Bayesian analysis of input uncertainty in hydrological modeling: 2. Application

Dmitri Kavetski, George Kuczera, and Stewart W. Franks

Received 20 June 2005; revised 11 October 2005; accepted 25 October 2005; published 8 March 2006.

The Bayesian total error analysis (BATEA) methodology directly addresses both input and output errors in hydrological modeling, requiring the modeler to make explicit, rather than implicit, assumptions about the likely extent of data uncertainty. This study considers a BATEA assessment of two North American catchments: (1) French Broad River and (2) Potomac basins. It assesses the performance of the conceptual Variable Infiltration Capacity (VIC) model with and without accounting for input (precipitation) uncertainty. The results show the considerable effects of precipitation errors on the predicted hydrographs (especially the prediction limits) and on the calibrated parameters. In addition, the performance of BATEA in the presence of severe model errors is analyzed. While BATEA allows a very direct treatment of input uncertainty and yields some limited insight into model errors, it requires the specification of valid error models, which are currently poorly understood and require further work. Moreover, it leads to computationally challenging highly dimensional problems. For some types of models, including the VIC implemented using robust numerical methods, the computational cost of BATEA can be reduced using Newton-type methods.


1. Introduction

Environmental models are a useful tool in understanding and predicting the behavior of hydrological systems. However, models of the current generation, whether conceptual or physical, are only approximations, possibly crude, of the very complex and heterogeneous natural environment. This makes their parameters difficult to determine a priori, necessitating calibration to make the model broadly consistent with the observations. Typically, hydrological models are calibrated by fitting the streamflow predicted by the model forced with observed precipitation to the observed discharge. However, both precipitation and discharge are subject to considerable sampling and measurement uncertainty. The situation is likely to be particularly severe for rainfall measurements, since rainfall fields are highly variable in time and space [Zawadski, 1973], yet are often agglomerated into a single areal average during calibration. We expect the significance of data and model errors to remain considerable in the foreseeable future.

The companion paper [Kavetski et al., 2006c] developed the comprehensive Bayesian total error analysis (BATEA) framework for the analysis of data uncertainty in hydrological modeling. In an earlier paper, Kavetski et al. [2002] also provided a critique of current parameter estimation methods for conceptual rainfall-runoff models, focusing on their performance when the accuracy of the observed forcing (precipitation) is poor. The chief motivation for BATEA is to systematically account for input uncertainty. The term “systematically” is used to emphasize that the BATEA techniques are not based on ad hoc modifications of fixed-regressor-type methods, but provide a consistent framework that allows the modeler to introduce as much detailed information about the data accuracy as is known. For example, BATEA allows (indeed, requires) different parameter estimation equations for catchments that are densely gauged versus sparsely gauged catchments. This cannot be achieved with current methods since they do not explicitly incorporate the modeler’s confidence in the forcing data. Moreover, BATEA provides a systematic framework where competing catchment and data uncertainty models are compared. In this paper, it is assumed that the precipitation errors are largely consistent within a single storm and take a multiplicative Gaussian form. This error model is then assessed empirically and some alternatives are suggested.

This paper applies BATEA to two North American catchments from the Model Parameter Estimation Experiment (MOPEX) database [Schaake et al., 2001]. The performance of BATEA is illustrated by means of case studies that focus on model predictions and inferred parameter distributions, as well as on posterior checks of the inference assumptions. Such checks are essential to identify any violations of the assumptions and error models underlying Bayesian inference, yet are often omitted in hydrological modeling. We also discuss the ability of BATEA-type methods to differentiate between failures of the catchment model and failures of the parameter...
estimation method and its underlying assumptions. Finally, since BATEA introduces a large number of additional latent variables to identify inconsistencies between the model and the observed data (both forcing and response), it requires more attention to computational details. An additional objective of this paper is to demonstrate that BATEA-type modeling with hundreds of storms is becoming realistic with standard computing resources, provided the model is implemented using stable numerical methods and the parameter distributions are analyzed using methods that exploit their smoothness.

2. Theoretical Background

2.1. Mathematical Form of Hydrological Models

Hydrological models, whether lumped or distributed, physical or conceptual, can be represented as

\[ Y = h(\theta, X) \]  

where the matrix \( Y \) contains the time series of model responses (e.g., runoff, saturated areas, groundwater levels, etc.) to forcing \( X \) (e.g., rainfall, ET, etc.). The model hypothesis is represented by a deterministic or stochastic function \( h : X \rightarrow Y \) closed by the parameter vector \( \theta \). We use the terms “input” and “forcing”, as well as “output” and “response”, interchangeably.

[5] Given the model (1), a parameter prior \( p(\theta) \) and observed data \( D = \{X, Y\} \), Bayesian analysis seeks the posterior probability distribution function (pdf) of the model parameters \( p(\theta | X, Y, h) \). For conciseness, we omit explicit dependencies on the model hypothesis \( h \) in the probability distributions in remainder of the paper.

2.2. Standard Least Squares (SLS) Regression

[7] Least squares methods are common in all areas of scientific and engineering modeling. When interpreted from a Bayesian viewpoint, Gaussian additive output errors lead to the following pdf

\[ p(\theta | \tilde{X}, \tilde{Y}) \propto SS_m(\theta | \tilde{X}, \tilde{Y})^{-\frac{N}{2}} \]  

where the sum-of-squared-errors kernel \( SS_m(\cdot) \) is

\[ SS_m(\theta | \tilde{X}, \tilde{Y}) = \sum_{n=1}^{N} [\tilde{y}_n - h_\theta(\tilde{X})]^2 \]  

where \( n \) indexes the observations (totalling \( N \)). The criterion (3) also generalizes to multiple responses [e.g., Theil, 1971].

