Kavetski, D., Franks, S. W. and G. Kuczera (2002) Confronting input uncertainty in environmental modelling.
in Duan, Q., Gupta, H. V., Sorooshian, S., Rousseau, A. N. and R. Turcotte (editors),
Calibration of Watershed Models, AGU Water Science and Applications Series Volume 6, pp. 49-68.

Confronting Input Uncertainty in Environmental Modelling

Dmitri Kavetski, Stewart W. Franks and George Kuczera

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

The majority of environmental models require calibration of some or all of their parameters before meaningful predictions of catchment behaviour can be made. Despite the importance of reliable parameter estimates, there are growing concerns about the ability of objective-based inference methods to adequately calibrate environmental models. The problem lies with the formulation of the objective or likelihood function, which is currently implemented using essentially ad-hoc methods. We outline limitations of current calibration methodologies, including least squares, multi-objective, GLUE and Kalman filter schemes and introduce a more systematic Bayesian Total Error Analysis (BATEA) framework for environmental model calibration and validation. BATEA imposes a hitherto missing rigour in environmental modelling by requiring the specification of physically realistic uncertainty models with explicit assumptions that can and must be tested against available evidence. Distinguishing between the various sources of errors will reduce the current ambiguity about parameter and predictive uncertainty and enable rational testing of environmental model hypotheses. A synthetic study demonstrates that explicitly accounting for forcing errors leads to immediate advantages over traditional least squares methods that ignore rainfall history corruption and do not directly address the sources of uncertainty in the calibration. We expect that confronting all sources of uncertainty, including data and model errors, will force fundamental shifts in the model calibration/verification philosophy.

INTRODUCTION

Hydrological and environmental modelling has benefited from significant developments over the last decade. The growing understanding of environmental physics, combined with dramatic increases in computing power, has allowed progressively more realistic representation and simulation of catchment dynamics, in many cases solving hitherto intractable analysis and prediction problems.

Paradoxically, these advances have increased the need for improved model calibration and validation methods. Hydrological models invariably require calibration before predictive use to ensure consistency with observed data. In recognition of the limitations of visual and manual model calibration, several methodologies have been developed for automatic calibration of hydrological models, including classical Bayesian methods (e.g., NLFIT of Kuczera [1994]); multi-objective calibration methods (e.g., MOCOM of Sorooshian et al. [1993] and Gupta et al. [1998]); the GLUE framework of Beven and Binley [1992]; and, less commonly, Kalman filters [*Bras and Rodriguez-Iturbe*, 1985]. As we shall see, these methodologies struggle to properly characterise the fundamental problem of calibration and validation, let alone cope with the ever-increasing number of competing environmental models (see, e.g., Beven and Binley [1992] and Singh [1995]). Due to a lack of widely accepted methods for addressing data uncertainty and model verification (i.e., hypothesis testing) in hydrology, it is difficult to rationally discriminate between competing models and assess trade-offs between model performance and complexity.

This chapter focuses on the calibration paradigms currently in use and on the paradigms that need to be developed in the future. We examine sources of uncertainty in hydrological modelling, survey current calibration methods and then outline a new systematic methodology for hydrological model calibration and validation. We expect that the ideas and concepts presented here are also applicable to other branches of environmental science and engineering, where observations of global system behaviour are inexact and are made on a local scale, and where the system itself is so complicated that its simulation requires considerable approximation.

MATHEMATICS OF ENVIRONMENTAL MODELS

Regardless of whether an environmental model M is conceptual or physical, it has the following functional form

$$\mathbf{Y} \leftarrow f_M(\mathbf{X}, \boldsymbol{\theta}) \quad (1)$$

The term $\mathbf{Y} = \{\mathbf{y}_n;\ n=1\ldots N\}$ is the response matrix of the catchment. It contains one or more directly observable hydrological quantities at one or more locations within the catchment at a series of times $\mathbf{t} = \{t_n;\ n=1\ldots N\}^T$. In the simplest and most common context, $\mathbf{y}_n$ is a vector of stream flows at several locations within the catchment. However, the definition of $\mathbf{y}_n$ is more general, e.g., it may also contain water table depths at selected locations, saturated areas, etc.

The catchment responds to forcing inputs denoted by the matrix $\mathbf{X} = \{\mathbf{x}_n;\ n=1\ldots N_x\}$, where the vector $\mathbf{x}_n$ contains one or more directly observable quantities at a series of times. The forcing vector typically comprises rainfall, but evapotranspiration, pumping and injection data can also be included. The dimension N_x need not equal N, but, for time stepping rainfall-runoff models, often $N_x = N$.

The function $f_M(\cdot)$ represents the hydrological model itself and describes the response $\mathbf{Y}$ of the system to the inputs $\mathbf{X}$, e.g., the routing of rainfall into streamflow. The vector $\boldsymbol{\theta} = \{\theta_n;\ n=1\ldots P\}^T$ contains the conceptual and physical hydrological parameters of the catchment model. Parameters are constants that quantify the hydrological behaviour of the catchment (given a particular mathematical model) and determine the response $\mathbf{Y}$ for a given forcing $\mathbf{X}$. We identify "physical" parameters as those parameters that can be inferred using procedures that are independent of observable catchment responses $\mathbf{Y}$, e.g., local permeability estimates obtained using core samples or slug tests. Conversely, "conceptual" parameters (e.g., discharge coefficients) have no formal physical interpretation and can only be inferred by matching the simulated catchment behaviour $f_M(\tilde{\mathbf{X}}, \boldsymbol{\theta})$ to the observed data $\{\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}\}$. The tilde over $\mathbf{X}$ and $\mathbf{Y}$ emphasises that these quantities are estimated and hence subject to sampling and measurement error. The procedure of matching observed data and simulated system behaviour by adjusting the parameters $\boldsymbol{\theta}$ is termed calibration and forms the basis for model validation and predictive use in hydrology.

Catchment models are commonly classified into conceptual or physical models. While this distinction is valuable for many purposes, the functional behaviour of successful hydrological models is mathematically similar, since they all simulate the same physical phenomenon (e.g., rainfall-runoff routing). In particular, all hydrological models include quickflow and slowflow simulators. The quickflow is invariably related to the rainfall in the immediately preceding time steps, while slowflow behaviour obeys storage-discharge relationships. The importance of the differentiation of slowflow and quickflow processes will become apparent when the error propagation properties of hydrological models and their impact on parameter estimation are considered.

UNCERTAINTY IN HYDROLOGICAL MODELLING

Although at first glance a simple exercise in optimisation, the calibration of hydrological models is nontrivial and subtle. Most hydrological models are nonlinear and contain multiple parameters. Reliable multi-dimensional nonlinear optimisation is challenging, since it is usually prohibitively difficult to exhaustively analyse the entire parameter space.

Considerable research has been dedicated to the development of robust optimisation methods (stochastic and deterministic, local and global search algorithms). One popular optimisation method in hydrology is the Shuffled Complex Evolution (SCE) algorithm, which makes use of a population of simplexes, the vertices of which are shuffled to improve (although not guarantee) its global convergence properties [*Duan et al.*, 1992].

However, the calibration of hydrological models is profoundly affected by sources of uncertainty completely unrelated to the numerical difficulties of multi-dimensional optimisation:

1. Uncertainty in observed system inputs and responses. For example, rain gauges offer only point estimates of precipitation, while the rating curves used to estimate streamflow are also inexact, particularly when the ratings are extended beyond the data. Observational uncertainty can be further split into two categories: a) uncertainty in forcing inputs (e.g., rainfall, evapotranspiration); and b) uncertainty in output responses (e.g., streamflow, piezometer responses). The convenience of separating observation error into two categories is due to the causal structure of environmental models and will become clear later, when parameter estimation methods that take account for uncertainty in observations are considered.
2. Inherent uncertainty in the model hypothesis. Indeed, even the most elaborate model is at best a simplification of the natural environment. Although most models are based on valid physical principles (typically derived at the laboratory scale), they nevertheless remain simplifications of reality, particularly if the grid scale is orders of magnitude larger than the laboratory scale.

Figure 1 shows a schematic of the propagation of errors through environmental models. In general the observations of external forcing (inputs) $\tilde{\mathbf{X}}$ are corrupted by measurement and sampling error, which propagates through the calibrated catchment model to corrupt the simulated responses (output) $\bar{\mathbf{Y}}$. In addition, $\bar{\mathbf{Y}}$ will be affected by model and response sampling error.

It is our view that these sources of uncertainty and their propagation characteristics are currently overlooked or misunderstood and a rigorous modelling framework is necessary to provide:

a) An ability to meaningfully account for observational uncertainty and model errors; and
b) Parameter estimates with realistic confidence limits, which can then be used for prediction with meaningful uncertainty bounds;

Accurate parameter inference is necessary for meaningful prediction of flows and parameter regionalisation (allowing transfer of parameters from gauged to ungauged catchments). We shall see that common calibration methods introduce un-predictable bias into the parameters estimates, confounding regionalisation attempts.

Moreover, it is perhaps ironic that, despite extensive research of environmental physics, relatively little is known about the uncertainty operating in environmental modelling. As a consequence, we often naively combine sophisticated environmental models with simplistic Gaussian error models. Such mismatch weakens the entire modelling process and undermines the validity of its predictions.

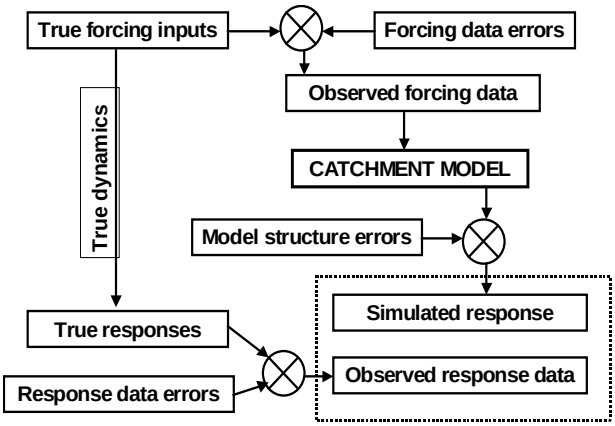

Figure 1. Error propagation in catchment modelling.

In the following sections the significance of data uncertainty in hydrological modelling is examined, with specific references to existing calibration techniques. We then articulate a conceptual framework for model calibration that addresses the two aims described above.

CALIBRATION: A CRITIQUE OF CURRENT PARADIGMS

The calibration of conceptual parameters (e.g., discharge coefficients) is required since, by definition, these parameters cannot be independently measured. In practice, despite advances in instruments and measuring devices, even physical parameters (e.g., soil permeability) often require calibration. The heterogeneity of environmental systems makes even the most accurate probe measurement only a point estimate, perhaps correct for a particular location, but often invalid as a representative average over the model grid cell, let alone the entire domain.

The difference between parameters estimated using various methods is often considerable. For example, Chappell et al. [1998] found orders of magnitude differences between permeability estimates based on core samples, inverted model and hillslope transect estimates. In the light of these differences, calibration is an important model verification procedure, since parameter estimates are site-specific and area-effective at the catchment scale. If a model cannot reproduce observed events, it is unlikely to provide accurate predictions of future events.

Least Squares Calibration

Traditionally, calibration is posed as an optimisation problem: obtain estimates of system parameters that, given the model at hand, achieve the best agreement between simulated and observed responses. Many calibration methods have been proposed and used, ranging from visual assessment of hydrographs to sophisticated search algorithms. Before providing a critique of current calibration paradigms, we consider the following case study [*Kavetski et al.*, 2000] based on the well-known hydrological model TOPMODEL [*Beven and Kirkby*, 1979; *Beven et al.*, 1995].

Figure 2 shows a portion of observed and simulated hydrographs (at this point we do not disclose the origin of the data – the reason for this will become apparent in a moment) obtained by calibrating TOPMODEL using the standard least squares (SLS) objective function

$$S_{SLS}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \left(\tilde{y}_n - f_n\left(\tilde{\mathbf{X}}, \boldsymbol{\theta}\right) \right)^2 \qquad (2)$$

where $\boldsymbol{\theta}$ are the parameters, $\tilde{y}_n$ is the observed streamflow at time t_n, $\tilde{\mathbf{X}}$ is the observed rainfall time series and $f(\cdot)$ denotes TOPMODEL. We used several years of data, perhaps more than normally employed in such calibrations, to ensure statistical averaging of parameter estimates.

