area α' between the *p*-value curve and the 1:1 line, and reflects the overall reliability of the PD. It varies between 0 (worst reliability, with all observed *p*-values equal to 0 or 1) and 1 (perfect reliability).

The index ξ is the complement of the fraction ξ' of observed *p*-values equal to 0 or 1, which correspond to x_t values outside the range of the PD. It varies between 0 (worst reliability, with all realizations outside their predictive range) and 1 (no incompatible realizations). Note that $\xi = 1$ does not imply perfect reliability. Consequently, this index is used primarily for detecting highly unreliable PDs.

For the rainfall PD these indices are denoted as α_r and ξ_R, while for the runoff PD, they are denoted as α_Q and ξ_Q.

6.4. Resolution

"Resolution" denotes the sharpness (effectively, the "average precision") of the PD. Note that two inferences can both yield reliable PDs, but with different resolutions.

In this paper, the resolution is quantified by indexes $\pi^{(abs)}$ and $\pi^{(rel)}$, defined, respectively, as the average absolute and relative precision of the predictions X_t:

$$\pi_x^{(abs)} = \frac{1}{N_x}\sum_{t=1}^{N_x}\frac{1}{\text{Sdev}[X_t]} \tag{25}$$

$$\pi_x^{(rel)} = \frac{1}{N_x}\sum_{t=1}^{N_x}\frac{\text{E}[X_t]}{\text{Sdev}[X_t]} \tag{26}$$

where E[] and Sdev[] are the expectation and standard deviation operators. In this paper, we use the index $\pi_R^{(abs)} = \pi_{x=\log(\phi)}^{(abs)}$ to assess the resolution of the rainfall PD, and the index $\pi_Q^{(rel)} = \pi_{x=\tilde{q}}^{(rel)}$ for the resolution of the observed runoff PD. The analysis of log-multipliers is based on the absolute measure because the multiplicative error model (14)-a already represents relative errors.

The data used in (23)-(27) can be pre-filtered. In order to focus on hydrologically significant events, the computation of indexes in this paper is restricted to observed rainfalls exceeding 10 mm/day and observed runoffs exceeding 1 mm/day.

7. Experiment A: Estimating input errors when the CRR model is exact

Experiment A examines the OI-3 calibration scheme (with weak prior knowledge of rainfall-error hyper-parameters) when the calibration data contains input/output errors but the model does not contain structural errors. This establishes the "best-case" scenario for parameter estimation, indicating what can be achieved when the model is accurate (indeed, exact), and provides a necessary benchmark for the comparison of more complicated calibration scenarios where structural errors are present.

7.1. Assessing well-posedness

MCMC convergence was readily achieved, suggesting that the inference is well-posed. This is consistent with previous synthetic studies focusing on input errors [e.g., *Kavetski et al.*, 2002; *Renard et al.*, 2009a].

7.2. Evaluating the predictive distribution of runoff

7.2.1. Reliability

The runoff PD shows a good agreement with the observed runoff (Figure 4a). The predictive QQ plot shown in Figure 4b confirms this observation, with the curve closely following the bisector. The reliability indexes $\alpha_Q = 0.92$ and $\xi_Q = 1$ further demonstrate that the PD of observed runoff is reliable.

7.2.2. Resolution

Figure 4a shows that the width of the prediction limits varies with the magnitude of the predicted runoff, which justifies the use of the relative precision measure $\pi_Q^{(rel)}$ for assessing the runoff PD. The resolution index $\pi_Q^{(rel)} = 4.87$ corresponds to an average coefficient of variation of about 20%.

7.3. Evaluating the predictive distribution of rainfall

7.3.1. Reliability

Figure 4c-d suggest that the true rainfall values are reliably estimated, with reliability indexes $\alpha_R = 0.92$ and $\xi_R = 1$. This is consistent with the results for runoff.

7.3.2. Resolution

Despite rainfall multipliers being reliably estimated, the precision of the individual estimates is not identical. Figure 5 shows that multipliers affecting large rainfalls can be identified much more precisely than multipliers affecting smaller rainfalls. The resolution index $\pi_R^{(abs)} = 7.52$, computed for rainfall values larger than 10 mm, corresponds to an average coefficient of variation of about 13%, which is relatively low.

Furthermore, Figure 6 shows the posteriors of some rainfall multipliers remain similar to the hyper-distribution. A given rainfall multiplier ϕ_τ affects the posterior pdf (19) both through the likelihood function and through the pdf of the hyper-distribution evaluated at ϕ_τ. Consequently, if the likelihood is only weakly dependent on ϕ_τ, as in condition (3), the posterior pdf will remain close to the hyper-distribution. Such multipliers are "weakly-identifiable".

It is stressed that weak identifiability of some individual rainfall multipliers does not imply that the entire hyper-distribution is non-identifiable. The estimated hyper-mean and hyper-SD of the rainfall multipliers was -0.215 (standard error ± 0.094) and 0.223 (standard error ± 0.018), which are close to the true values of -0.2 and 0.2 respectively. Hence, there is enough information in the identifiable multipliers to infer their hyper-distribution. The non-identifiability of some rainfall multipliers is effectively "benign" because it neither affects model predictions (since the hyper-distribution is properly identified), nor causes computational problems (MCMC sampling converges because the hyper-distribution constrains the rainfall multipliers).

8. Experiment B: Estimating input and structural errors using inaccurate CRR models

In this section, the nine inference schemes listed in Table 2 are used to calibrate the CRR model LogSPM using the synthetic dataset *D1* generated using GR4J. This experiment considers input, output and structural errors.

8.1. Achieving well-posedness using prior information

MCMC convergence was readily achieved for SLS, O, OI and OP, suggesting that these inferences are well-posed. However, convergence difficulties were encountered with OIP. This suggests the simultaneous inference of both input and structural errors may be ill-posed. This section examines the role of priors when attempting to decompose the total predictive uncertainty by estimating both input and structural errors.

8.1.1. Low precision priors (OIP-3)

As shown in Table 2, the OIP-3 scheme has a prohibitively slow rate of MCMC convergence – even after more than $3x10^6$ MCMC iterations, the Gelman-Rubin criterion still exceeded 5.0 for many quantities (including both latent variables and CRR parameters). This is symptomatic of an ill-posed inference. Since the inference was based on a vague prior, its ill-posedness can be attributed to non-identifiability, in particular of latent variables.

The MCMC samples from the OIP-3 posterior yield insights into the reasons for poor convergence. Figure 7c shows strong correlations between the latent variables characterizing input and structural errors affecting the same time steps. This yields a characteristic bloc-diagonal structure of the correlation matrix. This degeneracy is analogous to the simple example in section 3.5, where non-identifiable parameters θ_1 and θ_2 were almost perfectly correlated when a non-informative prior was used. The implications of this are discussed in section 10.1.2.

8.1.2. Medium and High precision priors (OIP-1 and OIP-2)

The MCMC sampling from the OIP-1 and OIP-2 posteriors was convergent, suggesting that the inference becomes well-posed when more precise priors on the rainfall multiplier hyper-parameters are used. However, the onset of ill-posedness is gradual: the posterior correlations for OIP-1 and OIP-2 (Figure 7a-b) display similar, though less pronounced, features as the OIP-3 case.

Note that since the non-identifiability criterion (2) depends solely on the likelihood but not on the prior, OIP-1 and OIP-2 methods are necessarily subject to the same non-identifiability issues as OIP-3. The MCMC convergence is due to a sufficiently precise prior restricting the size and improving the shape of the high-density regions of the posterior.

8.2. Evaluating the predictive distribution of runoff

The reliability and resolution runoff indexes obtained for the nine calibration schemes are reported in the second row of Figure 8.

8.2.1. Reliability

Figure 8 shows significant differences in the reliability of the runoff PDs between (i) standard calibration approaches SLS and O; vs. (ii) approaches OP, OI and OIP, which use Bayesian hierarchical inference for at least one source of uncertainty.

Approaches SLS and O lead to an unreliable quantification of predictive uncertainty, with low α_Q and ξ_Q values. In particular, about 40% and 25% of observed runoffs are outside the predictive range for SLS and O, respectively. This represents a significant underestimation of predictive uncertainty, especially for large runoff events.

Approaches OP, OI and OIP quantify predictive uncertainty much more reliably, with high α_Q values and no runoff values outside the predictive range in most cases. Scheme OI-1 is the only exception, with ξ_Q = 0.9 (i.e., 10% of observations outside the predictive range), corresponding to a mild underestimation of predictive uncertainty.

8.2.2. Resolution

Figure 8 shows that schemes SLS and O achieve a significantly higher resolution (with $\pi_Q^{(rel)} \approx 9$) than schemes OP, OI and OIP (with $\pi_Q^{(rel)} \approx 2-6$). However, section 8.2.1 demonstrated that the former schemes do not lead to a reliable estimation of the runoff PD. It follows that schemes SLS and O yield unduly optimistic estimates of predictive uncertainty: their higher resolution comes at the cost of an unacceptably low reliability, which can be misleading to a decision-maker.

On the other hand, schemes OP, OI and OIP yield similar results, with the exception of OI-1, which yields a higher resolution ($\pi_Q^{(rel)} \approx 6$). This causes the mild underestimation of predictive uncertainty noted in section 8.2.1.

8.3. Evaluating the predictive distribution of rainfall

The rainfall PD is evaluated only for OI and OIP. SLS, O and OP are not considered because they do not explicitly consider input errors, and hence do not produce a rainfall PD. The first row of Figure 8 summarizes the results using the indexes α_R, ξ_R and π_R.

8.3.1. Reliability

For OI and OIP with medium to high prior precision, the PD of true rainfall is inferred reliably (α_R and ξ_R are close to one in Figure 8). When only weak prior information is available (OI-3), the indexes α_R and ξ_R decrease to about 0.55 and 0.9 respectively, reflecting the deterioration of the inference as less prior knowledge is available. This deterioration is also reflected in the overestimation of the hyper-SD of the rainfall multipliers (Table 3, estimated value of 0.862 versus the true value of 0.2). Section 10.3.1 discusses the implications of this result.

8.3.2. Resolution

Two observations can be drawn from Figure 8:

(i) The resolution depends on the prior precision for both the OI and OIP methods. This implies that the prior exerts a significant influence on the inference;

(ii) For a given prior precision, OI yields a higher resolution than OIP.

Figure 9 offers insight about point (ii) above. In the OI case, the precision of the inferred rainfall multipliers increases with the observed rainfall. This is consistent with section 7.3.2. In the OIP case, this relationship is weaker, with the posterior precision of most multipliers remaining close to the precision of their posterior hyper-distribution. Indeed, the posterior distributions of the individual rainfall multipliers remain similar to the posterior hyper-distribution (similar to Figure 6). The implications of this are discussed in section 10.3.1.

9. Experiment C: Real-data study

In this experiment, LogSPM is calibrated to the observed runoff from the Abercrombie catchment. The aim is to investigate whether the conclusions drawn from synthetic experiment B hold in real-data applications. This is carried out by comparing experiments B and C in terms of (i) well-posedness of the inference and (ii) quantification of the predictive uncertainty in the runoff. Since we do not have information about the true rainfall, its PD cannot be assessed.

9.1. Achieving well-posedness using prior information

The MCMC sampler did not converge for OIP-2 and OIP-3, suggesting that the inference is ill-posed due to non-identifiability of some inferred quantities. In comparison with Experiment B (where OIP-2 was well-posed), the inference is ill-posed even when the prior contains appreciable information on the rainfall error hyper-parameters. The posterior correlation matrix of latent variables characterizing input and structural errors (Figure 10) exhibits the same bloc-diagonal structure as observed with Experiment B (section 8.1).