[8] Least squares methods seek the parameters that, given the model forced with observed data, give the closest fit to the observed responses in the Euclidean norm (2). Although attractive because of their simplicity, the major problem with SLS methods in hydrology is that the exact forcing data (primarily, precipitation, but also ET) is unknown. Many current alternatives to SLS are based on different objective functions, yet have the common feature that the model is always forced with the observed data. Whenever the latter is corrupt, this can severely bias parameter estimates [Kavetski et al., 2002].

2.3. Accounting for Forcing Uncertainty Using BATEA

[9] The BATEA formalism accounts for input uncertainty by introducing additional terms that describe the reliability of the observed forcing data. The general form of BATEA is derived in the companion paper as follows

\[ p(\theta, \phi, \beta_x, \beta_y | \tilde{X}, \tilde{Y}) \propto p(\tilde{Y} | \theta, \phi, \beta_x, \beta_y) p(\theta | \phi, \beta_x) p(\phi | \beta_y) p(\beta_y) \]  

[10] A key feature of BATEA is that additional latent variables \( \phi \) are included in the inference equations to map the observed forcings into “true” forcings. This allows posterior estimation of data uncertainty and, to the extent that the specified data uncertainty models are meaningful, reduces parameter bias occurring if the model is forced with the corrupt observed data. The modeler’s a priori confidence in the data is reflected in the prior distribution \( p(\phi | \beta_y) \); the error models, parameterized by \( \{\beta_x, \beta_y\} \), must be specified by modelers on the basis of their understanding of the data quality. The companion paper considered several input error models. In this paper, we examine the performance of the Gaussian storm multiplier model with unknown variance.

2.4. Gaussian Storm Multiplier Model With Unknown Variance

[11] Assuming that storm depth measurements are affected by systematic error different for each storm leads to the storm multiplier error model. If it is further assumed that the \( N_m \) multipliers \( m \) are approximately Gaussian with known mean \( \mu_m \) and unknown variance \( \sigma_m^2 \), described by a vague inverse gamma prior, the following objective function is obtained

\[ p(\theta, m | \tilde{X}, \tilde{Y}, \mu_m, \nu_0, \lambda_0^2) \propto SS_m(m) + \nu_0 \lambda_0^2 \nu_0^{-\frac{1}{2}} SS_y(\theta, X, Y)^{-\frac{\nu_0}{2}} \]  

where \( \sigma_m^2 \) has been integrated out to reduce the dimensionality of (5). The additional sum-of-squares term \( SS_m(\cdot) \) reflects the contribution of the input multipliers

\[ SS_m(m) = \sum_{i=1}^{N_m} (m_i - \mu_m)^2 \]  

[12] In this study the prior multiplier mean \( \mu_m = 1 \), reflecting the a priori belief that precipitation depth measurements are, on average, unbiased. The parameters \( \{\nu_0, \lambda_0^2\} \) of the inverse gamma prior reflect the modeler’s confidence in the forcing data; their selection is discussed in a later section. Note that the output sum-of-squares \( SS_y(\cdot) \) is based on the model forced with the estimated exact data \( X \), obtained from the observed precipitation \( X \) and the multipliers \( m \).

[13] Integrating pdf (5) over \( \theta \) yields the marginal posterior distribution of the multipliers \( \theta \), which plays an important role in model calibration and comparison using BATEA. A major advantage of the BATEA equations (4) and (5) is that the modeler can reflect the confidence in the forcing data directly and explicitly; conversely, the SLS equation (2) does not contain any terms reflecting the accuracy of the forcing series. Note that SLS estimates can be obtained as a special case of BATEA with \( \sigma_m^2 = 0 \) (or as \( \nu_0 \rightarrow \infty \)), implying the forcing is known exactly. BATEA calibration hence seeks the best fit of both the forcing and
response series; conflicts between the two are resolved by the user-prescribed input/output error models.

2.5. Computational Methods for BATEA

[14] In this study, the posterior distributions are analyzed using the Metropolis Monte Carlo algorithm [Kuczera and Parent, 1998]. Given the samples of the hydrological model parameters \( \mathbf{\theta} \) and latent variables (the multipliers \( \mathbf{m} \)), the prediction limits are computed by ranking the sampled discharges at each step (90% limits are used in this paper). The numerical results in this paper are based on 4 parallel Metropolis chains of 5000 samples each, retained after 1000 discarded “burn-in” samples. The reported modal values of \( \mathbf{\theta} \) and \( \mathbf{m} \) were computed using a Newton-type algorithm maximizing the logarithm of the posterior distributions (2) and (5). The gradients at the reported optima were verified to be small using adaptive finite difference approximations. For some runs, the optima were also verified using the SCE method, which is a stochastic simplex-based optimization algorithm that doesn’t assume continuity of the objective function [Duan et al., 1992]. These measures reduce the risk of converging to possible secondary local optima (which could delay the convergence of the Metropolis samples or yield incorrect results).

