Calibration of hydrological model GR2M using Bayesian uncertainty analysis

David Huard[1] and Alain Mailhot[1]

Received 5 February 2007; revised 1 November 2007; accepted 9 November 2007; published 19 February 2008.

[1] An outstanding issue of hydrological modeling is the adequate treatment of uncertainties in model calibration and prediction. The current paradigm is that the major sources of uncertainties, namely input, output and model uncertainty should be accounted for directly, instead of assuming they can be safely lumped into the output uncertainties. In this paper, Bayesian analysis is used to calibrate the conceptual hydrologic monthly model GR2M taking into account input, output, structural and initial state uncertainty through error models and priors. Calibration is performed under different error assumptions to study the influence of the initial state uncertainty, the consequences of large input errors, the impact of error assumptions on calibrated parameter posterior distributions and the definition of error models. It is shown how such an analysis can be used to separate, *a posteriori*, the different sources of errors, and in particular, to identify structural errors from data errors.

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

1. Introduction

[2] Uncertainty analysis currently enjoys a considerable amount of attention in hydrology [*Beven*, 2006a]. The recent literature hosts spirited debates about the value of uncertainty analysis methods [*Todini and Mantovan*, 2007; *Beven et al.*, 2007; *Mantovan and Todini*, 2006], discussions about the necessity of validating such methods [*Hall et al.*, 2007] and also about the very definition of the word uncertainty [*Montanari*, 2007]. With this renewal of interest in uncertainty analysis, it is legitimate to raise the question of how uncertainty estimates may undermine the confidence of stakeholders in science when, for example, bounds are large. Not presenting the uncertainty estimates is certainly not worthy of a scientific approach and, as pointed out by *Beven* [2006b], may be, as well, a sure way of undermining the hydrological sciences. Once one recognizes the necessity of uncertainty analysis, one is forced to admit that its application requires more rigor and consistency [*Hall et al.*, 2007]. The efforts needed for uncertainty assessment are certainly important from an operational point of view, when one has to present results to stakeholder, but is also crucial when one wants to assess overall model performance.

[3] The requirement for rigor and consistency can be achieved through Bayesian analysis. Within a Bayesian uncertainty assessment framework, a ''generalized'' model is put forward that grafts error models accounting for the various sources of uncertainties to the hydrological model itself. As the hydrological model encompasses our best understanding of hydrological processes occurring on a given watershed, error models should integrate and substantiate our knowledge of uncertainties affecting input and output data as well as those affecting the modeling process (structural uncertainties). The outcome of the calibration procedure, namely the posterior distribution of model parameters and, consequently, the predictive uncertainty [*Todini and Mantovan*, 2007] is conditioned by the error models and priors. Hence the overall ''performance''or ''reliability'' of uncertainty assessment (which determine acceptability of model results [*Beven*, 2006b]) hinges upon our ability to define as precisely as possible adequate error models.

[4] In a previous paper [*Huard and Mailhot*, 2006], a Bayesian framework inspired by the work of *Kavetski et al.* [2003] was used to calibrate ''abc'', a simple linear hydrological model [*Fiering*, 1967]. The linearity of the model and normality constraints on the error models allowed the derivation of an analytical solution to the parameter posterior distribution, allowing a better understanding of issues related to input errors. Linearity and normality constraints, however, made the resolution method unfit for common usage.

[5] In this paper, the framework is extended by taking explicitly into account structural and initial state uncertainties. The parameter posterior distribution is evaluated using Markov Chain Monte Carlo (MCMC) sampling [*Kavetski et al.*, 2006a, 2006b; *Kuczera and Parent*, 1998], thus removing linearity and normality constraints. A procedure is presented where error models and priors describing input, output, initial conditions and structural uncertainties are grafted to the hydrological model which is then calibrated. This approach enables the examination of issues related to: 1) the influence of initial state uncertainty (Under what conditions is it critical?), 2) overconditioning due to input error assumptions (What does calibration mean in this context?), 3) lumping of structural and output errors (When can it be done? What are the advantages?), 4) definition of error models (What are their impact on the posterior parameter distribution?) and 5) separation of model errors

[1]Institut National de la Recherche Scientifique, Centre Eau, Terre & Environnement, Québec (Qc), Canada.

0043-1397/08/2007WR005949$09.00

from data errors (How is it done? How can it be used?). The first four issues are examined through simple examples using deliberately crude error model assumptions and the fifth one by a case study where more care is given to the specification of realistic error models.

[6] Although this Bayesian uncertainty assessment framework is general and could be theoretically applied to any model, the numerical cost of evaluating the parameter posterior distribution is significant and grows with the length of the data series used for calibration. Hence given the available computational resources, its application to daily or hourly models appears unwieldy. The method employed in the paper should hence be considered as a formal exercise, rather than an operational solution to the calibration of hydrological models (for such methods, see *Moradkhani et al.* [2005]; *Vrugt et al.* [2005]; *Kavetski et al.* [2003]; *Kennedy and O'Hagan* [2001] and *Beven and Binley* [1992]). In this spirit, the model chosen to illustrate the issues presented above is the parsimonious monthly model GR2M [*Mouelhi et al.*, 2006]. GR2M's nonlinear components make it sensitive to input errors [*Paturel et al.*, 1995], and hence relevant in an uncertainty analysis context. With only two free parameters, the posterior parameter distribution can be easily visualized and interpreted.

[7] The model and data used for simulations are described in more detail in section 3, after the basic theoretical background and hypotheses underlying the proposed Bayesian uncertainty framework are explained in section 2. Section 4 discusses basic calibration and validation issues related to multiple sources of errors using simple error models. Section 5 presents a more realistic case study with carefully defined error models, along with a brief analysis of the results, including the *a posteriori* separation of the different types of errors. A brief discussion about the extension to daily model is included followed by the conclusion (section 6).

[8] This paper relies heavily on the theory and methods of Bayesian analysis. The authors suggest the textbooks of *Gelman et al.* [1995] for an application focused point of view and *Jaynes and Bretthorst* [2003] for more theoretical discussions on the foundations of probability theory.

2. Bayesian Uncertainty Framework

[9] This section introduces the proposed Bayesian uncertainty framework and derives equations to perform the calibration and validation of models in presence of multiple sources of uncertainties.

2.1. Basic Concept

[10] In a Bayesian calibration, the objective is to find the posterior distribution for the parameters $\boldsymbol{\theta}$ knowing the input data series $\tilde{\boldsymbol{x}}$ and output data series $\tilde{\boldsymbol{y}}$ as well as the model $\mathcal{M}$:

$$\mathrm{p}(\boldsymbol{\theta}|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}},\mathcal{M}). \tag{1}$$

[11] Solving equation (1) may be very simple or extremely complex, depending on the modeler's assumptions about the sources of uncertainty that affect the calibration process. In hydrological modeling, many different sources of uncertainties must be tackled simultaneously. Indeed, according to *Beven* [2006b] "... we have unknown errors in input and boundary conditions that get processed non-linearly through a model that has structural errors and which is then compared with observations that have unknown measurement and commensurability error characteristics." Although accounting for all these sources of errors realistically is a conceptual and technical challenge, not doing so can have undesirable consequences on modeling performance and the reliability of predictions [*Oudin et al.*, 2006; *Andréassian et al.*, 2004; *Kavetski et al.*, 2003; *Andréassian et al.*, 2001; *Nandakumar and Mein*, 1997; *Paturel et al.*, 1995; *Michaud and Sorooshian*, 1994; *Xu and Vandewiele*, 1994; *Troutman*, 1982]. In the following, a Bayesian approach is used to account for four different sources of uncertainty: the ignorance of the initial conditions, input data error, output data error and structural (model) error.

[12] To describe the effect of those four types of uncertainties on calibration and validation, conceptual true variables describing the initial condition ϕ_0, the true input series $\boldsymbol{x}$ and the true output series $\boldsymbol{y}$ are introduced in equation (1) as latent variables (idealized true variables ($\boldsymbol{x}$, $\boldsymbol{y}$) are what would be observed in the absence of data errors). Latent variables (or parameters) are variables useful to state a problem but whose effect is integrated out of the posterior of interest [*Jaynes and Bretthorst*, 2003]. Equation (1) then becomes

$$\mathrm{p}(\boldsymbol{\theta}|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}},\mathcal{M}) = \iiint \mathrm{p}(\boldsymbol{\theta},\boldsymbol{x},\boldsymbol{y},\phi_0|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}},\mathcal{M})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\phi_0, \tag{2}$$

where it is assumed that the integrals span the admissible domain of each variable. Using these conceptual true values, hypotheses concerning the nature of errors can be translated probabilistically and equation (2) brought under a form where it can be solved computationally. Schematically, the inference process is displayed in Figure 1: input and output data are linked to their respective true values by the input and output error models, and those true input and output variables are linked together by a structural error model, itself dependent on the model parameters and initial conditions.

2.2. Hypotheses, Priors and Error Models

[13] The first step is to use Bayes' theorem on equation (2) so that data is conditioned on the true variables:

$$\mathrm{p}(\boldsymbol{\theta}|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}},\mathcal{M}) = \iiint \mathrm{p}(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}|\boldsymbol{\theta},\boldsymbol{x},\boldsymbol{y},\phi_0,\mathcal{M}) \cdot \mathrm{p}(\boldsymbol{\theta},\boldsymbol{x},\boldsymbol{y},\phi_0|\mathcal{M})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\phi_0 \cdot \frac{1}{\mathrm{p}(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})}. \tag{3}$$

[14] Since the normalization constant $\mathrm{p}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ introduced by Bayes' theorem is of little relevance to the calibration process, it will be neglected in the following, with the equality relation replaced by proportionality sign.

[15] For the next step, two hypotheses are made: 1. data errors cannot be inferred from the initial conditions, the model parameters or the model, and 2. input and output errors are conditionally independent given the true input and true output [*Kavetski et al.*, 2003]. The first hypothesis is significant, in that it defines data error models as entities independent from the model simulations. Note however that data errors are only meaningful with respect to the true variables, themselves dependent on the model's spatial and

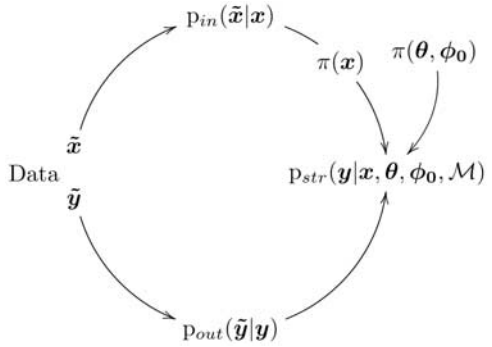

Figure 1. Inference process based on the theoretical uncertainty framework. Experimental input and output data are indicated by *tilded* variables $\tilde{x}$, $\tilde{y}$, and conceptual true values by non *tilded* variables: true input series x; true output series y; model parameters θ and initial model state ϕ_0. The input, output and structural error models are indicated by p_{in}, p_{out} and p_{str} respectively, $\pi(\theta, \phi_0)$ denotes the prior for the model parameters and the initial state, and $\pi(x)$ stands for the prior for the true input series.

temporal scales. Input and output error models are hence defined only on the basis of the experimental protocol, apparatus leading to data acquisition and the relation between the observed variables and the true variables. Formally,

$$p(\tilde{x}, \tilde{y}|\theta, x, y, \phi_0, \mathcal{M}) = p(\tilde{x}, \tilde{y}|x, y) = p(\tilde{x}|x)\, p(\tilde{y}|y), \quad (4)$$

where $p(\tilde{x}|x)$ is a statistical distribution defining the probability of measuring an input series $\tilde{x}$ knowing the true input series x. Similarly, $p(\tilde{y}|y)$ is a statistical distribution defining the probability of measuring $\tilde{y}$ knowing the true output series y. In the following, these particular distributions will be referred to as the input error model and the output error model, and denoted specifically by p_{in} and p_{out} as reminder of their meaning. The exact shape of these error models is up to the modeler and should capture the main sources of data uncertainty. Defining realistic error models is a difficult task since data errors typically combine both measurement errors and commensurability errors (the difference between the variable the model expects and what is actually measured in the field) in ways that are site specific.

[16] Using assumption (4), equation (3) becomes:

$$p(\theta|\tilde{x}, \tilde{y}, \mathcal{M}) \propto \iiint p_{in}(\tilde{x}|x)\, p_{out}(\tilde{y}|y) \cdot p(y|\theta, x, \phi_0, \mathcal{M})\, p(\theta, x, \phi_0|\mathcal{M})\, dx\, dy\, d\phi_0, \quad (5)$$

where $p(y|\theta, x, \phi_0, \mathcal{M})$ is a statistical distribution describing the probability of a true output series y knowing the simulated output series $\widehat{y} \equiv \mathcal{M}(\theta, x, \phi_0)$. In other words, this distribution describes the difference between model simulation and true output: the model error. It will hence be referred to as the structural error model, and denoted by $p_{str}(y|\widehat{y}) \equiv p(y|\theta, x, \phi_0, \mathcal{M})$ to shorten the notation.

[17] The last step consists in splitting the prior $p(\theta, x, \phi_0|\mathcal{M})$ into two independent priors $\pi_x(x|\mathcal{M})$ $\pi_{\theta,\phi_0}(\theta, \phi_0|\mathcal{M})$, where π is used, here and in the following, to denote priors (the subscript is dropped when there is no ambiguity about which prior is meant). While this separation is not mandatory, defining a joint multivariate prior distribution for the true input series, model parameters and state variables seems to be of a marginal interest given the length of the data series generally used in hydrology [*Mantovan and Todini*, 2006]. Equation (5) can then be written as

$$p(\theta|\tilde{x}, \tilde{y}) \propto \iiint p_{in}(\tilde{x}|x)\, p_{out}(\tilde{y}|y)\, p_{str}(y|\widehat{y}) \cdot \pi(x)\, \pi(\theta, \phi_0)\, dx\, dy\, d\phi_0, \quad (6)$$

where references to $\mathcal{M}$ were removed since the model is always assumed known.

