Calibration of a Distributed Flood Forecasting Model with Input Uncertainty Using a Bayesian Framework

Mingliang Li[1], Dawen Yang[1], Jinsong Chen[2], and Susan S. Hubbard[2]

1: State Key Laboratory of Hydroscience and Engineering, Department of Hydraulic Engineering, Tsinghua University, Beijing, China;

Email: liml06@mails.thu.edu.cn and yangdw@tsinghua.edu.cn

2: Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA;

Email: jchen@lbl.gov and sshubbard@lbl.gov

Abstract

In the process of calibrating distributed hydrological models, accounting for input uncertainty is important, yet challenging. In this study, we develop a Bayesian model to estimate parameters associated with a geomorphology-based hydrological model (GBHM). The GBHM model uses geomorphic characteristics to simplify model structure and physically based methods to represent hydrological processes. We divide the observed discharge into low- and high-flow data, and use the first order autoregressive model to describe their temporal dependence. We consider relative errors in rainfall as spatially distributed variables and estimate them jointly with the GBHM parameters. The joint posterior probability distribution is explored using Markov chain Monte Carlo methods, which include Metropolis-Hastings, delay rejection adaptive Metropolis, and Gibbs sampling methods. We evaluate the Bayesian model using both synthetic and field datasets. The synthetic case study demonstrates that the developed method generally is effective in calibrating GBHM parameters and in estimating their associated uncertainty. The calibration ignoring input errors has lower accuracy and lower reliability compared to the calibration that includes estimation of the input errors, especially under model structure uncertainty. The field case study shows that calibration of GBHM parameters under complex field conditions remains a challenge. Although jointly estimating input errors and GBHM parameters improves the continuous ranked probability score and the consistency of the predictive distribution with the observed data, the improvement is incremental. To better calibrate parameters in a distributed model, such as GBHM here, we need to develop a more complex model and incorporate much more information.

1. Introduction

The field of flood forecasting faces many challenges, including uncertainties in precipitation observations and predictions, in hydrological model parameters and structures, and in discharge observations [*Beven*, 2006; *Gupta et al.*, 2005; *Liu and Gupta*, 2007]. Most studies on uncertainty analysis have been focused on calibration of lumped conceptual rainfall-runoff models. Stochastic approaches, especially Bayesian models, have become the most commonly used methods for uncertainty analysis of hydrological models. For example, *Kuczera and Parent* [1998] used a Bayesian model and the Metropolis algorithm to assess parameter uncertainty in conceptual catchment models. *Bates and Campbell* [2001] developed a Bayesian approach to parameter estimation and inference in conceptual rainfall-runoff models, and used Markov chain Monte Carlo (MCMC) methods to explore the joint posterior distribution.

Recent studies have been focused on estimating input uncertainties and model structure uncertainties and their influence on model parameter calibration. Most of these studies rely on hierarchical Bayesian models to handle uncertainty in the forcing term (e.g., rainfall) by using latent variables, such as multipliers. *Kavetski et al.* [2003, 2006] developed a Bayesian total error analysis methodology (BATEA) to allow a modeler to incorporate, test, and refine the existing understanding of all sources of data uncertainty directly and transparently. *Ajami et al.* [2007] developed an integrated Bayesian uncertainty estimator (IBUNE) with inclusion of a Bayesian model averaging (BMA) scheme to account for input, parameter, and model structure uncertainties. *Vrugt et al.* [2008] developed a hierarchical Bayesian model and a novel MCMC sampler, a so-called differential evolution adaptive Metropolis (DREAM), to analyze input uncertainty. *Renard et al.* [2010, 2011] demonstrated the importance of prior information

for estimating the input and model structure errors and for decomposing predictive uncertainty. While all those methods use hierarchical Bayesian models, they differ in the handling of latent variables [*Ajami et al*, 2009; *Renard et al*, 2009] and in the computational methods used to sample from the posterior.

Although uncertainty analysis of distributed models is also important, few such investigations have been performed. Compared with lumped conceptual hydrological models, distributed models are expected to have lower model structure uncertainties because of the spatially distributed description of landscape conditions and the physically based representation of hydrological processes. However, distributed models introduce a large number of parameters that typically are difficult or impossible to observe directly and are difficult to estimate by calibration due to poor identifiability. Although physically based distributed hydrological models have become increasingly popular over the past two decades [*Singh and Woolhiser*, 2002], the uncertainty associated with those models has not received much attention. Current studies on calibration of distributed models mainly focus on utilizing multiple observations [*Immerzeel and Droogers*, 2008; *Khu et al.*, 2008] or multiple objective functions [*Marce et al.*, 2008; *Shrestha and Rode*, 2008].

In this study, we develop a hierarchical Bayesian model, following a similar approach to that given by *Kavetski et al.* [2003, 2006] and *Vrugt et al.* [2008], to calibrate a physically based distributed model for event-based mountain flood prediction. Within Gaussian likelihood functions, the discharge is split into two categories (i.e., low-flow and high-flow) [*Schaefli et al.*, 2007]. We use the first order autoregressive model to simulate temporal correlation in the residuals [*Sorooshian and Dracup*, 1980; *Thyer et al.*, 2002]. To achieve good sampling efficiency, we combine different MCMC sampling strategies, similar to those used by *Chen et al.* [2007]. We apply the developed Bayesian model first to synthetic datasets

for investigating the influence of input uncertainty on hydrological model calibration, and then to field datasets collected from a small catchment in the Yangtze River basin in China.

The remainder of this paper is organized as follows. Section 2 describes the distributed hydrological model and study area. Section 3 describes the parameterization and development of the Bayesian model based on the distributed hydrological model. Section 4 describes the MCMC sampling strategies for exploring the posterior distribution. The synthetic and field studies are given in Sections 5 and 6, respectively; discussion of the results and conclusions are provided in Sections 7 and 8, respectively.

2. Distributed Hydrological Model and Study Area

2.1. Geomorphology-Based Hydrological Model

In this study, we use a geomorphology-based hydrological model (GBHM) [*Cong et al.*, 2009; *Yang*, 1998; *Yang et al.*, 2004], which utilizes geomorphologic characteristics to simplify the model structure and physically based methods to represent hydrological processes. The GBHM differs from other physically based models, such as the Systeme Hydrologique Europeen [*Abbott et al.*, 1986], mainly in the discretization scheme and parameterization. The GBHM takes advantage of geomorphologic similarity to maintain hydrological similarity. The grid size in GBHM is therefore larger than that in the digital elevation model (DEM), leading to a significant reduction in spatial-structure complexity.

The GBHM uses the hillslope-stream formation to represent catchment topography. A catchment is divided into many sub-catchments, and the different sub-catchments are organized based on the Pfafstetter system [*Yang and Musiake*, 2003]. The grids within a sub-catchment are classified into flow intervals, taking into account the distances of the grids from the outlet. Each flow interval contributes to a segment of a main stream that is the simplification of a stream network. Flow routing in the stream

network is conducted using the kinematic wave method [*Chow et al.*, 1988]. Hillslopes within each grid are expected to be topographically similar; the topographical characteristics of the hillslopes are calculated using fine DEM [*Yang et al.*, 2002]. Figure 1 is a schematic diagram of the GBHM, where the hillslope is represented by a group of inclined parallel planes with slope β and length l. Based on the hypothesis of a linear slope, we introduce a slope-shape factor f_{ss} to account for the concave-convex effects on runoff generation and overland flow hydrographs. The factor f_{ss} is defined as $f_{ss} = l_s/l$, where l_s is the saturated outflow length (see Figure 2).

In the GBHM, water flow under the soil surface is simulated separately by a quasi-two-dimensional subsurface module along the vertical and parallel directions of the slope. The topsoil is divided into multiple layers according to their depths. Distribution of soil water content and hydraulic characteristics along the hillslope is assumed uniform. Vertical distribution of the saturated hydraulic conductivity in the soil is set to decrease exponentially with increasing depth [*Robinson and Sivapalan*, 1996]. The vertical flow module is described using the Richards equation and solved by an implicit numerical scheme. Subsurface flow along the hillslope occurs when soil water content surpasses the field capacity. Groundwater aquifers are discretized (corresponding to the grids) and treated as individual storage compartments. The exchange between groundwater and river water is expected to be a steady flow and is estimated by Darcy's law. GBHM parameters include vegetation, land surface, soil water, and river parameters [*Cong et al.*, 2009]. A detailed description of each parameter for calibration is provided in Section 3.1.

2.2. Study Area and Available Data

The Chuzhou catchment, which spans approximately 289 km^2 and is located in the Yangtze River

Basin of Southern China, is selected as the study area. Annual precipitation is approximately 1550 mm, with about 70% of the annual rainfall occurring from April to September. Floods occur frequently during the rainy season. The grid size of the GBHM in the present study is 1 km. DEM data were collected from the global topography database (http://www2.jpl.nasa.gov/srtm/) with a three arc-sec spatial resolution (approximately 90 m). Elevation in the Chuzhou catchment varies from 379 to 2090 m, with the majority of the catchment covered by forest. The soil is red loam with an estimated soil depth of 0.6 m. Since the catchment is relatively homogenous in soil type and land use, the GBHM parameters are treated uniformly in the Bayesian model. The developed method can be augmented to account for other conditions.

We used hourly averaged data collected from 32 large flood events occurring between 1980 and 2002 for the study, among which 16 flood events were used for calibration and another 16 flood events were used for validation. As shown in Figure 3, seven rainfall gauges and one discharge gauge can be found within the catchment. The rainfall input is interpolated using the angular distance-weighted (ADW) method [*New et al.*, 2000]. Daily pan evaporation data were obtained from a meteorological station located in Suichuan County, approximately 60 km away from the center of the catchment. A sub-catchment of Chuzhou was selected for the synthetic study. Three rainfall gauges and one synthetic discharge gauge were located in the synthetic basin. Data for two synthetic rainfall gauges were derived from nearby available gauges. Generation of synthetic discharge data is described in Section 5.1. Since the durations of those flood events in the study area do not overlap, we simulate each flood individually.

3. Bayesian Model

3.1. Model Structure and Parameterization

We develop a Bayesian model to calibrate parameters in the GBHM. They include the following: (1) the ratio of basin mean potential evaporation to gauged pan evaporation (k_{ep}); (2) the saturated hydraulic conductivity in the top soil layer (K_t), the bottom soil layer (K_b), and the unconfined aquifer (K_g); (3) the slope shape factor (f_{ss}); (4) the maximum surface storage (d_{surf}); (5) the specific yield (S_y); and (6) the roughness on the hillslope surface (n_s) and on the river channel (n_c). For simplicity, let vector $\boldsymbol{\theta} = (k_{ep}, K_t, K_b, K_g, f_{ss}, d_{surf}, S_y, n_s, n_c)^T$, where T is the transpose operator.

The simulated discharge of the GBHM model is represented by

$$q_t^{sim} = h(\mathbf{x}_t^{inp}, \boldsymbol{\theta}) = g(\mathbf{x}_t, \boldsymbol{\varphi}_t^*, \boldsymbol{\theta}), \qquad t = 1, ..., n \tag{1}$$

where q_t^{sim} denotes the simulated discharge at the time step t; $h(\mathbf{x}_t^{inp}, \boldsymbol{\theta})$ represents the mapping of the GBHM model with parameter $\boldsymbol{\theta}$, and input $\mathbf{x}_t^{inp}$, which will be described further by the measurement of input $\mathbf{x}_t$ and $\boldsymbol{\varphi}_t^*$ for estimation of the input errors; $g(\mathbf{x}_t, \boldsymbol{\varphi}_t^*, \boldsymbol{\theta})$ represents the mapping of the GBHM model with rainfall observation, estimated input errors, and model parameters. In the present study, we consider one discharge station only, although the model can be modified to accommodate more gauges under more complex conditions. Model inputs are specified as rainfall in the study, and errors $\boldsymbol{\varphi}_t^*$ are specified as relative rainfall errors, i.e., $\mathbf{x}_{t,j}^{inp} = \mathbf{x}_{t,j} / (1 + \varphi_{t,j}^*)$, where j is an index of rainfall gauges. Generally, $\varphi_{t,j}^*$ depends on gauge locations, equipment properties, and many other unknown factors.