2.6. BATEA: Using Versus Inferring Uncertainty Models

[15] BATEA offers, for the first time in hydrology, a formal Bayesian framework where the modeler can experiment with different types of error models for both the input and output data. Since it is the first method to have the ability to systematically account for input uncertainty, an inevitable difficulty in its early applications is the poor understanding of the input corruption processes. While BATEA does not require a precise model of precipitation uncertainty (vague priors on the variance of the multipliers can be used), a reduction in the amount and precision of the information about the data accuracy (particularly in the forcing data) generally leads to increased parameter uncertainty and a reduction in the confidence in the obtained results. This occurs because there is an increased reliance on the catchment model to provide feedback on the data accuracy and uncertainty models, which is not necessary when the data accuracy is known. In the latter case, it becomes possible to directly identify model error as the discrepancy between the observations and predictions (both forcing and response) that cannot be explained by data sampling uncertainty alone. This is fundamentally impossible when data accuracy is unknown. It is also envisaged that further work is required to develop realistic input uncertainty models that will supersede the simplistic stormwise Gaussian multiplier model introduced in this study. The current difficulties with applying BATEA must, however, be contrasted with treating the observed precipitation as exact, which is known to be a questionable assumption.

3. Hydrological Models: Variable Infiltration Capacity and Snow Models

[16] This paper employs the Variable Infiltration Capacity (VIC) model, a widely used conceptual rainfall-runoff model [Wood et al., 1992; Liang et al., 1994]. Here, a single-bucket version of the VIC model is cast into “saturated path model” (SPM) form to allow robust numerical implementation and facilitate its statistical analysis [Kavetski et al., 2003]

\[
\frac{dS}{dt} = P_L \left( 1 - \frac{S}{S_{\text{max}}} \right) - k_b \left( \frac{S}{S_{\text{max}}} \right)^\beta - e_p \left( 1 - \left( \frac{S}{S_{\text{max}}} \right)^\gamma \right)
\]

(7)

where \( S(t) \) is the catchment soil moisture storage [L] at time \( t \) [T]. The model is forced by the rainfall input \( P_L \) [L/T] and potential evapotranspiration (ET) \( e_p \) [L/T]. It has 5 parameters: the maximum storage \( S_{\text{max}} \) [L], the base flow time constant \( k_b \) [L/T], and the exponential parameters \( \alpha \), \( \beta \) and \( \gamma \), which introduce additional flexibility (and nonlinearity) into the model. The VIC model (7) is a nonlinear ODE and is implemented using the fixed-step implicit Euler method to ensure unconditional stability while maintaining model smoothness with respect to parameters and forcings [Kavetski et al., 2006a].

[17] Since the catchments analyzed in this study experience snowfall, the VIC model is supplemented with a snow preprocessor. The degree-day model is used, which assumes a linear dependence between the melting rate and the air temperature [e.g., Martinec, 1975; Hock, 2003]. Since the original degree-day model contains several internal thresholds that yield highly discontinuous objective functions, the following smoothed version is used [Kavetski et al., 2006a]

\[
\frac{dS_s}{dt} = P_s - M(T, T_{\text{mel}}, \theta_0, S_s)
\]

(8)

where \( S_s \) is the snow store [L], \( T \) is the air temperature [temp], \( M = \max(0,\theta_0(T - T_{\text{mel}}), 0) \) is the melting rate and \( P_s \) is the snow input [L/T] defined by a smoothed step function partitioning the total precipitation \( P \)

\[
P_s = \frac{P}{2} - \frac{P}{\pi} \arctan \left( \frac{T - T_{\text{mel}}}{m_1} \right)
\]

(9)

[18] The snow model (8)–(9) has 2 parameters: the melting temperature \( T_{\text{mel}} \) [temp], which controls whether snow accumulation or melting occurs, and the “degree-day” factor \( \theta_0 \) [L/T/temp] that defines the linear temperature melt relation. Since the precipitation data in this study has an hourly resolution, the snow model is also applied at an hourly step.

[19] The selection of the smoothing kernel \( f_s(S_s, m_2) \) in (8) and the smoothed step function (9) is discussed by Kavetski et al. [2006a]. The auxiliary parameters \( m_1 \) and \( m_2 \) control the numerical smoothness of the model and are not calibrated. The modified snow model (8)–(9) gives predictions that are consistent with the original model, but are smooth and more stable.

[20] The numerical smoothing of the catchment models is especially beneficial for BATEA, since the introduction of latent variables (the storm multipliers \( \mathbf{m} \)) leads to high-dimensional optimization and uncertainty estimation problems. For such problems, smoothness of the objective function is extremely beneficial, allowing efficient analysis using Newton- and Hessian-type derivative-based methods.

4. Case Study 1: French Broad River

[21] The French Broad River basin at Asheville, North Carolina has an area of about 2450 km², and is classified as
mainly closed shrubland (including cropland and wooded grassland), with minor mixed forests and 5% urbanization. The soil type is loam, as well as sandy/clay loam. Three progressively overlapping time periods beginning in 1/1/1960 are considered: period A is 300 days long, period B is 600 days long and period C is 1400 days long. These periods are chosen to examine the stability of parameter estimates and prediction limits obtained for the VIC model using SLS and BATEA. The precipitation data is hourly and obtained from 5 hourly gauges, whereas the discharge series are daily. A basic summary of the data for the three periods is listed in Table 1.