9.2. Evaluating the predictive distribution of runoff

The reliability of the runoff PD is summarized in Figure 11. Similar conclusions to those reached in Experiment B hold:

(i) Schemes SLS and O lead to a significant fraction of observed runoffs being outside their predictive range, with ξ_Q values of 0.83 and 0.68 respectively.

(ii) Scheme OI-1 has a high number of observations outside the predictive range (ξ_Q = 0.84), which is similar to findings in Experiment B. However, as discussed in section 10.2.5, the reasons for this may be different.

(iii) Schemes OI-2 and OI-3 have almost no observations outside the predictive range, (ξ_Q = 0.99 and 1 respectively). Moreover, the reliability of the runoff PD (α_Q values of 0.68 and 0.72) is acceptable, though far from perfect.

(iv) Schemes OP and OIP-1, which allow parameter stochasticity, have no observations outside the predicted range (ξ_Q = 1 in all cases). However, low α_Q values of 0.48 and 0.44 suggest that the reliability of the runoff PD is unsatisfactory - it considerably overestimates the predictive uncertainty. This is in contrast to Experiment B, which had higher values of α_Q around 0.8. The reasons for this difference are discussed in Section 10.2.5.

10. Discussion

This paper investigates the feasibility of decomposing the total predictive uncertainty into several components arising from input and structural errors. To achieve this, a calibration scheme must conform to the following progressive requirements:

(i) The inference is well-posed;

(ii) The total runoff PD is successfully quantified (i.e., with acceptable reliability and resolution);

(iii) Input and structural uncertainties are successfully decomposed.

This section discusses the results of sections 8 and 9 in the context of these requirements.

10.1. Well-posedness of the inference

10.1.1. Well-posed schemes

Schemes SLS, O, OI and OP lead to a well-posed inference in all experiments. Moreover, scheme OIP is also well-posed when sufficiently precise priors on rainfall errors are specified, though the required precision varied between experiments B and C.

This shows that direct modeling of multiple sources of error using hierarchical methods is not inherently ill-posed, but depends on the amount of prior knowledge relative to the number and complexity of the sources of uncertainty included in the analysis. Section 10.1.3 further discusses the relationship between dimensionality and well-posedness.

10.1.2. Ill-posed schemes

Experiments B and C show that when both input and structural errors are explicitly modeled using latent variables (OIP schemes) and only vague prior information on the input errors is available, the decomposition of input and structural errors becomes an ill-posed problem. This is due to interactions between latent variable representing input and structural errors. For example, an increase in log-multiplier $\phi_{\tau(t)}$ can be compensated by a decrease in the stochastic CRR parameter $\lambda_{\omega(t)}$ associated with the same time step. This results in large correlated subspaces within the inference space having near-constant likelihood values. This is the non-identifiability property

described in Section 3.1, which turns into ill-posedness in the absence of sufficient prior information.

Sufficient prior information on rainfall uncertainty is required for a well-posed inference (scheme OIP-1). It is stressed that the inference is then conditioned on this auxiliary information and it is crucial that the latter reflect actual knowledge rather than be viewed as a tuning parameter to achieve MCMC convergence. Section 10.3.2 outlines several avenues for obtaining adequate prior information.

The consistency of results of experiments B and C suggests that the strong interaction between input and structural errors is not an artifact due to the type of structural errors used in the synthetic case study (calibrating a CRR model *M1* with data generated from a different model *M0* in experiment B). Indeed, we encountered similar ill-posedness in case studies based on other catchments (not shown). This confirms that ill-posedness is not specific to experiments B and C, but reflects a general and intrinsic difficulty in separating multiple sources of error, especially with weak prior knowledge. These results are unsurprising - it is impossible to infer CRR parameters and individual input and structural errors using only a single rainfall-runoff dataset if the modeler has no idea about the accuracy of neither the CRR model nor the data.

Note that calibrating to longer time series may not necessarily help in identifying individual input errors or breaking their interaction with structural errors. In particular, due to the finite memory of the CRR model, the effect of a rainfall error decreases over time, such that, e.g,, additional data at step t+30 (days) will hardly improve the identifiability of a latent variable at step t.

Instead, independent estimates of data accuracy are required to formulate meaningful priors on the data errors. Whether these priors will be sufficient to achieve a well-posed inference is problem-specific. For example, a higher-precision prior was required to achieve well-posedness in

experiment C than in experiment B. From a practical perspective, an understanding of the data uncertainty needs to become an essential part of the CRR model calibration.

10.1.3. Well-posedness, non-identifiability and over-parameterization

The representation and inference of input and/or structural errors using stochastic parameters inevitably increases the dimensionality of the problem. Many hydrologists and practitioners instinctively shy away from high-dimensional inference problems, believing them to be invariably ill-posed or non-identifiable. However, high-dimensional problems are neither inherently non-identifiable nor inherently ill-posed - this depends on how the likelihood is formulated and what additional (prior) information is available.

It is stressed that identifiability, well-posedness and the dimensionality of the inference space are three distinct concepts. For example, section 3 shows that a simple 2-parameter problem is completely non-identifiable for any sample size. This non-identifiability may or may not lead to an ill-posed inference, depending on the strength of the prior distribution.

More generally, the notion of "model complexity" in Bayesian hierarchical models is non-trivial; in most cases, the number of inferred quantities is a poor measure of complexity [see *Spiegelhalter et al.*, 2002, for a detailed discussion]. In particular, different prior assumptions may affect the well-posedness of the inference. For example, the well-posedness of the OIP scheme in Experiment B varies with the prior precision even though the number of estimated quantities remained exactly the same.

10.2. Successful quantification of runoff predictive uncertainty

10.2.1. Effects of CRR parameter uncertainty on predictive distributions

Analysis of the posterior distributions in all experiments suggested that the uncertainty in the deterministic CRR parameters is relatively small (not shown) and its effect on predictive uncertainty is dominated by errors in the data and model structure. This is a consequence of posterior parametric uncertainty decreasing as more data is used [e.g., *Kuczera et al.*, 2006; *Stedinger et al.*, 2008]. Consequently, it is not considered in further detail in this paper (but see discussions by *Beven et al.* [2008] and *Mantovan and Todini* [2006]).

10.2.2. Traditional (non-hierarchical) schemes

Approaches SLS and O lead to an unreliable and underestimated predictive uncertainty, especially for high runoffs. This occurs because these calibration schemes lump several sources of errors (input/output/structural for SLS, input/structural for O) into the single remnant error term. Consequently, the majority of predictive uncertainty arises from remnant errors, which are assumed to have a Gaussian distribution. However, the Gaussian assumption is clearly not supported by the data: the standardized residuals are highly skewed and leptokurtotique (Figure 12). This violation of assumptions explains the underestimation of predictive uncertainty.

10.2.3. Hierarchical schemes: General comments

In Experiment B, approaches OP and OI quantified predictive uncertainty much more reliably than O and SLS. Method OI-1 is the only exception, with a mild underestimation of predictive uncertainty (see section 8.2.1). When well-posed due to sufficient prior precision (cases OIP-1 and OIP-2), approach OIP also improves the estimation of the runoff PD.

In all cases, the improvement is due to the use of latent variables for describing structural and/or input errors: most of the predictive uncertainty arises from stochastic parameters. Introducing stochastic parameters has two effects on remnant errors:

(i) it reduces their standard deviation σ (Figure 13). This is consistent with its expected behavior as the input/output/structural error models are improved (section 5.5).

(ii) the standardized residuals are more Gaussian (Figure 12). This suggests that the common observation that residuals of hydrological models are skewed and leptokurtotic [*Beven*, 2006] is probably caused by unduly simplistic lumped treatment of the different sources of uncertainty.

Note that the introduction of stochastic parameters did not significantly affect the autocorrelation of the residuals, with a lag-1 coefficient remaining between ~ 0.2-0.3 for all calibration schemes except scheme O (lag-1 coefficient ~ 0.5). While such low autocorrelation will not affect the conclusions of this study, much stronger autocorrelation may arise when modeling on a shorter time scale. Hence, simulations based on hourly rainfall may require specialized treatment to handle autocorrelation.

Overall, the results suggest that characterization of errors (input and/or structural) using stochastic parameters leads to a significant improvement over traditional additive-error approaches in terms of reliability of the predictive uncertainty.

10.2.4. Treating a single source of uncertainty hierarchically

Experiment B suggests that treating either input or structural error (but not both) with a single stochastic parameter can produce reliable runoff predictions (Figure 8, index α_Q in the range 0.78-0.9). However, this is only partially supported by experiment C (section 9.2), where the reliability index α_Q in the range 0.48-0.84 leaves room for improvement.

These results emphasize the importance of validating the predictive uncertainty [*Hall et al.*, 2007]: in its absence, there is no guarantee that the inferred predictive uncertainty is meaningful. The use of predictive distributions without comprehensive analysis of their reliability and resolution can lead to large prediction errors and misleading risk estimates.

Interestingly, representing either rainfall or structural errors using a stochastic parameter can lead to a reliable PD of the runoff (Figure 8) even though input and structural errors cannot be successfully decomposed. This is analogous to the simplified example in Section 3.6 – even though the individual parameters θ_1 and θ_2 were not inferable, the model still provided reliable predictions of the responses that it was calibrated to (but see Section 3.6 for very important caveats).

The approach of treating a single source of error (input or structural) using a stochastic parameter is not a complete solution. Even though it may produce more reliable predictions than SLS and additive errors models, the following problems remain:

(1) Interpretation of the stochastic parameter is problematic because it can encompass both input and structural errors. This provides no insight on whether the reduction of predictive uncertainty requires improving the input data (e.g., more rain gauges) or the model structure. While the need for more accurate and precise hydrological data (accompanied by uncertainty estimates) cannot be overstated, the ability to determine the relative contributions of input/structural uncertainties would strategically guide research efforts and experimental resources to reduce predictive uncertainty.

(2) Model extrapolation can be particularly unreliable. For example, the predictive ability of the model can deteriorate if forced with rainfall time series with different properties than those of the

calibration period. This can occur during climate change projections, flood forecasting, or simply when the number of raingauges changes.

10.2.5. Further comments on the OI-1 scheme

Scheme OI-1 deserves further comment. In both experiments B and C, OI-1 has a larger number of observations outside the predictive range ($\xi_Q = 0.9$ and $\xi_Q = 0.84$ respectively) than OI-2 and OI-3.

In experiment B, this occurs because the very precise prior used for the rainfall error hyper-parameters strongly constrains the latent variables, preventing them from compensating for the inadequate treatment of structural uncertainty. The structural uncertainty is accounted for by remnant errors, which in this case are highly skewed and non-Gaussian (Figure 12).

The interpretation of experiment C is more difficult. In addition to a poor remnant error model, the unreliable performance of the OI-1 methods is likely a consequence of inaccurate prior knowledge of rainfall and runoff data errors, moreover, specified using unduly precise priors (in particular, the generic 10% streamflow error model was fixed a priori). However, since additional data is not available for this catchment, it is impossible to verify either explanation. This highlights three key issues: (i) posterior scrutiny is essential to identify violations of underlying statistical hypotheses; (ii) reliable independent estimates of data accuracy are needed for meaningful statistical inference; and (iii) all hydrological data should be accompanied by error estimates.