The calibrated parameters are shown in Table 1. They have been obtained using the SCE algorithm with a tight convergence tolerance. The standard deviations have been estimated using the inverse negative Hessian matrix at the maximum of the objective function. The negative inverse Hessian matrix at the mode of probability density functions (pdfs) converges to the covariance matrix as the probability distribution converges to the Gaussian form [*Gelman et al.*, 1997]. As a large data set has been used, the distribution is sufficiently Gaussian.

Visually, the calibration is successful – the fit is generally good in all areas, the coefficient of determination R^2 of the

observed vs predicted responses is close to unity – what else could we hope for? In practical hydrology, far less appealing calibrations are routinely accepted.

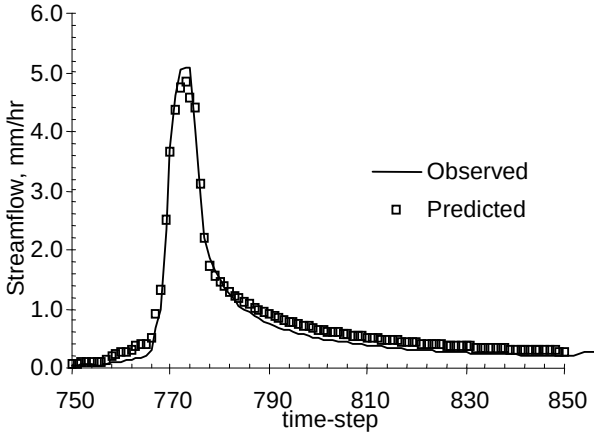

Figure 2. Portion of calibrated TOPMODEL hydrograph.

Table 1 – SLS calibration of TOPMODEL

Parameter	SLS value ± St.deviation
m-exponent, m	$0.00746 \pm 0.12\times10^{-3}$
Log-Transmissivity, ln(m)	0.879 ± 0.122
Root zone storage, m	$0.0555 \pm 0.18\times10^{-2}$
Stream velocity, m/hr	2840 ± 40.9

Appearances are deceptive, however, and it is now time to reveal the origin of the data. In fact, the streamflow was generated synthetically using TOPMODEL with a pre-determined set of parameters and no streamflow error. The set of "exact" parameters is listed in Table 2, which shows that the fitted parameters are actually quite far from the true values and, more disconcertingly, the uncertainty estimates significantly underestimate the actual errors. So what went wrong?

Table 2 – True TOPMODEL parameters and SLS errors, reported as $\mathbf{e}=(\theta_{true}-\theta_{SLS})/\sigma_{SLS}$.

Parameter	True value	SLS Error **e**
m-exponent, m	0.016	71.2
Log-Transmissivity, ln(m)	1.0	0.99
Root zone storage, m	0.1	24.7
Stream velocity, m/hr	3000	3.94

In fact, we have corrupted the observed rainfall depth of each storm used in the calibration using log-normally distributed error multipliers with an expected value of 1 and a coefficient of variation of 0.2. Note that the mean of the multipliers is 1, i.e., the rainfall estimates are on average unbiased. So why have the parameter estimates become biased? Could it be because the asymmetric lognormal distribution was used? It turns out that the symmetry of the error corruption is not at fault – using symmetrically distributed multipliers does not remove the bias in parameter estimates. The real culprit is the inadequate error model underlying the standard least squares objective function (2).

Indeed, although least squares fitting dates back to Gauss (who used the method to calculate solar system orbits), it is often misunderstood that the "objectiveness" of the criterion of the sum of squared residuals is not guaranteed. Instead, it corresponds to an assumption that the response series contains additive uncorrelated Gaussian noise,

$$\tilde{\mathbf{Y}} = f\left(\tilde{\mathbf{X}}, \boldsymbol{\theta}\right) + \boldsymbol{\varepsilon} \tag{3}$$

$$\boldsymbol{\varepsilon} \sim N\left(0, \sigma_y^2\right) \tag{4}$$

where σ_y^2 is a scalar covariance. Bayesian and maximum-likelihood estimators then obtain the most probable value of $\boldsymbol{\theta}$ by minimising the sum of squared residuals with respect to $\boldsymbol{\theta}$ [*Box and Tiao*, 1973]. Note that the simulated results are obtained using the observed forcing series $\tilde{\mathbf{X}}$.

Introducing Total Least Squares Calibration

It is not widely recognised that, if error is contained in both the *x* and *y* series, the fitting procedure must be modified. Fitting functions with errors in both *x* and *y* data is not a new research area and dates back decades (see, e.g., Deming [1943] and Macdonald and Thompson [1992]). Indeed, Jefferys [1980] shows that, even when fitting a straight line through the origin, the use of the SLS criterion (2) when both **X** and **Y** series are corrupt yields biased estimates of the slope coefficient:

$$\mathrm{E}\left[b_{SLS}\right] = b\left(1 - \left(\frac{\sigma_x}{\bar{x}}\right)^2\right) \tag{5}$$

where E[·] is the expectation, b_{SLS} is the SLS estimate of the true slope *b*, σ_x is the standard deviation of errors in the *x* data and $\bar{x}$ is the mean of the observations. The parameter bias is proportional to the error in *x*.

More importantly, however, equation (5) shows that the bias is independent of the quantity *N* of data used and, since $\mathrm{var}[b_{SLS}] \sim O(N^{-1})$, the SLS method yields progressively more misleading parameter estimates as more data is included in the analysis. This somewhat surprising result is a serious shortcoming of the SLS method and can be removed by adjusting the objective function, now termed the total least squares (TLS) function

$$S_{TLS}(\theta) = \sum_{n=1}^{N} \left[m_n \left(y_n - b x'_n \right)^2 + \left(x'_n - \tilde{x}_n \right)^2 \right] \qquad (6)$$

where x'_n is an estimate of the true value x_n and m_n is a weighting factor dependent on the relative uncertainty in x and y data (typically, $m_n = \sigma^2_{x(n)} / \sigma^2_{y(n)}$). The minimisation of (6) lies is at the heart of TLS methods, also referred to as the "Error in Variables Method" (EVM). The variance of the TLS estimates decays at the same asymptotic rate as the SLS variance, var[b_{TLS}] ~ $O(N^{-1})$, but unlike the SLS parameter estimates, the TLS estimates converge on the true values, i.e., E[b_{TLS}] = b.

Although (6) remains a sum of squared residuals, it is fundamentally different from (2), as it contains N additional unknowns $\mathbf{X}' = \{\mathbf{x}'_n; n = 1 \ldots N\}^T$, which are referred to as hidden or nuisance variables [*Gull*, 1989], latent variables, or incidental variables [*Zellner*, 1971]. Due to the form of the model $f(\cdot)$, the true values of x are necessary before the simulated values of y can be computed. This requirement has significant implications for the calibration algorithm, increasing the dimensionality of the problem.

We stress that the parameter bias (5) has been derived for the simplest 1-parameter linear model. Although analytical results are unavailable, it is unduly optimistic to expect that, when the model is nonlinear and contains many parameters, the bias in parameter estimates will disappear or decrease – indeed, the opposite is far more likely.

Structure of Hydrologic Models: The Slowflow Blues

Another factor that affects parameter inference is the structure of hydrological models – they represent physical systems with storage components. Baseflow Q_b (slow release of water from storage S) is governed by first-order ODEs, with solutions of a recursive auto-correlated form.

$$Q_b = f(S) \qquad (7)$$

and hence

$$\frac{dS}{dt} = f(S) \qquad (8)$$

A discretised solution of (8) takes the following form

$$Q_b(t_n) = f\left[Q_b(t_{n-1}) \right] \qquad (9)$$

All time stepping hydrological models implement some variation of (9) as they step through time from t_{n-1} to t_n. A first-order Taylor series approximation suggests that, due to the recursive form of (9), a perturbation of baseflow at t_n by ε_n will lead, after K time steps Δt, to an accumulated error ε_{n+k} of the form

$$\varepsilon_{n+K} = \varepsilon_n \prod_{k=1}^{K} \left(\frac{df}{dQ_b} \right)^{n+k} + O\left(\varepsilon_n^2\right) \qquad (10)$$

where $n + k$ denotes $t_n + k\Delta t$. The perturbation ε_n could arise, for example, due to observation error in the rainfall, which would propagate into the storage and hence alter the initial conditions for (10). Mathematically, the auto-correlation (10) vanishes at stationary points. In a physical context, this corresponds to the catchment drying out, when the storage is at a minimum and is, in a sense, "reset".

The auto-correlation in the residual error series violates the independence assumption of least squares schemes, decreases the informational content of the data accessible to SLS calibration and leads to additional distortion of the inferred parameters by portions of data dominated by baseflow. In contrast, quickflow processes, which rapidly respond to rainfall, induce little, if any, auto-correlation in the error series. The implication of (10) is that calibration of hydrological phenomena at the time scale of quickflow (within the storm) will not satisfy the independence assumption of least squares methods (2). It is emphasised that auto-correlated residuals present an additional problem (in addition to input errors) – even if there was no auto-correlation in the response series, the presence of input errors would lead to parameter bias similar to (5).

Finally, a third factor must be considered when calibrating environmental models – the environment is not (always) Gaussian. Although central limit theorems indicate convergence of arbitrary distributions to the Gaussian pdf, they remain limit theorems that may have limited applicability in realistic environmental modelling. Due to their relation to Gaussian error models (quadratic forms in (6)), even TLS methods cannot account for this problem. What is needed, therefore, is a general and systematic calibration framework for dealing with various sources of error in modelling that allows scrutiny of its various underlying assumptions. Although we cannot hope to develop an assumption-free framework for system analysis, we can make these assumptions maximally evident.

BAYESIAN METHODS IN HYDROLOGY

We have alluded to statistical methods when discussing least squares methods. Although calibration can be viewed as an optimisation problem, the interpretation of calibration as statistical analysis is arguably more useful, since, as well as identifying the most likely parameters, it is essential to assess the uncertainty associated with these estimates. Error estimates are not easily available within the optimisation paradigm, yet are a natural product of proper statistical

inference. In addition, Bayesian statistical analysis allows the combination of both quantitative and qualitative information, i.e., combine rainfall-runoff data with the intuitive knowledge and experience of hydrological practitioners (via prior distributions). Finally, there is a wealth of applications of Bayesian methods in areas ranging from artificial intelligence to pattern recognition and some techniques could migrate to hydrologic analysis.

When applied to rainfall-runoff modelling, Bayes equation yields the posterior pdf $p(\boldsymbol{\theta} \mid \tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$ of the model parameters, conditioned on the observed data

$$p\left(\boldsymbol{\theta} \mid \tilde{\mathbf{X}}, \tilde{\mathbf{Y}}\right) = \frac{p\left(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}} \mid \boldsymbol{\theta},\right) p\left(\boldsymbol{\theta}\right)}{p\left(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}\right)} \tag{11}$$

where the denominator $p(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$ ensures that the pdf integrates to 1. Typically, it is not necessary to explicitly evaluate $p(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$ and any factors independent of $\boldsymbol{\theta}$ in the likelihood function $p(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}} \mid \boldsymbol{\theta})$ can be absorbed into the proportionality constant, giving

$$p\left(\boldsymbol{\theta} \mid \tilde{\mathbf{X}}, \tilde{\mathbf{Y}}\right) \propto L\left(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right) \tag{12}$$

or, if a non-informative prior is imposed,

$$p\left(\boldsymbol{\theta} \mid \tilde{\mathbf{X}}, \tilde{\mathbf{Y}}\right) \propto L\left(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}} \mid \boldsymbol{\theta}\right) \tag{13}$$

Although the prior pdf has an important role in classic Bayesian analysis, hydrological applications tend to use non-informative priors, with the justification that this "lets the data speak for itself". Since prior knowledge of model parameters is usually case-specific, we will also use uniform priors, although nothing in the analysis precludes the use of informative prior distributions on $\boldsymbol{\theta}$.