[18] Note that although $\pi(\theta, \phi_0)$ and $\pi(x)$ are both priors, there is a notable difference between the two. Whereas θ, ϕ_0 are inferred from the whole data set, the inference on a given true value x_i is based mostly on $\tilde{x}_i$ as well as $\tilde{y}_{j \geq i}$ through inferred parameters. If the errors are assumed large (vague input and output error models), then there is very little information in the data about x_i and the prior plays a major role in the inferential process [*Gull*, 1989]. The prior for the true input must hence be chosen with care, and modelers should avoid using abusively vague priors [*Kavetski et al.*, 2006a; *Huard and Mailhot*, 2006].

[19] It is worthwhile to underline the conceptual differences between the different types of outputs that are defined in this Bayesian framework: simulations ($\widehat{y}$) are the model outputs computed from inferred true inputs and parameters, output measurements ($\tilde{y}$) are experimentally measured outputs, and true outputs (y) are the idealized true values that would be measured if no measurement and commensurability errors were present. Another useful output that will be discussed later is the simulated output measurement ($\tilde{Y}$), the output measurement predicted by the output error model given the inferred true output. In practice, different applications will require the use of different types of output. In water resource management contexts, decision makers will rather be interested in the inferred true outputs than the simulated outputs. In a model validation context however, field measurements should be compared to simulated output measurements to avoid mixing structural errors with measurement errors. Note that in many modeling studies, simulations are implicitly interpreted as the true values and compared directly with observations to assess model performance, with the ensuing risk of misinterpreting results.

2.3. Resolution Method

[20] In general, equation (6) has no analytical solution. One exception is when the model is linear with respect to the input variables and the error models as well as the true input prior $\pi(x)$ have a Gaussian shape. This particular case was studied by *Huard and Mailhot* [2006] using the "abc" model. However, typical hydrological models are nonlinear, equation (6) cannot be integrated analytically and a different resolution method is needed. Unfortunately, numerical integration methods such as quadratures become inefficient for more than three or four dimensions, due to the curse of dimensionality: the computational cost increases exponentially with the dimension of the problem [*Novak and Ritter*, 1997]. Quadrature methods are thus inappropriate for this kind of application where the integral spans hundreds of dimensions (one for each true input and true output). In

other words, the posterior probability cannot be evaluated directly. This is a frequent problem in Bayesian analysis, and a common solution is to sample parameters from the posterior distribution (6) using Monte Carlo Markov Chain (MCMC) algorithms [*Huang and Liang*, 2006; *Kavetski et al.*, 2006b; *Bates and Campbell*, 2001; *Kuczera and Parent*, 1998; *Gelman et al.*, 1995; *Neal*, 1993]. In this case the Metropolis algorithm is used, by which the parameter space is explored by making random steps. These steps are selected such that the distribution of the sampled parameters approaches the posterior distribution after a sufficient number of random steps (see *Chib and Greenberg* [1995] for an introduction to MCMC theory and more particularly Metropolis sampling). Hence to solve equation (6), one samples $(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y}, \phi_0)^i$ N times from $\mathrm{p}(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y}, \phi_0|\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$, and for N large enough (convergence can be assessed by different criteria [*Gelman et al.*, 1995]), the empirical distribution of the samples reproduces the probability distribution (superscripts ^{i} indicate sampled values). The integration over the latent variables is then performed by marginalizing over $\boldsymbol{x}$, $\boldsymbol{y}$ and ϕ_0, that is, by considering the sampled $\boldsymbol{\theta}^i$ independently from the latent variables.

2.4. Prediction and Validation

[21] *Hall et al.* [2007] suggest that "without validation, calibration is worthless and so is uncertainty estimation.[. . .] Validation is also needed [. . .] because the data do not contain full information about how the catchment will respond in the future. The same argument applies to uncertainty estimation." In other words, validation checks not only how the calibrated model performs outside the calibration period, but also if the error model assumptions hold.

[22] In a split-sample calibration/validation, a data series is split in a calibration series $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ and a validation series $(\tilde{\boldsymbol{x}}_+, \tilde{\boldsymbol{y}}_+)$. Calibration is performed under a set of assumptions defining the error models and priors, under which the posterior distribution $\mathrm{p}(\boldsymbol{\theta}|\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ for the parameters is computed. In validation, parameters can be seen as a source of uncertainty and thus treated as latent variables. Validation uses the entire posterior parameter distribution along with input $\tilde{\boldsymbol{x}}_+$ to infer the probable distribution of output measurements $\tilde{\boldsymbol{Y}}_+$, which can compared with the real observations $\tilde{\boldsymbol{y}}_+$ to validate the model and the calibration assumptions:

$$\begin{aligned}\mathrm{p}(\tilde{\boldsymbol{Y}}_+|\tilde{\boldsymbol{x}}_+,\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}) &= \iiint \\ &\quad \mathrm{p}(\tilde{\boldsymbol{Y}}_+,\boldsymbol{\theta},\phi,\boldsymbol{x}_+,\boldsymbol{y}_+|\tilde{\boldsymbol{x}}_+,\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})\ \mathrm{d}\boldsymbol{x}_+\,\mathrm{d}\boldsymbol{y}_+\,\mathrm{d}\phi\,\mathrm{d}\boldsymbol{\theta} \\ &= \iiint \mathrm{p}_{out}(\tilde{\boldsymbol{Y}}_+|\boldsymbol{y}_+)\,\mathrm{p}_{str}(\boldsymbol{y}_+|\hat{\boldsymbol{y}}_+) \\ &\quad \cdot \mathrm{p}_{in}(\tilde{\boldsymbol{x}}_+|\boldsymbol{x}_+)\,\pi(\boldsymbol{x}_+)\mathrm{p}(\boldsymbol{\theta},\phi|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}})\ \mathrm{d}\boldsymbol{x}_+\,\mathrm{d}\boldsymbol{y}_+\,\mathrm{d}\phi\,\mathrm{d}\boldsymbol{\theta}. \quad (7)\end{aligned}$$

[23] The simulated output measurements $\tilde{\boldsymbol{Y}}_+^i$ can then be compared with the observed outputs $\tilde{\boldsymbol{y}}_+$ to assess the global efficiency of the model, that is, the combined accuracy of the model, error models and priors.

3. Model and Data

3.1. GR2M

[24] Simulations in this paper use the parsimonious hydrological lumped monthly model GR2M [*Mouelhi et al.*, 2006], an improved version of the model developed by *Edijatno and Michel* [1989] and *Kabouya* [1990]. Despite having only two free parameters, the model has been shown to perform well when compared to similar monthly models; on a benchmark test consisting of 410 basins throughout the world, GR2M shows the best performance among nine models, some of them counting five free parameters [*Mouelhi et al.*, 2006].

[25] The two free parameters of GR2M are θ_1, the soil moisture storage maximum capacity, and θ_2, the water exchange term with neighboring catchments. The internal state variables consist of a soil moisture accounting store (S) and a quadratic reservoir (R). The model is forced by monthly rainfall ($\boldsymbol{r}$) and potential evapotranspiration ($\boldsymbol{e}$) and returns a monthly flow $\hat{\boldsymbol{y}}$. Readers should note that θ_1, the soil moisture store capacity, controls the model's response to rain event, and to a certain extent, the variability of the simulated flow. As θ_1 increases, the simulated flow depends less on the current rainfall and more on the store level, itself dependent on past rainfall. For small θ_1, more rainfall is directed as excess rainfall and instantaneously routed as output flow. In other words, θ_1 controls the model's low-pass filter behavior (its "memory" of past events). In the following, the GR2M model is denoted by $\hat{\boldsymbol{y}} = \mathcal{M}(\boldsymbol{\theta}, \boldsymbol{x}, \phi)$, with model parameters $\boldsymbol{\theta} = \{\theta_1, \theta_2\}$, input series $\boldsymbol{x} = \{\boldsymbol{r}, \boldsymbol{e}\}$ and internal state variables $\phi = \{R, S\}$.

[26] Sensitivity analyzes have determined that GR2M is sensitive to white noise errors on rainfall, but comparatively robust to random errors on potential evapotranspiration (PE) [*Paturel et al.*, 1995]. That is, the Nash-Sutcliffe efficiency (NSE) decreases rapidly as the magnitude of random errors over rainfall increases, but much more slowly in the case of random errors over PE. Although the study of *Paturel et al.* [1995] dates back to a previous version of GR2M, a similar analysis was carried out and the same conclusions hold for the current version.

3.2. MOPEX Data Sets

[27] The data used in this study is taken from the MOdel Parameter Estimation eXperiment (MOPEX) [*Schaake et al.*, 2006]. The database contains daily streamflow, rainfall, potential evapotranspiration, minimum and maximum temperatures. The rainfall data is produced by *Maurer et al.* [2002] and is the result of gridding point rainfall estimates from the United States, Canada and Mexico. The rainfall point estimates are daily totals taken from the National Oceanic and Atmospheric Administration (NOAA) Cooperative Observer (Co-op) stations. They are gridded using the synergraphic mapping system (SYMAP) [*Shepard*, 1984] as implemented by *Widmann and Bretherton* [2000]. The gridded data is then scaled to correct for complex topography in sparsly instrumented areas using long-term monthly averages from the Parameter-Elevation Regressions on Independent Slopes Model (PRISM) [*Daly et al.*, 1994, 1997]. The data is not corrected, however, for systematic gauge errors [*Maurer et al.*, 2002], which can significantly decrease the measured amount of rainfall.

[28] Generally speaking, the major sources of systematic errors in rainfall measurements are wind-induced undercatch, wetting losses and evaporation losses. The undercatch due to wind has been estimated to vary from 2% to 20% for liquid precipitations to over 100% for solid precipitation, wetting losses from 3% up to 10% and evaporation losses are generally below 2% [*Metcalfe et al.*, 1997; *Yang et al.*, 1999]. These estimates are highly

Table 1. Properties of the Eight Watersheds Used in the Study[a]

Code	Basin Name	Area, km^2	Mean Monthly Rainfall, mm	Mean Monthly PE, mm	Mean Monthly Flow, mm
05471500	South Skunk River near Oskaloosa, IA	4235	69.3	82.3	17.8
05570000	Spoon River at Seville, IL	4237	75.3	84.3	20.9
01562000	Raystown Branch Juniata River at Saxton, PA	1958	78.8	65.4	35.2
03011020	Allegheny River at Salamanca NY	4165	86.8	59.4	48.2
03303000	Blue River near White Cloud, IN	736	93.8	73.4	68.2
01321000	Sacandaga River near Hope NY	1272	102.9	57.6	64.7
02475500	Chunky River near Chunky, MS	956	117.2	87.9	38.5
03443000	French Broad River at Blantyre, NC	767	161.2	70.1	99.1

[a]All stations span between 53 and 54 years of uninterrupted data with no null rainfall at the monthly scale. Rainfall, PE and flow data are areal values.

dependent on the type of gauge, the local climatology, altitude, topography and the apparatus calibration.

[29] Another source of error is the ''representativeness'' of the data (commensurability). Lumped conceptual hydrological model assume the input variable is related to the areal rainfall over the entire catchment, whereas point rainfall estimates are often used [*Habib et al.*, 2001]. In the MOPEX database, the aerial adjustment is achieved by the spatial interpolation and catchment integration. However, it is difficult to discuss even roughly the relation between the MOPEX estimated rainfall and the ''true'' aerial rainfall, due to the complex manipulations separating the raw point data from the reanalysis results. Unless uncertainty estimates are provided, this is a setback to the use of reanalysis data for applications where uncertainty assessment is crucial.

[30] As for evapotranspiration, the MOPEX values are derived from the NOAA Evaporation Atlas [*Farnsworth et al.*, 1982]. More specifically, they are computed by fitting a Fourier series with only an annual cycle to evaporation pan monthly averages. These averages were obtained by gridding the Evaporation Atlas maps with a resolution of 1/6°. The parameters of the series were then computed on this grid and averaged over each basin to estimate the basin average annual evapotranspiration cycle.

[31] Since interannual random white noise PE errors do not seem to perturb hydrological models [*Andréassian et al.*, 2004; *Paturel et al.*, 1995], our concern is mostly with systematic or cyclic intraannual errors. One such cyclic source of error could be soil moisture availability. Indeed, PE strongly depends on the amount of moisture in the topsoil layer, a dependence that is not taken into account by the MOPEX estimates. If the soil moisture has a strong cyclic component (high in spring, low in September for instance), the resulting PE may show an intraannual variability not captured by pan evaporation estimates, with potentially detrimental influence on model simulation.

[32] Finally, MOPEX streamflow series are provided by the United States Geological Survey (USGS). Daily streamflows are taken at gauges unaffected by upstream regulation and where long time series are available. Stage measurements are converted into streamflow estimates through rating curves. Stage measurement errors are of the order of 3 mm while single discharge measurement errors typically range around 5–10%, depending on the protocol (number of vertical velocity profiles, depth of measurements) and apparatus (velocimeter, Doppler profiler) [*Hirsch and Costa*, 2004]. Moreover, rating curves are not static and may vary due to changes in bed roughness, accumulation of debris following a flood, vegetation growth, presence of ice, streambed scour and fill processes, bank erosion, etc [*Fenton and Keller*, 2001]. Some of these factors are accounted for by frequent updates to rating curves.

[33] The simulations presented in section 4 are run for all eight stations presented in Table 1 (although only results from the Chunky River watershed are shown). The basins chosen cover relatively dry to humid conditions found in the continental U.S.A. In the paper, units are systematically millimeters (mm) for rainfall (aerial) and monthly discharge (aerial).

4. Step by Step Bayesian Model Calibration

[34] Realistic calibration of a hydrological model in presence of data and model uncertainties using Bayesian analysis poses numerous conceptual and technical issues [*Kavetski et al.*, 2006b]. Among those, defining sensible error models and priors, integrating latent variables, validating the calibrated model and leveraging realism versus complexity are the most challenging. In a realistic calibration, many different issues are mingled and it is difficult to grasp how they interact and consequently, how to interpret the results. The objective of the current section is to isolate calibration issues in specific stand-alone examples, helpful in understanding the effects of calibration assumptions.