We use event-based methods, similar to those used by *Kavetski et al.* [2006] and *Vrugt et al.* [2008], to develop the Bayesian inference equations. We assume that the errors in precipitation are independent across storms and constant within storms. Let $\varphi_{k,j} = \varphi_{t,j}^*, \ t \in T_k$, where k is an index of the storm series and T_k is the set of time steps during the kth storm. We also assume that $\varphi_{k,j}$ at each gauge follows a truncated normal distribution [*Huard and Mailhot*, 2008] on the interval (-1, +∞), having a mean of μ_φ

and an inverse variance of τ_φ (see Appendix A) for different storms.

Following the method of *Schaefli et al.* [2007], we divide the discharge data into low-flow data **Y** and high-flow data **Z**. Such division is reasonable for the current study because the low-flow data are mainly dependent on groundwater storage, whereas the high-flow data are mainly determined by rainfall intensity and flow routing. Empirically, the separation of flow data makes parameter estimation more stable. We adopt a daily time step for the low flow periods since the associated variations in discharge are typically very small. For the high flow periods we use an hourly time step. Since the residuals between observed and simulated discharge are often temporally correlated, we employ the first order autoregressive model (AR(1)) used by *Bates and Campbell* [2001] and *Sorooshian and Dracup* [1980] to analyze the model residuals. Thus, the relationship between the observed and simulated discharge is given by

$$y_t = y_t^{sim} + \varepsilon_{y,t} = y_t^{sim} + \rho_y \varepsilon_{y,t-1} + w_{y,t} \qquad (2)$$

where y_t is the observed data, y_t^{sim} is the model simulation, $\varepsilon_{y,t}$ is the model residual, ρ_y is the correlation coefficient of the AR(1) model, and $w_{y,t}$ is the Gaussian noise with zero mean and unknown variance. Let τ_y denote the inverse variance of $\varepsilon_{y,t}$, and the inverse variance of $w_{y,t}$ is given by $\tau_y^w = \tau_y / (1 - \rho_y^2)$ [*Naidu*, 1996]. The likelihood of parameters associated with the given data could be derived in terms of τ_y. Similarly, the correlation coefficient and inverse variance for the high-flow discharge are denoted as ρ_z and τ_z, respectively. Consequently, we can write the Bayesian model for given inputs **X** and flow data (**Y**, **Z**) as follows:

$$
\begin{aligned}
& f(\boldsymbol{\theta}, \boldsymbol{\varphi}, \mu_{\varphi}, \tau_{\varphi}, \rho_{y}, \tau_{y}, \rho_{z}, \tau_{z} \mid \mathbf{X}, \mathbf{Y}, \mathbf{Z}) \\
& \quad \propto f(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\varphi}, \rho_{y}, \tau_{y}, \mathbf{X}) f(\mathbf{Z} \mid \boldsymbol{\theta}, \boldsymbol{\varphi}, \rho_{z}, \tau_{z}, \mathbf{X}) \\
& \quad \times f(\boldsymbol{\varphi} \mid \mu_{\varphi}, \tau_{\varphi}) f(\mu_{\varphi}) f(\tau_{\varphi}) \\
& \quad \times f(\rho_{y}) f(\tau_{y}) f(\rho_{z}) f(\tau_{z}) f(\boldsymbol{\theta})
\end{aligned}
\tag{3}
$$

where $f()$ denotes the probability distribution function (pdf) and "$\propto$" means "is proportional to." The left side of Equation 3 is the joint posterior pdf of all variables for estimation given measurements at the discharge station. The first two terms on the right side of the equation represent the likelihood functions of low- and high-flow data. The third term on the right side represents the hierarchical prior model of input error $\boldsymbol{\varphi}$ characterized by two hyper-parameters. The other terms are the prior distribution functions of the remaining parameters. Each prior function of those parameters is assumed independent of the others. Note that a normalizing constant was omitted in Equation 3, which does not affect the analyses in the present study because we use an MCMC sampling method.

3.2. Likelihood Function

We assume that the individual flood events are independent of each other. Consequently, the likelihood function of the entire dataset is the product of the likelihood functions of individual flood events. This assumption is justified in the present study because the duration of a flood event is no more than two days, but the 32 flood events are selected over 18 years, and the time gaps between the flood events within a year are 10–100 days. Under the assumptions listed in Section 3.1, the likelihood functions are multivariate normal with an AR(1) correlation structure. For the low-flow data of the kth flood $\mathbf{Y}_k$, the likelihood function is

$$f(\mathbf{Y}_k \mid \boldsymbol{\theta}, \boldsymbol{\varphi}_k, \rho_y, \tau_y, \mathbf{X}_k) \propto \left|\Sigma_k\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{Y}_k - \mathbf{Y}_k^{sim})^T \Sigma_k^{-1} (\mathbf{Y}_k - \mathbf{Y}_k^{sim})\right) \quad (4)$$

where Σ_k denotes the covariance matrix of the residuals for the kth flood event. The likelihood function is further written as (see Appendix B)

$$f(\mathbf{Y}_k \mid \boldsymbol{\theta}, \boldsymbol{\varphi}_k, \rho_y, \tau_y, \mathbf{X}_k) \propto \left|R_k\right|^{-\frac{1}{2}} \tau_y^{\frac{1}{2}n} \exp\left(-\frac{1}{2}\tau_y S_k\right) \quad (5)$$

where R_k denotes the correlation coefficient matrix of the model residuals, $\Sigma_k = R_k/\tau_y$, and $S_k = (\mathbf{Y}_k - \mathbf{Y}_k^{sim})^T R_k^{-1} (\mathbf{Y}_k - \mathbf{Y}_k^{sim})$.

3.3. Prior Models

We assume that each element in vector $\boldsymbol{\theta}$ is independent of others, since they represent different physical characteristics of the catchment. The prior function of $\boldsymbol{\theta}$ can be written as the product of each individual prior distribution. All elements (or the logarithmic transforms) of $\boldsymbol{\theta}$ are uniformly distributed over the given prior ranges (see Table 1) based on their physical properties and our calibration experience. We use the base 10 logarithmic transformations because elements K_t, K_b, K_g, f_{ss}, S_y, n_s, and n_c vary by several orders of magnitude.

The prior distribution of relative rainfall errors is assumed to be given by a truncated normal distribution (see Appendix A). Each component in $\boldsymbol{\varphi}$ is assumed to be independent of others. The prior distribution function of $\varphi_{.,j}$ (i.e., the component of $\boldsymbol{\varphi}$ at the jth gauge) can be obtained as

$$f(\varphi_{.,j} \mid \mu_\varphi, \tau_\varphi) \propto \begin{cases} \psi^{-n_k} \tau_\varphi^{\frac{1}{2}n_k} \exp\left(-\frac{1}{2}\tau_\varphi S_{\varphi,j} n_k\right) & \text{if } \varphi_{.,j} > -1; \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where n_k is the number of storms, $\psi = 1 + \mathrm{erf}((\mu_\varphi + 1)(\tau_\varphi / 2)^{1/2})$, $S_{\varphi,j} = \sum_{k=1}^{n_k} (\varphi_{k,j} - \mu_\varphi)^2 / n_k$, and erf() is the error function.

We assume that the prior distribution of the mean relative rainfall error, μ_φ, has a uniform distribution on (-0.05, 0.05), that is, a maximum of 5% systematic relative errors. As pointed out by *Renard et al.* [2010, 2011], the prior range of τ_φ (i.e., the precision of prior information about input errors) is critical for the inference of input errors. In the present study, we start from an informative prior in which the standard deviations of the truncated normal distributions of the input errors are between 5% and 15%. The upper bound (i.e., 15%) is the same as the one used by *Salamon and Feyen* [2010]. To investigate the effects of the upper limit on calibration, we also use other values as the upper bounds, including 10%, 20%, and 30%, for both synthetic and field case studies. Since the range between 5% and 15% in the standard deviation is equivalent to the range between 44.4 and 400 in the inverse variances, we use uniform distribution on (44.4, 400) as the prior distribution for τ_φ.

The prior distribution functions of the autocorrelation coefficients ρ_y and ρ_z are also set as the uniform distributions on (0, 1). Non-informative prior distributions are employed for inverse variances τ_y and τ_z as used by *Chen et al.* [2008], which are represented by gamma distributions $\Gamma(\alpha_0, \lambda_0)$ with shape parameter α_0 and inverse scale parameter λ_0 of 1e-3.

4. MCMC Sampling Method

We use MCMC methods to sample from the joint posterior distribution. In Section 4.1, we derive the conditional distributions for unknown variables. We describe the specific MCMC sampling methods in Section 4.2. The convergence diagnostics used for the present study are given in Section 4.3.

4.1. Conditional Distributions

We derive the conditional distribution of each parameter from Equation 3 given all the others. The conditional pdf of the GBHM parameters is given as

$$f(\boldsymbol{\theta}\,|\,\cdot) \propto f(\mathbf{Y}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi},\rho_y,\tau_y,\mathbf{X}) f(\mathbf{Z}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi},\rho_z,\tau_z,\mathbf{X}) f(\boldsymbol{\theta}) \tag{7}$$

The conditional pdfs of the other variables are similarly obtained.

The conditional pdfs of τ_y and τ_z are gamma distributions, because gamma distribution is a conjugate prior of the Gaussian likelihoods. Thus, the conditional distribution of τ_y is given by

$$\begin{aligned} f(\tau_y\,|\,\cdot) &= f(\mathbf{Y}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi},\rho_y,\tau_y,\mathbf{X}) f(\tau_y) \\ &\propto \tau_y^{\frac{1}{2}\sum_{k=1}^{m} n_k} \exp\left(-\frac{1}{2}\tau_y \sum_{k=1}^{m} S_k\right) \cdot \tau_y^{\alpha_0 - 1} \exp(-\lambda_0 \tau_y) \\ &\propto \tau_y^{\frac{1}{2}\sum_{k=1}^{m} n_k + \alpha_0 - 1} \exp\left(-\left(\frac{1}{2}\sum_{k=1}^{m} S_k + \lambda_0\right)\tau_y\right) \\ &\propto \Gamma\left(\frac{1}{2}\sum_{k=1}^{m} n_k + \alpha_0, \frac{1}{2}\sum_{k=1}^{m} S_k + \lambda_0\right) \end{aligned} \tag{8}$$

4.2. Sampling Strategies

MCMC methods provide a more powerful approach than conventional Monte Carlo methods in drawing samples from high-dimensional joint distribution functions [*Chen et al.*, 2008]. Previous studies have used various MCMC sampling methods, including Metropolis-Hastings (MH) methods [*Hastings*, 1970; *Metropolis et al.*, 1953], delayed rejection adaptive Metropolis (DRAM) algorithms [*Haario et al.*, 2006], and the Gibbs sampler [*Geman and Geman*, 1984].