In equations (12) and (13), $L(\cdot)$ is a likelihood function. This function must contain all functional dependencies of $p(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}} \mid \boldsymbol{\theta})$ on $\boldsymbol{\theta}$. It represents the likelihood of observing the data $\{\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}\}$ given the model parameters $\boldsymbol{\theta}$ and the model hypothesis. The form of $L(\cdot)$ must reflect the way error and uncertainty enter and propagate through the system. It is not necessary for $L(\cdot)$ to be a proper pdf and it does not have to integrate to unity – any constant factor can be absorbed into the proportionality relation in (12).

BAYESIAN ANALYSIS OF DATA UNCERTAINTY

Traditional Regression Methods: Hiding the Dirac Delta

Standard Bayesian regression ignores input uncertainty and lumps observed response error and model error into a single white noise term $\boldsymbol{\varepsilon}$ according to eqns (3) and (4). The likelihood function is then

$$L_{SLS}\left(\boldsymbol{\theta}, \sigma_y^2\right) = \frac{1}{\sigma_y^N} \exp\left(-\frac{1}{2\sigma_y^2} \sum_{n=1}^{N} \left[\tilde{y}_n - f_n\left(\tilde{\mathbf{X}}, \boldsymbol{\theta}\right) \right]^2 \right) \tag{14}$$

Assuming the non-informative invariant Jeffrey's prior $p(\sigma_y^2) \propto 1/\sigma_y^2$ and integrating σ_y^2 out of the posterior density yields the posterior pdf for the system parameters $\boldsymbol{\theta}$ [*Box and Tiao*, 1973]

$$p_{SLS}\left(\boldsymbol{\theta} \mid \tilde{\mathbf{X}}, \tilde{\mathbf{Y}}\right) \propto \left(\sum_{n=1}^{N} \left[\tilde{y}_n - f_n\left(\tilde{\mathbf{X}}, \boldsymbol{\theta}\right) \right]^2 \right)^{-\left(\frac{N-1}{2}\right)} \tag{15}$$

In practice, it is easier to maximize the logarithm of the posterior density, i.,e,

$$\log p_{SLS}\left(\boldsymbol{\theta} \mid \tilde{\mathbf{X}}, \tilde{\mathbf{Y}}\right) \propto -\left(\frac{N-1}{2}\right) \log\left(\sum_{n=1}^{N} \left[\tilde{y}_n - f_n\left(\tilde{\mathbf{X}}, \boldsymbol{\theta}\right) \right]^2 \right) \tag{16}$$

The variance of $\mathbf{Y}$ errors can then be estimated using

$$\sigma_y^2 \approx \frac{1}{N} \sum_{n=1}^{N} \left[\tilde{y}_n - f_n\left(\tilde{\mathbf{X}}, \hat{\boldsymbol{\theta}}\right) \right]^2 \tag{17}$$

where $\hat{\boldsymbol{\theta}}$ is the most probable parameter set.

Since the extrema of the posterior pdf (16) are identical to those of the SLS objective function (2), maximising (16) is equivalent to minimising the sum of squared errors. In fact, the input error model hidden in the classic least squares regression is the Dirac delta function $\delta(\cdot)$, which assigns all the probability mass to the observations $\tilde{\mathbf{X}}$:

$$p\left(\mathbf{X} \mid \tilde{\mathbf{X}}\right) = p\left(\tilde{\mathbf{X}} \mid \mathbf{X}\right) = \delta\left(\mathbf{X} - \tilde{\mathbf{X}}\right) = \begin{cases} \infty & \mathbf{X} = \tilde{\mathbf{X}} \\ 0 & \mathbf{X} \neq \tilde{\mathbf{X}} \end{cases} \tag{18}$$

or, in other words, $\mathbf{X} = \tilde{\mathbf{X}}$. Indeed, assuming that forcing uncertainty is statistically independent from the hydrological model parameters, i.e., $p(\tilde{\mathbf{X}} \mid \boldsymbol{\theta}) = p(\tilde{\mathbf{X}})$, substituting the Dirac function (18) and integrating over the support $\boldsymbol{\Omega}(\mathbf{X})$ of $\mathbf{X}$ yields

$$\begin{aligned} p\left(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}} \mid \boldsymbol{\theta}\right) &= p\left(\tilde{\mathbf{Y}} \mid \tilde{\mathbf{X}}, \boldsymbol{\theta}\right) p\left(\tilde{\mathbf{X}}\right) \\ &= C \int_{\boldsymbol{\Omega}(\mathbf{X})} p\left(\tilde{\mathbf{Y}} \mid \mathbf{X}, \boldsymbol{\theta}\right) p\left(\mathbf{X} \mid \tilde{\mathbf{X}}, \boldsymbol{\theta}\right) d\mathbf{X} \\ &= C \int_{\boldsymbol{\Omega}(\mathbf{X})} p\left(\tilde{\mathbf{Y}} \mid \mathbf{X}, \boldsymbol{\theta}\right) \delta\left(\mathbf{X} - \tilde{\mathbf{X}}\right) d\mathbf{X} \\ &= C p\left(\tilde{\mathbf{Y}} \mid \tilde{\mathbf{X}}, \boldsymbol{\theta}\right) = C p\left(\boldsymbol{\varepsilon} \mid \tilde{\mathbf{X}}, \boldsymbol{\theta}\right) \end{aligned} \tag{19}$$

where the fourth line follows from the properties of the Dirac function and $C = p(\tilde{\mathbf{X}})$ is independent of $\boldsymbol{\theta}$.

Since the response noise $\boldsymbol{\varepsilon}$ is assumed to be normally distributed according to (4), the likelihood function simplifies to (14). It follows that, since Bayesian SLS regression disregards input uncertainty via (18), it inherits all the weaknesses of SLS optimisation, in particular, parameter bias when error is present in both **X** and **Y** data, as well as non-robustness with respect to non-Gaussian errors. To a certain extent, however, the auto-correlation in the residuals can be addressed using AR methods.

Auto-Regressive (AR and ARMA) Methods

Auto-regressive (AR) and the more general ARMA models attempt to remove the auto-correlation structure from the residuals and reduce the latter to white Gaussian noise $\phi_n \sim N(0, \sigma_{AR}^2)$ by introducing additional parameters, e.g., for the AR1 model, $\varepsilon_n = a_1 \varepsilon_{n-1} + \phi_n$. *K*-order generalisations are readily obtained, e.g., the AR-*K* model

$$\varepsilon_n = \sum_{k=1}^{K} a_k \varepsilon_{n-k} + \phi_n \tag{20}$$

The correlation in the residuals can then be addressed by inferring the *K* parameters $\{a_k; k = 1 \dots K\}$ of the AR model in addition to the model parameters $\boldsymbol{\theta}$.

However, AR and ARMA models have the conceptual limitation that they do reflect the physical mechanism that induces the auto-correlation in model residuals in the first place. In rainfall-runoff modelling, the auto-correlation depends strongly on the process dominating the hydrological response (quickflow or slowflow) and it is cumbersome to introduce AR models that account for such distinctions (e.g., the AR parameters $\{a_k; k = 1 \dots K\}$ are effectively time- and process-dependent). Furthermore, since standard AR and ARMA methods do not explicitly introduce input error models, they suffer the same shortcomings as SLS methods when errors are present in both the **X** and **Y** data.

Total Least Squares Methods: A Bayesian Perspective

The Bayesian likelihood function corresponding to TLS methods is obtained by assuming Gaussian error models for both **X** and **Y** data

$$\tilde{\mathbf{X}} = \mathbf{X} + \boldsymbol{\varepsilon}_x \quad \varepsilon_x \sim N\left(0, \sigma_x^2\right) \tag{21}$$

$$\tilde{\mathbf{Y}} = \mathbf{Y} + \boldsymbol{\varepsilon}_y \quad \varepsilon_y \sim N\left(0, \sigma_y^2\right) \tag{22}$$

The limitation of TLS methods in hydrological modelling is that they assume that both input and output uncertainty can be described by independent Gaussian distributions, which is not always appropriate. Since the error models (21) and (22) are embedded in the TLS framework, the latter is not robust against deviations from normality and additive error forms (21) and (22). In addition, since TLS methods introduce one latent variable per each data point, the computational cost of a calibration may become prohibitive even with modern computer power.

However, TLS methods are significant in that they explicitly recognise input uncertainty and its impact on parameter estimation. A modern application of Bayesian TLS methods in pattern recognition [*Nestares et al.*, 2000] shows similarities to hydrological calibration, although the linearity of the models used in that study considerably simplifies the analysis and the approach is hence not readily applicable to nonlinear hydrological modelling.

Zellner [1971] offers an interesting Bayesian analysis of TLS methods (which he refers to as the Error-In-Variables Method, EVM). In particular, useful insights can be obtained by considering special cases of the EVM. Zellner [1971] shows that, even for linear models, the attempt to infer both the input variance and output variances leads to an ill-posed problem due to the likelihood function becoming unbounded. This implies that, in the hydrological context, TLS methods cannot estimate the variances of forcing and response errors σ_x^2 and σ_y^2. However, the inference of the ratio of the variances $\lambda = \sigma_x^2 / \sigma_y^2$ is well posed given an informative prior on either σ_x^2 or σ_y^2. These results show the subtlety of the apparently straightforward problem of data modelling, even using linear functions.

MULTI-OBJECTIVE CALIBRATION

An alternative approach to model calibration is to employ a composite objective function that contains terms corresponding to several distinct objectives. This approach mimics the manual calibration of hydrologists and is termed multi-objective (or multi-criteria) calibration. The following objectives are normally considered [*Sorooshian et al.*, 1993; *Gupta et al.*, 1998; *Madsen*, 2000]:

a) Correct flow volumes over the simulation, i.e., correct water balance;
b) Agreement in shape between the observed and predicted hydrographs;
c) Agreement of the peak flow characteristics: timing, rate, volume, etc.;
d) Agreement of recession limbs and low flow periods;

These objectives are expressed in numerical form and can be used to obtain Pareto solutions, reflecting various trade-offs between parameters and calibration criteria. Weighting factors can be used to obtain a composite objective function if a "globally optimal" single parameter set is required.

Multi-objective calibration is intuitive and conceptually simple. It recognises that different parts of the hydrograph can be subject to different error processes; hence it constitutes a conceptual advance compared with crude single-objective calibration methods such as SLS schemes.

However, we feel that multi-objective calibration methods have certain limitations:

1. The trade-off between various objectives is often unclear and there are no theoretical guidelines for the selection of weighting factors. The use of the entire Pareto front in a way circumvents the need for a single optimal parameter set, yet if predictions are necessary, particular parameter sets still have to be selected. A related pitfall is the possible correlation between the objectives. For example, minimising the discrepancy between observed and simulated responses would generally also improve the mass balance;
2. Multi-objective calibration is typically response focused and does not explicitly consider the influence of input errors. As a result, if the input history is corrupt and the model imperfect, this approach cannot in principle provide good fits and unbiased parameter estimates;
3. A fundamental limitation of multi-objective calibration from a Bayesian viewpoint is that it does not articulate an identifiable error model. It is therefore more difficult to appraise the validity of the inference procedure and provide uncertainty bounds on the parameter estimates (as these are strongly related to specific error models).

These shortcomings, in particular, the difficulty in obtaining confidence limits on parameters that can be used to obtain prediction limits on future events, undermine the range of application of multi-objective methods in hydrological modelling and forecasting.

GENERALISED LIKELIHOOD UNCERTAINTY ESTIMATION (GLUE)

The Generalised Likelihood Uncertainty Estimation (GLUE) methodology [*Beven and Binley*, 1992] was developed as a method for calibration and uncertainty estimation using generalised likelihood measures, and is related to the Generalised Sensitivity Analysis of Spear and Hornberger [1980]. The GLUE methodology explicitly recognises the fundamental limitations of simulating rainfall-runoff processes with contemporary hydrological models in data-sparse and data-corrupt applications. Its application so far has been predominantly in rainfall-runoff modelling [*Beven and Binley*, 1992], but GLUE has also been used to assess the uncertainty associated with predictions of land surface to atmosphere fluxes [*Franks and Beven*, 1997], geochemical modelling [*Zak et al.*, 1997] and flood inundation studies [*Romanowicz et al.*, 1994].