[35] The first issue discussed is the initial state uncertainty and its relevance in different calibration contexts. Then the first case is presented, an application of Bayesian analysis where only output errors are considered, highlighting the conceptual difference between simulations and measurements. The second case shows a calibration where errors are assumed to be only on input data, and discusses how the input error assumption impacts the very definition of model performance. The third case introduces structural errors and discusses their lumping with output errors. From these three cases follows a brief analysis of the effect of error models assumptions on the posterior parameters distribution.

[36] It should be noted that although the figures only show the case of the Chunky river watershed, identical simulations were run for all eight basins presented in table 1 with comparable results.

4.1. Initial State Uncertainty

[37] In GR2M, the internal state of the system is defined by the level of two reservoirs. Since they are conceptual reservoirs, their level cannot be measured experimentally and there is, *a priori*, no way to define exactly their initial values. A convenient solution to this problem is to warm-up the model, i.e., to set the internal state to arbitrary values

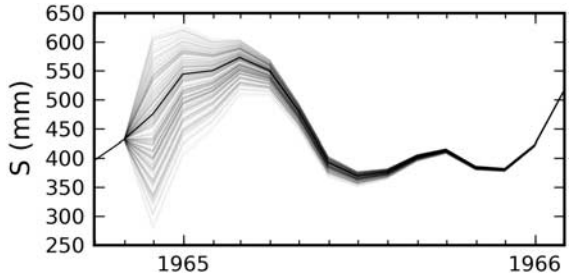

Figure 2. Once perturbed (thin lines), the internal state variable S describing the level (mm) of the soil moisture store returns to its unperturbed state (thick line). The other internal variable R (level of the routing store) displays a similar behavior. The figure shows results obtained with a large store capacity (θ_1 = 1000), i.e., a long memory.

and let the model run for a fixed period [*Shelton*, 1985]. The rationale behind this strategy is that the effect of the initial state ϕ_0 on the simulated flow decreases rapidly. Indeed, at time t the internal state ϕ_t depends on ϕ_{t-1} and on x_t, the current input. By induction this is equivalent to saying it depends on $\phi_0, x_1, x_2, \ldots, x_t$. Hence if t is large enough, the effect of ϕ_0 becomes negligible compared with the effect of the input series $\boldsymbol{x}$. The question then is, how long should this warm-up period be to minimize the effect of the initial state uncertainty without sacrificing too much data? For some catchments, rainfall data was logged long before flow measurements began, so the period where only rainfall data is available can be used as a warm-up period without concern about data expenditure. However, when this is not the case and only short rainfall and flow series are available, it may be worthwhile to minimize the warm-up period.

[38] An estimate of a sensible warm-up period can be found by artificially perturbing the internal state and measuring how fast it converges back to its unperturbed state (Figure 2). For GR2M, the rate of convergence depends closely on θ_1, the soil moisture store capacity; for large store capacity, the dependence of the model to its past state (its ''memory'') is stronger and internal state perturbations have a prolonged effect ($\sim$12 months). For small store capacities, the model has a short memory and perturbations vanish after as few as three months. Based on simulations on a number of catchments from the MOPEX database and different parameter sets, it was found that perturbations vanish with an average half-life of around 1.5 months. Warming-up the model during one year [*Paturel et al.*, 1995] hence leaves ample time to the internal state to reach its ''steady'' state. In the simulations shown in cases 1, 2 and 3, a warm-up period of two years was used since data availability is not an issue.

[39] Using Bayesian analysis, the warm-up period can be avoided entirely. The solution is to define a prior for the initial state $\pi(\phi_0)$ and consider ϕ_0 a latent variable, as in equation (6). The effect of ϕ_0 on simulations can then be integrated out and parameters calibrated without sacrificing the first months to warm-up the model. Simulations (not presented here for brevity) suggest that the advantages of such an approach over the warm-up strategy depend on the relative information content [*Wagener et al.*, 2003] of the warm-up period. This information content (loosely defined as the diversity of hydrological conditions) is itself dependent on the length of the time series, the presence of extreme events and the prior for the initial conditions. For example, if a long calibration series is available, the relative information content of the first months is low compared to the rest of the series and using those first months in the calibration has only a weak effect on the parameter distribution. However, if those first months contain a very large flood, the like of which is not observed during the rest of the calibration series, then including them in the calibration series may have a significant effect on the posterior parameter distribution. Finally, if a vague prior for the initial state is defined, then the uncertainty over flow simulations for the first few months is dominated by the initial state uncertainty rather than the parameter uncertainty, and only a weak inference can be conducted on the model's parameters. Therefore integrating the initial state uncertainty is advantageous only if a reasonably informative prior about the initial state can be defined.

4.2. Case 1: Assuming Output Errors Only

[40] In this first case, equation (6) is applied under the assumption that inputs are exact, or more pragmatically, that the effect of input errors can be safely lumped into output errors. This case is similar to a classical calibration, except that while a standard calibration aims at finding a unique set of parameters optimizing an objective function, a Bayesian analysis provides a parameter posterior probability distribution. Another difference is that objective functions can take any shape [*Gupta et al.*, 1998], whereas output error models are limited to formal statistical distributions (positive, piecewise continuous functions whose integral equals one). In this example, the calibration assumptions are given formally by:

$$\mathrm{p}_{in}(\tilde{\boldsymbol{x}}|\boldsymbol{x}) = \delta(\tilde{\boldsymbol{x}} - \boldsymbol{x}) \tag{8a}$$

$$\mathrm{p}_{str}(\boldsymbol{y}|\widehat{\boldsymbol{y}}) = \delta(\boldsymbol{y} - \widehat{\boldsymbol{y}}) \tag{8b}$$

$$\mathrm{p}_{out}(\tilde{\boldsymbol{y}}|\boldsymbol{y}) = \mathbf{L}(\tilde{\boldsymbol{y}}|\boldsymbol{\mu} = \ln \boldsymbol{y}, \sigma_{out}) \tag{8c}$$

$$\pi(\boldsymbol{\theta}, \phi_0) = \mathbf{U}(\boldsymbol{\theta}_1|0, 1000) \cdot \mathbf{U}(\boldsymbol{\theta}_2|0, 2) \cdot \boldsymbol{\delta}\left(\phi_0 - \widehat{\phi}_0\right) \tag{8d}$$

where δ stands for the Dirac delta, $\mathbf{L}$ for the lognormal probability density function (pdf, see equation (A3)) and $\mathbf{U}$ for the uniform pdf. The hypothesis motivating the choice of a lognormal distribution to model output errors is that streamflow errors are roughly proportional to the value they affect: errors are multiplicative rather than additive [*Yapo et al.*, 1996] (see Appendix A3 for details). This is a reasonable assumption if one considers that stage errors are transformed into discharge errors by a concave stage-discharge relationship (later explained in section 5.3). Note that the prior for the true input $\pi(\boldsymbol{x})$ is not defined here since it is just a constant (assumption (8a) implies that $\boldsymbol{x} = \tilde{\boldsymbol{x}}$). The prior for the model parameters is given by a uniform prior over intervals covering a broad range of watershed behavior. Initial conditions ϕ_0 are fixed to $\widehat{\phi}_0$ by warming-up the model for a period of 24 months.

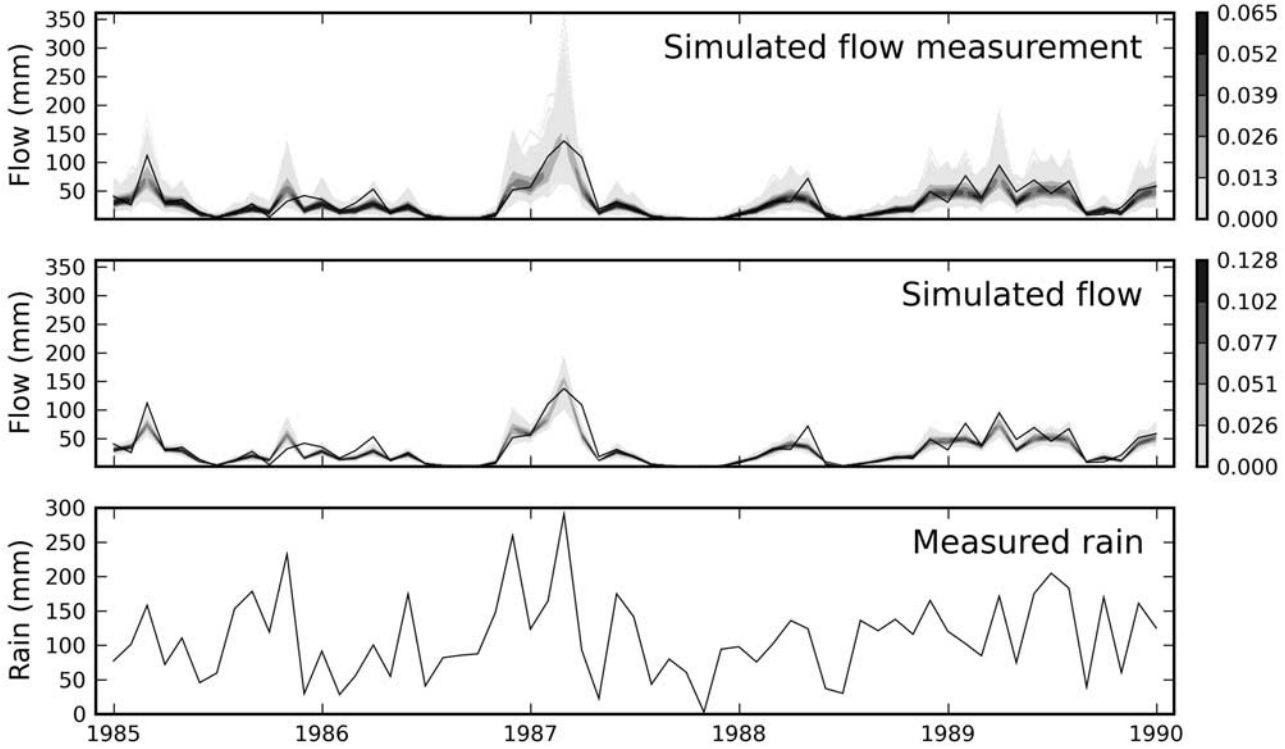

Figure 3. Case 1. Result of a calibration of the Chunky River watershed performed assuming that only output errors are present. The lower graph shows rainfall observations. The middle graph shows the pdf (contour plot) of the flow simulated using the parameters sampled during calibration compared with flow observations (full line). The upper graph shows the pdf of simulated flow measurements, including both parameter uncertainty and an explicit output observation error. Although calibration is performed over 20 years of data, only a subset of five years is shown here to avoid cluttering the figure. The scales on the right indicate the value of the pdf.

[41] Plugging the error model assumptions (8a)–(8d) into equation (6) and using the properties of the Dirac delta function to solve the three integrals yields a posterior distribution with a simple expression:

$$p(\boldsymbol{\theta}|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}) \propto p_{out}(\tilde{\boldsymbol{y}}|\widehat{\boldsymbol{y}})\pi(\boldsymbol{\theta}), \tag{9}$$

where $\widehat{\boldsymbol{y}} = \mathcal{M}(\boldsymbol{\theta}, \tilde{\boldsymbol{x}}, \widehat{\phi}_0)$ and $\pi(\boldsymbol{\theta}) = \mathbf{U}(\theta_1|0, 1000) \cdot \mathbf{U}(\theta_2|0, 2)$. With all integrals solved, equation (9) can be computed directly and there is no explicit need to use MCMC sampling. Nevertheless for the sake of the demonstration, 100000 values of $\boldsymbol{\theta}$ are drawn from the posterior distribution (9), half of which is discarded to remove transient samples. The shape parameter of the output error model is set to $\boldsymbol{\sigma}_{out} = .2$ (corresponding roughly to a 20% error on flow), and the calibration data $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ consists of 20 years of monthly rainfall, PE and flow from the Chunky River watershed (see Table 1).

[42] Figure 3 shows the results on a five years period. The lower plot shows the series of rainfall measurements from the MOPEX database. The middle plot shows the contoured probability density of the simulated flow along with the flow measurements from the MOPEX (full line). That is, the simulated flow $\widehat{\boldsymbol{y}}^i$ is computed for each $\boldsymbol{\theta}^i$ sampled, and a contour plot is drawn from the simulated flow histogram. Note that the variability of the simulated flow does not capture at all the variability seen in the observed flow. At first sight, it would appear that the parameter uncertainty is insufficient to account for the flow uncertainty. Although specifying a vaguer output error model (setting σ_{out} to some higher value) would increase the parameter uncertainty, it is not the real issue here and is not advisable either [*Refsgaard et al.*, 2006]. Indeed, the difference in variability is mainly due to the fact that the plot compares two conceptually different variables: the measured flow $\tilde{\boldsymbol{y}}$ and the true flow $\boldsymbol{y}$ (equal in this case to the simulated flow $\widehat{\boldsymbol{y}}$). To be meaningful, the plot must compare measured flow and simulated flow measurements. These simulated flow measurements $(\tilde{\boldsymbol{Y}})$ are computed by drawing random variates from the output error model $\tilde{\boldsymbol{Y}}^i \sim p_{out}(\boldsymbol{Y}|\widehat{\boldsymbol{y}}^i)$ with $\widehat{\boldsymbol{y}}^i$ computed for each $\boldsymbol{\theta}^i$ sampled from the posterior (9). The upper graph of Figure 3 shows a contour plot of the probability density of simulated flow measurements compared with flow measurements (full line). It can be seen that the simulated measured flow captures better the variability of the observed flow series.