We adopt a mixed sampling strategy by applying MH, Gibbs, and DRAM methods for different parameters considering their characteristics. Specifically, we use the DRAM algorithm for $\boldsymbol{\theta}$, the Gibbs sampler for τ_y and τ_z, and the MH method for other variables ($\boldsymbol{\varphi}$, μ_φ, τ_φ, ρ_y, and ρ_z). The main steps are outlined as follows:

(1) Initialize each variable as $\boldsymbol{\theta}^{(0)}$, $\boldsymbol{\varphi}^{(0)}$, $\mu_\varphi^{(0)}$, $\tau_\varphi^{(0)}$, $\rho_y^{(0)}$, $\tau_y^{(0)}$, $\rho_z^{(0)}$, and $\tau_z^{(0)}$, and then set $i = 1$.

(2) Update $\boldsymbol{\varphi}$, μ_φ, τ_φ, ρ_y, and ρ_z using MH methods, and refer to them as $\boldsymbol{\varphi}^{(i)}$, $\mu_\varphi^{(i)}$, $\tau_\varphi^{(i)}$, $\rho_y^{(i)}$, and $\rho_z^{(i)}$.

(3) Update $\boldsymbol{\theta}$ using the DRAM algorithm, and refer to them as $\boldsymbol{\theta}^{(i)}$.

(4) Update τ_y and τ_z using the Gibbs sampler, and denote them as $\tau_y^{(i)}$ and $\tau_z^{(i)}$. Let $i = i+1$.

(5) Repeat Steps 2 to 4 until the maximum number of allowable iterations is reached.

4.2.1. Metropolis-Hastings Sampling Method

We use the MH sampling method for $\boldsymbol{\varphi}$, μ_φ, τ_φ, ρ_y, and ρ_z. The MH sampling method, one of the most widely used MCMC sampling methods in hydrological studies [*Bates and Campbell*, 2001; *Kuczera and Parent*, 1998; *Marshall et al.*, 2004; *Schaefli et al.*, 2007], entails four steps. To sample μ_φ, a candidate value μ_φ^* is first generated from a proposed normal distribution $N(\mu_\varphi^{(i\text{-}1)}, \sigma_\mu^2)$, where σ_μ is the standard deviation. The probability for accepting the candidate value is then calculated based on its conditional distribution as

$$v_1 = \min\left\{1, \frac{f(\mu_\varphi^* \mid \cdot)}{f(\mu_\varphi^{(i-1)} \mid \cdot)}\right\} \tag{9}$$

Subsequently, a random value v is generated uniformly from an interval (0, 1). If $v < v_1$, we set $\mu_\varphi^{(i)} = \mu_\varphi^*$; otherwise, $\mu_\varphi^{(i)} = \mu_\varphi^{(i\text{-}1)}$. In MH sampling, the parameter σ_μ is important for convergence. If σ_μ is too large, the rejection rate of the proposed candidates will be too high, whereas if σ_μ is too small, the chain moves slowly and the tail regions of the joint posterior may not be sampled sufficiently [*Bates and Campbell*, 2001]. Hence, the MH method is suitable for sampling the variables that are relatively less sensitive to the value of σ_μ, and the DRAM algorithm is suitable for sampling those variables that are very sensitive to the value of σ_μ.

4.2.2. DRAM Algorithm

We use the DRAM algorithm for $\boldsymbol{\theta}$. The DRAM algorithm is developed based on a global adaptive

sampling strategy, Adaptive Metropolis (AM) [*Haario et al.*, 2001], and a local adaptive sampling strategy, Delayed Rejection (DR) [*Green and Mira*, 2001]. The AM algorithm initially calculates the correlation matrix of the variables from previously obtained samples and then provides a correlated proposal for Metropolis sampling methods. The AM algorithm appears to be efficient when the posterior correlation of variables is large, a common situation for hydrological parameters in calibration. The DR algorithm is also a variant of the MH algorithm. In the DR algorithm, when a proposed candidate is unwanted, it is not immediately rejected. The algorithm generates another candidate based on the first candidate at a second stage [*Green and Mira*, 2001]. In the present study, the two-stage DR is used considering the cost of additional computations. When combined, DR and AM complement each other. The adaptation of the AM algorithm will fail when the variance of the proposal is too large; the DR algorithm provides a remedy to this problem by reducing the variance at the second stage [*Haario et al.*, 2006]. Although the DRAM algorithm is a non-Markovian sampler, the ergodicity of the method has been proven by *Haario et al.* [2006] under weak conditions. To avoid the potential errors in the estimated posterior distributions caused by adaptive processes, we combine MH sampling methods with DRAM algorithm for a hybrid sampling strategy.

4.2.3. Gibbs Sampler

We use the Gibbs Sampler for τ_y and τ_z. Gibbs sampling methods [*Gelfand and Smith*, 1990; *Geman and Geman*, 1984] draw samples in succession from the conditional distributions of each parameter given the current values of all other parameters. The main advantage of Gibbs sampling is that all the samples are accepted, thereby avoiding wasting computing time on rejected samples. However, Gibbs sampling requires sampling directly from the conditional distribution, a requirement often difficult to

fulfill. In the present study, because the analytical form of the conditional distribution of τ_y and τ_z is the gamma distribution as given in Equation 8, $\tau_y^{(i)}$ and $\tau_z^{(i)}$ can be sampled directly.

4.3. Diagnostic of Convergence

Following the methods described in Section 4.2, many samples of parameters, that is, ($\boldsymbol{\theta}^{(i)}$, $\boldsymbol{\varphi}^{(i)}$, $\mu_\varphi^{(i)}$, $\tau_\varphi^{(i)}$, $\rho_y^{(i)}$, $\tau_y^{(i)}$, $\rho_z^{(i)}$, $\tau_z^{(i)}$, i=1, 2, ..., I_m), can be obtained. These samples are the approximations of the samples drawn from the target joint functions defined in Equation 3 after a long burn-in process [*Gilks et al.*, 1996]. The posterior marginal distributions of the parameters can be estimated by empirical distributions based on the drawn samples. A potential scale reduction factor (PSRF) developed by *Gelman and Rubin* [1992] is used to find the burn-in number and monitor the convergence of the chains. We run three to five MCMC chains from randomly generated initial values and accept them if the PSRF of each parameter is smaller than 1.2.

5. Synthetic Case Study

To evaluate the effectiveness of the developed Bayesian model, we generated synthetic datasets for calibration. We will first show how input uncertainty affects estimation of the GBHM parameters and then examine identifiability of the input uncertainty when model structure uncertainty is large.

5.1. Calibration Results of GBHM Parameters and Input Errors

As shown in Figure 3, we chose a sub-region of the Chuzhou catchment with three rainfall gauges and one synthetic discharge station for the study. We use four flood events that were generated by the GBHM following Equations 1 and 2 to calibrate the GBHM parameters. For each flood event, we first calculated flow rates as a function of time at the discharge station using the true GBHM parameters and the rainfall data at the three gauges, and then added AR(1) noise to the synthetic data. The AR(1) noise

was generated and added separately to the low- and high-flow data with the corresponding autocorrelation coefficients and inverse variances that are given in Table 2. We added synthetic input errors (**φ**) by drawing samples from the truncated normal distribution with the mean of $\mu_\varphi = 0$ and the inverse variance of $\tau_\varphi = 100$ (i.e., standard deviation of 10%).

Table 2 lists the true model parameters, prior ranges, and the estimated statistics of posterior marginal distributions for all the unknown parameters except the input errors (**φ**). Generally, the estimated medians and modes of the unknown parameters are close to their corresponding true values, and the 95% highest posterior density domains (HPDs) cover the true values. For the hyper-parameters of input errors (i.e., μ_φ and τ_φ), the HPDs are very close to their corresponding prior ranges. This means that the given datasets provide very limited information for updating those parameters. Additionally, the AR(1) model parameters related to the high-flow data (i.e., ρ_z and τ_z) are better resolved than those related to the low-flow data (i.e., ρ_y and τ_y). This is because we have used more information (hourly data) for the high flow than that for the low flow (daily data).

Figure 4 compares the estimated input errors with the corresponding true values. Overall, the estimated posterior medians of **φ** are close to the true values. However, uncertainty associated with the estimates is still high after calibration because the posterior distributions are widespread with most of the HPDs extending over 20% around the value of zero.

5.2. Effects of Input Uncertainty on Calibration

To illustrate the influence of input uncertainty on estimation of the GBHM parameters, we conducted comparative case studies with and without estimating **φ** in the Bayesian model. We generated synthetic data sets with two levels of input errors (i.e., 10% and 15%). However, we estimate the GBHM

parameters under two different situations, one ignoring the input errors and the other jointly estimating the input errors.

Table 3 summarizes the results of the comparative studies. We use the average of the absolute errors of the calibrated medians relative to their corresponding prior ranges as a measure of accuracy, and the average of the relative standard errors as a measure of uncertainty. We use the percentage of the true values covered by the estimated HPDs as a measure of reliability. It is evident that the calibration ignoring input errors has lower accuracy and lower reliability for the input errors of 10% and 15%. The negative effects on model calibration of ignoring input uncertainty increase with increasing of the input errors. Such results are consistent with the findings of many previous studies [e.g., *Kavetski et al.*, 2006; *Vrugt et al.*, 2008; *Thyer et al.* 2009] that found neglecting input uncertainty in calibrating hydrological models would cause biases and underestimate parameter uncertainty.

Figure 5 compares the estimated probability densities of each GBHM parameter with (black curves) and without (red curves) considering input errors for the case study having 15% input errors added. The black triangles are the true values, and all the priors are assumed to be uniform on the ranges shown in the figures. From the figure, we can see that although the influence varies depending on the parameters, ignoring input uncertainty clearly leads to worse estimates. As shown in Figure 5, the calibration of hydraulic conductivity in an unconfined aquifer (K_g) is biased for both situations. This is because the parameter is mainly determined by the low-flow data, which are very limited in the synthetic study considering the short daily time steps and high autocorrelation.

5.3. Effects of Model Structure Uncertainty

To investigate the robustness of the Bayesian calibration method under model structure uncertainty,

we conducted a synthetic study in which model structure errors were generated from the GBHM parameters with storm-dependent random biases. In accordance with the hypothesis proposed by *Kuczera et al.* [2006], the structural errors in the GBHM can be characterized by storm-dependent random variations of a subset of the model parameters. In this study, we let six GBHM parameters (i.e., k_{ep}, K_g, f_{ss}, d_{surf}, S_y, and n_s) vary through different storms. They were randomly sampled from the uniform distributions whose ranges were obtained by shrinking their original prior ranges around the mean by 40%. We calibrated the remaining three GBHM parameters (i.e., K_t, K_b, and n_c).

Table 4 is a summary of calibration results under various conditions. They include the cases where input errors are ignored by setting $\boldsymbol{\varphi}$=0 (see Row 2) and setting $\boldsymbol{\varphi}$ to their true values (see Row 3), and the cases jointly estimating the three GBHM parameters and the input errors with the lower prior bounds of the inverse variance from $1/(30\%)^2$ to $1/(10\%)^2$ (see Rows 4-7). The results show that in the presence of model structure errors, the joint estimation of GBHM parameters and input errors provides better calibration of the three GBHM parameters than estimating GBHM parameters alone. This is true even if we set the input errors to their true values in the latter case. The reason is that in the joint estimation, we fit the model to both the storm-dependent input errors and the unknown model structure errors.