10.2.6. Interaction between log-multipliers and structural errors

In the OI schemes, the latent variables (log-multipliers) are intended to represent input errors, whereas remnant errors are intended primarily for structural errors (Table 2). However, for these methods, the standard deviation (Figure 13) and the skewness (Figure 12) of the remnant errors

decrease as the precision of the priors on rainfall uncertainty decreases, while the estimated hyper-SD of log-multipliers increases.

This suggests that, in the absence of sufficient prior information on input uncertainty, the rainfall log-multipliers can be contaminated by structural errors. In other words, both sources of errors tend to be conflated in the input error model. This causes an overestimation of rainfall uncertainty (section 8.3). The implications of this behavior for practical applications that calibrate CRR models to rainfall data with no associated error estimates is further discussed in Section 10.3.2.

10.3. Successful decomposition of runoff predictive uncertainty in input/structural errors components

10.3.1. Reliability and resolution of input PD

In the absence of structural errors, no prior information on input errors appears to be required to achieve a well-posed and accurate inference. In particular, estimates of rainfall errors are reliable and precise (experiment A). This is not the case when structural errors are present (experiment B).

This section considers the estimation of the rainfall PD. In particular, it must be inferred reliably before a meaningful decomposition of predictive uncertainty can be obtained. In experiment B, only two approaches achieved this:

(i) Schemes OI with precise priors (OI-1, and to a lesser extent, OI-2) achieve high rainfall reliability ($\alpha_R \approx 0.9$, $\xi_R \approx 0.95$). This is an important result because it suggests that individual rainfall errors (and hence estimates of the true rainfall) can be retrieved from the data in the presence of structural errors, provided the properties of rainfall errors are well understood prior to the inference (i.e., precise priors for the hyper-parameters).

However, the reliability and resolution of the rainfall PD deteriorates rapidly when weaker prior information is supplied. In particular, the standard deviation of the hyper-distribution of input errors becomes progressively overestimated, up to by a factor of 4 for OI-3 (Table 3). Moreover, the improved reliability of rainfall PD achievable with high prior precision comes at the cost of a decreased reliability of the runoff PD (Section 10.2.5). Consequently, precise prior information on rainfall alone, without an appropriate representation of structural errors, appears insufficient for successfully decomposing the total predictive uncertainty.

(ii) Schemes OIP with precise priors (OIP-1 and OIP-2) also achieve high rainfall reliability ($\alpha_R \approx 0.9$, $\xi_R \approx 1$). Again, prior information plays a central role by controlling the well-posedness of the inference. However, although the rainfall PD is reliable, it remains similar to the hyper-distribution (see Figure 9 and section 8.3.2). This is a consequence of most multipliers being only weakly identifiable from the data; their inference is largely controlled by prior knowledge. In the language of probabilistic forecasting [*Atger*, 1999], the resulting rainfall PD is not "skillful" because it does not contain any information beyond that given by the prior hyper-distribution. The influence of the prior also emphasizes that meaningful uncertainty estimates are not an optional extra when collecting and reporting hydrological data.

As an aside, the point above also illustrates that reliability alone does not imply usefulness when the resolution is low. For example, climatologic predictions are reliable in the distributional sense, but are not useful for forecasting specific events. This is broadly analogous to the difference between the marginal versus conditional predictive distribution.

10.3.2. Perspectives on uncertainty in hydrological modeling

This study suggests that success of the inference (measured by the reliability of runoff predictions and successful decomposition of input and structural errors) is largely determined by the prior hypotheses describing the distributional properties of rainfall and runoff errors. It is therefore important that the priors used in the inference reflect actual knowledge, rather than be treated as mere mathematical tricks to ensure MCMC convergence. Indeed, a precise but inaccurate prior will simply yield fast convergence to the wrong posterior. The limiting case is SLS – it specifies the precise but incorrect prior that the observed rainfall is exact and yields a biased inference. This highlights the need to develop and implement reliable methods for estimating the accuracy and precision of measured environmental data at the data-collection and post-processing stages. Given the superior performance of methods exploiting accurate prior information, this would allow much deeper statistical inferences to be carried out than currently possible.

Contrary to widespread hydrological pessimism, formulating accurate prior hypotheses is not an impossible Herculean task, and several promising avenues are already apparent.

Useful prior distributions of rainfall errors and their hyper-parameters can be derived from spatial analysis of rainfall fields, e.g., using radar data and/or geostatistical analyses of raingauge networks [e.g., *Severino and Alpuim*, 2005]. Our preliminary research in this direction is very encouraging – using conditional rainfall simulation eliminated ill-posedness and significantly improved the reliability and resolution of predictive distributions [*Renard et al.*, 2009b].

Using data on other state variables can also be useful. For example, independent estimates of saturated areas [*Franks et al.*, 1998] may help identifying the variations of stochastic parameters controlling the catchment saturation. Additionally, isotope data can yield independent insights into residence times and internal model pathways [e.g., *Fenicia et al.*, 2008; *Fenicia et al.*, in press].

Further research is needed to derive meaningful probabilistic models for such additional data and will be reported in future work.

Finally, while this study focuses on lumped conceptual hydrological models, similar concerns hold for the identifiability of more complex physically-based distributed models. Indeed, given the increased data requirements necessary to support the identification and resolution of additional catchment processes represented in these models, we expect the role of reliable prior knowledge to become even more critical.

11. Conclusions

Bayesian total error analysis (BATEA) offers an inference framework that combines the estimation of rainfall-runoff dynamics with an honest accounting of errors in the observations and the hypothesized model structure. However, this study shows that sufficient independent information must be supplied to the inference before the total predictive uncertainty can be meaningfully decomposed into its contributing sources. Indeed, a key strength of the Bayesian paradigm is its ability to use independent prior knowledge to obtain a well-posed and useful inference even when the data alone may not be sufficient.

Empirical analysis suggests that a single set of rainfall-runoff data without sufficiently precise estimates of rainfall uncertainty is insufficient to infer more than one source of errors, even if the distribution of runoff errors is known. Non-identifiability problems arise when attempting to disaggregate input and structural errors; unless informative priors on rainfall uncertainty are used, this leads to an ill-posed inference. In this respect, priors on the hyper-parameters describing data uncertainty play a very different role to the priors on the CRR model parameters: while the latter

merely enhance the inference for short calibration datasets, the former control the overall well-posedness of the inference.

It was also demonstrated that ill-posedness of the inference can often be diagnosed from exceedingly slow MCMC convergence. In particular, when non-informative priors are used, poor MCMC convergence is symptomatic of inferred quantities (e.g., model parameters, data and structural errors, etc.) being poorly identifiable from the data.

In the broader hydrological context, this reflects the inherent limitations of using sparse data of unknown quality to make reliable statistical inference and meaningfully disaggregate multiple sources of uncertainty. If no independent estimates of data uncertainty are available, the discrepancy between observed and simulated responses only provides information about total errors. Without further information, it is impossible to decompose this error into its components. This is the fundamental reality confronting hydrologic modeling.

Another important conclusion is that hierarchical representation of input and/or structural errors produces more reliable runoff predictions than the traditional approach of a deterministic CRR model with an additive error model. While this results in an increased dimensionality of the problem, it remains computationally practical even on standard computers and laptops.

More specifically, synthetic and real data studies in this paper suggest that:

1. If only rainfall-runoff data are used and no independent data uncertainty estimates are available, only the total error can be analyzed. This can be accomplished using standard regression methods such as standard and weighted least squares schemes. The individual contributions of rainfall, runoff and structural errors to predictive uncertainty cannot be disaggregated. Moreover, in standard regression methods, unless the statistical properties of the total error are properly satisfied

by the residual error model - which is difficult to attain in practice, especially with multiple sources of error and large errors in the inputs - predictive uncertainty quantification is inadequate and predictions may be biased. Consequently, in the case where insufficient prior information is available, uncertainty analysis should be based on specialized statistical techniques [e.g., the semi-parametric approaches of *Krzysztofowicz*, 2002; *Montanari and Brath*, 2004], and the reliability of the predictive uncertainty should be thoroughly assessed. Yet attaining independent data uncertainty estimates is always preferable, and we strongly encourage experimentalists and data analysts to work towards this.

2. Adding independent knowledge to formulate an informative prior on the properties of runoff errors enables a meaningful inference of the combined distributional properties of rainfall and structural errors, and their combined contribution to predictive uncertainty. However, what may be identified as "input error" by the calibration scheme can also encompass a significant portion of structural error, and vice versa. In either case, the disaggregation of rainfall and structural errors is ill-posed.

3. Using independent knowledge to formulate precise priors for both runoff and rainfall hyper-parameters permits well-posed individual inference of rainfall and structural errors, including the distributional properties of the latter. In other words, the decomposition of the total predictive uncertainty into its three constituents requires precise priors for rainfall and runoff error hyper-parameters, with the rainfall-runoff data then providing closure on the remaining structural error. The resulting inference provides both (i) reliable estimates of total predictive uncertainty, with predictive precision dependent on the quality of the data and model; and (ii) reliable decomposition of the total uncertainty into its various sources. Along with a corresponding

improvement in the model representation, we consider this scenario to be a strategic goal for hydrologic model estimation.

These conclusions highlight inherent limitations of calibrating inaccurate CRR models to observed rainfall-runoff data of unknown quality. They also call for a more systematic reporting of errors affecting environmental data, both at the acquisition and post-processing stages. In particular, a reliable quantitative understanding of data uncertainty should not be viewed as some "esoteric" prior knowledge, but rather as an essential specification of the inference problem.

12. Acknowledgments

The authors thank Editor Praveen Kumar, Associate Editor Alberto Montanari, Keith Beven, Peter Reichert and the anonymous reviewers for helpful comments and suggestions, in particular, the ACF remark. This work was partially funded by the Australian Research Council grant DP00773000.

13. References

Atger, F. (1999), The skill of ensemble prediction systems, *Monthly Weather Review*, *127*(9), 1941-1953.
Beven, K. J. and A. M. Binley (1992), The future of distributed hydrological models: model calibration and uncertainty prediction., *Hydrological Processes*, *6*, 279-298.
Beven, K. J. (2006), A manifesto for the equifinality thesis, *Journal of Hydrology*, *320*, 18–36.
Beven, K. J., P. J. Smith and J. E. Freer (2008), So just why would a modeller choose to be incoherent?, *Journal of Hydrology*, *354*(1-4), 15-32.
Bras, R. L. and I. Rodriguez-Iturbe (1984), *Random Functions and Hydrology*, Addison-Wesley.
Bulygina, N. and H. Gupta (2009), Estimating the uncertain mathematical structure of a water balance model via Bayesian data assimilation, *Water Resources Research*, *45*.
Di Baldassarre, G. and A. Montanari (2009), Uncertainty in river discharge observations: a quantitative analysis, *Hydrology and Earth System Sciences*, *13*(6), 913-921.
Dottori, F., M. L. V. Martina and E. Todini (2009), A dynamic rating curve approach to indirect discharge measurement, *Hydrol. Earth Syst. Sci. Discuss.*, *6*(1), 859-896.
Duan, Q., N. K. Ajami, X. Gao and S. Sorooshian (2007), Multi-model ensemble hydrologic prediction using Bayesian model averaging, *Advances in Water Resources*, *30*(5), 1371-1386.