[43] This conceptual difference between model simulation and measurements has more profound implications than capturing the simulation variability. Indeed, depending on the output error model chosen, it may lead to apparent paradoxes. For instance, if the output error model assumes that flow measurements systematically underestimate the true value, the flow simulated from the calibrated parameters will systematically overestimate the flow measurements. Therefore a measure of calibration efficiency from a direct comparison of flow measure-

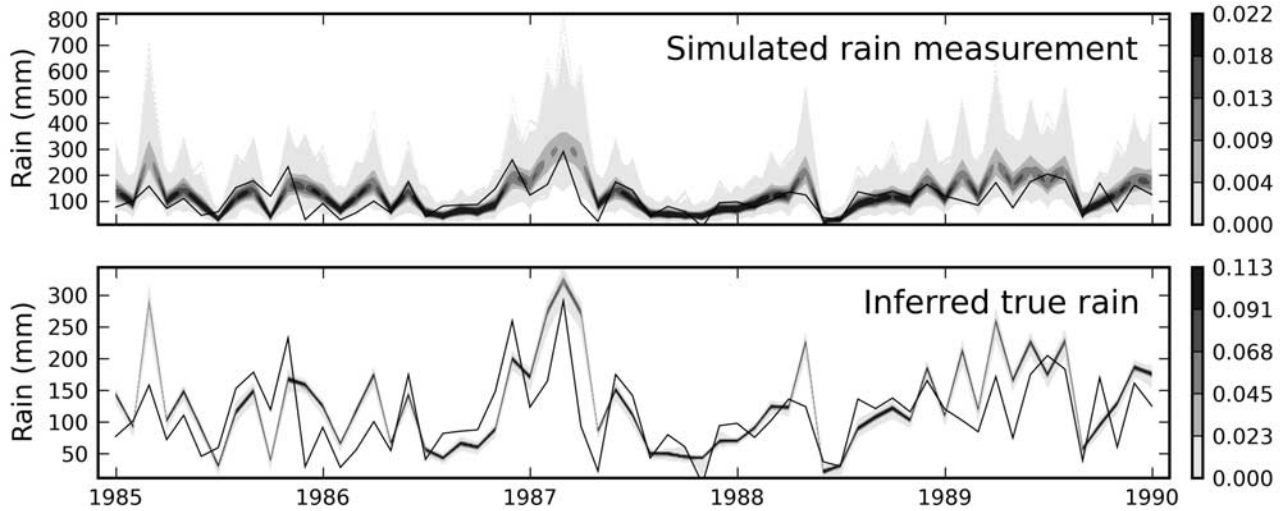

Figure 4. Case 2. Results of calibration of the Chunky River watershed under the assumption that only input errors are present. The lower graph shows a five years subset of the measured rainfall series (full line) used to calibrate the model, and a pdf (contour plot) of the sampled true rain. The upper graph shows the pdf of simulated rain measurements, including an explicit rain observation error. The scales on the right indicate the value of the pdf.

ments and flow simulations would be meaningless and misleading. This conceptual difference between simulations and measurements has already been noted in Kalman filtering theory [*Burgers et al.*, 1998] and in the Bayesian recursive parameter calibration context [*Thiemann et al.*, 2001].

4.3. Case 2: Assuming Input Errors

[44] The second case deals with the opposite situation, namely that outputs are assumed exact and precipitations uncertain. Although this is a rather peculiar situation, it illustrates some potential benefits and setbacks associated with taking input errors into account. Formally, the underlying hypotheses are given by:

$$\mathrm{p}_{in}(\tilde{\boldsymbol{x}}|\boldsymbol{x}) = \mathbf{L}(\tilde{\boldsymbol{r}}|\boldsymbol{\mu} = \ln \mathbf{r}, \sigma_{in} = .2)\, \delta(\tilde{\boldsymbol{e}} - \boldsymbol{e}) \tag{10a}$$

$$\pi(\boldsymbol{x}) = \mathbf{W}(\boldsymbol{r}|\alpha, \kappa, \mu, \sigma) \tag{10b}$$

$$\mathrm{p}_{out}(\tilde{\boldsymbol{y}}|\boldsymbol{y}) = \delta(\tilde{\boldsymbol{y}} - \boldsymbol{y}), \tag{10c}$$

where **W** stands for the exponentiated Weibull pdf (see equation (A5)). The parameters of the exponentiated Weibull distribution (α, κ, μ, σ) are estimated by fitting an historical rainfall series of 20 years, prior to the calibration period, from the same catchment the model is calibrated on. The exponentiated Weibull was chosen because it adequately captures the total monthly rainfall distribution over many catchments compared to other common semi-infinite distributions (see Appendix A4). The structural error model and the prior for the parameters and internal state are identical to those of case 1 (equations (8b) and (8d)).

[45] To solve this case using MCMC sampling, the Dirac delta of the output error model is approximated by a lognormal distribution with a small shape parameter, that is, equation (8c) with $\sigma_{out} \approx 0$. The integral over $\boldsymbol{x}$ must be performed in this case by Metropolis sampling of the following posterior

$$\mathrm{p}(\boldsymbol{\theta}, \boldsymbol{x}|\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \propto \mathrm{p}_{out}(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{y}})\, \mathrm{p}_{in}(\tilde{\boldsymbol{x}}|\boldsymbol{x})\, \pi(\boldsymbol{x})\pi(\boldsymbol{\theta}). \tag{11}$$

[46] The sampler now explores a space formed by the two model parameters $\boldsymbol{\theta}$ and the $20 \cdot 12 = 240$ true input values $\boldsymbol{x}$. However, as the shape parameter σ_{out} approaches zero, it becomes increasingly difficult to explore the $\boldsymbol{x}$-space because the values with a non-zero probability lie in a vanishingly small 240-dimensional volume, making random jumps very unlikely to be accepted by the Metropolis algorithm. With $\sigma_{out} = .02$, $3.0 \cdot 10^7$ samples had to be generated to obtain an approximately stable parameter distribution. That is, a change in the burn-in period has no effect on the mean of the distribution, but still has a small influence on its shape, even after ten million iterations. Ideally, after a given number of iterations, the shape of the distribution should show no dependence to the burn-in length.

[47] The calibration results are presented in Figure 4. The lower graph shows the probability density function (pdf) of the inferred true rainfall, while the upper graph shows the pdf of the simulated measured rain: true rain corrupted by random measurement errors drawn from the input error model. The simulated flow is not shown since it is, by definition, identical to the observed flow.

[48] There are two observations that deserve to be highlighted. First, by assuming the measured flow is exact, the true rain $\boldsymbol{x}$ is constrained to values that reproduce exactly the measured output. In other words, equation (11) infers the value of the true rain, based on the knowledge of the measured flow. Although this may seem artificial, some interesting observations can be derived from such an analysis. Indeed, in a case not shown here, a clear annual trend in the ratio true rain/measured rain was noted; during the summer months, the true rain systematically under-

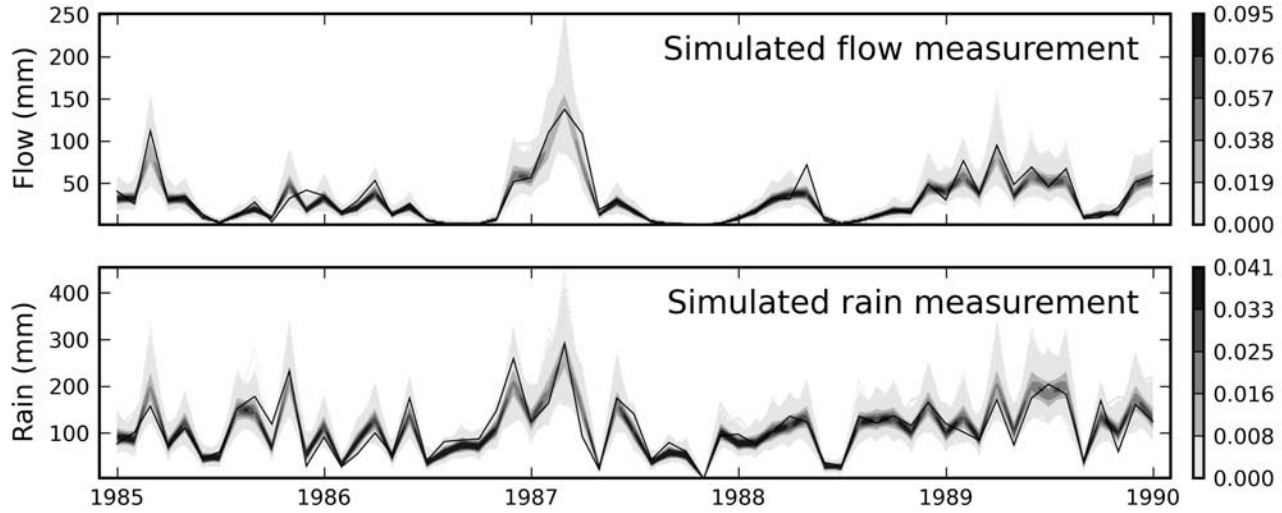

Figure 5. Case 3. Results from a calibration of the Chunky River watershed accounting for input, output and structural errors. The lower plot shows the pdf of simulated rain measurements (contour plot), compared with rainfall observations (full line), and the upper plot the pdf simulated flow measurement probability density (contour plot) compared with the observed flow series (full line). The scales on the right indicates the value of the pdf.

estimated the measured rain, suggesting that there was too much water input. Also, indeed, increasing the summer PE in those cases improved the overall model efficiency.

[49] The second observation is related to the assessment of model efficiency, often measured in hydrology by the Nash-Sutcliffe Efficiency criterion (NSE) (although *Schaefli and Gupta* [2007] point out that NSE does not provide a reliable basis for comparing results of different case studies). In this particular case, there is no difference between simulated and measured flow, and the NSE reaches, by design, its maximum value, 100%. Obviously, any efficiency measure based on the concordance of measured and simulated output will yield exceedingly good, yet meaningless results. In this particular case, the efficiency of the model should rather be measured by the concordance between measured inputs and simulated inputs (or rather simulated *input measurements*) since this really is the quantity "predicted" by the model.

4.4. Case 3: Lumping Output and Structural Errors

[50] The last few years have seen an increasing number of strategies designed to account for structural errors in calibration or prediction [*Refsgaard et al.*, 2006], taking advantage of Bayesian model averaging [*Ajami et al.*, 2007], ensembles simulations [*Vrugt et al.*, 2005] or Gaussian processes [*Kennedy and O'Hagan*, 2001]. These methods capture the uncertainty related to the model either by running a number of different models (model averaging and ensemble simulations), or by describing the model error by a mixture of functions (Gaussian processes). The proposed framework is not a replacement for those methods, but rather a way to include formally a probabilistic description of structural errors into the calibration and prediction process.

[51] Indeed, the common practice in hydrological modeling is to define an objective function which relates simulated values to observed ones, and assumes that uncertainties on input data are negligible. This objective function, even if it is not clearly stated, lumps together output and structural errors. This embedding of the structural error model into a "simple" objective function can lead to inconsistent results (e.g., residuals do not comply with the assumed output distribution). In this example, a structural error model is formally lumped with an output error model, yielding a response error model to be incorporated to the calibration process. Input errors are considered as well and the parameters posterior distribution is estimated.

[52] Given output and structural error models, a response error model p_{res} lumping both output and structural errors can be defined simply as

$$p_{res}(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{y}}) \equiv \int p_{out}(\tilde{\boldsymbol{y}}|\boldsymbol{y})\, p_{str}(\boldsymbol{y}|\hat{\boldsymbol{y}})\, d\boldsymbol{y}. \tag{12}$$

[53] In some cases, this integral can be solved analytically, greatly reducing the numerical cost of sampling over the true output variables. Indeed, replacing the integration over $\boldsymbol{y}$ in the right hand side of equation (6) by equality (12) yields

$$p(\boldsymbol{\theta}|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}) \propto \int p_{in}(\tilde{\boldsymbol{x}}|\boldsymbol{x})\, p_{res}(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{y}})\, \pi(\boldsymbol{x})\, d\boldsymbol{x} \cdot \pi(\boldsymbol{\theta}), \tag{13}$$

which can be solved by sampling only the parameters $\boldsymbol{\theta}$ and true input values $\boldsymbol{x}$ from

$$p(\boldsymbol{\theta}, x|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}) \propto p_{in}(\tilde{\boldsymbol{x}}|\mathbf{x})\, p_{res}(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{y}})\, \pi(\mathbf{x})\pi(\boldsymbol{\theta}) \tag{14}$$

and marginalizing over $\boldsymbol{x}$.

[54] As an example, this third case assumes that both p_{out} and p_{str} are expressed in terms of the lognormal distributions:

$$p_{out}(\tilde{\boldsymbol{y}}|\boldsymbol{y}) = \mathbf{L}(\tilde{\boldsymbol{y}}|\ln \boldsymbol{y}, \sigma_{out}) \tag{15a}$$

$$p_{str}(\boldsymbol{y}|\hat{\boldsymbol{y}}) = \mathbf{L}(\boldsymbol{y}|\ln \hat{\boldsymbol{y}}, \sigma_{str}). \tag{15b}$$

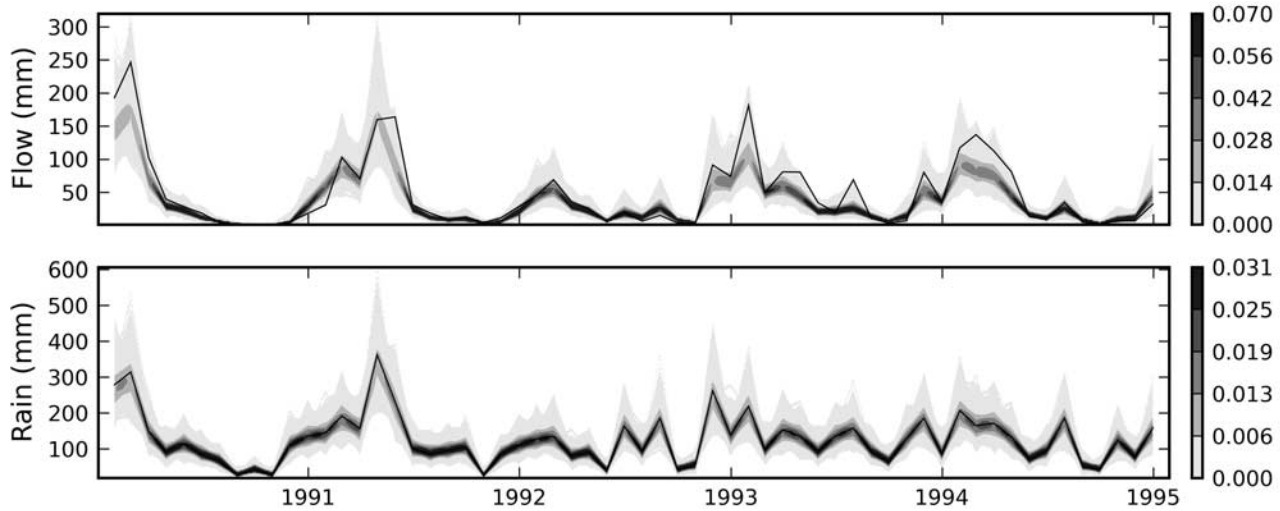

Figure 6. Validation of the model calibrated in case 3. The lower graph shows the pdf of the simulated measured rain $\tilde{\boldsymbol{X}}_+$ (contour plot) along with the observed rainfall (full line). The upper graph shows the pdf of the simulated measured flow $\tilde{\boldsymbol{Y}}_+$ (contour plot) compared with the observed flow series (full line) for the validation period. The scales on the right indicate the value of the pdf.