Although the use of latent variables (i.e., $\boldsymbol{\varphi}$) for describing input uncertainty has a beneficial effect on estimation of GBHM parameters under model structure errors, it makes estimation of input errors difficult. This is consistent with the finding of *Renard et al.* [2010]. To distinguish between the input errors and the model structure errors, we need to use other methods or other types of information to obtain good priors on the input errors. For example, *Renard et al.* [2011] demonstrated that geostatistical analysis of distributed rain gauge data can provide useful information for the priors. For this study, since

we do not have good information about rainfall, we assume that the standard deviation of the input errors is less than 15%. This is an assumed prior restriction introduced from generally heuristic considerations, in an attempt to prevent the input error latent variables from over-fitting the structural errors.

6. Field Case Study

We applied the Bayesian model to the field dataset collected from the Chuzhou catchment in the Yangtze River basin. We used a group of 16 flood events for calibration and another group of 16 flood events for validation.

6.1. Selection of the Warm-up Period

To avoid possible influence of the initial state on the GBHM simulations, we need to initiate the GBHM a number of days before the data period. The length of the warm-up period may have significant effects on the model calibration due to uncertainty in the initial state [*Huard and Mailhot*, 2008]. Since the Bayesian calibration of a physically based distributed model is time-consuming and the computing time is proportional to the number of flood events, we perform a sensitivity study by using the first two flood events in the training group to select the warm-up length.

Figure 6 compares the estimated posterior probability densities of the calibrated parameters using 7, 20, 30, and 40 days for the warm-up period. The results show that k_{ep}, K_t, K_g, d_{surf}, and S_y are more sensitive to the length of the warm-up period than other parameters. This is reasonable because those parameters are directly related to the runoff generation, which is sensitive to the initial conditions. From the figure, we can see that the results of using the warm-up lengths of 20, 30, and 40 days are close to each other. Because the simulation time increases with increasing of the warm-up length, we select 30

days as the length of the warm-up period in the subsequent studies.

6.2. Calibration Results of GBHM Parameters

Table 5 lists the calibration results of the GBHM parameters. The statistics of hydraulic conductivity in the top soil (K_t) and bottom soil (K_b) are close to the upper bounds of their prior ranges. The estimated infiltration capability at the soil surface (K_t) is much higher than the database value of the corresponding soil texture (i.e., 32.7 mm/h). Since the catchment is covered by forests, we consider the estimated high surface infiltration capability to be related to the humus layer that is not represented in the GBHM model. The calibrated slope shape factor (f_{ss}) is low, corresponding to the low concave-convex effect of the hillslope on the hydrograph. The HPD of roughness in the river channel is estimated to be (0.019, 0.027), which is slightly lower than the observed value at the Chuzhou gauging station (between 0.04 and 0.07). The calibrated results are not expected to be exactly equal to the on-site observations because the calibration may include model structure uncertainty.

To demonstrate the influence of input uncertainty on the calibration, we conducted another calibration by ignoring input errors. The posterior pdfs of the GBHM parameters estimated through the comparative study are illustrated in Figure 7. The results are similar to those obtained from the synthetic study (see Figure 5). Without including input uncertainty, the estimates of some parameters are different from those obtained with inclusion of input uncertainty.

6.3. Calibration Results of Input Errors

The calibration results for the hyper-parameters of the input errors are listed in Table 6. The posterior mode and median of the mean relative rainfall errors (μ_φ) are about -0.041, and the HPD of μ_φ is (-0.050, -0.014). These suggest that the observed rainfall frequently underestimates the "true" rainfall. The

estimated inverse variance in rainfall errors, τ_{φ}, is very close to the low limit of its prior range (i.e, 44.4). This means that prior information on the variation of input errors is important for model calibrations as reported by *Renard et al.* [2010, 2011].

We compared the inferred distributions of individual latent variables for different gauges during a flood that occurred on September 22, 1981 (see Figure 8) to analyze cross-correlation of the estimated input errors at different rainfall gauges. The absolute values of correlation coefficients between the estimated input errors were mostly smaller than 0.25, but the correlation coefficient between the estimated input errors at the gauges P240 and P260 was -0.395. This is reasonable, because the flow lengths from the two gauges to the catchment outlet are very close, and the same amount of rainfall occurring at either gauge would produce a similar amount of discharge at the catchment outlet. Similar correlation was found between the estimated input errors at gauges P180 and P200 during another storm. The cross correlations between the input errors of other gauges were generally weak.

Figure 9 illustrates the original and estimated spatial distributions of precipitation in the Chuzhou catchment for one time step during the flood on September 22, 1981; the latter are calculated based on three posterior samples of the input errors. The figure shows that the estimated precipitation in most areas is close among different samples, whereas those around gauges P240 and P260 apparently vary. Similar to Figure 8, a negative correlation is found between the estimated precipitations in the two regions. Although not frequently appearing, the negative correlation may be caused by the poor identifiability of the spatial input. Given that there is much rainfall observed at gauge P260, the estimated sample 2 may imply that the heavy rainfall could possibly occur at gauge P240 instead of gauge P260 given what is known about outlet discharge. The result demonstrates that some model

structure errors may be over-fitted by the estimated input errors, and therefore, more spatial information and better priors are needed to resolve the ambiguity.

6.4. Calibration and Validation Results of Discharge

To evaluate the calibrated discharge with associated uncertainty, we plotted the estimated median discharge of the GBHM model together with the predictive uncertainty bands. Figure 10 shows the estimated median discharge of the 16 floods and their corresponding measurements during the calibration period. The Nash-Sutcliffe coefficient [*Nash and Sutcliffe*, 1970] of the simulation is 0.97 in Figure 10(a). The model residuals in Figure 10(b) evidently exhibit heteroscedasticity, which has often been reported in other studies [*Schaefli et al.*, 2007], thus justifying the separation of discharge between low- and high-flow data. The normal probability of the high flow residuals is plotted in Figure 10(c). Here, the residuals approximately follow the normal distribution, although there are apparent deviations for those with large absolute values.

Figures 10(d)(e) show the autocorrelation functions (ACF) and partial autocorrelation functions (PACF) of high-flow residuals, respectively. We can see that the first-order autoregressive coefficient is mostly significant for the high-flow period. The partial autocorrelation functions of lags two and three are also notable, yet are smaller than 0.25. As shown in Table 6, the estimated autocorrelation coefficient for low flow data in a daily time step is about 0.76, while that for hourly high flow data in an hourly time step is about 0.65. These results confirm that the autocorrelation for the low-flow period is more significant than for the high-flow period, and that the two kinds of data should be treated separately in calibration.

Figure 11 illustrates the 95% credible predictive intervals of discharge obtained through the

comparative studies that estimate input errors (case (a)) or fix input errors to zeros (case (b)). The narrower uncertainty bands are the propagation of the uncertainties on the GBHM parameters and on model inputs. The wider uncertainty bands are the estimated total uncertainty bands with inclusion of residual uncertainty for the narrow bands. The efficiency of predictive uncertainty (EPU) and the continuous ranked probability score (CRPS) [*Gneiting et al.*, 2004] are introduced to evaluate the predictive uncertainty, with the EPU coefficient defined as the percentage of the observed discharge values bracketed by the 95% predictive uncertainty limits—the most desirable value of EPU is 95%. The CRPS score is defined as [*Gneiting et al.*, 2004]

$$\mathrm{CRPS} = \frac{1}{T}\sum_{t=1}^{T}\int_{-\infty}^{\infty}(F_t(q) - H(q - q_t))^2\, dq \qquad (10)$$

where $F_t(q)$ denotes the cumulative probability function (CDF) of the predicted discharge at time step t, $H(q\text{-}q_t)$ denotes the Heaviside function of observed discharge q_t. H takes the value 0 when $q < q_t$; the value is 1 otherwise. The CRPS score takes the same unit as discharge and is more desirable when it is smaller. Each of the values for EPU and CRPS given in the figures of the present study is for the high-flow period, since our focus is flood forecasting. In both cases (a) and (b) in Figure 11, the observed discharge is well bracketed by the total uncertainty bounds. Nevertheless, the predictive discharge distribution when fixing input errors to zero is worse (i.e., having relatively wider predictive uncertainty bounds and larger CRPS scores) than that when estimating input errors.

We follow three main steps to perform the validation: (1) pick one set of GBHM parameters from the MCMC samples obtained in the calibration period; (2) pick one set of input errors ($\boldsymbol{\varphi}$) from the MCMC samples of one randomly selected flood event (out of 16) in the calibration period; and (3) run the

forward model (i.e., GBHM) using the parameters and input errors chosen from steps (1) and (2). Because the input errors are fitted in calibration and used in validation, the input uncertainty in the validation period is typically larger than that in the calibration period.

Figure 12 shows the evaluation of predicted median discharge for the 16 flood events in the validation period. The calibrated model provides good predictions of the observed data, with a Nash-Sutcliffe coefficient of 0.90. The normality of the high-flow residuals in the validation period is worse than that in the calibration period. Figures 12 (d)(e) show the ACF and PACF of high-flow residuals in the validation stage. Compared to the calibration stage, the autoregressive coefficients of lags 1–7 in the high-flow period are important, as are the partial autocorrelation coefficients of lags 2-3. A possible explanation for the result is that the input errors in the calibration period are estimated while those in the validation period remain unknown and are sampled from the same input error distribution as used in the calibration period. The input errors and the potential model structure errors may lead to highly correlated residuals in prediction—this is an issue that we will address in future studies.

Figure 13 compares the estimated uncertainty with and without estimating input errors in the calibration and validation periods. In the case that includes estimated input errors, the uncertainty bands caused by input uncertainty are very wide and dominate the prediction uncertainty. EPU values of 100% for the selected flood events are larger than their most favorable value, 95%, which implies that the predictive distributions of discharge overestimated the uncertainty. One possible reason is that the input errors picked up for the validation period incorporated the inter-storm variance of input errors in the calibration period. Another possible reason is that the estimated input errors over-fit some model structure errors in the calibration period; consequently, the wide uncertainty bands may not be accurate.

Nonetheless, the wide uncertainty bands induced by large input uncertainty are more efficient in bracketing observed discharge than are the uncertainty bands estimated with zero input errors. Meanwhile, the CRPS scores in Figure 13 also suggest that most predictive CDFs estimated jointly with input uncertainty are better than those estimated without considering input uncertainty. In addition, the input and model structure uncertainties for flood forecasting are quite significant, although we have not explicitly accounted for the latter in the present study.

6.5. Evaluation of Predictive Uncertainty

The predictive quantile-quantile plot (PQQ) for critically evaluating predictive uncertainty was introduced by *Thyer et al.* [2009] and has been frequently adopted since then [*Renard et al.*, 2010; *Schoups and Vrugt*, 2010] to test the consistency of predictive distribution with observed data. The cumulated probability of observed discharge in the predictive CDF is called the observed p-value. The observed p-value theoretically has the uniform distribution on (0, 1) if the observed data are consistent with samples from the predictive PDF.

Figure 14 shows the influences of input uncertainty on predictive uncertainty in the calibration and validation periods using the PQQ methodology. Here we show only the results that are related to the high-flow data, since the focus of this study is flood forecasting. For the calibration period, the PQQ curves of most flood events are near the diagonal line when the hydrological parameters and input errors are jointly estimated (see Figure 14(a)). The large symbols that represent peak flow distribute evenly in the PQQ curves. This result suggests that the predictive distribution is consistent with observed data. When input errors are ignored, the PQQ curves in calibration are more dispersive (see Figure 14(b)). This means the estimated predictive distributions are more frequently overestimating or underestimating

the predictive uncertainty when ignoring input errors than considering input errors.