Eberly, L. E. and B. P. Carlin (2000), Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models, *Statistics in Medicine*, *19*, 2279-2294.
Fenicia, F., J. J. McDonnell and H. H. G. Savenije (2008), Learning from model improvement: On the contribution of complementary data to process understanding, *Water Resources Research*, *44*(6).
Fenicia, F., S. Wrede, D. Kavetski, L. Pfister, L. Hoffmann, H. Savenije and J. J. McDonnell (in press), Impact of mixing assumptions on mean residence time estimation, *Hydrological Processes (Special Issue on Residence Times and Preferential Flows)*.
Feyen, L., J. A. Vrugt, B. O. Nuallain, J. van der Knijff and A. De Roo (2007), Parameter optimisation and uncertainty assessment for large-scale streamflow simulation with the LISFLOOD model, *Journal of Hydrology*, *332*(3-4), 276-289.
Franks, S., P. Gineste, K. J. Beven and P. Merot (1998), On constraining the predictions of a distributed model: the incorporation of fuzzy estimates of saturated areas into the calibration process., *Water Resources Research*, *34*(4), 787-797.
Gelfand, A. E. and S. K. Sahu (1999), Identifiability, improper priors, and gibbs sampling for generalized linear models, *Journal of the American Statistical Association*, *94*, 247-253.
Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin (1995), *Bayesian data analysis*, 526 pp., Chapman & Hall.
Goldstein, M. and J. Rougier (2009), Reified Bayesian modelling and inference for physical systems, *Journal of Statistical Planning and Inference*, *139*(3), 1221-1239.
Haario, H., E. Saksman and J. Tamminen (2005), Componentwise adaptation for high dimensional MCMC, *Computational Statistics*, *20*(2), 265-273.
Hadamard, J. (1902), Sur les problèmes aux dérivées partielles et leur signification physique, *Princeton University Bulletin*, 49-52.
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), *Hydrological Processes*, *21*(7), 985-988.
Huard, D. and A. Mailhot (2008), Calibration of hydrological model GR2M using Bayesian uncertainty analysis, *Water Resources Research*, *44*.
Jacquin, A. and A. Y. Shamseldin (2007), Development of a possibilistic method for the evaluation of predictive uncertainty in rainfall-runoff modeling, *Water Resources Research*, *43*.
Kavetski, D., S. Franks and G. Kuczera (2002), Confronting Input Uncertainty in Environmental Modelling in Calibration of Watershed Models, in *Water Science and Application Series 6*, edited by Q. Y. Duan, et al., pp. 49-68, American Geophysical Union, Washington DC.
Kavetski, D., G. Kuczera and S. W. Franks (2003), Semidistributed hydrological modeling: A "saturation path" perspective on TOPMODEL and VIC, *Water Resources Research*, *39*(9).
Kavetski, D., G. Kuczera and S. W. Franks (2006a), Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, *Water Resources Research*, *42*(3).
Kavetski, D., G. Kuczera and S. W. Franks (2006b), Calibration of conceptual hydrological models revisited: 2. Improving optimisation and analysis, *Journal of Hydrology*, *320*(1-2), 187-201.
Kennedy, M. C. and A. O'Hagan (2001), Bayesian Calibration of Computer Models, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *63*(3), 425-464.
Krzysztofowicz, R. (2002), Bayesian system for probabilistic river stage forecasting, *Journal of Hydrology*, *268*(1-4), 16-40.

Kuczera, G. and B. J. Williams (1992), Effect of Rainfall Errors on Accuracy of Design Flood Estimates, *Water Resources Research*, *28*(4), 1145-1153.
Kuczera, G. and E. Parent (1998), Monte Carlo assessment of parameter uncertainty in conceptual catchment models: The Metropolis algorithm, *Journal of Hydrology*, *211*, 69-85.
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, *Journal of Hydrology*, *331*(1-2), 161-177.
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 Congress, Christchurch, New Zealand.
Laio, F. and S. Tamea (2007), Verification tools for probabilistic forecasts of continuous hydrological variables, *Hydrology and Earth System Sciences*, *11*(4), 1267-1277.
Mantovan, P. and E. Todini (2006), Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, *Journal of Hydrology*, *330*(1-2), 368-381.
Marshall, L., D. Nott and A. Sharma (2007), Towards dynamic catchment modelling: a Bayesian hierarchical mixtures of experts framework, *Hydrological Processes*, *21*(7), 847-861.
Montanari, A. and A. Brath (2004), A stochastic approach for assessing the uncertainty of rainfall-runoff simulations, *Water Resour. Res.*, *40*.
Montanari, A. (2007), What do we mean by 'uncertainty'? The need for a consistent wording about uncertainty assessment in hydrology, *Hydrological Processes*, *21*, 841-847.
Moradkhani, H., S. Sorooshian, H. V. Gupta and P. R. Houser (2005), Dual state–parameter estimation of hydrological models using ensemble Kalman filter, *Advances in Water Resources*, *28*, 135-147.
Moyeed, R. A. and R. T. Clarke (2005), The use of Bayesian methods for fitting rating curves, with case studies., *Advances in Water Resources*, *28*, 807-818.
Neppel, L., B. Renard, M. Lang, P. A. Ayral, D. Coeur, E. Gaume, N. Jacob, O. Payrastre, K. Pobanz and F. Vinet (2009), Flood frequency analysis using historical data: accounting for random and systematic errors, *Hydrological sciences Journal. In Press.*
Oudin, L., C. Perrin, T. Mathevet, V. Andreassian and C. Michel (2006), Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models, *Journal of Hydrology*, *320*(1-2), 62-83.
Perrin, C., C. Michel and V. Andreassian (2003), Improvement of a parsimonious model for streamflow simulation, *Journal of Hydrology*, *279*(1-4), 275-289.
Refsgaard, J. C., J. P. van der Sluijs, J. Brown and P. van der Keur (2006), A framework for dealing with uncertainty due to model structure error, *Advances in Water Resources*, *29*(11), 1586-1597.
Reichert, P. and J. Mieleitner (2009), Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters, *Water Resources Research*, *45*.
Reitan, T. and A. Petersen-Overleir (2009), Bayesian methods for estimating multi-segment discharge rating curves, *Stochastic Environmental Research and Risk Assessment. In press.*
Renard, B., D. Kavetski and G. Kuczera (2009a), Comment on "An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction." by N. K. Ajami, Q. Y. Duan, and S. Sorooshian, *Water Resources Research*, *45*.
Renard, B., G. Kuczera, E. Leblois, D. Kavetski, M. Thyer and S. Franks (2009b), Characterizing areal rainfall errors: application to uncertainty quantification and decomposition in hydrologic

modelling., paper presented at 32nd Hydrology and Water Resources Symposium 2009, Engineers Australia, Newcastle, Australia.
Severino, E. and T. Alpuim (2005), Spatiotemporal models in the estimation of area precipitation, *Environmetrics*, *16*, 773-802.
Spiegelhalter, D. J., N. G. Best, B. R. Carlin and A. van der Linde (2002), Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society Series B-Statistical Methodology*, *64*, 583-616.
Stedinger, J. R., R. M. Vogel, S. U. Lee and R. Batchelder (2008), Appraisal of the generalized likelihood uncertainty estimation (GLUE) method, *Water Resources Research*, *44*.
Tarantola, A. (2005), *Inverse Problem Theory and Methods for Model Parameter Estimation*, Society for Industrial and Applied Mathematics.
Thiemann, M., H. Trosset, H. Gupta and S. Sorooshian (2001), Bayesian recursive parameter estimation for hydrologic models, *Water Resources Research*, *37*(10), 2521-2535.
Thyer, M., B. Renard, D. Kavetski, G. Kuczera, S. Franks and S. Srikanthan (2009), Critical evaluation of parameter consistency and predictive uncertainty in hydrological modelling: a case study using bayesian total error analysis, *Water Resources Research*, *45*.
Tonkin, M., J. Doherty and C. Moore (2007), Efficient nonlinear predictive error variance for highly parameterized models, *Water Resources Research*, *43*.
Wagener, T., D. P. Boyle, M. J. Lees, H. S. Wheater, H. V. Gupta and S. Sorooshian (2001), A framework for development and application of hydrological models, *Hydrology and Earth System Sciences*, *5*(1), 13-26.
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*.
Young, P. (1998), Data-based mechanistic modelling of environmental, ecological, economic and engineering systems, *Environmental Modelling & Software*, *13*(2), 105-122.

14. Appendix A: Description of LogSPM

This paper uses a modified version of the LogSPM model [*Kavetski et al.*, 2003; *Kuczera et al.*, 2006]. The model simulates runoff (q) using rainfall (r) and potential evapotranspiration (pet) (here, all in mm). The model has three stores and six parameters (shown in bold below):

Soil store:

$$\begin{aligned}
&\frac{dh_s}{dt} = r - q_{quick} - q_{rge} - q_{et} && \text{(a) Surface water balance} \\
&q_{quick} = f(h_s) \times r && \text{(b) Quickflow} \\
&q_{rge} = f(h_s) \times \boldsymbol{rge_{Max}} && \text{(c) Groundwater recharge} \\
&q_{et} = pet \times (1 - \exp(-h_s \times \boldsymbol{k_{Et}})) && \text{(d) Actual evapotranspiration} \\
&f(h_s) = \frac{2}{1 + \exp(-h_s \times \boldsymbol{k_S})} - 1 && \text{(e) Fraction of saturated area}
\end{aligned} \tag{27}$$

Groundwater store:

$$\begin{aligned}
&\frac{dh_{gw}}{dt} = q_{rge} - q_b - q_{deep} && \text{(a) Groundwater balance} \\
&q_b = h_{gw} \times \boldsymbol{k_{Gw}} && \text{(b) Baseflow} \\
&q_{deep} = h_{gw} \times \boldsymbol{k_{Dp}} && \text{(c) Percolation to deep aquifers}
\end{aligned} \tag{28}$$

Stream store:

$$\begin{aligned}
&\frac{dh_{stream}}{dt} = q_{quick} + q_b - q && \text{(a) Stream store balance} \\
&q = h_{stream} \times \boldsymbol{k_{Stream}} && \text{(b) River runoff}
\end{aligned} \tag{29}$$

The prior parameter distributions are given in Table 1.

List of captions

Tables

Table 1. Description of LogSPM parameters and their prior distributions.

Table 2. Summary of calibration schemes in experiments A-C and run details of Experiment B.

Table 3. Estimated hyper-parameters of rainfall data errors (log-multipliers). The first number is the marginal posterior mean of the hyper-parameter, the number in brackets is the marginal posterior standard deviation.

Figures

Figure 1. Posterior distributions for the didactic example of Section 3.4. (a) with prior π_1; (b) with prior π_2; (c) posterior distribution of $\theta_1+\theta_2$ with prior π_2.

Figure 2. Evolution of two parallel MCMC chains for parameters θ_1 (left), θ_2 (center) and $\theta_1 + \theta_2$ (right) for the didactic problem of Section 3.4. Top row = prior π_1, bottom row = prior π_2.

Figure 3. Schematic of the predictive QQ plot and derived indexes.

Figure 4. Experiment A: Diagnostic plots for calibration scheme OI-3. (a) observed vs. simulated runoff (validation period); (b) predictive QQ-plot of runoffs exceeding 1 mm (validation period); (c) true, observed and estimated rainfall; (d) predictive QQ-plot of true rainfall. The size of the bubbles in (b) and (d) is proportional to the observed runoff and rainfall, respectively.

Figure 5. Experiment A: Dependence of the posterior precision of estimated log-multipliers on the magnitude of observed rainfall. The horizontal line denotes the precision of the posterior hyper-distribution.