### 4.1. Standard Least Squares Analysis

[22] The standard least squares (SLS) calibration of the VIC model to period A is shown in Figure 1. The parameter values and their uncertainty estimates are shown in Table 2 and were computed using the Metropolis algorithm. Similar results would be obtained from alternative objective functions solely on the basis of response errors. For example, the Nash-Sutcliffe efficiency often used in hydrology is a linear function of $SS_r(-)$ and hence yields identical optima to SLS calibration. By varying the factor and power of $SS_y$, (e.g., as in GLUE [Franks and Beven, 1997]), it is possible to widen or contract the prediction limits, yet the modal and median (and sometimes the mean) predictions are not affected by such transformations. Figure 1 shows a considerable mismatch of the model predictions and the observed data, particularly for the first and third peaks ($t \approx 50$ and $t \approx 220$ days), which is hardly solved by merely expanding the prediction limits to envelope the observations. The qualitative mismatch could be explained by (1) discharge errors, (2) model errors, (3) precipitation errors, or (4) combinations of these factors. The SLS method is specifically geared toward error sources (1) and (2) provided there are no strong outliers in the output series. Inspection of the discharge data does not reveal any points that could be classified as obvious single outliers (unless one counts entire storms as outliers), yet the residual plots are highly non-Gaussian. The remaining explanation for the mismatch, that is, that the precipitation data contains significant errors, cannot be assessed using calibration methods that assume the inputs are known, but can be explored using the BATEA methodology.

[23] Another possible weakness of SLS can be seen in Figure 1, where the prediction limits are obtained by adding Gaussian noise (with variance computed from residual errors) to the model predictions (which incorporate parameter uncertainty). Whenever the process and error models are poorly specified, this results in the prediction limits not corresponding to realistic model runs (for example, the upper bound seems much too high). GLUE rectifies this problem to some extent, since GLUE prediction limits are obtained from model simulations directly, with no additional white noise. However, since it lumps all uncertainty onto the parameters, it can give misleading results when the poor model performance is simply a consequence of poor data (particularly forcing). It is emphasized that the problems with SLS in this case are a consequence of the error model being inappropriate, and analysis of residual errors clearly shows significant non-Gaussian structure. In addition, the magnitude of the output residuals arguably exceeds the sampling uncertainty associated with streamflow measurements (e.g., there are errors of almost 50% in the prediction of peak flows for the major storms).

[24] The next step in the analysis consists of doubling the length of data used in the calibration. Table 2 shows the results of SLS calibration to 600 days of data. The fit is qualitatively similar, with the Nash-Sutcliffe index remaining constant at 0.76–0.74. Although this value is relatively high for hydrological simulations, the quality of the fit remains poor for the majority of the storm peaks. Since the VIC model is mass conservative and hence requires an increased input in order to generate an increase in the output, it is reasonable to suspect that these problems arise because of precipitation sampling errors. A regression method disregarding input uncertainty then tries to accommodate the discrepancy by varying (biasing) the model parameters, particularly those affecting the mass balance of the model. VIC-type conceptual models have several such components, including the ET functions and the snow model. While for the 300-day simulation, the base flow parameter $k_b$ is estimated at its upper bound $10^5$ mm/hr, which is an extreme value. The exponential parameters are also quite sensitive; for example, it can be seen from Table 2 that the base flow parameter $\beta$ is pushed to its upper bound for both the 300- and 600- day simulations. Although in principle the upper bound of 10 ($1.0 \text{ in } \log_{10} \text{ space}$) is not necessary, it is useful to identify degenerations in the estimation of the model structure, since simulations with $\beta = 10$ and, for example, $\beta = 100$ are virtually indistinguishable (the effect of $\beta$ is asymptotically small for very large or very small $\beta$). Similar comments apply to the quick flow exponent $\alpha$. A fairly common occurrence with the VIC model observed in these and other studies is its tendency to use its exponential parameters to increase the variability in its fluxes. It remains to be seen whether such behavior is reasonable, and whether it is due to input errors, or to model inadequacy, or both.

[25] It is stressed that poor model identifiability is a different issue to poor model structure. For example, the extreme behavior of the exponents in the VIC model does not necessarily imply that the model itself is definitely wrong. Rather, it may indicate that a single forcing-response series is unable to provide sufficient information about two nonlinear stores (the snow store in the degree-day model and the SPM store in the rainfall-runoff model). The situation is particularly problematic with the SPM store, since three pathways (quick flow, base flow and ET) combine to produce a single system state (total catchment storage) and response (streamflow). Analysis of the Monte Carlo samples shows a strong correlation of 0.97 between

---

**Table 1. Summary of Data for the French Broad River Catchment**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Period A</th>
<th>Period B</th>
<th>Period C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total precipitation depth</td>
<td>1195.0</td>
<td>2307.0</td>
<td>5279.0</td>
</tr>
<tr>
<td>Total discharge depth</td>
<td>702.0</td>
<td>1295.0</td>
<td>2901.0</td>
</tr>
<tr>
<td>Total ET (potential depth)</td>
<td>760.0</td>
<td>1429.0</td>
<td>3224.0</td>
</tr>
</tbody>
</table>

*All depths in mm.*
Table 2. French Broad River SPM-VIC Parameters Obtained by SLS Calibration to Periods A, B, and C