[55] The lognormal distribution is chosen both for its simplicity and because it is heteroscedastic, a typical feature for hydrologic model residuals [*Xu*, 2001; *Yapo et al.*, 1996]. Inserting (15a), (15b) into (12) and performing the integration yields

$$\mathrm{p}_{res}(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{y}}) = \mathbf{L}\left(\tilde{\boldsymbol{y}}|\ln\hat{\boldsymbol{y}}, \sqrt{\sigma_{str}^2 + \sigma_{out}^2}\right), \tag{16}$$

a lognormal distribution lumping both output and structural errors. Again, the integration of the output and structural error models is by no means mandatory, and is done here out of optimization considerations. Given ample computing resources, the analytical integrability of the output and structural error models is not an issue.

[56] Using the response error model (16) with $\sigma_{out} = .05$ and $\sigma_{str} = .1$, the input error model (10a) with $\sigma_{in} = .1$ and priors (10b), (8d), $1.0 \cdot 10^6$ samples $(\boldsymbol{\theta}, \boldsymbol{x})$ are drawn from equation (14) using MCMC sampling and the first half is discarded. The lower graph of Figure 5 shows the simulated rain measurement probability density and the upper graph the simulated flow measurement probability density. Although this example has no pretense to realism, the overall calibration results appear satisfying, both for the inferred rain and flow. One thing deserving improvement though is the uncertainty over high flows, apparently overestimated by the response error model.

[57] This example shows that from a conceptual point of view, taking structural uncertainty into account is relatively simple: define a structural error model and integrate, analytically or numerically, over the true output series $\boldsymbol{y}$. However, as Bayesians know too well, specifying a prior or an error model is not difficult, but specifying a good one is. The situation is further complicated since the proposed definition of structural errors $(\boldsymbol{y} - \hat{\boldsymbol{y}})$ combines two types of structural errors: model inadequacies and the inherent stochasticity of the process (due to explaining variables not taken into account). Theoretically, with a long "error free" data set, it would possible to identify both types of structural errors, rectify the model to remove the inadequacies and describe probabilistically the remaining randomness of the process. Without such a perfect data set however, the expertise of hydrologists seems the best way to define sensible structural error models.

[58] To validate the model and error models, the procedure detailed in section 2.4 is applied and the results shown in Figure 6. Although the error models are rather crude, the resulting uncertainty for the flow (contour plot) appears consistent with the flow observations (full line).

4.5. Impact of Error Models on Posterior Parameter Distribution

[59] Values of calibrated parameters and their associated uncertainties are conditioned by the choice of the errors models. To illustrate this, the parameters calibrated from the three cases discussed above (with identical calibration data sets) are shown in Figure 7. The sensitivity of parameters on error model assumptions can be explained by noting that

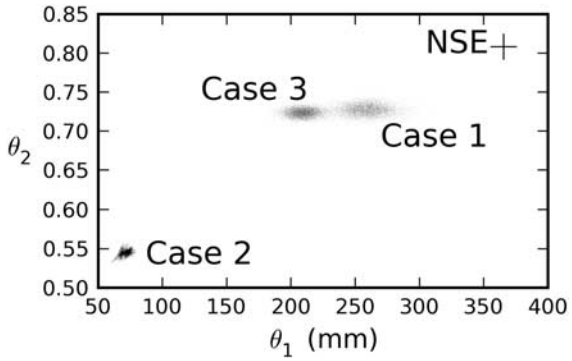

Figure 7. Parameter distributions sampled in the three cases presented. Parameter θ_1 describes the capacity (mm) of the soil accounting store and θ_2 the water exchange coefficient. As θ_2 decreases, more water is lost to neighboring catchments. The NSE parameter is computed by optimizing the NSE criteria computed over the square root transformed flow.

parameters are estimated through the inferred true input and output values, which depend directly on the error models. In other words, the parameter distribution is inseparable from the choice of error models, and calibration results can only be understood by looking at the entire set of assumptions.

[60] A feature of Figure 7 deserving to be highlighted is that lumping all sources of errors on the output (Case 1) yields a larger value for θ_1 (the moisture store capacity) than when input errors are accounted for (Case 2 and 3). By increasing this parameter, the model reduces the potential for excess rainfall and, at the same time, the flow variability. In other words, calibration selects parameters that enhance the low-pass filter behavior of the model and reduce its sensitivity to input error.

[61] Finally, readers may wonder about the significant difference between the parameters estimated with NSE and those obtained from case 1, since both methods assume that only the outputs are corrupted. The authors explanation for this difference is that it is largely due to the difference in heteroscedasticity assumed by the two methods. Generally speaking, a heteroscedastic error model tolerates large errors on large flows and estimates parameters that simulate well the low flows. On the other side, a homoscedastic error model considers all errors equal and estimates parameters that somehow capture the average behavior. For GR2M, this average behavior is obtained by a higher exchange coefficient θ_2 (less water is lost to neighboring catchments, increasing the base flow) and a higher soil moisture capacity θ_1, reducing the model's sensitivity to rainfall. In Figure 7, the NSE parameters are estimated by maximizing the NSE criterion computed on the flows square root [*Xu and Vandewiele*, 1994], which is akin to a heteroscedasticity assumption [*Sorooshian and Dracup*, 1980]. However, this square root heteroscedasticity is weaker than the lognormal heteroscedasticity assumed by the output error model (Case 1), explaining why the NSE parameters are shifted to the upper right. If the parameters are estimated again by an NSE criterion computed this time directly on the flows (homoscedasticity assumption), one obtains an even greater shift with $\hat{\boldsymbol{\theta}} = (453, 0.88)$. This artificial favoritism of a low-pass filter behavior may have serious impacts if the model is used in a predictive mode as it can lead to underestimation of peak flows, as well as the associated uncertainties on these estimates.

[62] One may suggest that shifts in parameters could possibly be traced back to distinct model failures. For instance, the large difference in the soil moisture capacity could be explained by the lack of an interception process in the model [*Savenije*, 2004] or an incorrect direct runoff parameterization. Although the compensation mechanisms of parameters for un- or mis-accounted sources of errors are interesting in their own right, it is not clear at this point whether such shifts can provide more than cursory evidence to identify structural inadequacies. Moreover, for hydrological models counting dozens of parameters, such an analysis would lose much of its intuitiveness.

5. Case Study

[63] The objective of this section is to demonstrate the use of the Bayesian uncertainty framework under more realistic calibration assumptions. For each source of error, hypotheses are formulated in probabilistic terms and error models constructed on the basis of these hypotheses. Although the GR2M model may be too rudimentary for this application to be really realistic, it shows what can be done in terms of error models, and serves as a test case before tackling more complex daily models. In the following sections, error models and priors are defined and the simulation setup is presented. Results are then discussed, along with the necessary modifications for an extension to daily models.

5.1. Rainfall Error Model

[64] The rainfall error model is inspired by *Weerts and El Serafy* [2006] who describe hourly rainfall errors by a normal distribution with a variance $\sigma^2_{\delta r}$ that has both a proportional component (15%) and fixed one (.2 mm):

$$\sigma_{\delta r} = 0.15r + 0.2\,\text{mm},$$

where r stands for rainfall (mm). The fixed component is due to the finite resolution of rain gauges (for example the volume of the tipping bucket). For this application, the fixed component is set at .254 mm (.01 in), the default volume for tipping bucket gauges [*Habib et al.*, 2001]. It is assumed that this error affects hourly measurements, so that once scaled to monthly values, the standard variation due to the finite resolution of the gauge is estimated by $\sqrt{30 \cdot 24} \cdot .254$ mm $\approx$ 7 mm. The proportional component roughly represents commensurability (aerial representativity) errors and measurements errors such as evaporation and wetting losses as well as wind-induced undercatch. Once aggregated to monthly values, the proportional component reduces to about 7% (see Appendix B). Since these measurement errors cause an underestimation of the true rainfall, the rainfall input error model is assumed to be also biased by around 7% [*Metcalfe et al.*, 1997]; that is, it is assumed that rainfall measurements are on average 93% of the true rainfall.

[65] One should note that the smallest error standard deviation (7 mm) is of the order of the smallest monthly precipitation. Hence there is a normalization issue due to the fact that for small rainfalls, the normal distribution assigns non zero probability to negative rainfall. To deal with this problem, a truncated normal distribution $\mathbf{N}_T$ (see equation (A1)) is used instead of a standard normal distribution. Formally, the input error model is defined as:

$$\mathrm{p}_{in}(\tilde{\boldsymbol{r}}|\boldsymbol{r}) = \mathbf{N}_T(\tilde{\boldsymbol{r}}|.93\boldsymbol{r}, .07\boldsymbol{r} + 7\ \text{mm}, 0, \infty). \tag{17}$$

5.2. PE Error Model

[66] The error model for PE is based primarily on the assumption that white noise error on PE have only a weak influence on model performance [*Paturel et al.*, 1995]. It is also based on the authors observation that PE computations performed with the McGuinness-Bordne model [*McGuinness and Bordne*, 1972] yields values similar within a multiplicative factor to MOPEX PE series (results not shown here for the sake of brevity). In other words, the relevant uncertainty seems to be related mostly to intrannual rather than interannual variability. Therefore it is assumed here that the true PE ($\boldsymbol{e}$) is obtained by multiplying the MOPEX PE series ($\tilde{\boldsymbol{e}}$) by a unique multiplier (ϵ): $\boldsymbol{e} = \epsilon\tilde{\boldsymbol{e}}$, thus rescaling the PE series. This hypothesis enormously

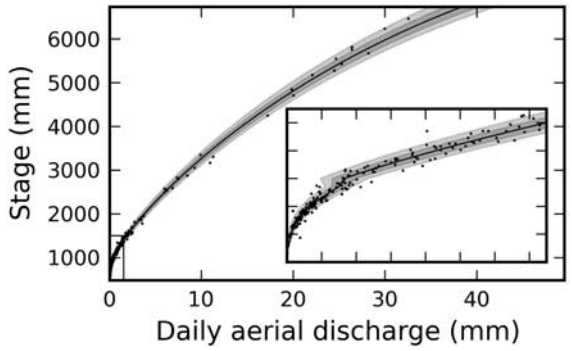

Figure 8. Daily output (discharge) error model for the Chunky River watershed. Dots indicate stage-discharge measurements taken by the USGS since 1938. The black line is a piecewise polynomial fit of the stage-discharge relation. The light gray area indicates the span of the first standard deviation (68% confidence) of the discharge error model and the darker gray the second deviation (95% confidence) for daily measurements. The rectangle in the lower-left corner shows the boundaries of the close-up view at the right.

simplifies computations as it replaces an integration over N true PE values by an integration over only one multiplier ϵ. Since the multiplier along with measured PE completely determine the conceptual true PE, there is no need for a genuine PE error model. However, a prior for the multiplier still has to be defined:

$$\pi(\epsilon) = \mathbf{L}(\epsilon|0, .5). \tag{18}$$

[67] The lognormal is chosen because it is the most general hypothesis in terms of entropy (see Appendix A3) that can be made for a positive variable of fixed mean and variance. The hyperparameters are set such that the multiplicative factor has a median of 1 and a standard variation of about .5.

5.3. Output Error Model

[68] The output error model describes the probability of an error on the flow knowing the conceptual true flow. Flows $\tilde{y}$ are commonly obtained from stage measurements $\tilde{h}$ and a rating curve (stage-discharge) relationship $\tilde{y} = f(\tilde{h})$, defined from previous simultaneous stage and discharge measurements. The analysis of discharge errors has received substantial attention [*Shiklomanov et al.*, 2006; *Schmidt*, 2002; *Pelletier*, 1988; *Herschy*, 1985; *Dymond and Christian*, 1982] and multiple methods have been proposed to define the nature and structure of discharge errors. In the following, a simplified approach is chosen in order to reduce the complexity of the output error model.

[69] Discharge errors originate from many different sources: stage measurement errors, single discharge measurements, rating curve sampling errors (a limited amount of data is available to fit the curve) and the effect of explaining variables other than stage (river slope, bed roughness, presence of ice, vegetation growth, etc). Although it is possible to describe different source of errors independently [*Dymond and Christian*, 1982], it requires a considerable amount of information about single discharge measurements (number and type of velocity vertical profiles, apparatus, etc) and may be too complex in the case of aggregated monthly discharges. To reduce the complexity of the error model, it is assumed that all sources of errors can be lumped into a single term: a stage measurement error δ_h. Stage errors can then be converted into discharge errors, $\delta_y \approx f'\delta_h$, where $f(h)$ is the function relating stage h to discharge y. Visually, this error structure seems to capture the variability of stage-discharge measurements (see Figure 8).

[70] The rating curve f is defined using stage-discharge data, courtesy of the USGS (waterdata.usgs.gov/nwis). Following the suggestion of *Fenton and Keller* [2001], it is fitted by a polynomial computed over the square root of the flow. In many instance, a single curve does not capture reliably the behavior over the entire range, so piecewise polynomials are used. In the particular case of the Chunky river shown in Figure 8, two piecewise polynomials of order 3 were used:

$$y = f(h) \equiv \begin{cases} \sum_{i=0}^{3} a_i \sqrt{h^i} & h \le h_k \\ \sum_{i=0}^{3} b_i \sqrt{h^i} & h > h_k \end{cases}, \tag{19}$$

where h_k is the position of the kink separating both polynomials.