GLUE is based on Monte Carlo simulation, generating a large number of model runs with parameter sets sampled from a uniform probability distribution on prior parameter bounds. The likelihood of the parameter sets is then evaluated using a user-defined pseudo-likelihood measure, parameters with a pseudo-likelihood below a threshold are rejected, and the remaining likelihoods normalised to add up to 1. Next, at each time step, the predicted output from the retained runs are likelihood-weighted and ranked to form a cumulative distribution of response variables, from which quantiles can be selected to represent predictive uncertainty.

While GLUE is based on Bayesian conditioning, it does not articulate a specific error model structure - instead it embeds an unknown implicit error model within a suitably lenient pseudo-likelihood measure. As a result, all sources of uncertainty in GLUE manifest themselves as parameter uncertainty, giving rise to the concept of parameter equifinality, which admits multiple disjoint parameter sets that fit the observed data equally well. In the GLUE framework it is difficult to scrutinise and improve the uncertainty model underlying the inference. GLUE therefore lacks the conceptual rigour to address the challenges posed by error structures typified in Figure 1.

KALMAN FILTERS

The Kalman filter is a method widely used in electrical engineering and system analysis, primarily for linear Gaussian dynamics [*West and Harrison*, 1997], and has seen some use in hydrology [*Bras and Rodriguez-Iturbe*, 1985]. Kalman filters, at least in principle, explicitly specify the uncertainty in the system states that arises from imperfect process approximation and from data uncertainty.

The state-space formulation underlying Kalman filters is general and applicable to almost arbitrary models and uncertainty distributions. However, to obtain analytical closed-form solutions to the state estimation equations, the process is assumed to be linear with respect to the state variables and all errors are assumed to have a Gaussian distribution. A Bayesian interpretation of the classic Kalman filter follows.

Consider the discrete time stepping state-space model

$$\boldsymbol{\psi}_{n+1} = \mathbf{A}_n \boldsymbol{\psi}_n + \mathbf{L}_n \mathbf{u}_n + \mathbf{v}_n \qquad (23)$$

where $\boldsymbol{\psi}_n$ is the state vector at step n, $\mathbf{A}_n$ is the transfer matrix, $\mathbf{u}_n$ is a control vector, $\mathbf{L}_n$ is the control matrix and $\mathbf{v}_n$ represents model error. When the Kalman filter is used for simulation, $\boldsymbol{\psi}$ contains model state variables, e.g., simulated streamflow and internal fluxes. When used for calibration, $\boldsymbol{\psi}$ is augmented with the model parameters $\boldsymbol{\theta}$, giving rise to

extended Kalman filters. Often $\boldsymbol{\psi} \equiv \boldsymbol{\theta}$, with the transfer matrix reduced to the identity matrix.

In addition, consider the observation equation

$$\mathbf{z}_{n+1} = \mathbf{H}_{n+1}\boldsymbol{\psi}_{n+1} + \mathbf{w}_n \tag{24}$$

where $\mathbf{z}_{n+1}$ contains system observations (e.g., observed forcing inputs and responses), $\mathbf{H}_{n+1}$ is the observation matrix and $\mathbf{w}_{n+1}$ represents observation error.

If the observation history up to and including step n is stored in $\mathbf{Z}_n = \{\mathbf{z}_i; i = 1 \ldots n\}$ and a (prior) pdf $p(\boldsymbol{\psi}_n \mid \mathbf{Z}_n)$ of the state vector along with the pdf of the noise terms $\mathbf{v}_n$ and $\mathbf{w}_n$ are known at step n, the posterior pdf of the state vector at consequent steps can be constructed in two stages:

Prediction Step

Using total probability, the pdf of the state $\boldsymbol{\psi}_{n+1}$ is

$$p(\boldsymbol{\psi}_{n+1} \mid \mathbf{Z}_n) = \int_{\Omega(\boldsymbol{\psi}_n)} p(\boldsymbol{\psi}_{n+1} \mid \boldsymbol{\psi}_n, \mathbf{Z}_n)\, p(\boldsymbol{\psi}_n \mid \mathbf{Z}_n)\, d\boldsymbol{\psi}_n \tag{25}$$

which, given the Markovian property of (23), simplifies to

$$p(\boldsymbol{\psi}_{n+1} \mid \mathbf{Z}_n) = \int_{\Omega(\boldsymbol{\psi}_n)} p(\boldsymbol{\psi}_{n+1} \mid \boldsymbol{\psi}_n)\, p(\boldsymbol{\psi}_n \mid \mathbf{Z}_n)\, d\boldsymbol{\psi}_n \tag{26}$$

Conditioning Step

The observation can be processed using Bayesian updating

$$p(\boldsymbol{\psi}_{n+1} \mid \mathbf{Z}_{n+1}) = \frac{p(\mathbf{z}_{n+1} \mid \boldsymbol{\psi}_{n+1}, \mathbf{Z}_n)\, p(\boldsymbol{\psi}_{n+1} \mid \mathbf{Z}_n)}{p(\mathbf{z}_{n+1} \mid \mathbf{Z}_n)} \tag{27}$$

The Markovian property of the observation equation yields

$$p(\boldsymbol{\psi}_{n+1} \mid \mathbf{Z}_{n+1}) = \frac{p(\mathbf{z}_{n+1} \mid \boldsymbol{\psi}_{n+1})\, p(\boldsymbol{\psi}_{n+1} \mid \mathbf{Z}_n)}{p(\mathbf{z}_{n+1} \mid \mathbf{Z}_n)} \tag{28}$$

Analytical Solution of State Estimation Equations

In general, (26) and (28) do not possess closed-form analytical solutions. However, if the model and observation equations are linear in the state vector $\boldsymbol{\psi}$

$$\frac{\partial \mathbf{A}_n}{\partial \boldsymbol{\psi}_{n+1}} = \frac{\partial \mathbf{A}_n}{\partial \boldsymbol{\psi}_n} = \frac{\partial \mathbf{H}_n}{\partial \boldsymbol{\psi}_{n+1}} = \mathbf{0} \tag{29}$$

and the random variables follow a multi-Gaussian pdf

$$\begin{bmatrix} \boldsymbol{\psi}_n \\ \mathbf{v}_n \\ \mathbf{w}_n \end{bmatrix} \sim N\left(\begin{bmatrix} \hat{\boldsymbol{\psi}}_n \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_n & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R}_n \end{bmatrix} \right) \tag{30}$$

it is straightforward to derive stepwise exact solutions to (26) and (28) [*West and Harrison*, 1997].

Since data uncertainty in both forcing and response observations can be specified through the observation noise $\mathbf{w}$, while model error can be specified through $\mathbf{v}$, Kalman filters offer a very general framework for the estimation of system parameters accounting for all sources of uncertainty. Moreover, the Kalman filter requires only matrix multiplications and inversions and is consequently computationally fast. However, the derivation of the classic Kalman filter based on assumptions (29) and (30) highlights the limitations of this estimation scheme in rainfall-runoff model calibration.

Limitation 1 – Hydrological Models are Nonlinear

Selecting sufficiently small time steps in principle reduces the effect of model nonlinearity with respect to time and forcing – this is the approach used to control discretisation error in numerical DE solvers. However, it is the nonlinearity of the model with respect to the parameters $\boldsymbol{\theta}$ within the state vector $\boldsymbol{\psi}$ that undermines the validity of Kalman filtering in hydrological calibration. Although statistical and Taylor series linearisations of Kalman filters have been proposed, they are susceptible to nonlinear divergence, when the state estimates converge to incorrect values [*Bras and Rodriguez-Iturbe*, 1985].

Limitation 2 – Modelling Errors may not be Gaussian

The Kalman filter equations are derived by substituting Gaussian kernels into the forecast-update equations. The use of alternative error models requires the re-derivation of the Kalman filter equations and analytical solutions are unlikely in most cases.

Limitation 3 – Implementation of Kalman Filters

Casting hydrological models into the stepwise Kalman matrix forms is not always trivial, complicating the development of general model-independent analysis software. When using extended linearised Kalman filters, the derivatives are typically unavailable in closed form and numerical differentiation must be used.

Due to these and other problems, classic Kalman filters do not offer a complete solution to the calibration problem in hydrological modelling. However, recent developments in Monte Carlo and particle filters [*Carter and Kohn*, 1994;

Fruhwirth-Schnatter, 1994; *Cargnoni et al.*, 1997] suggest that there is scope for the application of these recent methods in hydrological parameter estimation. Indeed, the Monte Carlo computational tools allowing the generalisation of Kalman filters to nonlinear non-Gaussian models are related to the Monte Carlo Markov Chain tools that we will use to implement the generalised parameter inference with arbitrary error models latter in this chapter (albeit not in the stepwise iterated form characteristic of Kalman filters).

BAYESIAN TOTAL ERROR ANALYSIS (BATEA)

We have now surveyed all the major current calibration philosophies and see that, although inherent uncertainty in environmental systems is partially recognised (explicitly by Bayesian SLS and TLS methods and Kalman filters, implicitly by GLUE and MOCOM), limited attempt has been made to rigorously define this uncertainty using realistic error models. The (Bayesian) SLS regression avoids the specification of input error and suffers a parameter bias, TLS schemes account for special cases of input error but do not allow for non-Gaussian error models. GLUE offers a Monte Carlo algorithm for sampling from posterior parameter probability distributions, but does not make explicit the assumptions used to derive the likelihood function. Multi-objective calibration methods recognise the multi-criteria nature of calibration, yet lack the ability to provide confidence limits on parameter values and model predictions. Kalman filters can, at least in principle, explicitly specify error structure, yet linearisations and Gaussian assumptions undermine their stability properties and hence their suitability to environmental modelling. However, armed with an understanding of current limitations and insights provided by the analysis of TLS and Kalman methods, we can use the very general equation (11) and its simplified variant (13) to develop a family of methods that explicitly account for the various sources of system uncertainty. Whilst we present these methods from a Bayesian perspective, the parameter estimation equations become identical to those obtained from general maximum-likelihood theory when non-informative priors on the hydrologic parameters are employed.

Crude Input Error-Sensitive Approaches

One possible approach in rainfall-runoff modelling to extend the traditional regression framework is to augment the fitted parameter vector with rainfall depth multipliers. In the early approach of Kavetski et al. [2000], no likelihood function was specified for the fitted multipliers m, equivalent to setting

$$m \sim U(a,b) \tag{31}$$

where $U(\cdot)$ denotes the uniform distribution, and a and b are positive bounds, e.g., 0.1 and 10.0, reflecting uncertainty in the magnitude of precipitation depth errors.

Error models such as (31) are not new; a similar approach is implicitly used by the PDM model [*Lamb*, 1999] under the guise of a hydrological parameter (a single rainfall multiplier r_f for all time steps). Equation (31) implies that the informational content of rain gauges is limited to hyetograph shape only and that depth measurements have little or no influence on the hydrological parameter estimates, which is clearly an extreme statement. However, by making error structure assumptions such as (31) explicit, it becomes possible (and necessary) to scrutinise them, motivating our understanding of data corruption mechanisms.

Expected Likelihood Approach

When input uncertainty is included into the likelihood function, the following expression can be obtained

$$\begin{aligned} p\left(\tilde{\mathbf{X}},\tilde{\mathbf{Y}} \mid \boldsymbol{\theta}\right) &\propto L\left(\tilde{\mathbf{Y}} \mid \tilde{\mathbf{X}},\boldsymbol{\theta}\right) p\left(\tilde{\mathbf{X}} \mid \boldsymbol{\theta}\right) \\ &= \int_{\Omega(\mathbf{X})} L_p\left(\tilde{\mathbf{Y}} \mid \mathbf{X},\boldsymbol{\theta}\right) p\left(\mathbf{X} \mid \tilde{\mathbf{X}}\right) d\mathbf{X} \cdot p\left(\tilde{\mathbf{X}} \mid \boldsymbol{\theta}\right) \end{aligned} \tag{32}$$

where $\mathbf{X}$, the unknown true input, has been integrated out. As shown in (18) and (19), omitting the input error model is a special case of (32), equivalent to assigning $p(\mathbf{X} \mid \tilde{\mathbf{X}}) = \delta(\mathbf{X} - \tilde{\mathbf{X}})$. It leads to the SLS scheme and biased parameter estimates.