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean (A) ± SD (Mode)</th>
<th>Mean (B) ± SD (Mode)</th>
<th>Mean (C) ± SD (Mode)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \log_{10} S_{\text{max}} )</td>
<td>+0.19 ± 0.11 (+0.26)</td>
<td>−0.21 ± 0.052 (−0.15)</td>
<td>−0.22 ± 0.056 (−0.16)</td>
</tr>
<tr>
<td>( \log_{10} \alpha )</td>
<td>−0.66 ± 0.14 (−0.60)</td>
<td>−1.44 ± 0.035 (−1.46)</td>
<td>−1.24 ± 0.065 (−1.32)</td>
</tr>
<tr>
<td>( \log_{10} k_b )</td>
<td>+3.24 ± 0.97 (+4.00)</td>
<td>−0.56 ± 0.034 (−0.56)</td>
<td>−0.12 ± 0.11 (−0.20)</td>
</tr>
<tr>
<td>( \log_{10} \beta )</td>
<td>+0.95 ± 0.045 (+1.00)</td>
<td>+0.93 ± 0.058 (+1.00)</td>
<td>+0.87 ± 0.077 (+0.97)</td>
</tr>
<tr>
<td>( \log_{10} \gamma )</td>
<td>−0.50 ± 0.14 (+0.54)</td>
<td>−0.17 ± 0.034 (−0.20)</td>
<td>+0.016 ± 0.065 (−0.063)</td>
</tr>
<tr>
<td>( \log_{10} b_{def} )</td>
<td>−1.14 ± 0.037 (−1.15)</td>
<td>−1.16 ± 0.056 (−1.18)</td>
<td>−1.16 ± 0.063 (−1.17)</td>
</tr>
<tr>
<td>( T_{\text{melt}} )</td>
<td>+2.28 ± 0.26 (+2.18)</td>
<td>+2.08 ± 0.41 (+2.12)</td>
<td>+1.72 ± 0.47 (+1.75)</td>
</tr>
</tbody>
</table>

*Period A: 300 days, Nash–Sutcliffe (NS) index 0.76; period B: 600 days, NS index 0.74; period C: 1400 days, NS index 0.73.*

The base flow parameter \( k_b \) and the soil depth parameter \( S_{\text{max}} \), indicating the inference becoming ill posed. Generally, most of the parameters tend to be correlated with \( S_{\text{max}} \) by the virtue of the influence of \( S_{\text{max}} \) on every single model component: quick flow, base flow and ET. The snow store is also behaving somewhat anomalously, with the estimated melting temperature \( T_{\text{melt}} \) ~ 2°C. However, this could also be due to the location of the temperature gauges.

Adding another year of data to the analysis changes some of the findings yet again (Table 2). The exponential parameters \( \alpha \) and \( \beta \) become somewhat less extreme, no longer pushed against their bounds (although their values remain quite large); however, the ET exponent \( \gamma \) remains on the bound. There is also little change in the estimated melting temperature (particularly considering the posterior uncertainty), which differs from its a priori expected value \( T_{\text{melt}} \) ~ 0°C. The fit remains generally acceptable, with the Nash–Sutcliffe around 0.73, but with some storms severely underpredicted (e.g., the storm at 600 days is underestimated by nearly a factor of two).

The behavior of the prediction limits of the SLS methods is generally poor, as seen in Figure 1. The bounds are very wide for the 300-day simulation, but still do not envelope the major storms. For the 1400-day simulation (not shown), the situation is more severe; for example, an observed storm at 600 days lies largely outside the prediction limits.

4.2. BATEA Analysis

4.2.1. Specification of Input Uncertainty Model

In order to use BATEA, it is necessary to make explicit assumptions about the reliability of the data used in the calibration. The synthetic study of Kavetski et al. [2002] proceeded on the assumption that the mean \( \mu_m \) and variance \( \sigma_m^2 \) of the storm multipliers are known. Assuming unbiased (on average) gauges (\( \mu_m = 1 \)) is reasonable as a first approximation, but a priori estimation of the variance \( \sigma_m^2 \) is not straightforward. As discussed in the companion paper, it is possible to relax the assumption of known input error variance and specify a vague prior on \( \sigma_m^2 \). In this study, the inverse gamma prior \( p(\sigma_m^2|\nu_0, s_0) \) with \( \nu_0 = 1 \) and \( s_0 = 0.2 \) is used, which corresponds to a vague prior on \( \sigma_m^2 \) favoring values in the vicinity of 0.1–0.3. These values reflect an a priori expectation of likely corruption of precipitation data. Setting \( \nu_0 = 1 \) (interpreted as a single degree of freedom) makes the information content of the prior roughly equivalent to a single additional storm. Since the analysis involves the order of 100 storms, the prior has a limited influence on the posterior distribution. Sensitivity analysis performed by varying the prior parameters \( \nu_0 \) and \( s_0 \) suggests that, provided the prior is sufficiently vague to allow standard deviations in excess of 0.2–0.3, the results are consistent, whereas specifying priors concentrated around \( \sigma_m \sim 0 \) leads back to known input methods.

4.2.2. Parameter Estimates and Prediction Limits

Figure 2 shows the VIC response fit obtained using the storm multiplier BATEA model. The hourly precipitation series have been partitioned into discrete storms separated by at least 16 consecutive hours without rain. This critical value was determined empirically so that major bursts of rain had their own multipliers. In addition, some manual adjustment of individual partitions was performed (e.g., removing partitions associated with very small precipitation events).

The difference between the SLS and BATEA results for the discharge is qualitatively significant, with the Nash–Sutcliffe efficiency raised to \( \approx 0.95 \) for all three time periods. Importantly, there are no missing storms in the modeled hydrograph and the peaks are fitted much better. Since the observed data appears to have insufficient precipitation in this time period, it is problematic to obtain good fits using a calibration method that treats the observed input as a fixed known quantity. BATEA, on the other hand, faces a tradeoff between fitting the precipitation or the discharge (since they appear incompatible for some portions of the hydrograph series). The companion paper shows that it is impossible to resolve this tradeoff without some explicit prior assumptions about the data errors (effectively defining the relative weights of the observed forcing and response data). Such information cannot be incorporated into the calibration methods currently used in hydrology.