Figure 14(c)(d) show that the consistency of the predictive distributions with observed data is worse in validation than in calibration, which is consistent with the detailed predictive uncertainty bounds shown in Figures 11 and 13. The uncertainty of observed data with relative small magnitude (i.e., having small size) tends to be overestimated more often, since many of them have p-values within (0.2, 0.8). This is similar to the finding of *Thyer et al.* [2009]. Observed data with relative large magnitude tend to have p-values near 0 or 1, and off the 1:1 line, which means that the uncertainty of flood peaks is still underestimated or overestimated in spite of the wide predictive uncertainty limits shown in Figure 13. Comparatively, underestimation of flood peaks is more frequent than overestimation, because the large symbols have more p-values near 1 than near 0, especially when input errors are not considered. Figure 14(d) shows that neglecting input errors results in a low EPU coefficient, high CPRS score, and frequent underestimation of flood peaks.

The results presented in the previous subsections are based on the upper bounds of 15% for the standard deviation of input errors. We also carried out analysis for the field study using the upper bounds of 10%, 20%, and 30% for the standard deviation of input errors. We found that most parameters are not very sensitive to the prior upper bounds of input errors, except for the evaporation parameter (k_{ep}) and the slope shape factor (f_{ss}). However, the effects on estimation of input errors are large, and this is consistent with the findings by *Renard et al.* [2010, 2011].

7. Discussion

7.1. Likelihood Function

The construction of a suitable likelihood function is crucial for Bayesian analysis because the

likelihood function determines how unknown parameters are linked to data. In this study, the separation between low- and high-flow data and employment of autoregressive models are two key assumptions for the derivation of likelihood functions.

The present method of classifying discharge data is a variation of the finite mixture model presented by *Schaefli et al.* [2007]. The advantage of classifying discharge data is that we can adopt different statistical models to different datasets, such as different time steps and different transformations. We use an hourly time-step for the high-flow period and daily time-step for the low-flow period. Moreover, classifying the hydrological parameters in accordance with different classifications of datasets is possible because some parameters are related to low-flow data and some are primarily related to high-flow data. We could classify such parameters in the future when we find a data transformation method to reduce the error propagation between the low-flow and the high-flow periods. The classification method is limited by the need to maintain independence among different classifications. Thus, it should be conducted based on a physical rationale. To avoid confusion, we abandon some discharge datasets with a high-flow period mixed with a low-flow period in the current study.

The results of both synthetic and field case studies show consistency of the AR(1) model with model residuals (see Figure 10). Nevertheless, validation results from the field case study reveal that autocorrelation functions of higher lags are also important in prediction (see Figure 12). In addition to adopting autoregressive models based on time series analysis, there is potentially an alternative way of treating a flood event in its entirety and constructing likelihood functions based on the internal structure of residual series.

Spatially distributed data can be adopted for calibration and diagnosis of distributed hydrological

models [*Marce et al.*, 2008]. Under the Bayesian framework, further research is expected to combine different likelihood functions from varied information to calibrate spatially distributed parameters. More general likelihood functions other than normal distributions are also expected to improve the inferences of the parameters [*Schoups and Vrugt*, 2010].

7.2. Influence of Input Uncertainty

The input uncertainty of the GBHM model is represented by spatially distributed relative rainfall errors in the developed Bayesian model. Results from the synthetic case study show that the integrated estimation of input errors leads to a more reasonable calibration of the GBHM parameters, including the posterior medians and the uncertainty bands. The field case study also reveals apparent differences between the posterior distributions of certain GBHM parameters with and without considering input uncertainty. These differences emphasize the significance of considering input uncertainty in calibration, as reported by the BATEA and DREAM approaches [*Kavetski et al.*, 2006; *Vrugt et al.*, 2008]. The most evident influence of input uncertainty is in the validation period of the field case study, where the consideration of input uncertainty expands the predictive uncertainty bands to bracket a number of observed discharge values (see Figure 13) and improve the performance of predictive distributions (see Figure 14), especially for flood peak periods. However, it is still challenging to fully cope with the effects of input uncertainty, because of the compound effects of model structure uncertainty and the poor identifiability of spatial input errors.

In the prior model, the input errors of the GBHM model at different gauges are specified as independent. The spatial correlation of the input errors for different computation grids is propagated by the interpolation of the estimated precipitation at the gauges. The calibration result of the field case

study reveals that the estimated input errors are sometimes negatively correlated when the flow distances are close (see Figures 8 and 9), which displays the limitation of the current Bayesian inference of using only outlet discharge for estimating spatially distributed variables. Spatially distributed information is important for better inference because of this limitation. Meanwhile, both the synthetic and the field case studies demonstrate that the estimated input errors are restricted by the specified prior ranges of their hyper-parameters. The estimated input uncertainty would be larger if we release the restrictions. Therefore, prior knowledge of rainfall observation and distribution is crucial for more accurate estimation of input errors. Studies on quantitative estimation of spatially distributed precipitation [*Clark and Slater*, 2006; *Renard et al.*, 2011] can help improve existing knowledge of rainfall uncertainty.

7.3. Influence of Model Structure Uncertainty

Model structure uncertainty is not represented explicitly in the current study, although we have shown the interaction between the input and the model structure errors in the synthetic study (see Table 4). The estimated input errors may fit the model structure errors because they are both storm-dependent, which is termed as compensatory behavior by *Kavetski et al.* [2006]. Therefore, the prior ranges of the hyper-parameters of input errors are specified to prevent the trend of over-fitting. The difficulty of identifying input and model structure errors together has been discussed extensively by *Renard et al.* [2010], based on the BATEA framework [*Kuczera et al.*, 2006]. We would emphasize that the estimation of model structure uncertainty is another challenging topic that deserves further investigation.

In the present study, the effect of the model structure errors on model residuals is characterized by the AR(1) model, because model structure errors are usually autoregressive due to internal storage processes [*Bates and Campbell*, 2001; *Yang et al.*, 2007]. The input errors also lead to autocorrelation in model

residuals. Therefore, the estimated autocorrelation of model residuals is a combined effect of both model structure uncertainty and input uncertainty.

Nevertheless, we believe that model limitations are partially caused by the residual influence of model structure uncertainty. For example, some estimated physical parameters are evidently different from their database or point observation values, the estimated inverse variance of rainfall errors is close to the low limit of its prior range, and the predictive distributions of the peak period discharge in the validation period tend to underestimate the predictive uncertainty, even if the input uncertainty is considered. Further studies are certainly needed to improve our understanding of the physical parameters of catchments, the inversion of input errors, and the performance of the predictive distributions.

7.4. Sensitivity of Calibration Results to Prior Information

Acknowledging that the accuracy of prior distributions is crucial for calibration results, we have conducted sensitivity analyses for both the synthetic and field case studies. Sensitivity is low for the synthetic study and high for the field case study. For the field case study, the influence is relative lower for the calibration results of GBHM parameters, but higher for the inferred input errors and for the predictive uncertainty. This confirms the findings of *Renard et al.* [2010] that the inference of input errors is ill-posed when the prior information is not sufficiently accurate. An alternative way to determine prior information is needed for both the inference of input errors and for the decomposition of predictive uncertainty.

8. Conclusions

We have developed a Bayesian model to aid in predicting mountain floods and have investigated calibration of hydrological parameters with consideration of input errors in a distributed hydrological

model. The synthetic case study shows that the developed method can be used to calibrate GBHM parameters and to estimate the associated uncertainty. The calibration ignoring input errors has lower accuracy and lower reliability compared to the calibration that includes estimation of the input errors, especially under model structure uncertainty.

However, calibration of GBHM parameters under complex field conditions remains a challenge. Although jointly estimating input errors and GBHM parameters improves the continuous ranked probability score and the consistency of the predictive distribution with the observed data, the improvement is incremental. To better calibrate parameters in a distributed model, such as GBHM here, we need to develop a more complex model and incorporate much more information.

Appendix A. Probability Density Function of Truncated Normal Distribution

When variable x is distributed normally within an interval (a, b), the probability density function can be written in the following form [*Huard and Mailhot*, 2008]:

$$f(x) = \frac{\frac{1}{\sigma} \cdot \phi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} \tag{A1}$$

where μ is the mean, σ is the standard deviation, $\phi()$ is the pdf of standard normal distribution, and $\Phi()$ is its cumulative distribution function. Here, we assume that $\varphi_{k,j}$ for different storms at the same gauge follows a truncated normal distribution within (-1, +∞), with mean μ_φ and inverse variance τ_φ. Thus, the pdf of $\varphi_{k,j}$ is given as

$$\begin{aligned} f(\varphi_{k,j} \mid \mu_\varphi, \tau_\varphi) &= \tau_\varphi^{\frac{1}{2}} \cdot \phi\left(\tau_\varphi^{\frac{1}{2}}(\varphi_{k,j} - \mu_\varphi)\right) \cdot \left(1 - \Phi\left(\tau_\varphi^{\frac{1}{2}}(-1 - \mu_\varphi)\right)\right)^{-1} \\ &\propto \tau_\varphi^{\frac{1}{2}} \exp\left(-\frac{1}{2}\tau_\varphi(\varphi_{k,j} - \mu_\varphi)^2\right)\left(1 - \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{1}{\sqrt{2}}\tau_\varphi^{\frac{1}{2}}(-1 - \mu_\varphi)\right)\right)\right)^{-1} \\ &\propto \tau_\varphi^{\frac{1}{2}} \exp\left(-\frac{1}{2}\tau_\varphi(\varphi_{k,j} - \mu_\varphi)^2\right)\left(1 + \mathrm{erf}\left(\frac{1}{\sqrt{2}}\tau_\varphi^{\frac{1}{2}}(\mu_\varphi + 1)\right)\right)^{-1} \end{aligned} \tag{A2}$$

where erf() is the error function (http://en.wikipedia.org/wiki/Error_function).

Appendix B. Derivation of Likelihood Functions Based on the AR(1) Model

The original form of the likelihood function is given in Equation 4. Assuming the residuals following the AR(1) model [*Naidu*, 1996], the terms in the covariance matrix, Σ_k, are

$$\mathrm{cov}(\varepsilon_{y,t}, \varepsilon_{y,t-c}) = \frac{(\rho_y)^c}{\tau_y}, \qquad c = 1, \ldots, n. \tag{B1}$$

Therefore, $\Sigma_k = R_k/\tau_y$, and

$$R_k = \begin{pmatrix} 1 & \rho_y & \cdots & (\rho_y)^{n-1} \\ \rho_y & \ddots & & \vdots \\ \vdots & & \ddots & \rho_y \\ (\rho_y)^{n-1} & \cdots & \rho_y & 1 \end{pmatrix}, \tag{B2}$$

where R_k denotes the correlation coefficient matrix of the model residuals. The likelihood function is thus written as

$$\begin{aligned} f(\mathbf{Y}_k \mid \boldsymbol{\theta}, \boldsymbol{\varphi}_k, \rho_y, \tau_y, \mathbf{X}_k) &\propto |R_k|^{-\frac{1}{2}} \tau_y^{\frac{1}{2}n} \exp\left(-\frac{1}{2}\tau_y (\mathbf{Y}_k - \mathbf{Y}_k^{sim})^T R_k^{-1} (\mathbf{Y}_k - \mathbf{Y}_k^{sim}) \right) \\ &= |R_k|^{-\frac{1}{2}} \tau_y^{\frac{1}{2}n} \exp\left(-\frac{1}{2}\tau_y S_k \right) \end{aligned} \tag{B3}$$

where

$$S_k = (\mathbf{Y}_k - \mathbf{Y}_k^{sim})^T R_k^{-1} (\mathbf{Y}_k - \mathbf{Y}_k^{sim}) \tag{B4}$$

Using the Cholesky decomposition of matrix R_k, we can have

$$R_k^{-1} = U_k^T U_k \tag{B5}$$

where U_k is an upper triangular matrix. Therefore, S_k can be evaluated by

$$\begin{aligned} S_k &= (\mathbf{Y}_k - \mathbf{Y}_k^{sim})^T U_k^T U_k (\mathbf{Y}_k - \mathbf{Y}_k^{sim}) \\ &= (U_k (\mathbf{Y}_k - \mathbf{Y}_k^{sim}))^T (U_k (\mathbf{Y}_k - \mathbf{Y}_k^{sim})) \end{aligned} \tag{B6}$$

Acknowledgments

The research was supported by the National Natural Science Foundation of China (project no. 50939004 and 51025931), and by the US Department of Energy, Biological and Environmental Research Program, Contract No. DE-AC02-05CH11231 to Lawrence Berkeley National Laboratory. The first author was also partially supported by a scholarship from the China Scholarship Council. We wish to thank Hoshin Gupta, Jasper Vrugt, and three anonymous reviewers for valuable comments and suggestions. We also wish to thank Daniel Hawkes and Michael Kowalsky from Lawrence Berkeley

National Laboratory for providing detailed comments on the manuscript.