Figure 6. Experiment A: Comparison of the posterior distributions of individual log-multipliers (thin lines) with the true (solid thick line) and the estimated (dashed thick line) hyper-distribution. For readability, only 11 log-multipliers are displayed.

Figure 7. Experiment B: Correlation matrix of latent variables representing structural errors λ_ω and input errors ϕ_τ as a function of the prior precision of the input-error hyper-parameters (OIP-1 assumes the highest prior precision). For readability, only latent variables affecting time step 1 to 58 of the calibration period are displayed.

Figure 8. Experiment B: Summary of the reliability and resolution of the predictive distribution of rainfall (first row) and runoff (second row) inferred using the nine calibration schemes. The indices are defined in section 6. The star denotes the non-convergent OIP-3 case.

Figure 9. Experiment B: Dependence of the posterior precision of individual log-multipliers on the magnitude of observed rainfall. The horizontal line denotes the precision of the posterior hyper-distribution.

Figure 10. Experiment C: Correlation matrix of latent variables representing structural errors λ_ω and input errors ϕ_τ as a function of the prior precision of the input-error hyper-parameters (OIP-1 assumes the highest prior precision).

Figure 11. Experiment C: Summary of the reliability and resolution of the predictive distribution of runoff inferred using the nine calibration schemes. The indices are defined in section 6. The stars denote the non-convergent OIP-2 and OIP-3 cases. Since the true rainfall is unknown, its PD cannot be assessed.

Figure 12. Experiment B: Skewness and excess kurtosis of standardized residuals.

Figure 13. Experiment B: Reduction of remnant errors as more sources of uncertainty are treated explicitly. Note the logarithmic scaling of the y-axis.

Table 1. Description of LogSPM parameters and their prior distributions.

Parameter	Description	Prior
rge_{Max}	Groundwater recharge at full saturation	$\log(rge_{Max}) \sim N(3,3^2)$
k_{Et}	Evapotranspiration (ET) coefficient	$\log(k_{Et}) \sim N(0,4^2)$
k_S	Saturated area function parameter	$\log(k_S) \sim N(-2,4^2)$
k_{Gw}	Baseflow coefficient	$\log(k_{Gw}) \sim N(-6,6^2)$
k_{Dp}	Percolation coefficient	$\log(k_{Dp}) \sim N(0,5^2)$
k_{Stream}	Stream coefficient	$\log(k_{Stream}) \sim N(-1,2^2)$

Table 2. Summary of calibration schemes in experiments A-C and run details of Experiment B.

Name [†]	Handles input errors	Prior precision of $p(\mu_r)$ and $p(\sigma_r)$	Handles output errors	Stochastic CRR model	Interpretation of remnant errors [€]	Treatment of structural errors	Experiment B details		
							Inferred quantities	MCMC iterations N [£] (x 10^3)	Total CPU time[$] (hours)
SLS	no	n/a	no [¥]	no	OIS	Additive, lumped with IO	7	1.8	0.04
O	no	n/a	yes	no	IS + F	Additive, lumped with I	7	0.3	0.04
OP	no	n/a	yes	yes	I + F	P	165	91.5	0.55
OI-1	yes	high	yes	no	S + F	Additive	260	48.6	0.73
OI-2	yes	medium	yes	no	S + F	Additive	260	62.4	0.82
OI-3	yes	low	yes	no	S + F	Additive	260	205.0	1.31
OIP-1	yes	high	yes	yes	F	P	418	176.6	2.45
OIP-2	yes	medium	yes	yes	F	P	418	624.8	5.41
OIP-3	yes	low	yes	yes	F	P	418	∞	∞

[†] Name is constructed as follows: SLS = standard least squares method, O = uses the (heteroscedastic) output error model, I = recognizes input uncertainty, P = uses a stochastic parameter to characterize structural errors. The numbers 1, 2, 3 denote decreasing prior precision

[€] Described as follows: O = denotes ignored output errors, I = denotes ignored input errors, S = denotes ignored structural errors, F = denotes errors remaining from imperfect error models (as opposed to ignored sources of uncertainty)

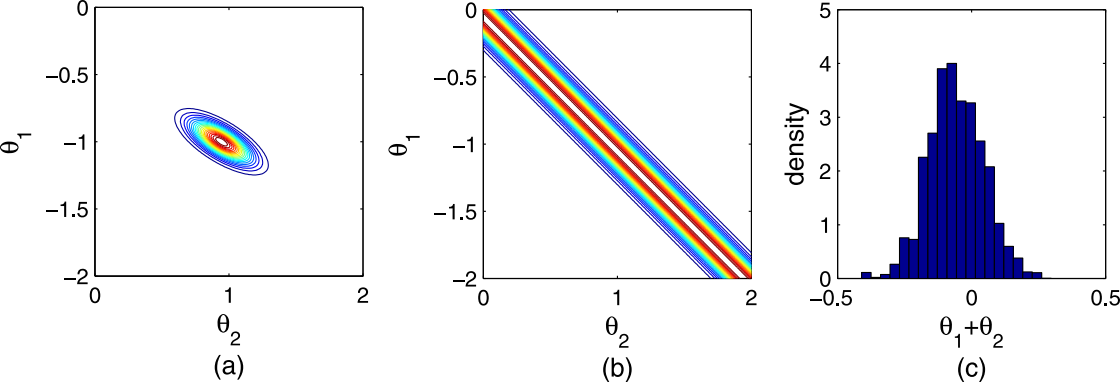
[£] Number of MCMC iterations needed for a max Gelman-Rubin criterion below 1.2 in Experiment B

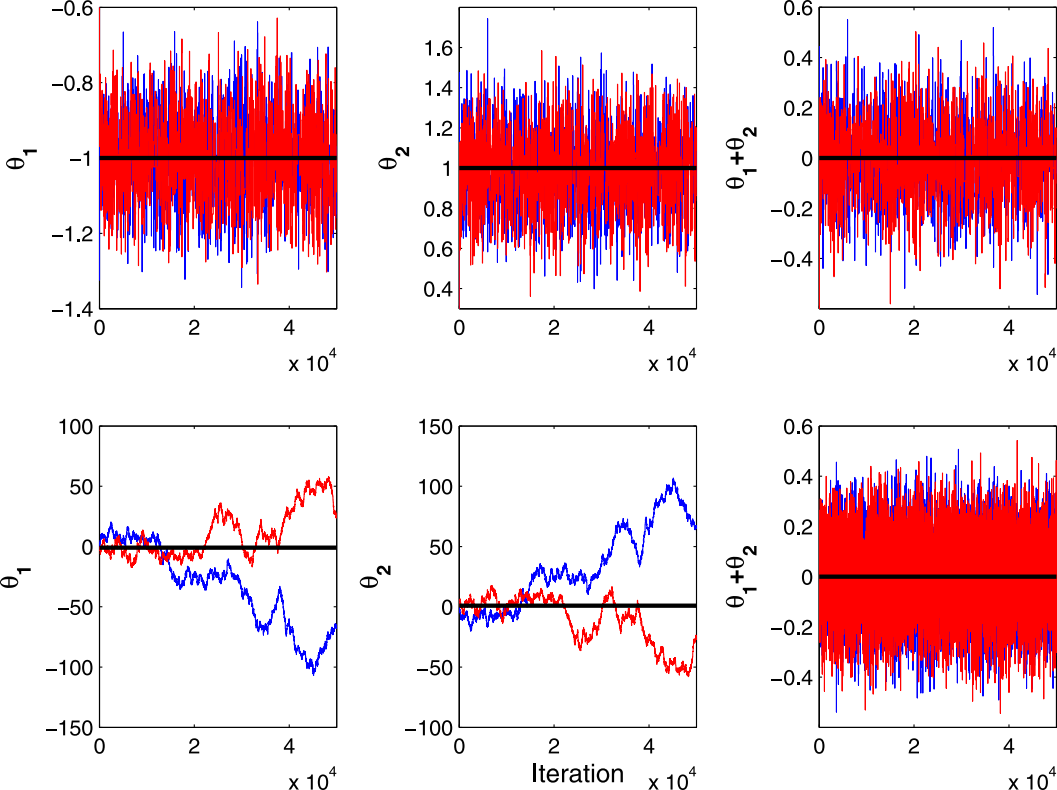
[$] Standard desktop 2GHz CPU for Experiment B

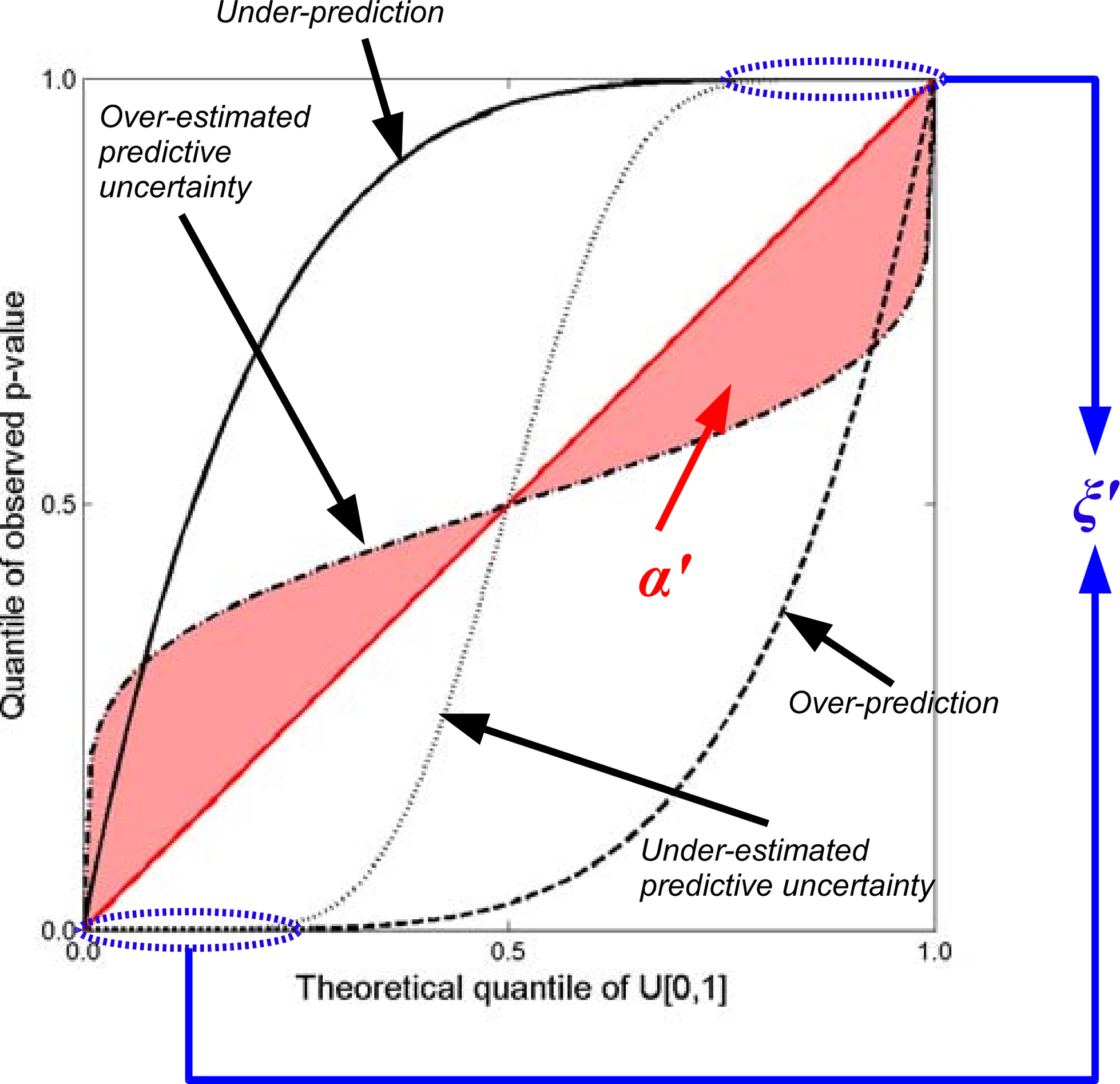
[¥] SLS does not distinguish between output and structural errors

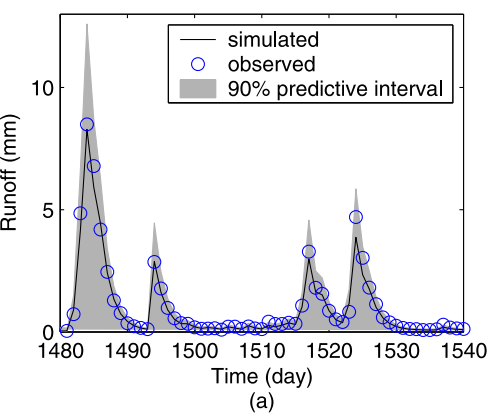
Table 3. Estimated hyper-parameters of rainfall data errors (log-multipliers). The first number is the marginal posterior mean of the hyper-parameter, the number in brackets is the marginal posterior standard deviation.