Table 3 shows that the differences between the parameter estimates computed using SLS and BATEA are generally moderate (although the posterior uncertainty in the BATEA parameters is smaller then in the SLS parameters). Yet there are appreciable differences in the melting temperature parameter \( T_{\text{melt}} \) depending on whether precipitation is treated as an exact known quantity or as a random variable. The BATEA estimate appears more realistic, \( \approx –0.6°C \) – +0.2°C versus \( \approx 2.0°C \) for the SLS method. This difference can translate into significant changes in predictions, since in the snow model small differences in melting temperatures can control whether snowmelt or precipitation takes place. Finally, there are orders-of-magnitude differences in the exponential VIC parameters \( \{\alpha, \beta, \gamma\} \) for the 300-day simulations. Such differences could confound attempts to regionalize the model parameters.

The prediction limits obtained using BATEA are also more reasonable than those obtained using SLS, mainly...
because of the model performing better for the major storms. However, it is clear that input adjustment alone cannot ensure totally acceptable model performance. For example, the prediction limits in Figure 2 show some unreasonably fast base flow behavior. Whereas traditional calibration methods could attribute qualitatively poor model performance to either precipitation or model errors, the incorporation of input uncertainty makes the BATEA results more specific; the poor behavior is likely due to deficiencies in the VIC model hypothesis. Therefore the next logical step in improving the calibration of hydrological models is a more robust handling of model errors, while retaining systematic handling of forcing and response uncertainties.

4.2.3. Computational Performance of BATEA

The computational feasibility of BATEA for desktop-type computing (as opposed to large-scale computing) is of practical importance. In this paper, BATEA is implemented by first optimizing the parameters and multipliers (i.e., maximizing the posterior density (5) with respect to \( \{Q, m\} \)) using a Newton-type method, followed by Metropolis Monte Carlo sampling of the parameter distributions [Kuczera and Parent, 1998]. Since the 1400-day period C contained approximately 200 storms, computational efficiency becomes critical. In particular, SCE-type analysis is infeasible for such high-dimensional problems: SCE optimization of the VIC model to even 10–20 storms using the BATEA objective function (5) can exceed several hours, whereas 100-storm problems could take weeks or even months. Conversely, certain Newton-type methods do not even require a single linear solution for the entire optimization [Nocedal and Wright, 1999]. While discussion of these methods is tangential to this paper (for a recent application in hydrology, see, e.g., Kavetski et al. [2006b]), their efficiency is essential for the BATEA computational problem. In the French Broad River study, the 200-storm 1400-day period C was preoptimized in approximately 20 minutes on a laptop. The smoothness of the hydrological model, allowing analytical or numerical differentiation, is essential to achieve such efficiencies. However, as the dimension of the problem increases further, BATEA does become more severely computationally constrained. At this stage, simulations with more than 200–300 storms appear infeasible, unless more specialized numerical techniques are invoked, such as nonlinear conjugate-gradient-type methods [Nocedal and Wright, 1999]. In addition, automatic partitioning of long data series into discrete storm events can also be problematic and need further work.

4.2.4. Performance of ET Model Components During Calibration

Within SPM, the actual ET is controlled by the saturation fraction of the catchment and the exponential parameter \( \gamma \). Under these circumstances, the ET components
of the model are used to correct mass balance errors (e.g., arising from corrupt precipitation data) and may not meaningfully capture actual ET processes. This problem is exacerbated by the fact that actual ET data is seldom if ever used in calibration. Consequently, a model with two outputs (runoff and actual ET) is calibrated to a single observed output series, which may not sufficiently constrain the ET components and may explain the extreme values of ET parameters often encountered in calibration. Although BATEA is able to modify the uncertain input to make it more consistent with the model and the observed discharge, the fundamental limitation of fitting a relatively complex (and yet still inaccurate) model to a single input-output series is not straightforward to overcome unless more data and/or a better model is available. Similar problems have also been encountered elsewhere in parameter estimation [e.g., Bates and Watts, 1988].

### 4.2.5. A Posteriori Checks of Inference Assumptions

A posteriori verification of the modeling assumptions is an important step in modeling. For SLS methods, analysis of residual plots should be undertaken. It is fairly clear from Figures 1 and 2 that the SLS residuals in this case are anything but Gaussian. In particular, there is significant autocorrelation in the errors. Kavetski et al. [2002] discuss several possible origins of such behavior, including the propagation of input errors into the storage and subsequently the discharge. Conversely, the response residuals in the BATEA runs are much smaller (although certain problems in simulating the recession for the first storm were noted). However, a good response match does not necessarily imply that the BATEA calibration succeeded. Traditional regression methods, such as SLS, can be viewed as response data adjustment mechanism: given the model hypothesis and the exact forcing data, adjust the observed response data. A posterior diagnostic then assesses whether this adjustment is “reasonable”. BATEA can also be viewed from the “data adjustment” perspective. However, since it is designed for cases where both the output and input data are measured with errors, BATEA adjusts both the response and the forcing. Consequently, it is also necessary to examine the consistency of the forcing adjustment with the assumed input uncertainty model. Indeed, under the assumption that precipitation uncertainty dominates discharge errors, one should expect more adjustment in the input than in the output. This occurs naturally in BATEA if a sufficiently loose input error variance is prescribed a priori.