The numerical cost of evaluating (32) can be considerable, due to the high dimensionality of the integral. In Bayesian image analysis [*Nestares et al.*, 2000], a similar integral is evaluated by taking advantage of the multi-normal kernel in the equivalent of $p(\mathbf{X} \mid \tilde{\mathbf{X}})$ and a linear likelihood function $L_p(\cdot)$. However, since hydrological models are highly nonlinear, there is little hope in obtaining a closed-form solution of the integral (32), certainly not for arbitrary models. Although classical quadrature schemes based on hyper-lattices (e.g., the trapezoidal rule) could be employed to obtain approximate solutions of (32), these suffer an exponential deterioration of convergence rates in high-dimensional spaces and become computationally infeasible [*Evans and Swartz*, 2000]. Equi-distributed (quasi-random) numbers such as Sobol and Halton sequences raise the rate of convergence to $O([\ln K]^D/K)$, where K is the number of function evaluations and D is the number of dimensions of integration. Although asymptotically this rate is almost as fast as $O(K^{-1})$, it remains impractical in high dimensions [*Geweke*, 1996].

Instead, Monte Carlo (MC) schemes could be employed, since their asymptotic convergence rate is only weakly dependent on the dimensionality of the problem, $O(K^{-1/2})$. A simple MC integration algorithm for (32) is

$$L\left(\tilde{\mathbf{Y}} \mid \boldsymbol{\theta}, \tilde{\mathbf{X}}\right) = \frac{1}{K}\sum_{k}^{K} L_p\left(\tilde{\mathbf{Y}} \mid \boldsymbol{\theta}, \mathbf{X}_k\right) + O\left(K^{-1/2}\right) \qquad (33)$$

where $\mathbf{X}_k$ is the k^{th} sample from $p(\mathbf{X} \mid \tilde{\mathbf{X}})$. The effective dimension D of integration (33) depends on the input error model and is not necessarily equal to the number of data points. For example, if a single multiplier is used for each storm, then the dimension of integration will be equal to the number of storms in the calibration dataset.

As vividly illustrated by Hammersley and Handscomb [1964], it is possible to reduce the leading constant of the Monte Carlo error term by factors of 100,000's through modifications such as antithetic variates and sampling from orthonormal basis functions. Unfortunately these acceleration techniques exploit special features in the integrand and are cumbersome, if not impossible, in high dimensions. Further, there are hybrids of Monte Carlo schemes with quadrature methods that achieve the highest possible order of convergence, $O(K^{-G/D})$, where G is the highest bounded derivative of the integrand on the domain of integration [*Bahvalov*, 1959], but their efficacy depends on a set of continuity and variational constraints on the integrand that are hard to verify *a priori*. In spite of these challenges, adaptive Monte Carlo algorithms are available (e.g., the Vegas scheme of Lepage [*Press et al.*, 1992] and the stratification scheme of Press and Farrar [1990]) that could be more efficient than (33). Evans and Swartz [2000] and Fishman [1996] offer good summaries of stochastic integration methods.

Finally, although (32) allows the use of explicit input error models, it has the disadvantage that, after integrating the true input history out of the posterior pdf, it remains difficult to assess the suitability of the selected input error model $p(\mathbf{X} \mid \tilde{\mathbf{X}})$, since estimates of the true input history are not be explicitly available for inspection. As rigorous calibration demands *a posteriori* verification of its assumptions, we are led to reject the use of (32) in favour of a different method of incorporating input uncertainty.

Latent Variables as Subjects of Inference

An alternative to the integration of the true input history via (32) is its Bayesian estimation with the same logical status as the hydrological parameters. Consider the Bayesian inference equation, closely related to (11)

$$p\left(\boldsymbol{\theta}, \mathbf{X} \mid \tilde{\mathbf{X}}, \tilde{\mathbf{Y}}\right) \propto L\left(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}} \mid \boldsymbol{\theta}, \mathbf{X}\right) p\left(\boldsymbol{\theta}, \mathbf{X}\right) \qquad (34)$$

where $L(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}} \mid \boldsymbol{\theta}, \mathbf{X})$ is the joint likelihood of observing $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ given a hydrological parameter vector $\boldsymbol{\theta}$ and the true input history $\mathbf{X}$. The advantage of using (34) instead of (32) is that the posterior pdf of the true forcing data (e.g., rainfall) becomes available and can be used to verify the data and the assumed probability models.

The joint likelihood $L(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}} \mid \boldsymbol{\theta}, \mathbf{X})$ can be re-formulated to maximally separate the probability models for input and response error. Conditional probability yields

$$L\left(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}} \mid \boldsymbol{\theta}, \mathbf{X}\right) \propto L\left(\tilde{\mathbf{Y}} \mid \tilde{\mathbf{X}}, \boldsymbol{\theta}, \mathbf{X}\right) p\left(\tilde{\mathbf{X}} \mid \boldsymbol{\theta}, \mathbf{X}\right) \qquad (35)$$

Two further simplifications can be made, based on the following assumptions:

1. $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ are statistically independent, i.e., $\tilde{\mathbf{Y}}$ depends only on the observation error affecting the response and on the <u>true</u> forcing (see Figure 1). The errors affecting the forcing do not causally affect the true response and its associated observation error;
2. $\tilde{\mathbf{X}}$ is statistically independent of $\boldsymbol{\theta}$, e.g., rainfall sampling errors are uncorrelated with the hydrological model parameters.

Subject to the above assumptions and allowing for un-normalised probability models of forcing uncertainty, (35) simplifies to

$$L\left(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}} \mid \boldsymbol{\theta}, \mathbf{X}\right) \propto L\left(\tilde{\mathbf{Y}} \mid \boldsymbol{\theta}, \mathbf{X}\right) L\left(\tilde{\mathbf{X}} \mid \mathbf{X}\right) \qquad (36)$$

Substituting (36) into the posterior pdf (34) yields

$$p\left(\boldsymbol{\theta}, \mathbf{X} \mid \tilde{\mathbf{X}}, \tilde{\mathbf{Y}}\right) \propto L\left(\tilde{\mathbf{Y}} \mid \boldsymbol{\theta}, \mathbf{X}\right) L\left(\tilde{\mathbf{X}} \mid \mathbf{X}\right) p\left(\boldsymbol{\theta}, \mathbf{X}\right) \qquad (37)$$

Pdf (37) contains the inference of the model hydrological parameters <u>and</u> the true rainfall history $\mathbf{X}$. It requires the specification of the input error model $L(\tilde{\mathbf{X}} \mid \mathbf{X})$ and the response error model $L(\tilde{\mathbf{Y}} \mid \boldsymbol{\theta}, \mathbf{X})$. The analysis of pdf (37) is computationally intensive due to the large dimensionality of the parameter space (now augmented with the latent variables of the forcing error model).

Computational Implementation – The Metropolis Algorithm

A couple of decades ago the formulation of a posterior pdf such as (32) or (37) would have been satisfying from a theoretical viewpoint yet practically useless – there were no effective approaches to examine such distributions, let alone determine their moments and other characteristics. We suspect that it is precisely the formidable computational aspect of (32) and (37) that has hindered the development of a rigorous parameter estimation framework with realistic uncertainty models. Indeed, classic sampling methods such as acceptance-rejection and importance schemes are

difficult and inefficient for complicated and high-dimensional pdfs [*Fishman*, 1996].

Fortunately, the development of Monte Carlo Markov Chain (MCMC) methods offers practical ways to sample from probability distributions of considerable complexity and dimensionality. Fishman [1996], Gelman et al. [1997] and Evans and Swartz [2000] provide a good overview of MCMC methods, which originated in nuclear physics (see Metropolis et al. [1953] for historical background) and, in the 1990's, enjoyed an explosive growth in areas as diverse as econometrics [*Geweke*, 1996], biology [*Gelman et al.*, 1997] and hydrology [*Kuczera and Parent*, 1998; *Bates and Campbell*, 2001].

In this work, we implement the Metropolis algorithm [*Chib and Greenberg*, 1995; *Gilks et al.*, 1996; *Gelman et al.*, 1997] in order to generate samples from pdf (37) and summarise the posterior parameter pdf using moments and histograms. It must be emphasised, however, that any other sampler can be used, e.g., importance schemes.

The theory behind MCMC methods is complex and still evolving. It has been described in hydrological context by Kuczera and Parent [1998] and Bates and Campbell [2001]. Our implementation of the Metropolis algorithm follows

1. Sample $\{\boldsymbol{\theta}, \mathbf{X}\}_i$ from a pre-specified symmetric jump distribution, here the multi-normal pdf centred on the current sample location;
2. Evaluate $p(\{\boldsymbol{\theta}, \mathbf{X}\}_i \mid \tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$ up to a constant;
3. Accept the new sample $\{\boldsymbol{\theta}, \mathbf{X}\}_i$ with probability given by $r = \min[\, p(\{\boldsymbol{\theta}, \mathbf{X}\}_i \mid \tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) \big/ p(\{\boldsymbol{\theta}, \mathbf{X}\}_{i-1} \mid \tilde{\mathbf{X}}, \tilde{\mathbf{Y}}), 1]$.

Multiple Markov chains are started at the posterior mode of (37) found using the SCE algorithm. After sufficient samples have been collected (the termination criteria of Gelman et al. [1997] were used), the posterior distribution of the hydrological parameters $\boldsymbol{\theta}$ and the true forcing parameters $\mathbf{X}$ can be examined using histograms and scatter plots, or summarised using moments and quantiles.

We stress that although the entire input history appears as the random variable in the posterior pdf, the actual number of additional latent variables will depend on the particular error model used. For example, if storm-wise multipliers are assumed, the number of latent variables will equal the number of storms in the calibration data. It follows that the dimensionality of the pdf support is not necessarily prohibitive. The ability to control, to a certain extent, the number of latent variables and hence the dimensionality of the state space, is an important advantage of the generalised BATEA framework over classical TLS schemes, which always introduce N additional latent variables.

Block updating of the sampled variables is advantageous when strong correlation is suspected between these variables [*Gelman et al.*, 1997; *Bates and Campbell*, 2001]. We suspect that considerable computational optimisation can be carried out by a judicious selection of jump distributions and updating sequences. Given enough sampling, however, all these algorithmic variations converge to the same target distribution (37).

CASE STUDY: THE ABC MODEL

The theoretical arguments calling for the more systematic calibration formalism BATEA can be illustrated by numerical experimentation.

Consider the ABC model, a simple time stepping hydrological model with three parameters $\{a, b, c\}$, two state variables (discharge Q and storage S, both per unit catchment area) and one forcing term (rainfall r). For any time step i, the ABC equations are given as

$$Q_i = (1-a-b)\, r_i + c S_i \tag{38}$$

$$S_{i+1} = (1-c)\, S_i + a r_i \tag{39}$$

Parameter a represents the fraction of rainfall entering the groundwater storage, b is the evaporation fraction and c is a constant of linear proportionality between storage and discharge. In addition, the initial value S_1 is required.

While the ABC model is simplistic and certainly uncompetitive with more complex models, it does have an elementary groundwater store and a quickflow component. We employ ABC to illustrate some of the issues common to all hydrological models. In addition, we use synthetic data to establish the "bona fides" of the calibration methods. If real data were used, there would be no way of checking whether the parameter estimates converge to the true values and whether the confidence limits reflect the actual errors.

The "true" rainfall data $\mathbf{X}$ was generated using the DRIP algorithm [*Heneker et al.*, 2001] using parameters corresponding to the Sydney region in Australia. The "true" streamflow $\mathbf{Y}$ was generated using the ABC model with the "true" parameter set $\boldsymbol{\theta}$ = {0.6 0.15 0.2}. The $\mathbf{X}$, $\mathbf{Y}$ and $\boldsymbol{\theta}$ data represent the "truth" that will be used to gauge the success of the parameter estimation scheme.