[71] Formally, the output error model is based on the assumption that daily discharge errors are normally distributed with a variance dependent on stage:

$$\sigma_{y_d} = (10 \text{ mm} + .02h)\, f'(h). \tag{20}$$

[72] Although the numerical parameters of 10 mm and 2% were chosen empirically to capture the variability observed in the data (see Figure 8), they agree reasonably well with physically realistic values. Indeed, the accuracy of a stage mesurement has been estimated at 9 mm [*Schmidt and Garcia*, 2003] (although it is more commonly estimated at 3 mm [*Hirsch and Costa*, 2004]). The accuracy of single discharges is known to vary between 5% and 10% [*Hirsch and Costa*, 2004], which correspond in the case of the Chunky river to stage accuracies between 1.2% and 2.4%.

[73] To link the daily discharge error model with GR2M monthly flows, the total monthly discharge is converted into an average daily discharge. Assuming that daily errors can be approximated by a first-order autoregressive process with a fixed variance σ^2_{yd} and correlation between successive measurement given by ϕ, then the variance of the mean over n samples is given by

$$\sigma_y^2 = \frac{\sigma^2_{y_d}}{n} \frac{1+\phi}{1-\phi} \left(1 - \frac{2\phi(1-\phi^n)}{n(1-\phi^2)}\right). \tag{21}$$

[74] Setting $n = 30$ and $\phi = .8$ in equation (21) yields a standard variation of $\sigma_y \approx 0.5\sigma_{y_d}$ for the monthly average daily streamflow error, with σ_{y_d} given by equation (20). Note that since the daily variance σ_{y_d} is not constant but depends on the stage h, equation (21) constitutes an approximation. However, the intramonthly variability of the stage is sufficiently low for this approximation to hold (see Appendix B). The value of $\phi = .8$ is chosen to reflect

the belief that daily streamflow errors are moderately correlated over a span of about one week.

[75] Summing up, the output error model is defined by a normal distribution centered on the true discharge with a variance depending on the stage associated to the mean daily discharge and the derivative of the rating curve at that point:

$$\sigma_y = (5\text{ mm} + .01h)\, f'(h). \tag{22}$$

5.4. Structural Error Model

[76] Model residuals are often found to be heteroscedastic, auto-correlated and non-normal [*Yang et al.*, 2007; *Xu*, 2001]. Although it is custom to use Box Cox or similar transformations to stabilize the variance and reduce heteroscedasticity, using transformed variables may have undesirable consequences on model calibration [*Schaefli et al.*, 2007]. Hence we follow *Beven and Freer* [2001] suggestion that residuals could be described by multiplicative autoregressive process and define the structural error model by a multivariate lognormal distribution, with covariance matrix Σ chosen to reproduce an autoregressive process of order one [*Sorooshian and Dracup*, 1980]. The log likelihood of such a distribution is given

$$L(\boldsymbol{y}|\boldsymbol{\mu} = \widehat{\boldsymbol{y}}, \rho = .6, \sigma = .15\text{ mm})$$

with L defined by equation (A4) in Appendix (A3). The hyperparameters of the structural error model correspond roughly to a relative error of 15% and a moderate auto-correlation (for daily models, the auto-correlation is sometimes found to be around .9 [*Høbye and Rosbjerg*, 1999]).

5.5. Priors

[77] The prior for the initial conditions is defined by fitting a bivariate normal distribution (see equation (A2)) to a series of "historical" internal states. This series of internal state is computed by running GR2M over 20 years of data prior to the calibration period using different sets of parameters. These parameter sets are chosen randomly around the classical estimate (minimizing NSE on the square root of discharges). The mean $\boldsymbol{\mu}_\phi$ and covariance matrix Σ_ϕ of the internal variables series are then computed and used as hyperparameters for the bivariate normal, thus defining the prior for the initial model state:

$$\pi(\phi_0) = \mathbf{N}\left(\phi_0|\boldsymbol{\mu}_\phi, \Sigma_\phi\right). \tag{23}$$

[78] The prior for the parameter is chosen to be a uniform distribution and is defined by

$$\pi(\boldsymbol{\theta}) = \mathbf{U}(\theta_1|0, 1000) \cdot \mathbf{U}(\theta_2|0, 2). \tag{24}$$

[79] One reason for this choice is simplicity; with 20 years of calibration data, the prior for the parameter is not expected to play a significant role. The other motivation for a uniform distribution is to enable a fair comparison between the SLS and the Bayesian results.

[80] The prior for the true rainfall $\pi(\boldsymbol{r})$ is defined as in equation (10b), that is, by an exponentiated Weibull distribution whose parameters are estimated by maximizing the likelihood of an "historical" precipitation series spanning 20 years prior to the calibration period (see section 4.3). The importance of this prior must not be underestimated, as it plays an important role in the inference process [*Huard and Mailhot*, 2006]. One way to improve the accuracy of this prior could be to define seasonal priors, i.e., different prior for each season or month when strong climatic seasonalities are expected.

5.6. Simulation Setup

[81] The posterior distribution is now given by:

$$\begin{aligned} \mathrm{p}(\boldsymbol{\theta}|\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{e}}, \tilde{\boldsymbol{y}}) &= \iiint \mathrm{p}(\boldsymbol{\theta}, \boldsymbol{r}, \boldsymbol{y}, \epsilon, \phi_0|\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{e}}, \tilde{\boldsymbol{y}})\, \mathrm{d}\boldsymbol{r}\, \mathrm{d}\boldsymbol{y}\, \mathrm{d}\boldsymbol{\epsilon}\, \mathrm{d}\phi_0 \\ &= \iiint \mathrm{p}_{in}(\tilde{\boldsymbol{r}}|\boldsymbol{r})\mathrm{p}_{out}(\tilde{\boldsymbol{y}}|\boldsymbol{y})\mathrm{p}_{str}(\boldsymbol{y}|\widehat{\boldsymbol{y}}) \\ &\quad \cdot \boldsymbol{\pi}(\boldsymbol{\theta})\pi(\phi_0)\boldsymbol{\pi}(\boldsymbol{r})\boldsymbol{\pi}(\boldsymbol{\epsilon})\, \mathrm{d}\boldsymbol{r}\, \mathrm{d}\boldsymbol{y}\, \mathrm{d}\boldsymbol{\epsilon}\, \mathrm{d}\phi_0 \end{aligned} \tag{25}$$

derived under hypotheses similar to those presented in section 2.2. The integrand on the right hand side of equation (25) is again sampled using Markov Chain Metropolis using 2.0×10^6 iterations and discarding the first half. Validation is performed using the procedure described in section 2.4 with the same error models used in calibration.

5.7. Results and Analysis

[82] The results of the calibration and validation of the Chunky watershed are now presented. Figure 9 shows contour plots of the posterior pdf for simulated flow, true flow and simulated flow measurements, along with the inferred true rain and simulated rain measurements (the plots show the last five of the twenty years series used for calibration to avoid cluttering the graph). Note that while the true flow and simulated flow measurements are very similar, the distribution of the simulated rain measurements shows more dispersion than the pdf for the true rain. Also, the agreement between true flow and observed flow (full line) appears very good. Both observations are direct consequences of the fact that the error models assume small flow errors and large rainfall errors. As discussed in section 4.3, the true rainfall is strongly conditioned on the observed flow, and model performance should not be judged on the agreement between output observations and simulations.

[83] A clearer picture of the different errors is presented in Figure 10, showing the posterior pdf for the rainfall errors ($\tilde{\boldsymbol{r}} - \boldsymbol{r}$), structural errors ($\boldsymbol{y} - \widehat{\boldsymbol{y}}$) and output errors ($\tilde{\boldsymbol{y}} - \boldsymbol{y}$). As discussed earlier, the output errors are the smallest since the output error model assumes relatively small measurement errors. Rainfall errors are almost always negative, reflecting the assumption that rainfall measurements underestimate the true value. Structural errors, on the other hand, are mostly positive, a feature that is not imposed by the structural error model. One possible explanation could be the underestimation of overland flow by the model, reducing the amount of instant runoff during particularly rainy periods.

[84] This separation of errors into distinct components, especially the identification of structural errors, is expected to have useful applications both as a diagnostic tool to identify model deficiencies and as a measure of model performance. Indeed, model performance in calibration

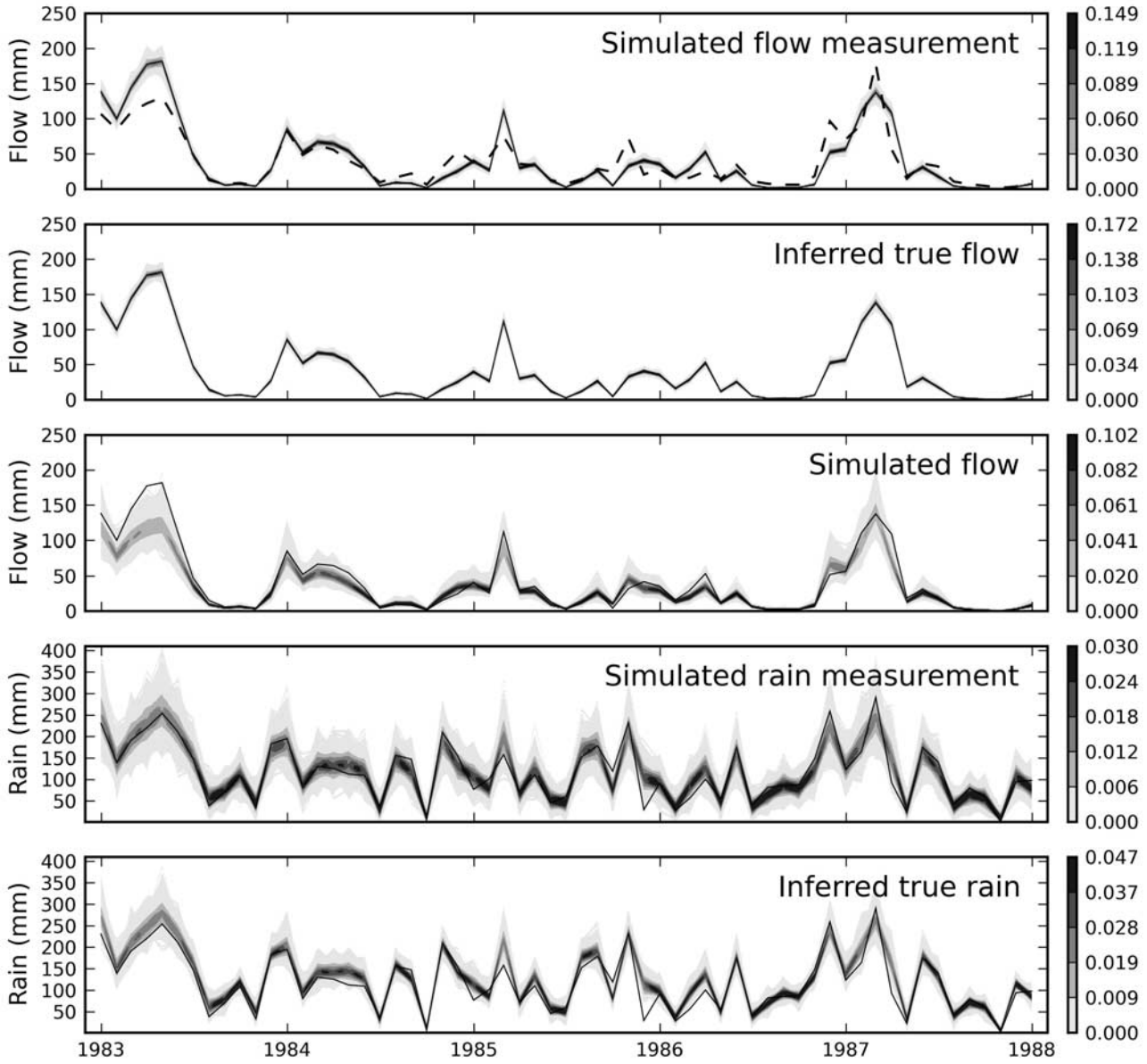

Figure 9. Calibration results for the Chunky River watershed using realistic priors and error model assumptions. Pdfs for each variable are shown as contour plots while observations are indicated by full lines. In the upper graph, the dashed line shows the flow simulated by parameters fitted using a least squares criteria (NSE) on the square root of the flows. The scales on the right indicate the value of the pdf.

could be estimated, irrespective of data error, by measuring the dispersion and autocorrelation of the posterior structural error distribution. In terms of diagnosis, analyzing dependencies between structural errors and environmental factors might also reveal variables that are relevant to the modeling process but unaccounted for by the model. One should be careful, however, with conclusions drawn from inferred error series. Indeed, with inferred error series as long as data series, the influence of error model assumptions (statistical distribution, hyperparameters) on inferred errors is bound to be significant. Again, this calls for further research on the definition of accurate error models as well as their validation.

[85] Figure 11 shows the probability distribution for the model parameters, compared to the parameters estimated by minimizing the NSE criteria computed on the flows square root (a variation of the standard least squares (SLS)). When superposed with the parameter distributions of Figure 7, the mode of the sampled parameters fall in between those of cases 1 and 3, but the dispersion is larger.

[86] Figure 12 shows the model simulations during the validation period compared with the SLS estimate (dashed line). Although there is not much difference between the SLS prediction and the modes of the Bayesian distributions, the uncertainty assessment appears reasonable in the sense that observed flows fall within the high density regions of the pdf. It shall be enlightening to pursue this type of investigation for daily or hourly models to see whether larger differences in prediction accuracy exist between standard optimization and a Bayesian analysis.