A normal probability plot of the multipliers is shown in Figure 3. It shows that the posterior mean of the multipliers is almost identically 1.0, suggesting on average unbiased forcing measurements and agreeing with the prior mean $\mu_m = 1.0$. However, while the data clearly does not support the hypothesis of no errors in the inputs, the posterior multiplier distribution also deviates from the assumed Gaussian distribution, resembling instead a mixture of two (possibly Gaussian) distributions, or a single fat-tailed distribution. This behavior is not unexpected, since Gaussian distributions are thin-tailed and may therefore underestimate the probability of significant deviations from the mean. In precipitation sampling, having a compact but intense storm cell miss a rain gauge can readily result in very large errors that fundamentally exceed those expected from the sampling variability inside a storm cell. This observation suggests that the Gaussian multiplier distribution should be replaced with a more fat-tailed distribution (or a mixture distribution) to account for such cases. The multiplicative error model is also limited in its ability to modify near-zero input, which can be necessary when the storm missed the gauge almost entirely or entirely. Neither can a multiplier model meaningfully account for timing errors, which can be very troublesome since even a small time shift between the observed and predicted responses can yield very large penalties in the sum-of-squared-errors sense. Finally, it would be beneficial to undertake a more in-depth analysis of the relation between the storm multipliers and the storm characteristics. It is stressed that only explicit input error sensitive methods such as BATEA can be used to incorporate such information into hydrological analysis.

The input error model can be further scrutinized by comparing the multiplier values estimated from overlapping time series of different length. In particular, Figure 3 suggests moderate differences in the multipliers for the time periods A, B and C. Moreover, since the same model is used, the estimates cannot be considered totally independent. It is therefore reasonable to compare the multipliers estimated using different models. In this case, however, it becomes possible for differences in multiplier estimates to be explained by differences in the model structure.

### 5. Case Study 2: Potomac River

The second case study considers the South Branch Potomac River near Springfield, West Virginia, with basin area of $\approx 2250$ km$^2$ and 8 hourly precipitation gauges. The soil is a mix of loam, bedrock, as well as sandy and silt loam. The vegetation is primarily deciduous (mixed forest). The data for a 200-day period is used, where the total precipitation is 575 mm, the cumulative potential ET is

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean (A) ± SD (Mode)</th>
<th>Mean (B) ± SD (Mode)</th>
<th>Mean (C) ± SD (Mode)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\log_{10} S_{\text{max}}$</td>
<td>$-0.57 \pm 0.064 (-0.57)$</td>
<td>$-0.35 \pm 0.005 (-0.31)$</td>
<td>$-0.48 \pm 0.009 (-0.47)$</td>
</tr>
<tr>
<td>$\log_{10} \alpha$</td>
<td>$-1.20 \pm 0.011 (-1.20)$</td>
<td>$-1.60 \pm 0.005 (-1.60)$</td>
<td>$-1.42 \pm 0.006 (-1.43)$</td>
</tr>
<tr>
<td>$\log_{10} \beta_0$</td>
<td>$-0.15 \pm 0.009 (-0.15)$</td>
<td>$-0.41 \pm 0.003 (-0.41)$</td>
<td>$-0.27 \pm 0.003 (-0.27)$</td>
</tr>
<tr>
<td>$\log_{10} \beta_3$</td>
<td>$+0.54 \pm 0.008 (+0.53)$</td>
<td>$+0.99 \pm 0.003 (+1.00)$</td>
<td>$+0.98 \pm 0.011 (+1.00)$</td>
</tr>
<tr>
<td>$\log_{10} \gamma$</td>
<td>$+0.21 \pm 0.013 (+0.20)$</td>
<td>$-0.20 \pm 0.004 (-0.19)$</td>
<td>$-0.49 \pm 0.010 (-0.50)$</td>
</tr>
<tr>
<td>$\log_{10} \beta_4$</td>
<td>$-1.24 \pm 0.013 (-1.25)$</td>
<td>$-1.18 \pm 0.002 (-1.18)$</td>
<td>$-1.13 \pm 0.008 (-1.13)$</td>
</tr>
<tr>
<td>$T_{\text{mf}}$</td>
<td>$-0.64 \pm 0.015 (-0.64)$</td>
<td>$-0.61 \pm 0.004 (-0.61)$</td>
<td>$+0.16 \pm 0.030 (+0.16)$</td>
</tr>
</tbody>
</table>

*Period A: 300 days, Nash–Sutcliffe (NS) index 0.93; period B: 600 days, NS index 0.92; period C: 1400 days, NS index 0.94.*
442 mm and the total runoff is 272 mm (47% of total precipitation input).

Although the VIC model used in the French Broad River study is a highly simplified conceptualization of the catchment dynamics, it does not contain any “catastrophic” deficiencies in the sense of missing essential dominant processes (at least on the basis of our current understanding of hydrological dynamics). In the Potomac case study, the VIC model is deliberately corrupted by removing the degree-day snow submodel, thus introducing a known and serious structural error into the simulation. The purpose of this section is to examine the diagnostics that BATEA can yield in conditions where the model contains severe flaws. This is then discussed and compared to the diagnostics available from current methods such as SLS and GLUE.

The least squares (SLS) fit of VIC (no snow model) is shown in Figure 4. Unlike the French Broad River case, where the SLS fit was generally meaningful, the Potomac SLS fit is extremely poor for the entire time period, with the Nash–Sutcliffe efficiency as low as 0.30. This problem is not specific for the VIC model used in this study; it is very doubtful, for example, that the largest discharge (at t ≈ 90 days) could be matched by any reasonable mass conservative model, since the corresponding precipitation is very small and comparable to events that did not even register on the outflow hydrograph. In this catchment, a significant portion of the total precipitation takes place during freezing temperatures and is thus immobilized in the snow store (possibly for a long time) until melting occurs. Moreover, when it melts it can significantly compound the discharge due to any concurrent storm events.