The "observed" rainfall $\tilde{\mathbf{X}}$ was obtained by corrupting each storm depth r within $\mathbf{X}$ using normally distributed storm multipliers, yielding the corrupted depth $\tilde{r}$

$$\tilde{r}_j = m_j r_j; \qquad m \sim N\left(0, \sigma_m^2\right) \forall j \tag{40}$$

where j indexes the storms within the rainfall series.

The "observed" streamflow was obtained as

$$\tilde{y}_i = y_i + \varepsilon_i; \qquad \varepsilon_i \sim N\left(0, \sigma_y^2\right) \forall i \tag{41}$$

where i indexes the time steps within the simulation.

The additive response error model (41) is the same as that used in the SLS and TLS schemes. The multiplicative input

error model (40), although related to the additive TLS input error model, is nonetheless different. It offers an attractive way to parsimoniously parameterise rainfall errors. For example, if the storm largely misses a gauge, its overall temporal pattern will register on the pluviograph, but the true strength will be underestimated. Conversely, if the core of the storm passed directly over the rain gauge, the effective catchment-averaged precipitation would be overestimated.

It is relatively straightforward to accommodate various probabilistic error models within the BATEA model analysis formalism. In contrast SLS and TLS methods, as well as classic Kalman filters, have embedded error models that can not be easily modified.

Unless stated otherwise, $N = 1000$ time steps with $\Delta t = 1$ hr were used (42-day runs with 5 storms). Quantitative results for $\sigma_m^2 = 0.05$ and $\sigma_y^2 = 0.01$ are presented; in general, qualitatively similar behaviour occurred for other range of σ_m^2 and σ_y^2. In addition, we found no major qualitative differences in calibration behaviour when using alternative input error distributions, e.g., log-normal, instead of normal, multipliers.

The ABC model is calibrated to the "observed" data using a) the SLS scheme; and b) the BATEA formalism. Comparisons are then made, focusing on i) the accuracy of the parameter estimates; and ii) the relation between the estimated uncertainty and the actual parameter errors. Unless synthetic data were used, these assessments, which in our opinion must be applied to any parameter estimation scheme, would have been impossible.

Although typical implementations of SLS schemes are limited to fitting the model parameters using an optimisation scheme, we have performed the additional step of sampling from the posterior SLS parameter distribution (15) using the Metropolis algorithm. This allows a more comprehensive comparison of the posterior parameter distributions inferred by the SLS and BATEA schemes, since the mode of the posterior pdf corresponds to the best-fit parameters, while the shape and spread of the distribution quantify the parameter uncertainty.

Five Metropolis chains with 40,000 samples in each were generated and the first 10,000 samples were discarded. The *r*-statistic of Gelman et al. [1997] was monitored to ascertain the chains' convergence to the target posterior pdf. Unless stated otherwise, the posterior covariance matrices of the parameters were approximated from the Metropolis samples, since inverse finite difference Hessian approximations at the posterior mode are more "local" and hence less informative.

We also define a dimensionless error measure $\eta(\boldsymbol{\theta})$ for calibrated parameters.

$$\eta(\boldsymbol{\theta}) = \frac{\boldsymbol{\theta}_{true} - \mathrm{E}[\boldsymbol{\theta}]}{\sigma[\boldsymbol{\theta}]} \tag{42}$$

where $\mathrm{E}[\boldsymbol{\theta}]$ and $\sigma[\boldsymbol{\theta}]$ are the mean and standard deviation of the posterior parameter pdf.

Bayesian SLS Calibration

The SLS-fitted response for a 50-hr segment of the data is shown in Figure 3. Visually, the fit is good in practically all sections of the hydrograph. Undoubtedly the parameter estimates must be good! However, the reader, by now perhaps sceptical of the suitability of SLS schemes in the presence of input error, will not be surprised to learn that, while parameter *c* has been fitted quite well, parameters *a* and especially *b* are poorly identified. Figures 4-5 and Table 3 summarise the SLS calibration results, showing that the SLS estimates of *a* and *b* are ten standard deviations away from the true values, well outside any meaningful confidence limits.

We stress that in this study no model error was introduced and only moderate unbiased data corruption was used (e.g., the actual response noise corresponded exactly to the additive Gaussian noise in the SLS response error model). The fact that poor estimates were obtained for *a* and b, while good results were obtained for *c*, illustrates the unpredictability of parameter inference based on inappropriate models. We are forced to conclude that, given 42 days of data (1000 data points), the SLS calibration fails to identify the correct parameters even when model error is not present.

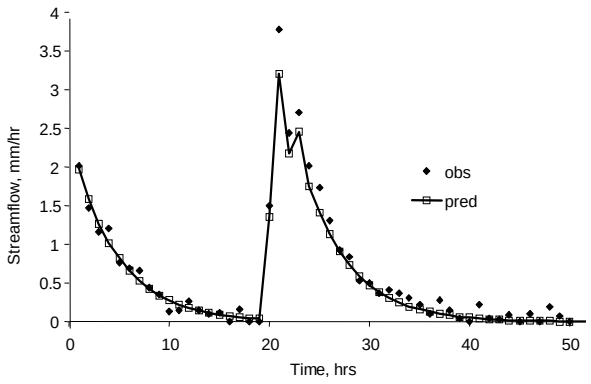

Figure 3. Portion of SLS-calibrated ABC hydrograph.

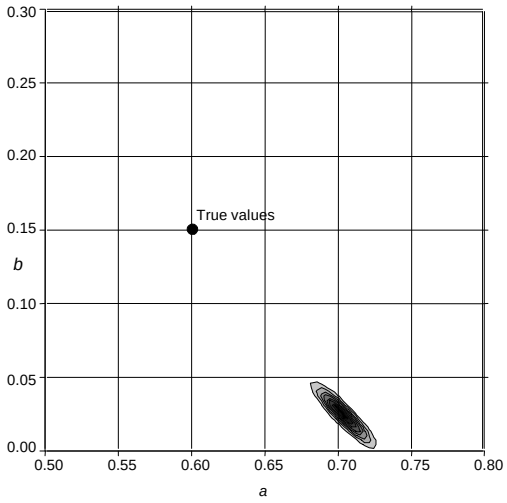

Figure 4. SLS posterior pdf of parameters *a* and *b*.

Table 3 – SLS calibration, *N*=1000 (42 days).

Parameter	Mean ± St. deviation	Error $\eta(\boldsymbol{\theta})$
a	0.702 ± 0.0106	-9.62
b	0.0250 ± 0.0105	11.9
c	0.198 ± 0.00413	0.484

Another aspect of the ABC model can be seen in Figure 4, which shows a moderately strong correlation between the fitted parameters *a* and *b*. In general, correlation indicates that the model is over-determined. A reasonable model validation question then is: *is the correlation between* a *and* b *an intrinsic feature of the model or a consequence of data errors*? Due to the simple functional form of the ABC model, this question can be answered analytically. However, for a more complicated model, the solution to this question must be found using some approximate numerical or experimental means – i.e., calibration. But if the SLS scheme does not even admit the possibility of input errors, how can it meaningfully tell the modeller whether parameter correlation is due to model structure or data uncertainty?

SLS Calibration – More Data for Accuracy?

Parameter errors are often blamed on the limited amount of data available for calibration. Indeed, the large parameter errors in Table 3 could have been explained by the fact that only one month of rainfall-runoff data was used. In this synthetic case study, however, we are free to test this statement, since the "truth" is completely known. Extending the data series to 7 months (5000 time steps) and then to 10 yrs (100,000 time steps) and calibrating ABC using the SLS scheme yield the posterior distributions of parameters *a* and *b* shown in Figures 6 and 7.

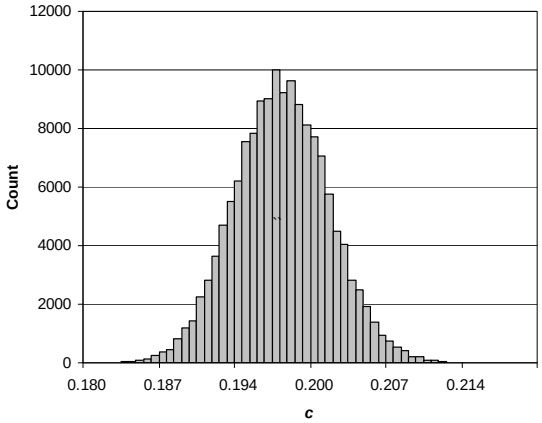

Figure 5. SLS posterior pdf of parameter *c*.

Figures 6-7 suggest that large data sets are unlikely to resolve the input error problem in SLS calibration. While the variance of all three parameters has contracted by two orders of magnitude (following the asymptotic $O(N^{-1})$ variance decay rate of SLS schemes), the convergence is to incorrect values. Although the optimal parameter values calibrated using 10 years of data are closer to the true values than those obtained using 1 and 7 months of data, the uncertainty estimates are extremely misleading and the parameter estimates converge to biased values. This bias is a serious shortcoming of the SLS scheme – if the approach cannot handle simple synthetic data error in such a simple hydrologic model (with $N = 10^5$ data points), we must be wary of applying it in the real world. However, the non-robustness of SLS schemes with respect to input error is not a mystery – analytic results dating back to the 1940's show parameter bias even for a 1-parameter straight line fit!

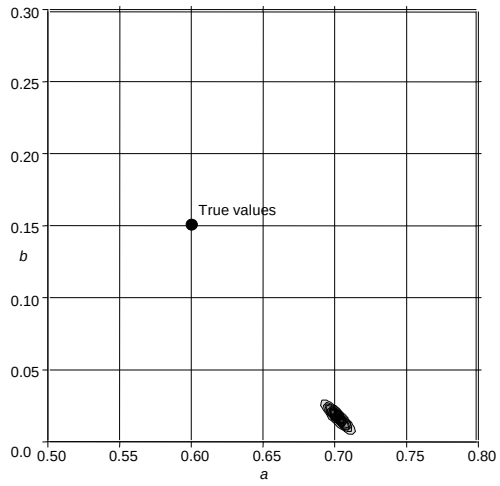

Figure 6. SLS posterior pdf of *a* and *b* with *N*=5000.

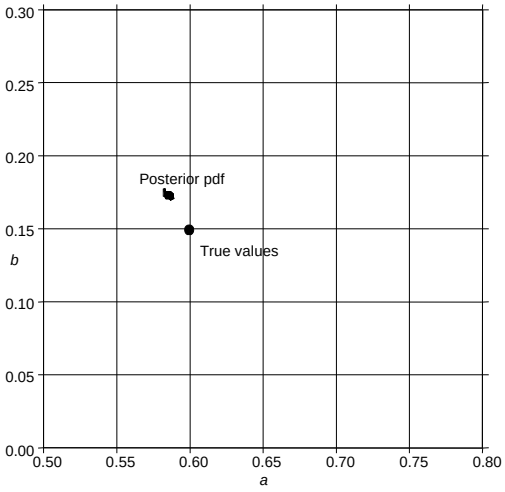

Figure 7. SLS posterior pdf of *a* and *b* with *N*=100000.

BATEA Calibration Using Diffuse Multiplier Error Model

In this part of the case study, we deliberately mis-specify the input error model, assuming a diffuse uniform likelihood for all values of storm multipliers between 0.1 and 10, as in (31). The results of the parameter inference are shown in Table 4. The Metropolis analysis of the posterior distribution failed to converge and the standard deviation reported in Table 4 was computed using the inverse finite difference Hessian approximation at the posterior mode.

Table 4 – BATEA calibration, diffuse multiplier model, *N*=1000.

Parameter	Mean ± St. deviation	Error $\eta(\boldsymbol{\theta})$
a	0.636 ± 0.00872	-4.13
b	0.094 ± 0.00981	5.71
c	0.199 ± 0.00300	0.153

It can be seen that the introduction of a different incorrect input error model led to little if any reduction of the actual parameter error for all three parameters of the ABC model (although the posterior mode has shifted). These results illustrate potential limitations of ad-hoc approaches that introduce additional model parameters in an attempt to account for rainfall data errors (e.g., the PDM model [*Lamb*, 1999]).