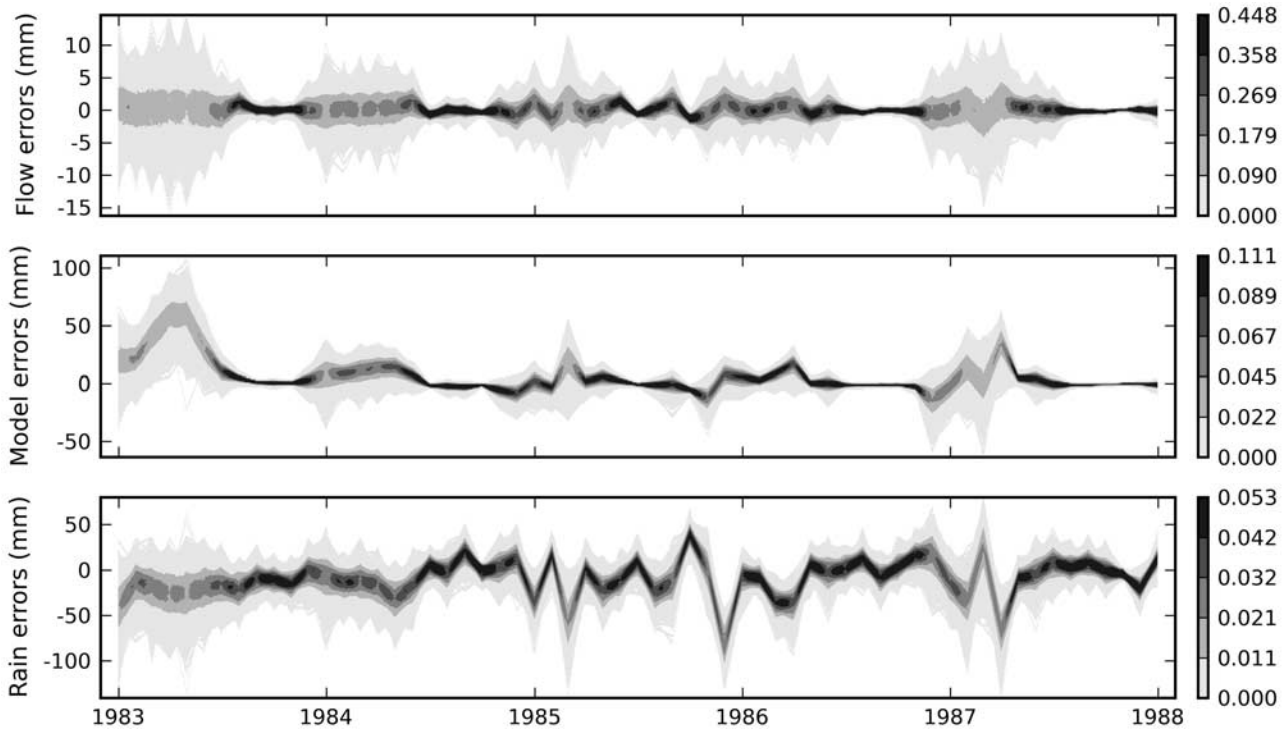

Figure 10. Inferred posterior pdf for the three main sources of errors: input (rainfall) errors (mm), structural or model errors (mm), and output (flow) errors (mm). The graphs show the last five years of the calibration period. The scales on the right indicate the value of the pdf.

5.8. Extension to Daily Models

[87] There is a certain contradiction of purpose in applying an elaborate uncertainty assessment method to a simple two parameter monthly model. On the one hand, this choice makes the application of the method and interpretation of the results certainly simpler than with complex daily models. On the other hand, the added value of uncertainty assessment is lower because simple monthly models are expected to be less sensitive to input errors than their daily or hourly counterpart. To apply the method to daily models, however, a number of issues have to be addressed.

[88] The prior for the true input, for instance, has to deal gracefully with null rainfall. One possibility would be to use statistical distributions that have both continuous and discrete components. The Tweedy distribution is one of those, designed to yield a finite probability to null values and a probability density on the positive domain [*Dunn and Smyth*, 2005]. Another issue that will have to be dealt with is the proper description of autocorrelations in the rainfall, flow and structural error series. Instead of being a liability, the dependence between errors is an asset for a modeler because it reduces the effective number of degrees of freedom in the calibration problem. Hence from an inferential point of view, error autocorrelation is a useful feature and should be considered more carefully. Whatsmore, reducing the effective number of degrees of freedom might help overcoming the main practical challenge, namely the increasing dimensionality of the problem. Indeed, as the number of dimensions increases, so do the number of samples necessary for the posterior distributions to reach convergence. The approach taken in this paper is adequate for monthly series, but it is not yet clear how well it will scale for daily series, two order of magnitudes longer.

6. Conclusion

[89] In calibrating hydrological models to empirical data, hydrologists face the challenge of identifying model parameters that reasonably reproduce a watershed's behavior. As this behavior is quantified by a limited set of field observations corrupted by errors, hydrologists run the risk of fitting the model to the noise instead of the real process. Accounting for this noise, however, requires assumptions about its nature, that is, a plausible description of input and output errors. Moreover, the model itself does not emulate all the processes that actually occur in reality, introducing discrepancies between input and output that also need to be

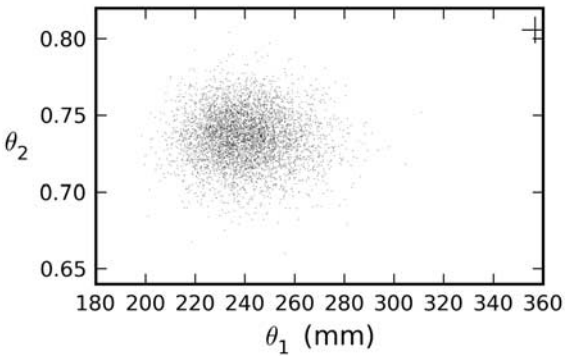

Figure 11. Parameters sampled during calibration under realistic error model assumptions. The cross indicates the optimal parameters found by maximizing the NSE criteria.

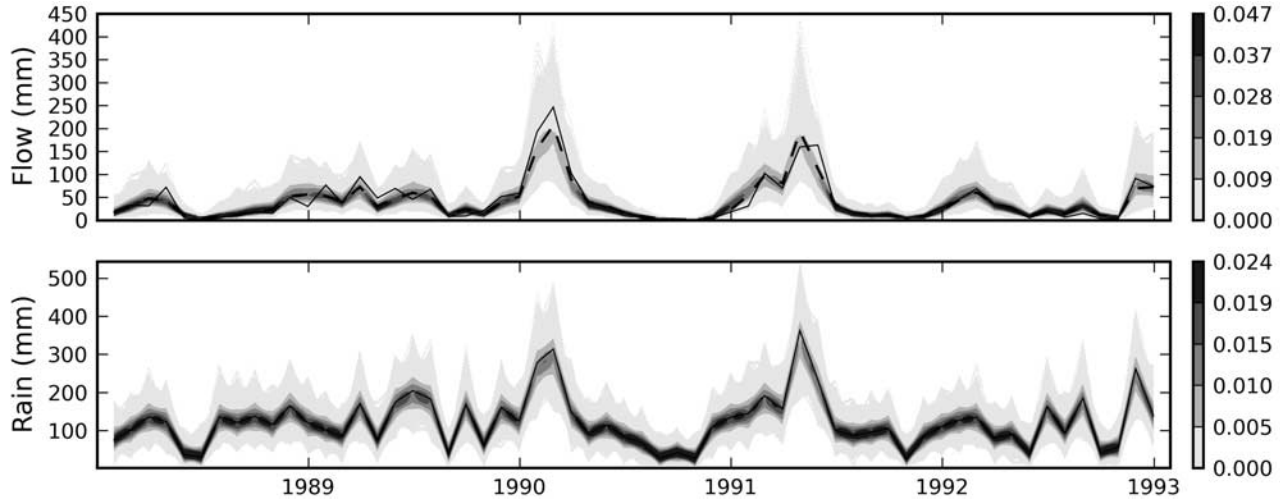

Figure 12. Flows simulated during the validation period. The lower graph shows a contour plot of the simulated rain measurements density, with actual measurements indicated by a full line. The upper plot displays the pdf of simulated flow measurements (contour plot), obtained by running the model over the parameters sampled during calibration and adding structural and measurement noise. The dashed line shows the flows estimated by the SLS method and the full line indicates flow observations. The scales on the right indicate the value of the pdf.

described. In this paper, Bayesian analysis is used to merge hypotheses about input, output and model errors into the calibration and validation of a model. Using the parsimonious monthly hydrological model GR2M, simulations are run to answer a number of questions about the assessment of multiple sources of uncertainties.

[90] The first question concerned the effect of the initial state uncertainty on calibration efficiency. Simulations show that for GR2M, the initial state uncertainty lasts only a few months, after which the internal state reaches a ''steady state''. Hence for long time series, accounting for initial state uncertainty has only marginal effects. Input uncertainty, however, plays a far more important role. In fact, when input errors are considered large and output errors small, model calibration proceeds in ''reverse'', conditioning the input on the output. This results in simulations reproducing almost perfectly the output observations, rendering traditional efficiency measures (NSE) meaningless. Moreover, calibrated parameters are completely different from those obtained under an ''output error only'' hypothesis. This emphasizes how important it is to analyze parameter distributions only in the light of the error model assumptions.

[91] In a calibration where multiple sources of uncertainties are accounted for, validation plays the sensitive role of assessing the hydrological model efficiency as well as the accuracy and validity of the error model assumptions. With so many factors influencing predictions, it is difficult to identify possible causes of failure just from the comparison between output observations and simulations. By looking directly at the inferred pdf for the input, output and structural error series, however, model failures or error model inadequacies are easier to detect and diagnosis easier to perform. Moreover, analyzing structural error series might provide the means to evaluate model performance, without interference from data quality issues.

[92] The obvious interest in accounting explicitly for different sources of uncertainty affecting hydrological modeling must be weighted against the interest of stakeholders in realistic uncertainty assessment and the efforts required to specify accurate error models. Data model errors should reflect as accurately as possible the errors introduced in the acquisition of data. Ideally, those error models would evolve with changes in the gauge station network (number, location and type of stations) and changes in rating curves following new discharge measurements. Structural error models appear yet more difficult to define rigorously, and progress might only be made through the prior/posterior Bayesian learning process. Hopefully, the knowledge gained from the next experiments in model calibration and the rapid increase in computing power will eventually make uncertainty assessment a routine exercise and an integral part of hydrological modeling.

Appendix A: Probabilistic Distributions

A1. Truncated Normal Distribution

[93] Assuming variable x is distributed normally but bounded to the interval $[a, b]$, then the probability density function of x is given by

$$\mathbf{N}_T(x|\mu,\sigma,a,b) = \frac{\phi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}, \tag{A1}$$

where ϕ and Φ stand for the probability density function and cumulative density function of the standard normal distribution.

A2. Multivariate Normal Distribution

[94] The probability density function of the multivariate normal distribution is given by

$$\mathbf{N}(\boldsymbol{x}|\boldsymbol{\mu},\Sigma) = \frac{1}{|2\pi\Sigma|}\exp\left\{-(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2\right\} \tag{A2}$$

where $\boldsymbol{\mu}$ is the vector mean and Σ the covariance matrix.

A3. Lognormal Distribution

[95] Assuming y is normally distributed with mean μ and variance σ^2 and $x = \exp(y)$, then x has a lognormal distribution with a probability density function given by

$$\mathbf{L}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}. \quad \text{(A3)}$$

[96] A rationale for the use of the lognormal in science is given by *Limpert et al.* [2001], along with some of its properties. For instance, the lognormal is the distribution maximizing entropy on the semi-infinite interval given a fixed mean and variance [*Goodman*, 1987]. It is also the limiting distribution of the product of random variates, and hence, ideally suited to model multiplicative errors.

[97] If y is generated by an auto-regressive process of order one: $y_i = \rho\, y_{i-1} + \epsilon_i$ with $\epsilon \sim N(0, \sigma)$ for $i = 1, 2, \ldots, n$ and $x_i = \exp(\mu_i + y_i)$, then the log likelihood of $\boldsymbol{x}$ is given by

$$\begin{aligned} L(\boldsymbol{x}|\boldsymbol{\mu},\rho,\sigma) = &-\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln\frac{\sigma^{2n}}{1-\rho^2} - \sum_{i=1}^{n}\ln x_i \\ &- \frac{1}{2\sigma^2}\left\{(1-\rho^2)(\ln x_1 - \mu_1)^2\right. \\ &\left. + \sum_{i=2}^{n}\left[(\ln x_i - \mu_i) - \rho(\ln x_{i-1} - \mu_{i-1})\right]^2\right\}. \end{aligned} \quad \text{(A4)}$$

A4. Exponentiated Weibull Distribution

[98] The exponentiated Weibull distribution, introduced by *Mudholkar and Hutson* [1996] has the following pdf:

$$\mathbf{W}(x|\alpha,k,\mu,\sigma) = \frac{\alpha k}{\sigma} z^{k-1} e^{-z^k}\left[1 - e^{-z^k}\right]^{\alpha-1}, \quad z = \frac{x-\mu}{\sigma} \quad \text{(A5)}$$

with $z > 0$ and $k > 0$. It is a generalization of the Weibull distribution, with an additional shape parameter α. The Weibull distribution is frequently used in hydrology to model rainfall characteristics [*Zhang and Singh*, 2007; *Sharma*, 1996; *Wilks*, 1989].

Appendix B: Aggregation of Proportional Daily Errors

[99] Given a daily series d_i with $i = 1, \ldots, n$ with proportional errors $\delta_{d_i} = f_d d_i$, one wishes to determine the corresponding proportionality factor f_m for aggregated monthly values $m = \sum_{i=1}^{n} d_i$, such that

$$f_m^2 m^2 = \sum_{i=1}^{n} f_d^2 d_i^2. \quad \text{(B1)}$$

[100] Isolating f_m and taking the expectation on both sides of equation (B1) yields

$$E[f_m] = f_d E\left[\frac{\sqrt{\sum_{i=1}^{n} d_i^2}}{m}\right]. \quad \text{(B2)}$$

[101] This equation shows that the monthly proportionality factor increases with the intramonthly variability of the series, in other words, that errors on large daily events strongly contribute to the error on the monthly totals. For example, using daily rainfall series from the Chunky watershed, taking $f_d = .15$ yields a monthly proportionality factor $f_m \approx .07$. With a daily rainfall equal to its mean value, i.e., no variability in the series, equation (B2) reduces to $f_m = f_d/\sqrt{n}$ and $f_m \approx .03$.