The BATEA fit of VIC (no snow model) is shown in Figure 5 and represents a substantial improvement over the SLS fit. The Nash–Sutcliffe efficiency is increased to 0.80, the large deviations between observed and predicted responses disappeared and the prediction limits envelope the majority of the observations. These improvements are due to BATEA modifying the uncertain forcing series to make them more consistent with the model and the observed responses. The BATEA parameter estimates (Table 4) are, unsurprisingly, quite different to the SLS results.

In order to assess how the SLS and BATEA methods have coped with the serious model error in this study (the lack of the snow model), the VIC model with snow preprocessor was calibrated using SLS and BATEA.
and in the prediction limits. SLS fit of VIC (with snow model) to the

Potomac River Parameters Inferred Using SLS and BATEA With and Without the Snow Model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>SLS (No Snow)</th>
<th>BATEA (No Snow)</th>
<th>SLS (Snow)</th>
<th>BATEA (Snow)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nash-Sutcliffe efficiency</td>
<td>0.30</td>
<td>0.80</td>
<td>0.85</td>
<td>0.97</td>
</tr>
<tr>
<td>$\log_{10} S_{\text{max}}$</td>
<td>$-0.96 \pm 0.04$</td>
<td>$-0.90 \pm 0.12$</td>
<td>$-0.87 \pm 0.03$</td>
<td>$-1.05 \pm 0.07$</td>
</tr>
<tr>
<td>$\log_{10} \alpha$</td>
<td>$-1.45 \pm 0.16$</td>
<td>$-1.57 \pm 0.19$</td>
<td>$-1.54 \pm 0.11$</td>
<td>$-1.72 \pm 0.11$</td>
</tr>
<tr>
<td>$\log_{10} k$</td>
<td>$-1.36 \pm 0.06$</td>
<td>$-0.66 \pm 0.05$</td>
<td>$-0.48 \pm 0.04$</td>
<td>$-0.33 \pm 0.03$</td>
</tr>
<tr>
<td>$\log_{10} \gamma$</td>
<td>$-0.54 \pm 0.29$</td>
<td>$-0.73 \pm 0.15$</td>
<td>$+0.98 \pm 0.01$</td>
<td>$+0.88 \pm 0.08$</td>
</tr>
<tr>
<td>$\log_{10} \theta_d$</td>
<td>NA</td>
<td>NA</td>
<td>$-0.12 \pm 0.04$</td>
<td>$-0.31 \pm 0.11$</td>
</tr>
<tr>
<td>$T_{\text{melt}}$</td>
<td>NA</td>
<td>NA</td>
<td>$+2.31 \pm 0.23$</td>
<td>$+2.10 \pm 0.37$</td>
</tr>
</tbody>
</table>

NA, not applicable.

6. Conclusions

The paper presented two case studies illustrating the application of the Bayesian Total Error Analysis (BATEA) methodology for exploring data and model uncertainty in hydrological modeling. The empirical assessment demonstrates the sensitivity of hydrological model predictions to the precipitation data: even fairly modest modifications of the precipitation, well within expected uncertainty, can lead to significant changes in the modelled hydrographs and estimated parameters. This sensitivity is ignored by the majority of current calibration methods; conversely, BATEA has the potential to identify and correct these errors. However, this ability hinges on the user providing uncertainty models that are not grossly wrong (e.g., timing errors cannot be handled using the storm depth multiplier model). Since BATEA introduces additional parameters of the input error model, the interplay between forcing errors and model errors should be examined by considering the fit of both (1) the observed and simulated forcing series and (2) the observed and simulated response series. In contrast to current model assessment methods, this comparison does not assume the input data is exact. The second case study illustrated the misuse of an inference method by specifying a grossly wrong model. Unlike cases where the instrumental accuracy is known and a model can be rejected if it results in discrepancies with observations far exceeding expected sampling uncertainty, calibrations that assume unknown instrumental accuracy yield less insight into model errors. However, BATEA provides the ability to experiment with different data uncertainty models for both input and output data to assess

Figure 6. SLS fit of VIC (with snow model) to the Potomac data.
if the model can be reconciled with the observed data, whereas fixed-regressor methods (e.g., SLS and current versions of GLUE and MOCOM) do not have this capability.

The elimination or reduction of input error bias in estimated hydrological parameters is a significant step forward in several directions:

1. Parameter regionalization: Bias is caused by input errors obscures potential relationships between the calibrated parameters and catchment characteristics.

2. Assessing parameter stability in time: Input errors may lead to spurious parameter changes for different time periods, or, conversely, mask genuine alterations in catchment properties.

3. Model assessment: Currently, poor model performance can be blamed on a variety of reasons, for example, precipitation errors, model errors, ET errors, etc. Disaggregating the sources of uncertainty using BATEA facilitates a more systematic model assessment and comparison.

4. Model sensitivity analysis: Cases where BATEA suggests strong correlation between input errors (reflected in the latent variables $\phi$) and model parameters are precisely those where traditional methods assuming known inputs can be misleading (since the highly uncertain forcing data then has particularly strong leverage over the model parameters).

5. Data quality monitoring: The need for large (and similar) input adjustments by different models is suggestive of data quality problems.

We argue that these potential advantages warrant further development and assessment of the BATEA methodology and outweigh the current limitations outlined in this paper.

Acknowledgments. The authors are grateful to Qingyun Duan and John Schaake for their assistance with the MOPEX data. This work was supported by a grant from the Australia Research Council.

References


S. W. Franks and G. Kuczera, Civil, Environmental Engineering and Surveying, University of Newcastle, Callaghan, NSW 2308, Australia. (stewart.franks@newcastle.edu.au; george.kuczera@newcastle.edu.au)

D. Kavetski, Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ 08542, USA. (kavetski@princeton.edu)