The poor performance of the diffuse multiplier model can be understood by noting that statistical inference extracts information (parameter values) from data (here, rainfall-runoff series) using some set of rules (likelihood functions and prior distributions). Specifying a diffuse uniform likelihood on the storm multipliers instructs the inference scheme to disregard all information on precipitation depth contained in $\tilde{\mathbf{X}}$ and limits the use of the rainfall series to relative hyetograph shape only. The multipliers then become completely unconstrained degrees of freedom that at best merely intensify the computational effort and at worst destabilise the inference algorithm.

The insight of Zellner [1971] into the inability of TLS schemes to operate with non-informative prior distributions on the parameters of the error models is also valuable. Setting a diffuse uniform likelihood on the storm multipliers is equivalent to using the (correct) error model (40) with $\sigma_m^2 = \infty$ and is similar to using a non-informative prior on σ_x^2 in the Bayesian TLS scheme. Since the prior on response noise variance σ_y^2 is also non-informative, it is little surprise that the parameter accuracy is poor and the Metropolis scheme fails to converge – the inference is ill-conditioned.

Power of BATEA: Correct Uncertainty Characterisation

The preceding empirical analysis confirms that simplistic treatment of data uncertainty fails to produce accurate and reliable parameter estimates. The Bayesian SLS schemes assumed input uncertainty is described by the Dirac function and the diffuse multiplier model was also incorrect. Both these schemes are specific cases of BATEA, but with incorrect error models for the particular problem. It was this error mis-characterisation that was responsible for the poor calibration results – not the data errors themselves, nor the quantity of data used, nor the model (which we know is exact in this synthetic case study).

To demonstrate the performance of the BATEA formalism, we specify the correct input error model – the Gaussian multiplicative depth error at each storm with the correct value of σ_m^2 in the input error model (some additional comments on selecting and verifying σ_m^2 are made later). The results are shown in Figures 8 and 9, with a summary in Table 5. Figure 10 and Table 6 show the results of applying BATEA to the 7-month data set.

Table 5 – BATEA calibration, correct error model, *N*=1000.

Parameter	Mean ± St. deviation	Error $\eta(\boldsymbol{\theta})$
a	0.593 ± 0.0528	0.125
b	0.155 ± 0.0751	-0.0625
c	0.199±0.00305	0.141

Table 6 – BATEA calibration, correct error model, *N*=5000.

Parameter	Mean ± St. deviation	Error $\eta(\boldsymbol{\theta})$
a	0.604 ± 0.0267	-0.135
b	0.146 ± 0.0376	0.120
c	0.199 ± 0.00150	1.38

Figures 8-10 confirm that errors and biases in parameter estimates obtained earlier were not a product of poor data or model inadequacy – they were direct results of applying calibration schemes that did not correctly represent the way errors enter and propagate through the system. As soon as

the correct error model is specified, the bias in the parameter estimates disappears. In fact, Table 5 indicates that the parameter estimates inferred by BATEA with 40 days of data are considerably closer to the true values than the parameters calibrated with the SLS scheme using 10 years of data. Although the BATEA posterior parameter uncertainty has increased compared to the SLS case, the mean and mode of the posterior pdf are close to the true values and the error statistic is of the order of 1.0, much below the SLS case. BATEA is hence more "honest" than SLS in reporting parameter uncertainty. At least for this test case, the trade-off between SLS vs BATEA could be described as "precisely wrong vs probably correct"!

As discussed previously, the shift in calibration philosophy from ignoring data errors to rigorously defining and treating them incurs a substantial computational cost – the necessity to estimate the latent variables of the error model (in this case, the storm multipliers). This computational effort is not at all worthless, however, since we chose to implement (37) rather than (32). Following the calibration using (37), the posterior distribution of the latent variables is available for explicit inspection. In particular, fitted distribution of the latent variables should resemble the error model that was used to describe the corruption process. In our synthetic data case, we can go further and compare the posterior distribution of the multipliers to the known "true" values.

Figure 11 shows a 2D histogram of two of the five storm depth multipliers. The true values were $m_4 = 1.34$ and $m_5 = 1.11$. Figure 11 shows that the multipliers have been estimated with a notable degree of accuracy. The remaining multipliers were also estimated with small errors and with realistic uncertainty bounds. In a practical context, the values of the multipliers (and all other latent variables) must be examined to determine whether they are realistic. If the posterior pdf of latent variables is unreasonable, then either the calibration scheme has failed, or the rainfall-

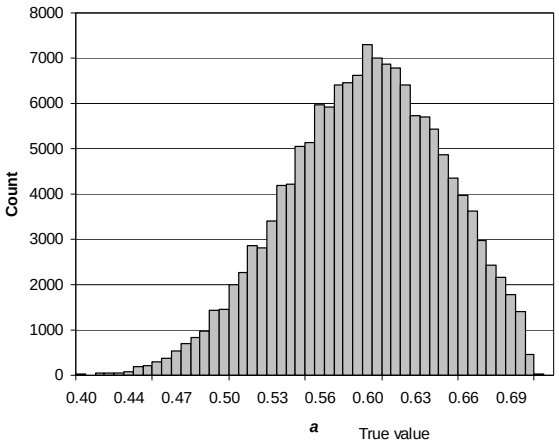

Figure 8. BATEA posterior pdf of *a* with *N*=1000.

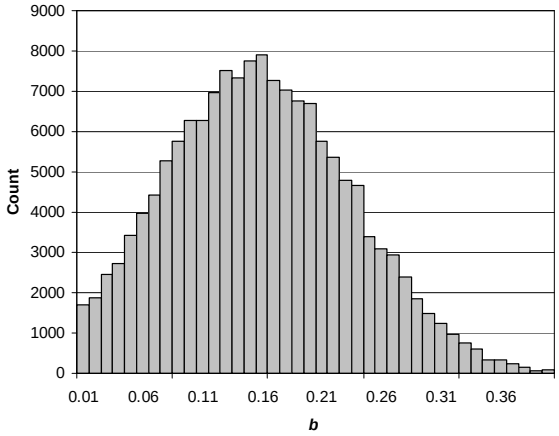

Figure 9. BATEA posterior pdf of *b* with *N*=1000.

sampling network has produced data inadequate for the calibration of rainfall-runoff models, or the model itself is inadequate.

Another interesting point arises by considering that rainfall depth corruption must necessarily disturb the water balance of the catchment model. Input error insensitive methods, including SLS, GLUE and multi-objective schemes, will attempt to restore the mass balance by adjusting (biasing) the parameters. By including the depth error into the analysis, it becomes possible to directly address the cause of such mass balance errors, in a sense implicitly giving the BATEA calibration methodology multi-objective characteristics.

We also stress that good fits obtained using the SLS scheme (e.g., Figure 3) do not imply that the model predictions in response to future forcing will be correct. Since rainfall data errors were ignored by the SLS scheme, the SLS parameter estimates have been adjusted to cancel

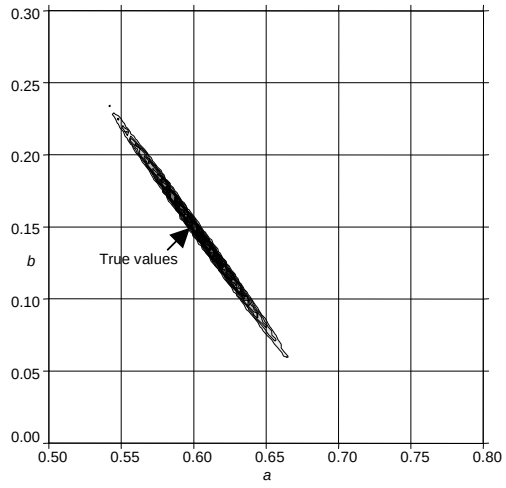

Figure 10. BATEA posterior pdf of *a* and *b* with *N*=5000.

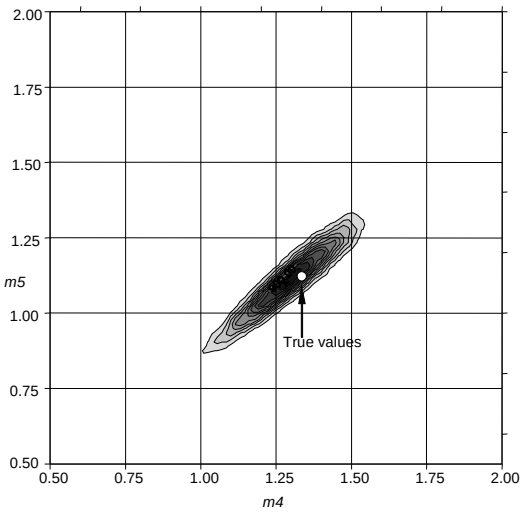

Figure 11. BATEA posterior pdf of multipliers with *N*=5000.

out (to the largest extent possible) the input errors in the particular calibration data set. However, it would be naïve to expect that future predictions in response to different forcing would benefit from such "cancellation". It is preferable to explicitly address (in a probabilistic sense) all sources of data uncertainty to maximally remove the influence of data errors from the parameter estimates. The reason for biased SLS parameter estimates is not the model itself, but rather unaccounted rainfall errors distorting the mass balance of the calibrated model.

We are also now in a position to empirically answer one of the earlier questions on the ABC model – is the correlation between *a* and *b* an intrinsic feature of the model or an artefact of data error? SLS schemes could not reliably answer that question, as they do not admit input errors. Observing that the correlation between *a* and *b* does not vanish when the correct error models are specified (Figure 10), we conclude that the correlation is an intrinsic feature of the model. Analytic assessment confirms this empirical observation.

Validation of Data Uncertainty Models

The reader by now might be asking: *the results in Tables 5 and 6 are promising, but how do we tell when the chosen data error model is appropriate?* Indeed, *a posteriori* verification of modelling assumptions is an essential step in model calibration: omitting it throws us back to ad-hoc modelling techniques. Since there are criteria that can be used to reject hypotheses, data uncertainty models can and must be scrutinised through seeking invalidating evidence, not necessarily used in the inference itself. For example, if the rainfall data was collected in a densely gauged catchment, but the inference suggests large rainfall errors, then a flawed data error model must be suspected.

A powerful rejection criterion is the compatibility between the posterior pdf of the latent variables and their corresponding likelihood functions. For example, in the final BATEA case study we assumed that the multipliers can be described by the Gaussian pdf with mean of 1.0 and variance of 0.05. We should therefore inspect the posterior distribution of the multipliers. Although it is hard to confirm trends from 5 and 28 storm multiplier samples, it is reasonable to expect some basic compatibility. The posterior mean and variance of the fitted multipliers are 1.01 and 0.052 for the 1000-step 5-storm BATEA simulation and 0.99 and 0.057 for the 5000-step 28-storm BATEA calibration. These results are consistent with the rainfall error model used in the calibration. A more rigorous test is to examine the distribution of the multipliers, e.g., using probability plots. Care must be exercised, however, since probability plots based on small sample sizes can depart from linearity even if the data indeed came from that distribution. Figure 12 presents a normal probability plot for the posterior means of the multipliers in the 28-storm calibration. Its comparison with the assumed Gaussian multiplier likelihood function is satisfactory.

The error models do not exist in a vacuum – if the rainfall data came from a densely gauged catchment, we would reduce the variance of the multipliers σ_m^2 and inspect the results. In the limit as the rainfall data becomes more exact, the multiplier likelihood would approach the Dirac function and we would arrive at the original SLS scheme. It is not the SLS scheme itself that is faulty; rather, it is its use to calibrate models to corrupt input data that is inappropriate.