References

Abbott, M. B., J. C. Bathurst, J. A. Cunge, P. E. O'Connell, and J. Rasmussen (1986), An introduction to the European Hydrological System -- Systeme Hydrologique Europeen, "SHE", 1: History and philosophy of a physically-based, distributed modelling system, *J. Hydrol.*, *87*(1-2), 45-59, doi:10.1016/0022-1694(86)90114-9.

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

Ajami, N. K., Q. Duan, and S. Sorooshian (2009), Reply to Comment by B. Renard et al. on "An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction", *Water Resour. Res.*, *45*, W03604, doi:10.1029/2008WR007215.

Bates, B. C., and E. P. Campbell (2001), A Markov chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modeling, *Water Resour. Res.*, *37*(4), 937-947, doi:10.1029/2000WR900363.

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

Chen, J. S., G. M. Hoversten, D. Vasco, Y. Rubin, and Z. S. Hou (2007), A Bayesian model for gas saturation estimation using marine seismic AVA and CSEM data, *Geophysics*, *72*(2), WA85-WA95, doi:10.1190/1.2435082.

Chen, J. S., A. Kemna, and S. S. Hubbard (2008), A comparison between Gauss-Newton and Markov-chain Monte Carlo-based methods for inverting spectral induced-polarization data for

Cole-Cole parameters, *Geophysics*, *73*(6), F247-F259, doi:10.1190/1.2976115.

Chow, V. T., R. M. David, and W. M. Larry (1988), *Applied Hydrology*, McGraw-Hill Inc.

Clark, M. P., and A. G. Slater (2006), Probabilistic quantitative precipitation estimation in complex terrain, *J. Hydrometeorol.*, *7*(1), 3-22.

Clark, M. P., A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener, and L. E. Hay (2008), Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, *Water Resour. Res.*, *44*, W00B02, doi:10.1029/2007WR006735.

Cong, Z. T., D. W. Yang, B. Gao, H. B. Yang, and H. P. Hu (2009), Hydrological trend analysis in the Yellow River basin using a distributed hydrological model, *Water Resour. Res.*, *45*, W00A13, doi:10.1029/2008WR006852.

Gelfand, A. E., and A. F. M. Smith (1990), Sampling-based approaches to calculating marginal densities, *J. Am. stat. Assoc.*, *85*(410), 398-409, doi:10.2307/2289776.

Gelman, A., and D. Rubin (1992), Inference from iterative simulation using multiple sequences, *Statist. Sci.*, *7*(4), 457-472, doi:10.1214/ss/1177011136.

Geman, S., and D. Geman (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*(6), 721-741, doi:10.1109/TPAMI.1984.4767596.

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

Gneiting, T., A. Westveld, A. Raferty, and T. Goldman (2004), Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, (Technical Report no. 449), Department of Statistics, University of Washington, Washington.

Gotzinger, J., and A. Bardossy (2008), Generic error model for calibration and uncertainty estimation of hydrological models, *Water Resour. Res.*, *44*, W00B07, doi:10.1029/2007WR006691.

Green, P., and A. Mira (2001), Delayed rejection in reversible jump metropolis-hastings, *Biometrika*, *88*, 1035-1053.

Gupta, H., K. Beven, and T. Wagener (2005), Model calibration and uncertainty estimation, in *Encyclopedia of Hydrological Sciences*, edited by M. G. Anderson, John Wiley & Sons.

Haario, H., E. Saksman, and J. Tamminen (2001), An adaptive Metropolis algorithm, *Bernoulli*, *7*(2), 223-242.

Haario, H., M. Laine, A. Mira, and E. Saksman (2006), DRAM: Efficient adaptive MCMC, *Stat. Comput.*, *16*(4), 339-354, doi:10.1007/s11222-006-9438-0.

Hastings, W. K. (1970), Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, *57*(1), 97-109.

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

Immerzeel, W. W., and P. Droogers (2008), Calibration of a distributed hydrological model based on satellite evapotranspiration, *J. Hydrol.*, *349*(3-4), 411-424, doi:10.1016/j.jhydrol.2007.11.017.

Kavetski, D., S. W. Franks, and G. Kuczera (2003), Confronting input uncertainty in environmental modelling, in *Calibration of Watershed Models*, *Water Sci. Appl.*, vol. 6, edited by Q. Duan et al., pp. 49–68, AGU, Washington, D. C., doi:10.1029/WS006p0049.

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

Khu, S. T., H. Madsen, and F. di Pierro (2008), Incorporating multiple observations for distributed hydrologic model calibration: An approach using a multi-objective evolutionary algorithm and clustering, *Adv. Water Resour.*, *31*(10), 1387-1398.

Kuczera, G. (1988), On the validity of first-order prediction limits for conceptual hydrologic models, *J. Hydrol.*, *103*(3-4), 229–247.

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

Kuczera, G., D. Kavetski, S. Franks, and M. Thyer (2006), Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters, *J. Hydrol.*, *331*(1-2), 161-177, doi:10.1016/j.jhydrol.2006.05.010.

Liu, Y. Q., and H. V. Gupta (2007), Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, *Water Resour. Res.*, *43*,W07401, doi:10.1029/2006WR005756.

Marce, R., C. E. Ruiz, and J. Armengol (2008), Using spatially distributed parameters and multi-response objective functions to solve parameterization of complex applications of semi-distributed hydrological models, *Water Resour. Res.*, *44*, W02436, doi:10.1029/2006WR005785.

Marshall, L., D. Nott, and A. Sharma (2004), A comparative study of Markov chain Monte Carlo methods for conceptual rainfall-runoff modeling, *Water Resour. Res.*, *40*,W02501, doi:10.1029/2003WR002378.

McMillan, H., B. Jackson, M. Clark, D. Kavetski, and R. Woods (2011), Rainfall uncertainty in hydrological modelling: An evaluation of multiplicative error models, *J. Hydrol.*, *400*(1-2), 83-94, doi:10.1016/j.jhydrol.2011.01.026.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953), Equation of State Calculations by Fast Computing Machines, *J. Chem. Phys.*, *21*(6), 1087-1092.

Naidu, P. S. (1996), *Modern Spectrum Analysis of Time Series*, CRC Press.

Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models, part I, A discussion of principles, *J. Hydrol.*, *10*, 398-409.

New, M., M. Hulme, and P. Jones (2000), Representing twentieth-century space-time climate variability. Part II: Development of 1901-96 monthly grids of terrestrial surface climate, *J. Climate*, *13*(13), 2217-2238.

Reichert, P., and J. Mieleitner (2009), Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters, *Water Resour. Res.*, *45*, W10402, doi:10.1029/2009WR007814.

Renard, B., D. Kavetski, and G. Kuczera (2009), Comment on "An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction" by Newsha K. Ajami et al., *Water Resour. Res.*, *45*, W03603, doi:10.1029/2007WR006538.

Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks (2010), Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, *Water Resour. Res.*, *46*, W05521, doi:10.1029/2009WR008328.

Renard, B., D. Kavetski, E. Leblois, M. Thyer, G. Kuczera, and S. W. Franks (2011), Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation, *Water Resour. Res.*, *47*(11), W11516, doi: 10.1029/2011WR010643.

Robinson, J. S., and M. Sivapalan (1996), Instantaneous response functions of overland flow and

subsurface stormflow for catchment models, *Hydrol. Processes*, *10*(6), 845-862.

Salamon, P., and L. Feyen (2010), Disentangling uncertainties in distributed hydrological modeling using multiplicative error models and sequential data assimilation, *Water Resources Research*, *46*, W12501, doi:10.1029/2009WR009022.

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

Schoups, G., and J. A. Vrugt (2010), A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, *Water Resour. Res.*, *46*, W10531, doi:10.1029/2009WR008933.

Shrestha, R. R., and M. Rode (2008), Multi-objective calibration and fuzzy preference selection of a distributed hydrological model, *Environ. Modell. & Softw.*, *23*(12), 1384-1395, doi:10.1016/j.envsoft.2008.04.001.

Singh, V. P., and D. A. Woolhiser (2002), Mathematical modeling of watershed hydrology, *J. Hydrologic Eng.*, *7*(4), 270-292, doi:10.1061/(ASCE)1084-0699(2002)7:4(270).

Thyer, M., G. Kuczera, and Q. J. Wang (2002), Quantifying parameter uncertainty in stochastic models using the Box-Cox transformation, *J. Hydrol.*, *265*(1-4), 246-257.

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

Vrugt, J. A., C. J. F. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson (2008), Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte

Carlo simulation, *Water Resour. Res.*, *44*, W00B09, doi:10.1029/2007WR006720.

Yang, D. W. (1998), Distributed hydrological model using hillslope discretization based on catchment area function: development and applications, Ph.D. thesis, University of Tokyo, Tokyo.

Yang, D. W., and K. Musiake (2003), A continental scale hydrological model using the distributed approach and its application to Asia, *Hydrol. Processes*, *17*, 2855-2869, doi:10.1002/hyp.1438.

Yang, D. W., S. Herath, and K. Musiake (2002), A hillslope-based hydrological model using catchment area and width functions, *Hydrol. Sci. J.*, *47*(1), 49-65.

Yang, D. W., T. Koike, and H. Tanizawa (2004), Application of a distributed hydrological model and weather radar observations for flood management in the upper Tone River of Japan, *Hydrol. Processes*, *18*, 3119-3132, doi:10.1002/hyp.5752.

Yang, J., P. Reichert, K. C. Abbaspour, and H. Yang (2007), Hydrological modelling of the chaohe basin in china: Statistical model formulation and Bayesian inference, *J. Hydrol.*, *340*(3-4), 167-182, doi:10.1016/j.jhydrol.2007.04.006.

Tables

Figure Captions

Figure 1: Schematic diagram of the geomorphology-based hydrological model (GBHM).

Figure 2: Definition of saturated outflow length l_s in an actual hillslope and a GBHM hillslope.

Figure 3: Description of the Chuzhou catchment and the synthetic basin, where "Q" and "P" refer to discharge gauges and rainfall gauges respectively, and "syn" refers to synthetic gauges. One actual rainfall gauge and two synthetic rainfall gauges are used for the synthetic study.

Figure 4: Estimated $\boldsymbol{\varphi}$ versus the true values of $\boldsymbol{\varphi}$ in the synthetic study with 10% input errors. Each point represents an input error for an individual storm at one particular gauge (there are a total of four storms and three rainfall gauges). The solid circles represent the estimated medians, and vertical line segments show the 95% highest posterior density domains (HPD).