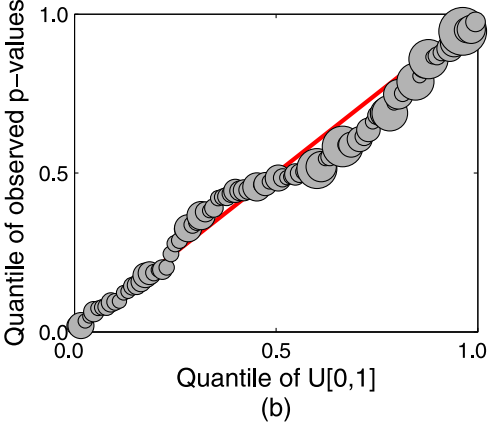
BATEA Model	Hyper-mean μ_r	Hyper-SD σ_r
OI-1	-0.200 [0.001]	0.205 [0.003]
OI-2	-0.203 [0.011]	0.499 [0.038]
OI-3	-0.500 [0.069]	0.862 [0.074]
OIP-1	-0.200 [0.001]	0.201 [0.003]
OIP-2	-0.200 [0.009]	0.349 [0.059]
OIP-3	Did not converge	

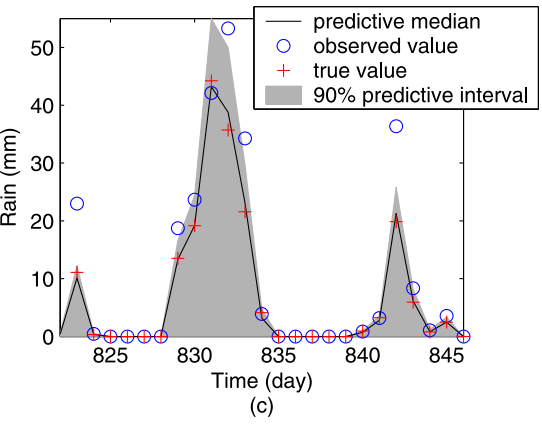

0
−0.5
−1
−1.5
−2
θ1
0
1
2
θ2
(a)
0
−0.5
−1
−1.5
−2
θ1
0
1
2
θ2
(b)
5
4
3
2
1
0
density
−0.5
0
0.5
θ1+θ2
(c)

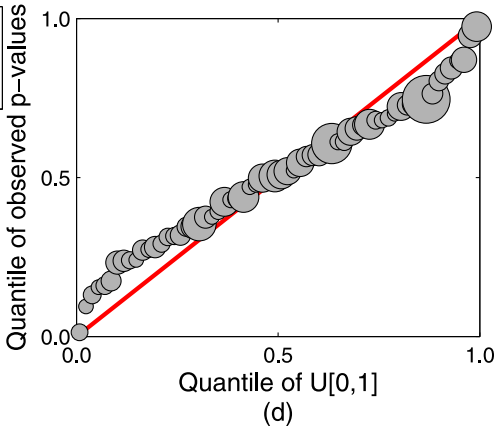

θ1
θ2
θ1+θ2
x 10^4
Iteration

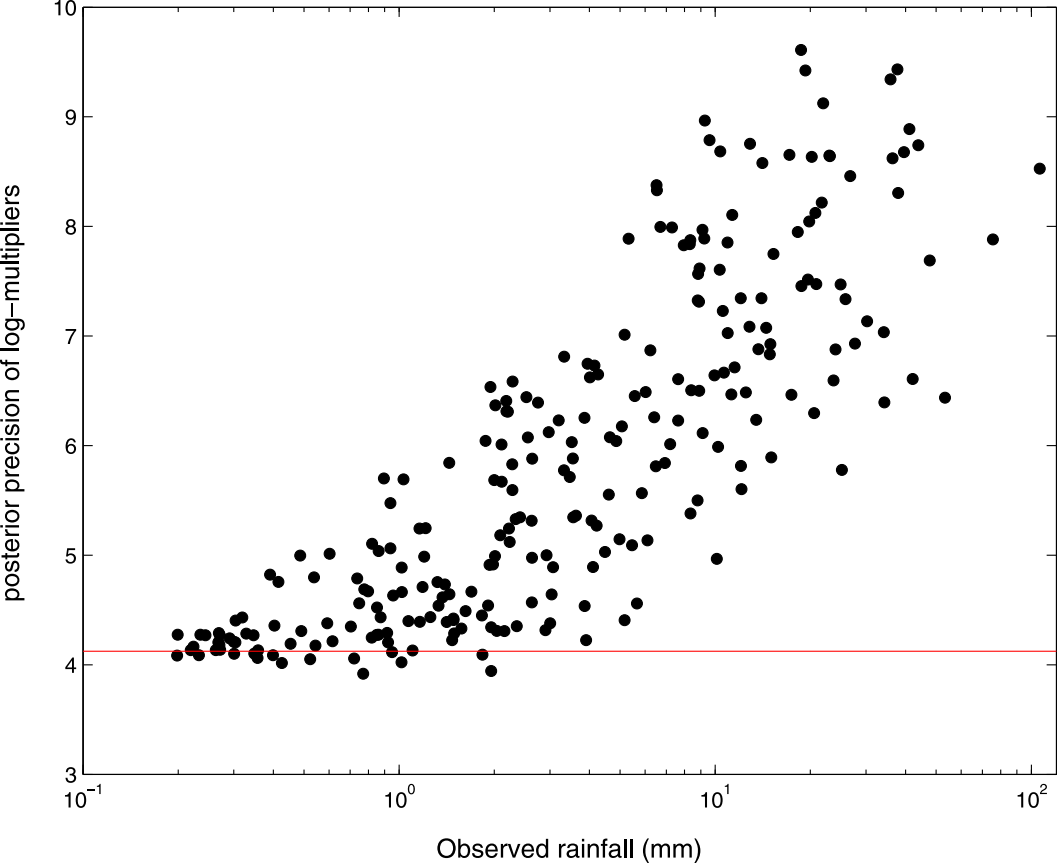

Under-prediction
Over-estimated predictive uncertainty
1.0
0.5
0.0
Quantile of observed p-value
α′
ξ′
Over-prediction
Under-estimated predictive uncertainty
0.0
0.5
1.0
Theoretical quantile of U[0,1]

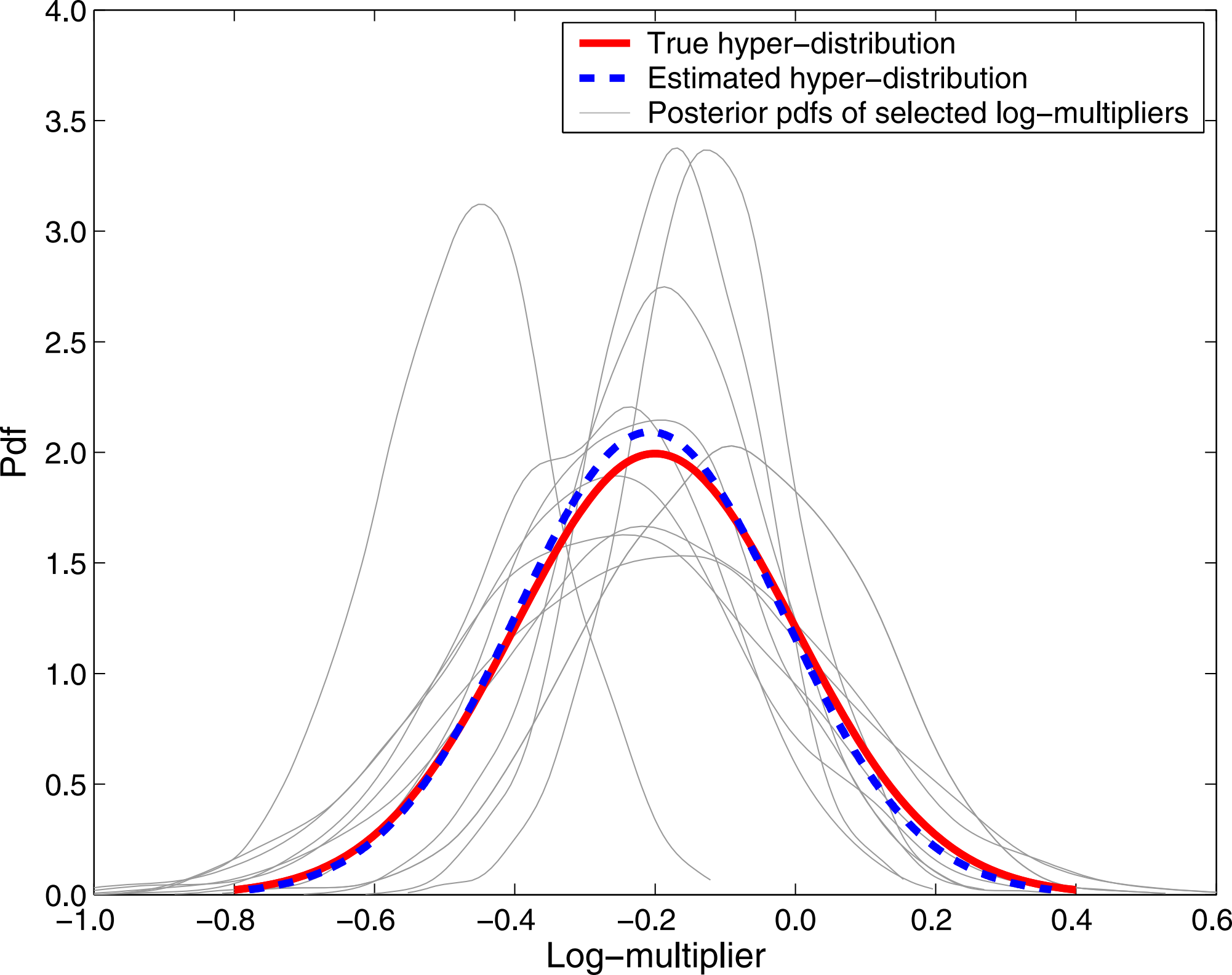

simulated
observed
90% predictive interval
Runoff (mm)
Time (day)
(a)