Finally, we note that the input error model is not the only model that needs to be assessed. The response error model embedded in SLS and also employed in our ABC case study using BATEA is additive white noise, which is not necessarily appropriate. The verification of the response uncertainty model is also essential in a holistic calibration and can be carried out, e.g., via graphic analysis of

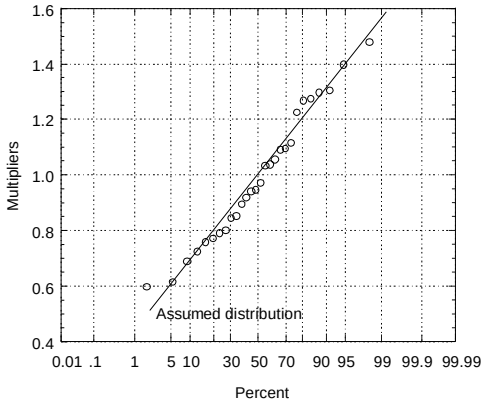

Figure 12. Normal probability plot of multipliers, *N*=5000.

residuals. The convenience of the additive Gaussian noise for response is largely mathematical, as it allows analytical integration of the output error variance σ_y^2 from the posterior pdf. Since BATEA employs MCMC methods to infer latent variables, the analytical convenience of the Gaussian residual assumption is useful yet no longer essential. However, since the literature describes many response error models (e.g., AR and ARMA models), but largely overlooks input uncertainty (with the exception of the TLS scheme, which is not used in hydrology and, at least in principle, Kalman filters, which are rarely used), we have focused on the latter to restore the balance.

A VIEW INTO THE FUTURE

We recognise that BATEA is still in its nascency and a range of outstanding issues remain:

a) The current ability to specify probabilistic data and model error models is limited: uncertainty functions must be chosen to reflect our physical understanding of processes contributing to input, response and model corruption. Such functions do not yet exist.
b) Parametric approaches for model error quantification are currently lacking. Nonetheless, the recent work by Gaganis and Smith [2001] evidences the growing recognition that explicit analysis of model error can and should be incorporated into environmental studies.
c) The numerical performance of the inference algorithm is of evident practical importance. Whilst the increased number of latent variables can be managed using Monte Carlo methods, the theory and practice of these algorithms are themselves still evolving. Much work remains to be done to refine sampling convergence criteria and develop more efficient jump rules and state update algorithms.

However, these issues must be kept in perspective – we argue that fundamental and overriding benefits arise from developing a framework that explicitly recognises data and model uncertainty and that produces greater insight about how uncertainty affects our worldview.

A common concern with hydrological modelling is that there appears to be a limit to the model complexity supported by rainfall-runoff data. However, the informational insufficiency of data series to support specific model hypotheses should be demonstrated via excessively wide confidence limits on calibrated parameters, preferably when data uncertainty and its effect on the inference is minimised. Although the parameter confidence limits in Bayesian SLS methods and GLUE are also indirectly related to model and data uncertainty, the lack of explicit error models in these approaches makes that link ambiguous, since confidence limits obtained using incorrect error models will have erroneous significance levels. BATEA, however, is in no way limited to calibration using rainfall-runoff series alone. The data matrices **X** and **Y** contain all the observations of the system – saturated areas, evapotranspiration, piezometer heads and so on. The inclusion of this data into the inference tends to constrain parameter uncertainty as it provides independent information on the model's internal state variables [*Franks et al.*, 1998]. The inclusion of these variables is, in fact, an avenue for more stringent model assessment [*Beven*, 1993; *Kuczera and Franks*, 2002].

CONCLUSIONS

Reliable parameter inference is critical for meaningful prediction using environmental models. Yet the calibration of these models is currently accomplished using essentially ad-hoc approaches. For example, standard least squares calibration, although computationally convenient, disregards input uncertainty and suffers from a bias in the estimated parameters. GLUE recognises the inherent uncertainty in modelling, yet does not treat it explicitly, complicating the verification of its underlying assumptions. Multi-objective methods attempt to mimic the intuitive multi-criteria approach of hydrologists, yet, without specifying explicit error models, suffer from subjectivity in the selection of objective functions and cannot produce meaningful confidence limits on their predictions. Currently, only Kalman filters allow specification of model and data error, yet their reliance on system linearisation and Gaussian error models undermines their suitability in hydrological modelling.

Recognising the necessity to account for all sources of uncertainty in environmental modelling, a general Bayesian approach for total error analysis (BATEA) was introduced. It makes explicit the probabilistic error models used to describe the uncertainty in the observed data, notably, in forcing inputs such as rainfall (overlooked by current calibration schemes), as well as in the observed responses. The fundamental benefit of the BATEA formalism is that it forces the modeller to explicitly specify uncertainty models that can and must be verified against available evidence. Ignoring data and model uncertainty during calibration not only subtracts from the veracity of parameter estimates and predictions, but also may lead to stagnation in rainfall-runoff hydrology if hypotheses of catchment dynamics are only weakly challenged, with any discrepancies between observations and predictions attributed to the currently nebulous concept of "data and model error". Conversely, accounting for all sources of error leads to an honest assessment of parameter and predictive uncertainty of hydrologic models and paves the way for rational discrimination between competing models on the basis of explicit criteria, e.g., model error and susceptibility to data uncertainty.

REFERENCES

Bahvalov, N. S., On approximate calculation of multiple integrals, *Vestnik Moscov. Univ. Ser. Mat. Meh. Astr. Fiz. Him.*, 4, 3-18, 1959.

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

Beven, K. J., Prophecy, reality and uncertainty in distributed hydrological modelling, *Advances in Water Resources*, 16, 41-51, 1993.

Beven, K. J. and A. M. Binley, The future of distributed models: model calibration and uncertainty prediction, *Hydrological Processes*, 6, 279-298, 1992.

Beven, K. J. and M. J. Kirkby, A physically based variable contributing area model of basin hydrology, *Hydrological Science Bulletin*, 24, 43-69, 1979.

Beven, K. J., R. Lamb, P. F. Quinn, R. Romanowicz and J. Freer, TOPMODEL, *in Computer Models of Watershed Hydrology*, edited by V. P. Singh, pp. 627-628, Water Resources Publication, 1995.

Box, G. E. P. and G. C. Tiao, *Bayesian Inference in Statistical Analysis*. Wiley, New York, 1973.

Bras, R. L. and I. Rodriguez-Iturbe, *Random Functions in Hydrology*. Addison-Wesley Publishing Co, 1985.

Cargnoni, C., P. Muller and M. West, Bayesian forecasting of multinomial time series through conditionally Gaussian dynamic models, *Journal of the American Statistical Association*, 92, 587-606, 1997.

Carter, C. K. and R. Kohn, On Gibbs sampling for state-space models, *Biometrika*, 81, 541-553, 1994.

Chappell, N. A., S. W. Franks and J. Larenus, Multi-scale permeability estimation for a tropical catchment, *Hydrological Processes*, 12, 1507-1523, 1998.

Chib, S. and E. Greenberg, Understanding the Metropolis-Hastings Algorithm, *The American Statistician*, 49, (4), 327-335, 1995.

Deming, W. E., *Statistical Adjustment of Data*. Wiley, NY, 1943.

Duan, Q., S. Sorooshian and V. K. Gupta, Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resources Research*, 28(4), 1015-1031, 1992.

Evans, M. and T. Swartz, *Approximating Integrals Via Monte Carlo and Deterministic Methods*. Oxford University Press, Oxford, 2000.

Fishman, G. S., *Monte Carlo: Concepts, Algorithms, and Applications*. Springer-Verlag, 1996.

Franks, S. W. and K. J. Beven, Bayesian estimation of uncertainty in land surface-atmosphere flux predictions, *Journal of Geophysical Research*, 102(D20), 23, 23991-23999, 1997.

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

Fruhwirth-Schnatter, S., Data augmentation and dynamic linear models, *Journal of Time Series Analysis*, 15, 183-202, 1994.

Gaganis, P. and L. Smith, A Bayesian approach to the quantification of the effect of model error on the predictions of groundwater models, *Water Resources Research*, 37(9), 2309-2322, 2001.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin, *Bayesian Data Analysis*. Chapman and Hall, London, 1997.

Geweke, J., Monte Carlo simulation and numerical integration, pp. *in Handbook of Computational Economics*, edited by H. Amman, D. Kendrick and J. Rust, North-Holland, Amsterdam, 1996.

Gilks, W. R., S. Richardson and D. J. Spiegelhalter (ed.), *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London, 1996.

Gull, S. F., Bayesian data analysis: straight-line fitting, *in Maximum entropy and Bayesian methods*, edited by J. Skilling, pp. 511-518, Kluwer Academic Publishers, Dordtrecht, 1989.

Gupta, H. V., S. Sorooshian and P. O. Yapo, Towards improved calibration of hydrologic models: multiple and non-commensurable measures of information, *Water Resources Research*, 34(4), 751-763, 1998.

Hammersley, J. M. and D. C. Handscomb, *Monte Carlo Methods*, New York, 1964.

Heneker, T. M., M. F. Lambert and G. Kuczera, A point rainfall model for risk-base design, *Journal of Hydrology*, 247, 54-71, 2001.

Jefferys, W. H., On the method of least squares, *The Astronomical Journal*, 85(2), 177-181, 1980.

Kavetski, D., S. W. Franks and G. Kuczera, Calibration of hydrologic models: The role of input errors, *in XIII International Conference on Computational Methods in Water Resources*, edited by L. R. Bentley, J. F. Sykes, C. A. Brebbia, W. G. Gray and G. F. Pinder, pp. 503-510, A. A. Balkema, Calgary, Canada, 2000.

Kuczera, G. and E. Parent, Monte Carlo assessment of parameter uncertainty in conceptual catchment models: The Metropolis algorithm, *Journal of Hydrology*, 211, 69-85, 1998.

Kuczera, G. A. 1994. NLFIT: A Bayesian nonlinear regression program suite. Dept. of Civ. Eng. and Surv, University of Newcastle, Australia.

Kuczera, G. A. and S. W. Franks, Testing hydrologic models: Fortification or falsification?, *in Mathematical Modelling of Large Watershed Hydrology*, edited by V. P. Singh and D. K. Frevert, pp. 141-185, Water Resources Publications, Littleton, Co., 2002.

Lamb, R., Calibration of a conceptual rainfall-runoff model for flood frequency estimation by continuous simulation, *Water Resources Research*, 35(10), 3103-3114, 1999.

Macdonald, J. R. and W. J. Thompson, Least-squares fitting when both variables contain errors: Pitfalls and possibilities, *American Journal of Physics*, 60(1), 66-73, 1992.

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

Metropolis, N., A. W. Rosenbluth, A. H. Teller and E. Teller, Equations of state calculations by fast computing machines, *Journal of Chemical Physics*, 21, 1087-1091, 1953.

Nestares, O., D. J. Fleet and D. J. Heeger, Likelihood functions and confidence bounds for total-least-squares-problems, *in IEEE Conference on Computer Vision and Pattern Recognition*, pp. 523-530, 2000.

Press, W. H. and G. R. Farrar, Recursive stratified sampling for multidimensional Monte Carlo integration, *Computers in Physics*, 4, 190-195, 1990.

Press, W. H., B. P. Flannery, S. A. Teukolsky and W. T. Vetterling, *Numerical Recipes in Fortran-77: The Art of Scientific Computing*. Cambridge University Press, 1992.

Romanowicz, R., K. J. Beven and J. Tawn, Evaluation of predictive uncertainty in non-linear hydrological models using a Bayesian approach, *in Statistics for the Environment, II, Water Related Issues*, edited by V. Burnett and K. F. Turkman, pp. 297-317, Wiley, New York, 1994.

Singh, V. P. (ed.), *Computer Models of Watershed Hydrology*, Water Resources Publication, 1995.

Sorooshian, S., Q. Duan and V. K. Gupta, Calibration of rainfall-runoff models: Application of global optimization to the Sacramento Soil Moisture Accounting Model, *Water Resources Research*, 29(4), 1185-1194, 1993.

Spear, R. C. and G. M. Hornberger, Eutrophication in Peel Inlet, II: Identification of critical uncertainties via Generalised Sensitivity Analysis, *Water Resources Research*, 14, 43-49, 1980.

West, M. and J. Harrison, *Bayesian Forecasting and Dynamic Models, 2nd Edition*. Springer-Verlag, New York, 1997.

Zak, S., K. J. Beven and B. Reynolds, Uncertainty in the estimation of critical loads: A practical methodology, *Soil, Water and Air Pollution*, 98, 297-316, 1997.

Zellner, A., *An Introduction to Bayesian Inference in Econometrics*. John Wiley and Sons, New York, 1971.