Figure 5: Comparison of the posterior distributions of the GBHM parameters for the synthetic study with 15% input errors. The black curves are the results obtained by jointly estimating rainfall errors, and the red curves are those obtained by ignoring rainfall errors (i.e., fixing $\boldsymbol{\varphi}$=0). The black triangles are the true values. All the priors are assumed to be uniform on the ranges shown in the figures.

Figure 6: Comparison of the effects of different warm-up lengths on the model calibration. The black solid, red dashed, green solid and blue dashed curves are the estimated posterior probability densities of GBHM parameters using the warm-up lengths of 7, 20, 30, and 40 days, respectively.

Figure 7: Comparison of the posterior distributions densities of the GBHM parameters for the field case study given by considering (black curves) and ignoring (red curves) rainfall errors.

Figure 8: Scatter-plots and histograms of the estimated input errors for different rainfall gauges during the flood on September 22, 1981.

Figure 9: Observed and estimated spatial distributions of precipitation over the Chuzhou catchment for a time step during the flood on September 22, 1981. The angular distance-weighted method is used for the spatial interpolation.

Figure 10: Evaluation of the estimated discharge medians for the calibration period: (a) medians of the estimated discharges (black curves) and observed discharges (red circles); (b) plots of residuals between the simulated and observed discharges; (c) normal probability plot of residuals for the high flow data; (d) autocorrelation functions (ACF) of high flow residuals; and (e) partial autocorrelation functions (PACF) of high flow residuals. For readability, 16 isolated floods are plotted together in (a) and (b).

Figure 11: Simulated and observed discharges for the calibration period: (a) Bayesian model that estimates $\boldsymbol{\varphi}$ together with GBHM parameters in calibration. The darker shadow areas show the 95% model simulation uncertainty related to the uncertainty of GBHM parameters and input errors. The lighter shadow areas show the total uncertainty with the additional influence of model residuals. Within a flood event, discharges with relatively wide uncertainty ranges are those classified into high flow data; others are classified into low flow data. (b) Fixing $\boldsymbol{\varphi} = 0$ in calibration. Thus, the dark shadow areas are related exclusively to parameter uncertainty.

Figure 12: Evaluation of the estimated discharge medians for the validation period: (a) medians of the estimated discharges (black curves) and observed discharges (red circles); (b) plots of residuals between the simulated and observed discharges; (c) normal probability plot of residuals for the high flow data; (d) autocorrelation functions (ACF) of high flow residuals; and (e) partial autocorrelation functions (PACF) of high flow residuals. For readability, 16 isolated floods are plotted together in (a) and (b).

Figure 13: Predicted and observed discharges for the validation period: (a) Bayesian model that

estimates $\boldsymbol{\varphi}$ together with GBHM parameters in prediction. The darker shadow areas show the 95% model prediction uncertainty related to the uncertainty of GBHM parameters and input errors. The lighter shadow areas show the total uncertainty with additional influence of model residuals; and (b) $\boldsymbol{\varphi} = 0$ fixed in prediction. Thus, the dark shadow areas are related exclusively to parameter uncertainty.

Figure 14: Predictive quantile-quantile plot (PQQ) of the estimated predictive uncertainty for both calibration and validation periods. The PQQ curves of 16 flood events are plotted together. The symbol size is proportional to the magnitude of the observed data.

Table 1. Prior ranges of parameters for calibration of the GBHM

No.	Parameter	Description of parameter	Unit	Distribution
1	k_{ep}	Coefficient of potential evaporation		U(0.5, 1.5)
2	$\log(K_t)$	Saturated hydraulic conductivity of top soil	mm/h	U(1, 3)
3	$\log(K_b)$	Saturated hydraulic conductivity of bottom soil	mm/h	U(0, 2)
4	$\log(K_g)$	Saturated hydraulic conductivity of unconfined aquifer	mm/h	U(0, 1)
5	$\log(f_{ss})$	Slope shape factor		U(-1.5, -0.5)
6	d_{surf}	Maximum surface storage	mm	U(3, 15)
7	$\log(S_y)$	Specific yield		U(-2, -1)
8	$\log(n_s)$	Roughness on the hillslope surface		U(0, 1)
9	$\log(n_c)$	Roughness in the river channel		U(-2.5, -1)

Table 2. True values, prior ranges, and estimated statistics of the calibrated variables for the synthetic study with 10% input errors, where "HPD" refers to the 95% highest posterior density domains

	True value	Prior range	Mode	Median	95% HPD
k_{ep}	1	(0.5, 1.5)	0.678	0.770	(0.502,1.274)
$\log(K_t)$	2	(1, 3)	2.206	2.173	(1.683,2.592)
$\log(K_b)$	1	(0, 2)	0.863	0.949	(0.471,1.604)
$\log(K_g)$	0.5	(0, 1)	0.578	0.565	(0.381,0.725)
$\log(f_{ss})$	-1	(-1.5, -0.5)	-1.139	-1.106	(-1.421,-0.780)
d_{surf}	9	(3, 15)	11.976	11.772	(8.606,14.855)
$\log(S_y)$	-1.5	(-2, -1)	-1.240	-1.268	(-1.566,-1.000)
$\log(n_s)$	0.5	(0, 1)	0.466	0.462	(0.327,0.570)
$\log(n_c)$	-1.75	(-2.5, -1)	-1.745	-1.740	(-1.944,-1.521)
μ_φ	0	(-0.05, 0.05)	-0.005	-0.003	(-0.049, 0.044)
τ_φ	100	(44.4, 400)	114.9	140.6	(44.8, 331.6)
ρ_y	0.8	(0, 1)	0.825	0.795	(0.478, 0.973)
ρ_z	0.8	(0, 1)	0.807	0.807	(0.710, 0.919)
τ_y	100	(0, +∞)	105.3	113.5	(24.3, 218.8)
τ_z	25	(0,+∞)	26.2	26.3	(11.3, 40.2)

Table 3. Calibration results of the GBHM parameters with and without estimating input errors in the synthetic case study

Specified input errors [a]	Input errors in calibration	Mean absolute error (%) [b]	Mean St. dev. (%) [c]	Percentage located in HPD (%) [d]
10%	Estimated	11.35	13.57	100.0
	Not estimated	12.23	12.62	88.9
15%	Estimated	8.07	13.06	88.9
	Not estimated	18.52	11.02	55.6

[a] Standard deviation of input errors used for generating synthetic data

[b] Mean of absolute errors of calibrated medians of GBHM parameters relative to their corresponding prior ranges

[c] Mean of standard deviations of the posterior distributions of the GBHM parameters relative to their corresponding prior ranges

[d] Percentage of true values of GBHM parameters located in the estimated 95% highest posterior density domain for nine parameters

Table 4. Calibration results of the three GBHM parameters with and without estimating input errors when the synthetic data include model structure errors. The standard deviation of input errors used for generation of synthetic dataset is set to 10%.

Input errors in calibration	Mean absolute error (%) [a]	Mean St. dev. (%) [b]	Percentage located in HPD (%) [c]
Not estimated (set $\boldsymbol{\varphi} = 0$)	9.09	6.41	66.7
Not estimated (set $\boldsymbol{\varphi}$ to their true values)	8.65	4.76	33.3
Estimated with $\tau_{\varphi} \sim [1/(10\%)^2, 1/(5\%)^2]$	2.81	5.34	100.0
Estimated with $\tau_{\varphi} \sim [1/(15\%)^2, 1/(5\%)^2]$	2.80	5.46	100.0
Estimated with $\tau_{\varphi} \sim [1/(20\%)^2, 1/(5\%)^2]$	2.62	5.48	100.0
Estimated with $\tau_{\varphi} \sim [1/(30\%)^2, 1/(5\%)^2]$	2.91	5.55	100.0

[a] Mean of absolute errors of calibrated medians of GBHM parameters relative to their corresponding prior ranges

[b] Mean of standard deviations of the posterior distributions of the GBHM parameters relative to their corresponding prior ranges

[c] Percentage of true values of GBHM parameters located in the estimated 95% highest posterior density domain for three parameters

Table 5. Estimated posterior statistics of the GBHM parameters

Parameter	Mode	Median	Mean	Low HPD	Up HPD
k_{ep}	1.192	1.196	1.193	0.960	1.439
K_t (mm/h)	956.185	914.086	859.605	613.440	999.986
K_b (mm/h)	95.424	90.367	85.962	63.661	100.000
K_g (mm/h)	3.665	3.669	3.647	3.070	4.489
f_{ss}	0.063	0.066	0.066	0.047	0.080
d_{surf} (mm)	3.112	3.248	3.333	3.000	4.036
S_y	0.010	0.011	0.011	0.010	0.013
n_s	2.400	2.387	2.376	1.914	3.036
n_c	0.024	0.023	0.023	0.019	0.027

Table 6. Posterior statistics of variables in the Bayesian model other than GBHM parameters

	μ_φ	τ_φ	ρ_y	ρ_z	τ_y	τ_z
Mode	-0.046	45.007	0.779	0.651	293.938	8.640
Median	-0.041	45.612	0.757	0.651	311.487	8.628
Mean	-0.038	46.213	0.746	0.652	321.400	8.628
Low HPD	-0.050	44.440	0.571	0.568	151.660	6.468
Up HPD	-0.014	49.921	0.903	0.735	511.983	10.703

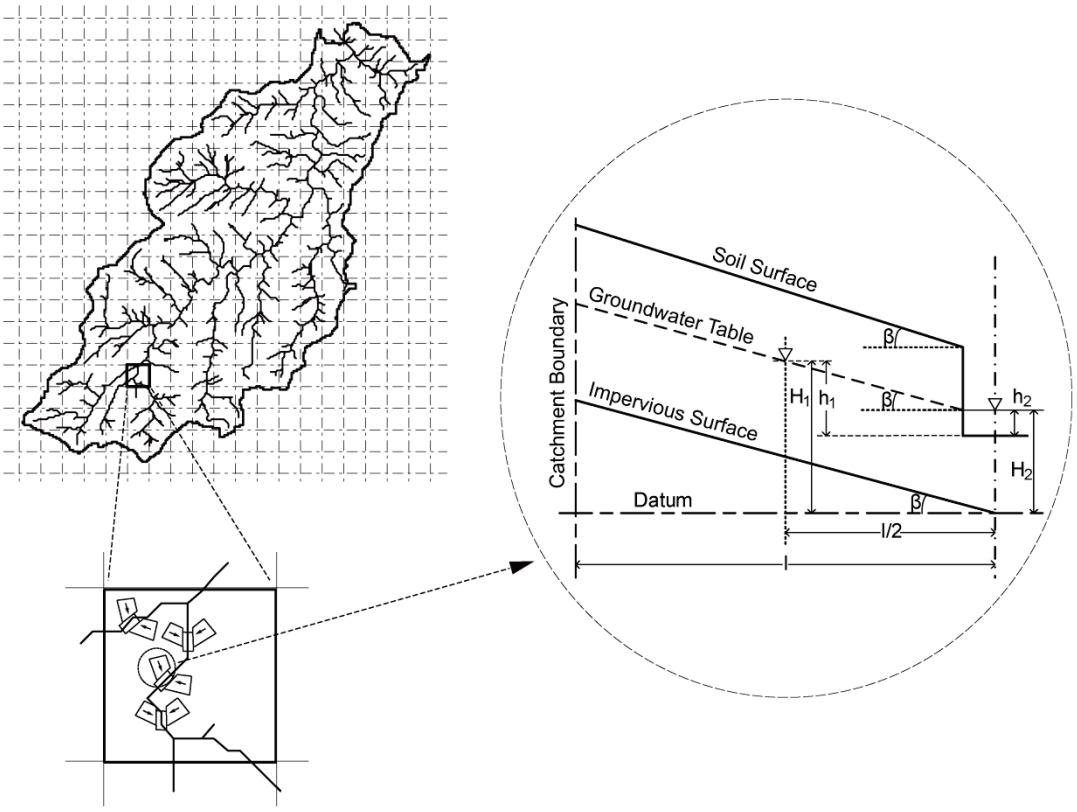

Figure 1. Schematic diagram of the geomorphology-based hydrological model (GBHM).