[102] **Acknowledgments.** David Huard is grateful for the financial support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

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

Andréassian, V., C. Perrin, C. Michel, I. Usart-Sanchez, and J. Lavabre (2001), Impact of imperfect rainfall knowledge on the efficiency and the parameters of watershed models, *J. Hydrol.*, *250*(1-4), 206–223, doi:10.1016/S0022-1694(01)00437-1.

Andréassian, V., C. Perrin, and C. Michel (2004), Impact of imperfect potential evapotranspiration knowledge on the efficiency and parameters of watershed models, *J. Hydrol.*, *286*, 19–35, doi:10.1016/j.jhydrol.2003.09.030.

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

Beven, K. (2006a), A manifesto for the equifinality thesis, *J. Hydrol.*, *320*(1-2), 18–36, doi:10.1016/j.jhydrol.2005.07.007.

Beven, K. (2006b), On undermining the science?, *Hydrol. Processes*, *20*, 3141–3146, doi:10.1002/hyp.6396.

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

Beven, K., and J. Freer (2001), Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, *J. Hydrol.*, *249*, 11–29, doi:10.1016/S0022-1694(01)00421-8.

Beven, K., P. Smith, and J. Freer (2007), Comment on "Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology" by Pietro Mantovan and Ezio Todini, *J. Hydrol.*, *338*(3–4), 315–318, doi:10.1016/j.jhydrol.2007.02.023.

Burgers, G., P. J. van Leeuwen, and G. Evensen (1998), Analysis scheme in the ensemble Kalman filter, *Mon. Weather Rev.*, *126*, 1719–1724.

Chib, S., and E. Greenberg (1995), Understanding the Metropolis-Hastings algorithm, *Am. Stat.*, *49*(4), 327–335.

Daly, C., R. P. Neilson, and D. L. Phillips (1994), A statistical—topographic model for mapping climatological precipitation over mountainous terrain, *J. Appl. Meteorol.*, *33*, 140–158.

Daly, C., G. H. Taylor, and W. P. Gibson (1997), The PRISM approach to mapping precipitation and temperature, in *10th Conference on Applied Climatology*, pp. 10–12, American Meteorological Society, Reno, NV.

Dunn, P. K., and G. K. Smyth (2005), Series evaluation of Tweedie exponential dispersion model densities, *Stat. Comput.*, *15*, 267–280, doi:10.1007/s11222-005-4070-y.

Dymond, J. R., and R. Christian (1982), Accuracy of discharge determined from a rating curve, *Hydrol. Sci. J.*, *4*(12), 493–504.

Edijatno, and C. Michel (1989), Un modèle pluie-débit journalier à trois paramètres, *La Houille Blanche*, *2*, 113–122.

Farnsworth, R., E. S. Thompson, and E. Peck (1982), Evaporation atlas for the contiguous 48 United States, *Technical Report NWS 33*, NOAA, Washington, DC.

Fenton, J. D., and R. J. Keller (2001), The calculation of streamflow from measurements of stage, *Tech. rep.*, Cooperative research center for catchment hydrology.

Fiering, M. B. (1967), *Streamflow Synthesis*, Harvard Univ. Press, Cambridge, Mass.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995), *Bayesian Data Analysis*, Texts in Statistical Science, Chapman & Hall/CRC Press, London.

Goodman, J. (1987), A comment on the Maximum Entropy principle, *Risk Analysis*, *7*(2), 269–272.

Gull, S. F. (1989), Bayesian Data Analysis: Straight-line fitting, in *Maximum Entropy and Bayesian Methods*, edited by J. Skilling, pp. 511–518, Kluwer Academic Publishers.

Gupta, H. V., S. Sorooshian, and P. Yapo (1998), Towards improved calibration of hydrologic models: Multiple and non-commensurable measures of information, *Water Resour. Res.*, *34*(4), 751–764.

Habib, E., W. F. Krajewski, and A. Kruger (2001), Sampling errors of tipping-bucket rain gauge measurements, *J. Hydrol. Eng.*, *6*(2), 159–166.

Hall, J., E. O'Connell, and J. Ewen (2007), On not undermining the science: coherence, validation and expertise. Discussion of Invited Commentary by Keith Beven. Hydrological Processes 20, 3141_3146 (2006), *Hydrol. Processes*, *21*(7), 985–988, doi:10.1002/hyp.6639.

Herschy, R. W. (1985), *Streamflow measurement*, Elsevier.

Hirsch, R. M., and J. E. Costa (2004), U.S. stream flow measurement and data dissemination improve, *EOS*, *85*(21), 197,203.

Høybye, J., and D. Rosbjerg (1999), Effect of input and parameter uncertainties in rainfall-runoff simulations, *J. Hydrol. Eng.*, *4*(3), 214–224.

Huang, M., and X. Liang (2006), On the assessment of the impact of reducing parameters and identification of parameter uncertainties for a hydrologic model with applications to ungauged basins, *J. Hydrol.*, *320*(1–2), 37–61, doi:10.1016/j.jhydrol.2005.07.010.

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

Jaynes, E. T., and G. L. Bretthorst (2003), *Probability theory: The logic of science*, Cambridge University Press.

Kabouya, M. (1990), Modélisation pluie-débit au pas de temps mensuel et annuel en Algérie Septentrionale, Ph.D. thesis, Orsay-Paris Sud.

Kavetski, D., S. W. Franks, and G. Kuczera (2003), Confronting input uncertainty in environmental modeling, in *Calibration of Watershed Models, Water Science and Applications*, vol. 6, edited by Q. Duan, H. V. Gupta, S. Sorooshian, A. N. Rousseau, and R. Turcotte, pp. 49–68, AGU Water Science and Applications Series.

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

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

Kennedy, M. C., and A. O'Hagan (2001), Bayesian calibration of computer models, *J. R. Stat. Soc. Series B-Statistical Methodology*, *63*, 425–464.

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

Limpert, E., W. A. Stahel, and M. Abbt (2001), Log-normal distributions across the sciences: Keys and clues, *BioScience*, *51*(5), 341–352.

Mantovan, P., and E. Todini (2006), Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, *J. Hydrol.*, *330*, 368–381, doi:10.1016/j.jhydrol.2006.04.046.

Maurer, E. P., A. W. Wood, J. C. Adam, and D. P. Lettenmaier (2002), A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States, *Am. Meteorol. Soc.*, *15*(22), 3237–3251.

McGuinness, J., and E. Bordne (1972), A comparison of lysimeter-derived potential evapotranspiration with computed values, *Technical Bulletin 1452*, Department of Agriculture.

Metcalfe, R., B. Routledge, and K. Devine (1997), Rainfall measurement in Canada: Changing observational methods and archive adjustment procedures, *J. Clim.*, *10*, 92–101.

Michaud, J. D., and S. Sorooshian (1994), Effect of rainfall-sampling errors on simulations of desert flash floods, *Water Resour. Res.*, *30*(10), 2765–2776.

Montanari, A. (2007), What do we mean by 'uncertainty'? The need for a consistent wording about uncertainty assessment in hydrology, *Hydrol. Processes*, *21*, 841–845, doi:10.1002/hyp.6623.

Moradkhani, H., K.-L. Hsu, H. Gupta, and S. Sorooshian (2005), Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter, *Water Resour. Res.*, *41*(5), W05012, doi:10.1029/2004WR003604.

Mouelhi, S., C. Michel, C. Perrin, and V. Andréassian (2006), Stepwise development of a two-parameter monthly water balance model, *J. Hydrol.*, *318*, 200–214, doi:10.1016/j.jhydrol.2005.06.014.

Mudholkar, G., and A. Hutson (1996), The exponentiated Weibull family: Some properties and a flood data application, *Commun. Stat.: Theory Methods*, *25*(12), 3059–3083.

Nandakumar, N., and R. G. Mein (1997), Uncertainty in rainfall-runoff model simulations and the implications for predicting the hydrologic effects of land_use change, *J. Hydrol.*, *192*, 211–232, doi:10.1016/S0022-1694(96)03106-X.

Neal, R. M. (1993), Probabilistic inference using Markov Chain Monte Carlo methods, *Tech. rep.*, Department of Computer Science, University of Toronto.

Novak, E., and K. Ritter (1997), The curse of dimension and a universal method for numerical integration, in *Multivariate Approximation and Splines, ISNM*, edited by G. Nürnberger, J. W. Schmidt, and G. Walz, pp. 177–188, Birkhäuser.

Oudin, L., C. Perrin, T. Mathevet, V. Andréassian, and C. Michel (2006), Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models, *J. Hydrol.*, *320*, 62–83, doi:10.1016/j.jhydrol.2005.07.016.

Paturel, J., E. Servat, and A. Vassiliadis (1995), Sensitivity of conceptual rainfall-runoff algorithms to errors in input data – case of the GR2M model, *J. Hydrol.*, *168*, 111–125, doi:10.1016/0022-1694(94)02654-T.

Pelletier, P. M. (1988), Uncertainties in the single determination of river discharge: A literature review, *Can. J. Civ. Eng.*, *15*, 834–850.

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, *Adv. Water Resour.*, *29*(11), 1586–1597, doi:10.1016/j.advwatres.2005.11.013.

Savenije, H. H. G. (2004), The importance of interception and why we should delete the term evapotranspiration from our vocabulary, *Hydrol. Processes*, *18*(8), 1507–1511, doi:10.1002/hyp.5563.

Schaake, J., Q. Duan, V. Andréassian, S. Franks, A. Hall, and G. Leavesley (2006), The model parameter estimation experiment (MOPEX), *J. Hydrol.*, *320*(1–2), 1–2, doi:10.1016/j.jhydrol.2005.07.054.

Schaefli, B., and H. V. Gupta (2007), Do Nash values have value?, *Hydrol. Processes*, *21*(15), 2075–2080, doi:10.1002/hyp.6825.

Schaefli, B., D. B. Talamba, and A. Musy (2007), Quantifying hydrological modeling errors through a mixture of normal distributions, *J. Hydrol.*, *332*(3–4), 303–315, doi:10.1016/j.jhydrol.2006.07.005.

Schmidt, A. R. (2002), Analysis of stage-discharge relations for open-channel flows and their associated uncertainties, Ph.D. thesis, University of Illinois.

Schmidt, A. R., and M. H. Garcia (2003), Theoretical examination of historical shifts and adjustments to stage-discharge rating curves, in *World Water and Environmental Resources Congress 2003 and Related Symposia 2003*, edited by P. Bizier and P. DeBarry, American Society of Civil Engineers.

Sharma, T. C. (1996), A Markov-Weibull rain-sum model for designing rain water catchment systems, *Water Resour. Manage.*, *10*(2), 147–162.

Shelton, M. L. (1985), Modeling hydroclimatic processes in large watersheds, *Ann. Assoc. Am. Geogr.*, *75*(2), 185–202.

Shepard, D. S. (1984), *Spatial Statistics and Models*, chap. Computer mapping: the SYMAP interpolation algorithm, pp. 133–145, D. Reidel.

Shiklomanov, A. I., T. I. Yakovleva, R. B. Lammers, I. P. Karasev, C. J. Vorosmarty, and E. Linder (2006), Cold region river discharge uncertainty–estimates from large Russian rivers, *J. Hydrol.*, *326*(1–4), 231–256, doi:10.1016/j.jhydrol.2005.10.037.

Sorooshian, S., and J. A. Dracup (1980), Stochastic parameter estimation procedures for hydrologic rainfall-runoff models: Correlated and heteroscedastic error cases, *Water Resour. Res.*, *16*(2), 430–442.

Thiemann, M., M. Trosset, H. V. Gupta, and S. Sorooshian (2001), Bayesian recursive parameter estimation for hydrologic models, *Water Resour. Res.*, *37*(10), 2521–2535.

Todini, E., and P. Mantovan (2007), Comment on: 'On undermining the science?' by Keith Beven, *Hydrol. Processes*, *21*, 1633–1638, doi:10.1002/hyp.6670.

Troutman, B. M. (1982), An analysis of input errors in precipitation-runoff models using regression with errors in the independent variables, *Water Resour. Res.*, *18*(4), 947–964.

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

Wagener, T., N. McIntyre, M. J. Lees, H. S. Wheater, and H. V. Gupta (2003), Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis, *Hydrol. Processes*, *17*, 455–476, doi:10.1002/hyp.1135.

Weerts, A. H., and G. Y. H. El Serafy (2006), Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual

rainfall-runoff models, *Water Resour. Res.*, *42*, W09403, doi:10.1029/2005WR004093.
Widmann, M., and C. Bretherton (2000), Validation of mesoscale precipitation in the NCEP reanalysis using a new gridcell dataset for the northwestern United States, *J. Clim.*, *13*, 1936–1950.
Wilks, D. S. (1989), Rainfall intensity, the Weibull distribution, and estimation of daily surface runoff, *J. Appl. Meteorol.*, *28*, 52–58.
Xu, C.-Y. (2001), Statistical analysis of parameters and residuals of a conceptual water balance model – Methodology and case study, *Water Resour. Manage.*, *15*(2), 75–92.
Xu, C.-Y., and G. Vandewiele (1994), Sensitivity of monthly rainfall-runoff models to input errors and data length, *Hydrol. Sci. J.*, *39*(2), 157–176.
Yang, D., S. Isida, B. E. Goodison, and T. Gunther (1999), Bias correction of daily precipitation measurements for Greenland, *J. Geophys. Res.*, *104*(D6), 6171–6181.
Yang, J., P. Reichert, and K. C. Abbaspour (2007), Bayesian uncertainty analysis in distributed hydrologic modeling: A case study in the Thur River basin (Switzerland), *Water Resour. Res.*, *43*, W10401, doi:10.1029/2006WR005497.
Yapo, P. O., H. V. Gupta, and S. Sorooshian (1996), Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data, *J. Hydrol.*, *181*(1–4), 23–48, doi:10.1016/0022-1694(95)02918-4.
Zhang, L., and V. P. Singh (2007), Bivariate rainfall frequency distributions using Archimedean copulas, *J. Hydrol.*, *332*(1–2), 93–109, doi:10.1016/j.jhydrol.2006.06.033.

D. Huard and A. Mailhot, Institut National de la Recherche Scientifique, Centre Eau, Terre & Environnement, Québec (Qc), Canada, G1K 9A9. (david.huard@ete.inrs.ca; alain.mailhot@ete.inrs.ca)