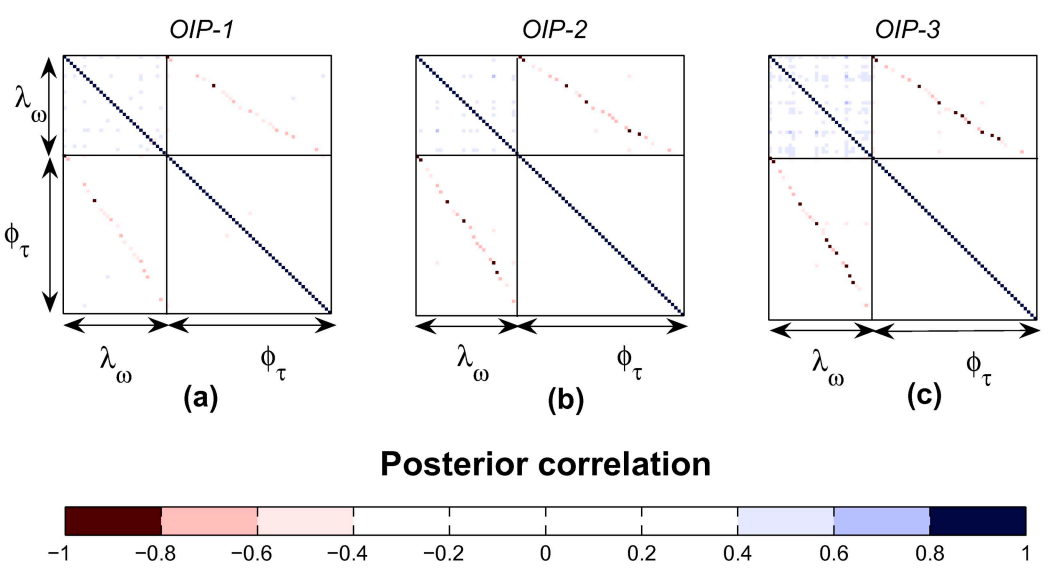

Quantile of observed p−values
Quantile of U[0,1]
(b)

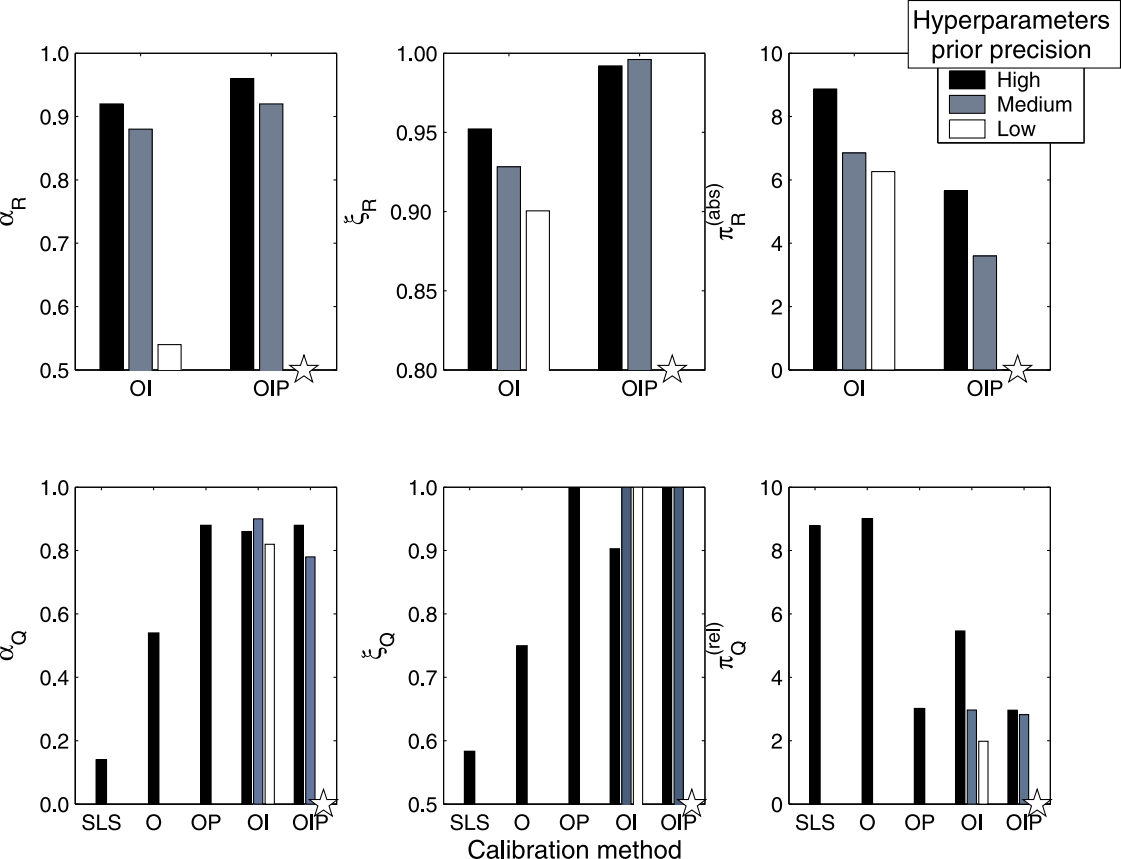

predictive median
observed value
true value
90% predictive interval
Rain (mm)
Time (day)
(c)

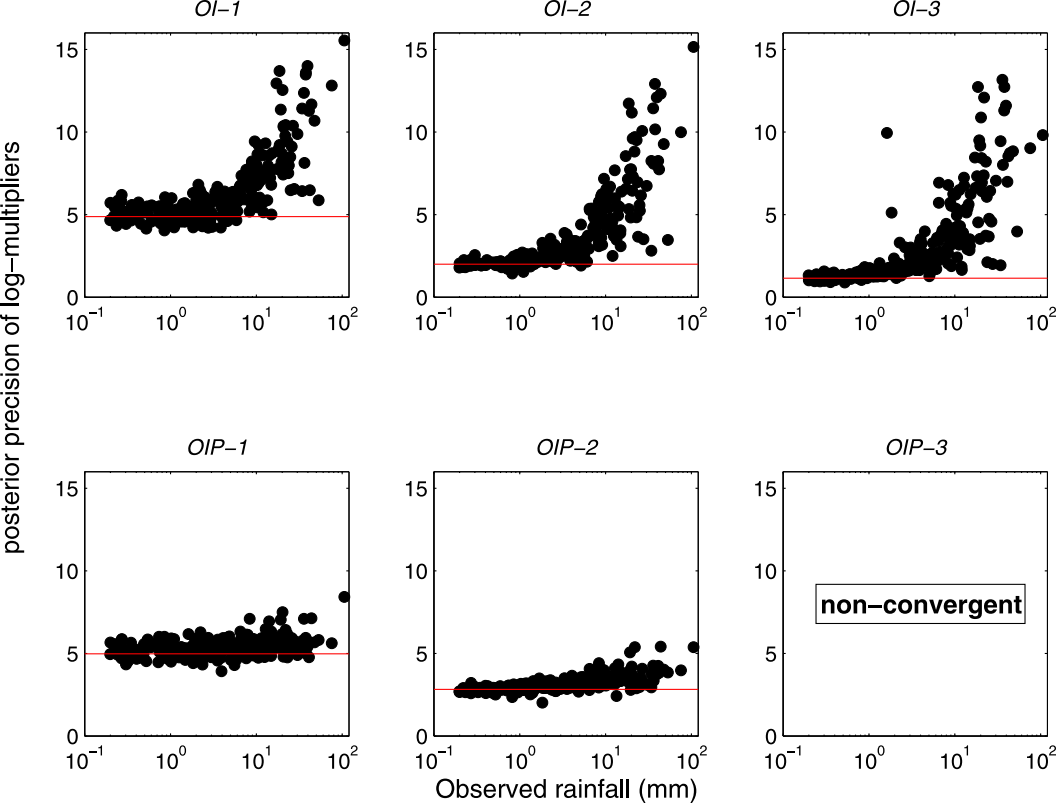

Quantile of observed p−values
Quantile of U[0,1]
(d)

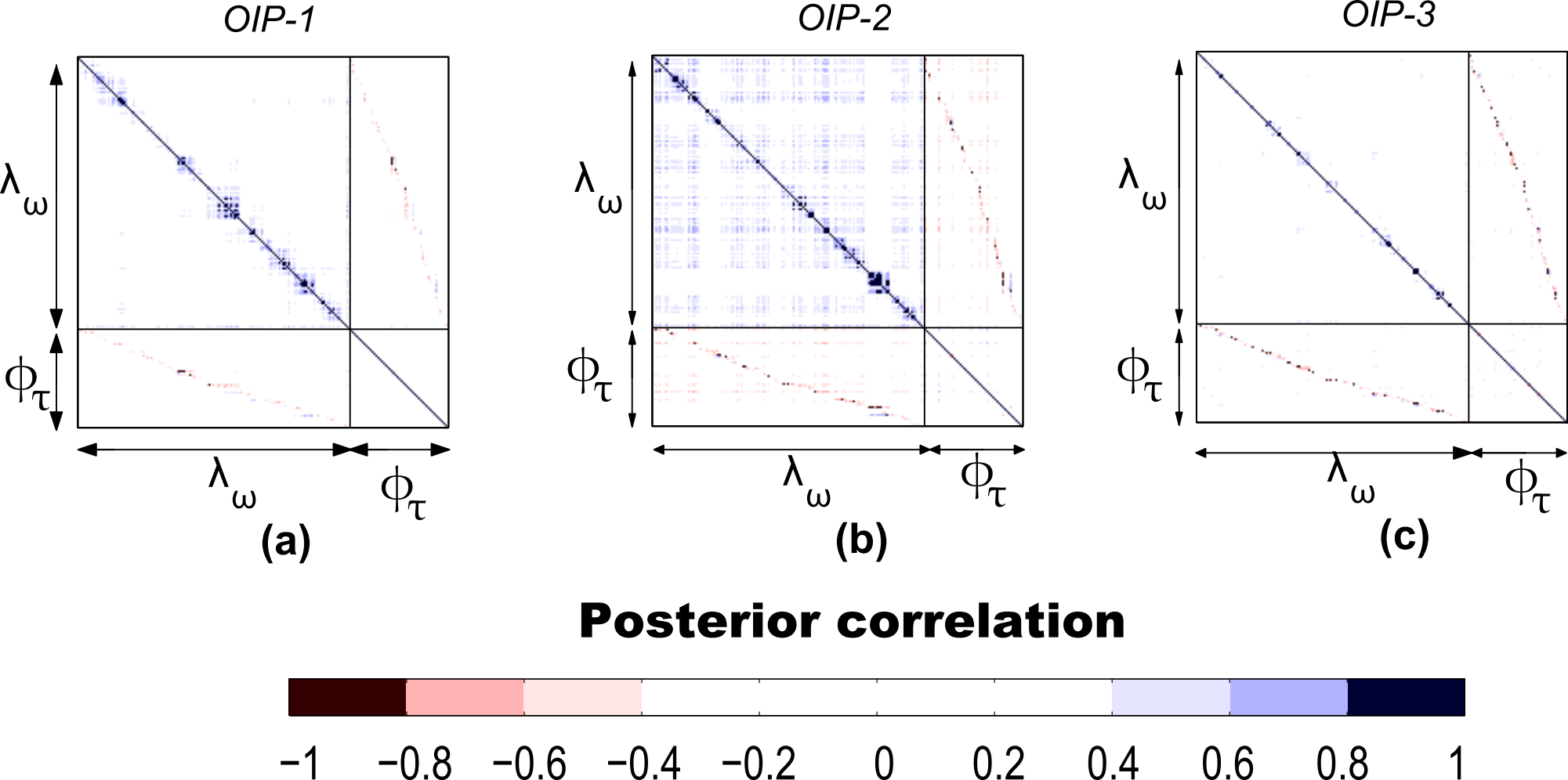

posterior precision of log−multipliers
Observed rainfall (mm)
10
9
8
7
6
5
4
3
10^{-1}
10^{0}
10^{1}
10^{2}

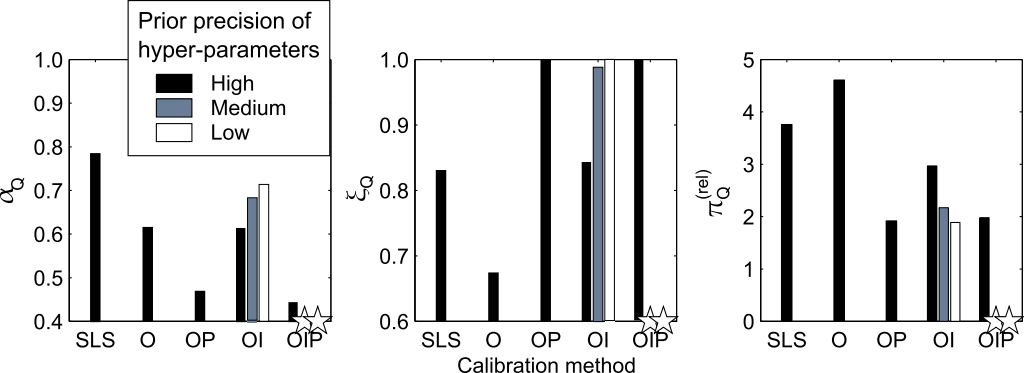

True hyper−distribution
Estimated hyper−distribution
Posterior pdfs of selected log−multipliers
Pdf
Log−multiplier
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6

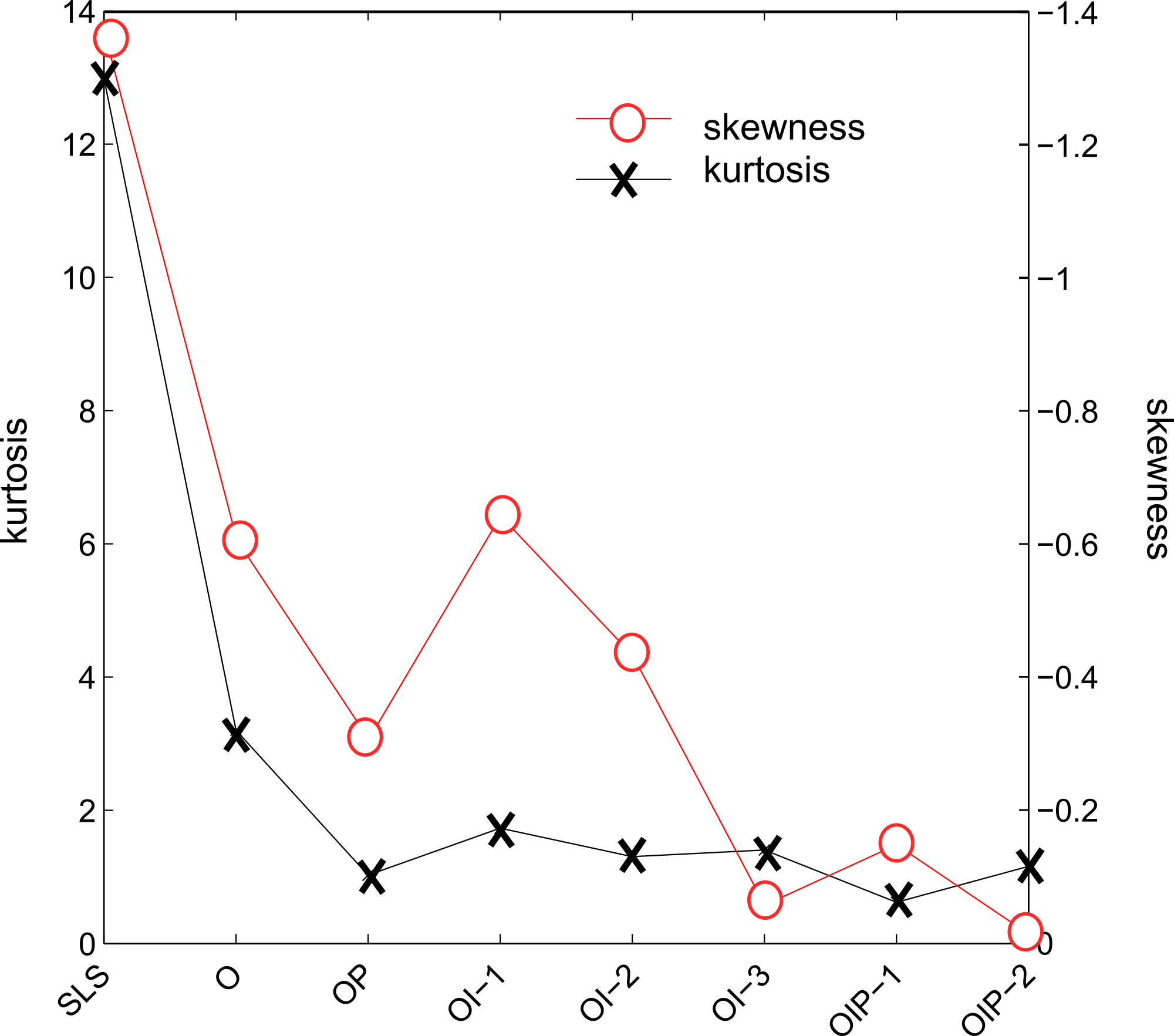

OIP-1
OIP-2
OIP-3
λω
ϕτ
λω
ϕτ
(a)
λω
ϕτ
(b)
λω
ϕτ
(c)
Posterior correlation
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1

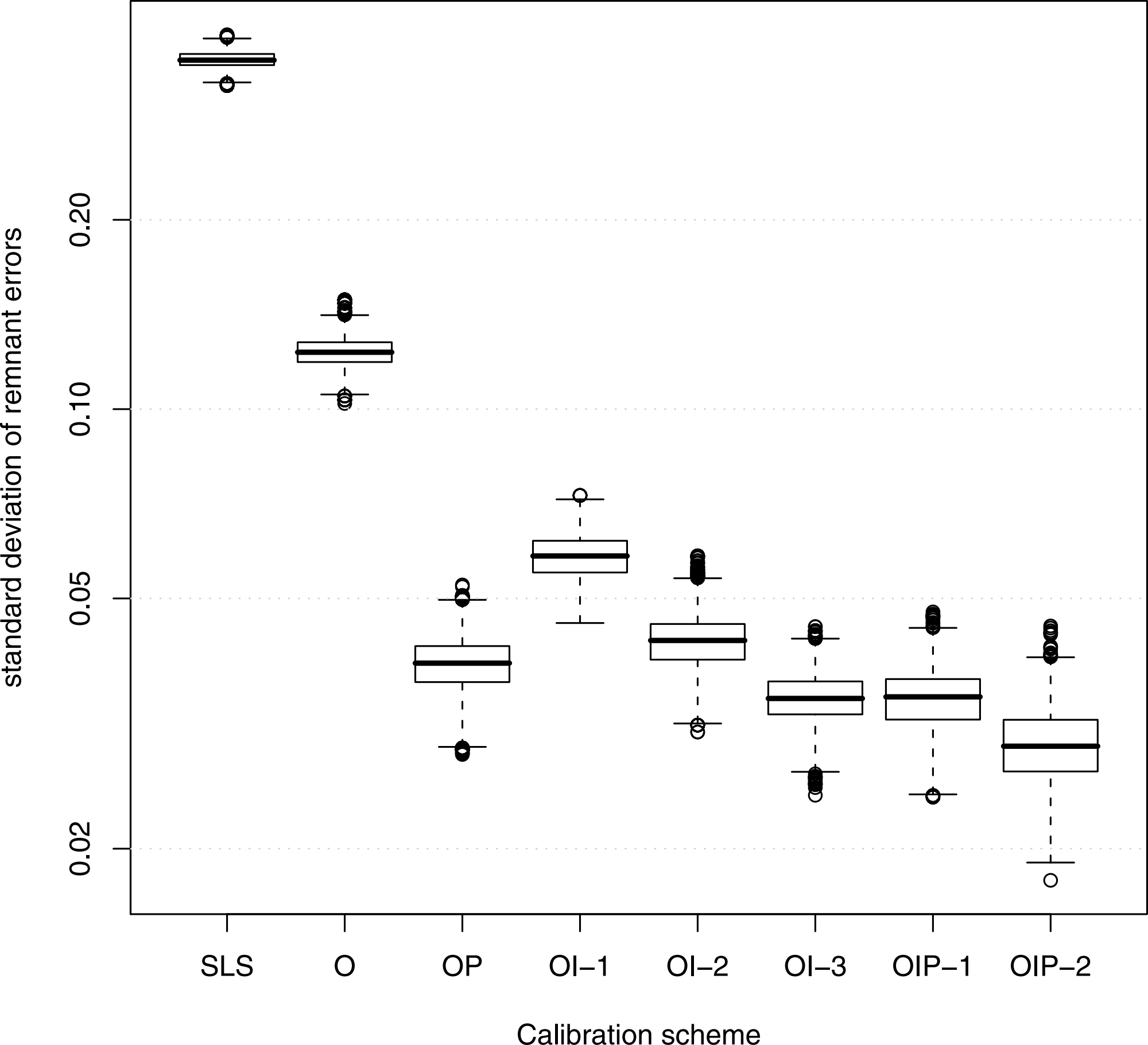

Hyperparameters prior precision
High
Medium
Low
α_R
ξ_R
$\pi_R^{(abs)}$
α_Q
ξ_Q
$\pi_Q^{(rel)}$
OI
OIP
SLS
O
OP
Calibration method


OI−1
OI−2
OI−3
OIP−1
OIP−2
OIP−3
non−convergent
posterior precision of log−multipliers
Observed rainfall (mm)


OIP-1
OIP-2
OIP-3
λω
ϕτ
(a)
(b)
(c)
Posterior correlation
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1


Prior precision of hyper-parameters
High
Medium
Low
α_Q
ξ_Q
$\pi_Q^{(rel)}$
1.0
0.9
0.8
0.7
0.6
0.5
0.4
5
4
3
2
1
0
SLS
O
OP
OI
OIP
Calibration method


skewness
kurtosis
kurtosis
skewness
14
12
10
8
6
4
2
0
−1.4
−1.2
−1
−0.8
−0.6
−0.4
−0.2
0
SLS
O
OP
OI−1
OI−2
OI−3
OIP−1
OIP−2


standard deviation of remnant errors
0.02
0.05
0.10
0.20
SLS
O
OP
OI–1
OI–2
OI–3
OIP–1
OIP–2
Calibration scheme