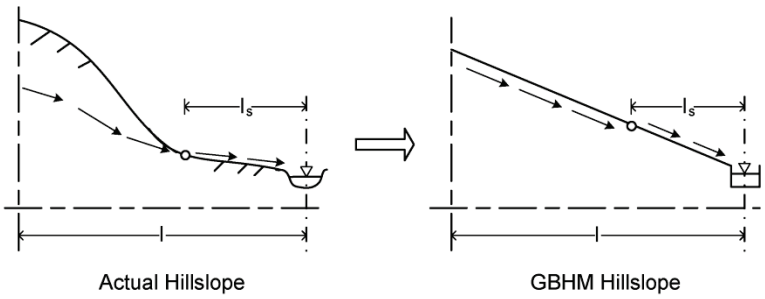

Figure 2. Definition of saturated outflow length l_s in an actual hillslope and a GBHM hillslope.

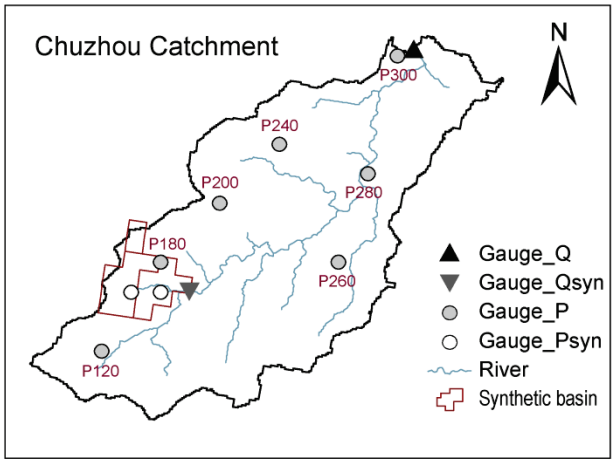

Figure 3. Description of the Chuzhou catchment and the synthetic basin, where "Q" and "P" refer to discharge gauges and rainfall gauges respectively, and "syn" refers to synthetic gauges. One actual rainfall gauge and two synthetic rainfall gauges are used for the synthetic study.

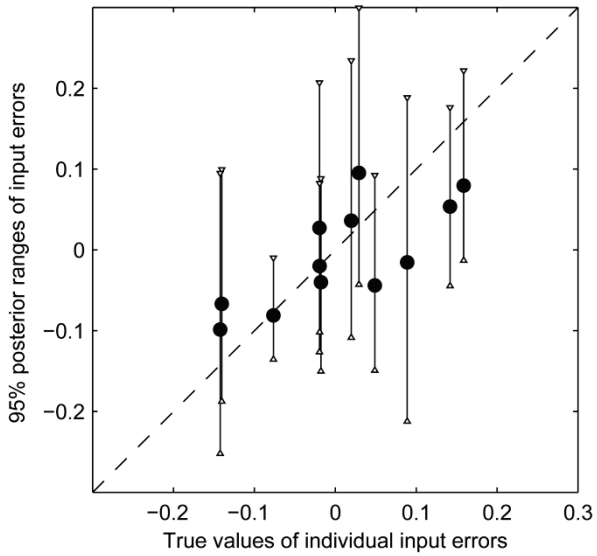

Figure 4. Estimated **φ** versus the true values of **φ** in the synthetic study with 10% input errors. Each point represents an input error for an individual storm at one particular gauge (there are a total of four storms and three rainfall gauges). The solid circles represent the estimated medians, and vertical line segments show the 95% highest posterior density domains (HPD).

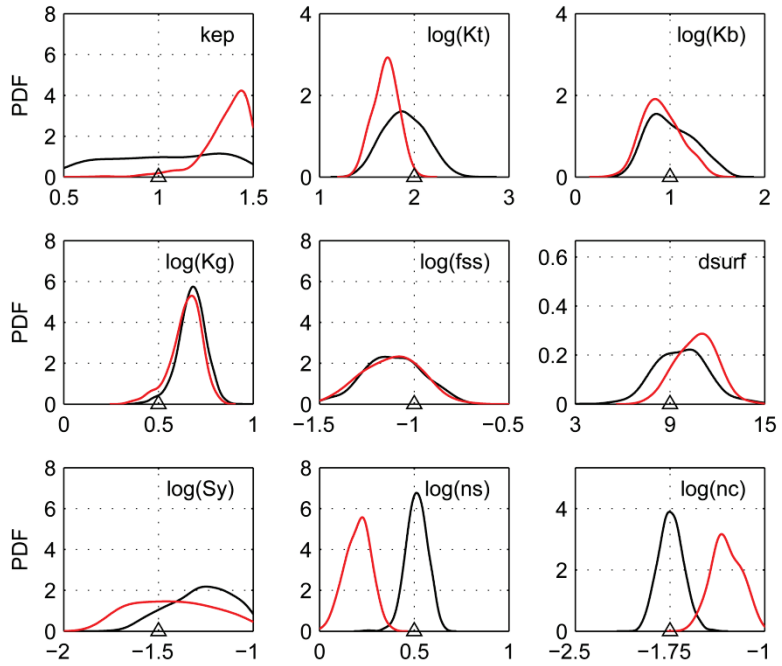

Figure 5. Comparison of the posterior distributions of the GBHM parameters for the synthetic study with 15% input errors. The black curves are the results obtained by jointly estimating rainfall errors, and the red curves are those obtained by ignoring rainfall errors (i.e., fixing $\boldsymbol{\varphi}$=0). The black triangles are the true values. All the priors are assumed uniform on the ranges shown in the figures.

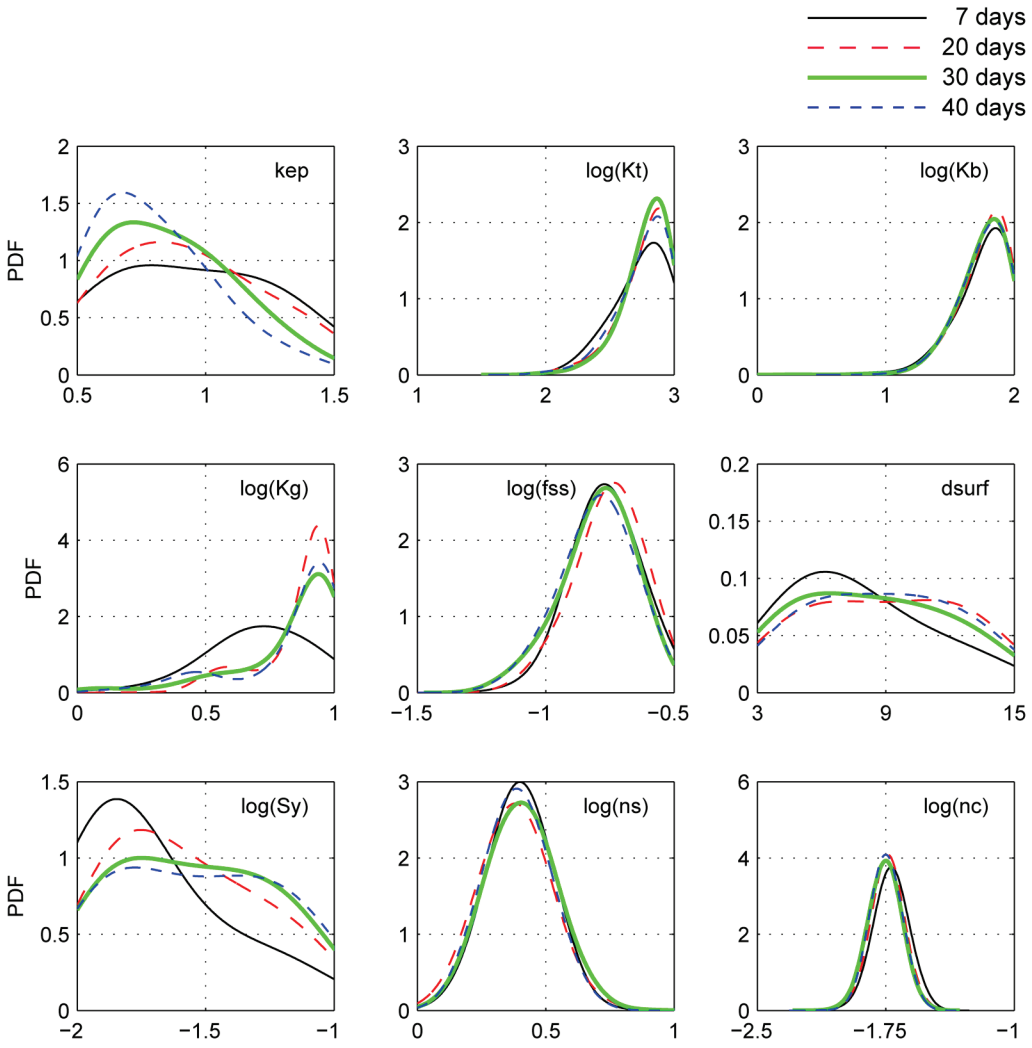

Figure 6. Comparison of the effects of different warm-up lengths on the model calibration. The black solid, red dashed, green solid and blue dashed curves are the estimated posterior probability densities of GBHM parameters using the warm-up lengths of 7, 20, 30, and 40 days, respectively.

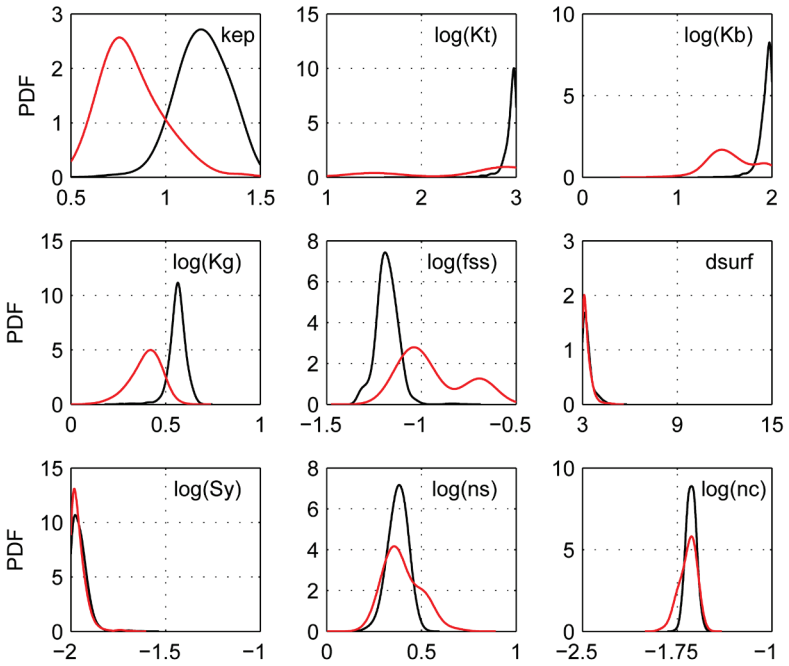

Figure 7. Comparison of the posterior distributions densities of the GBHM parameters for the field case study given by considering (black curves) and ignoring (red curves) rainfall errors.

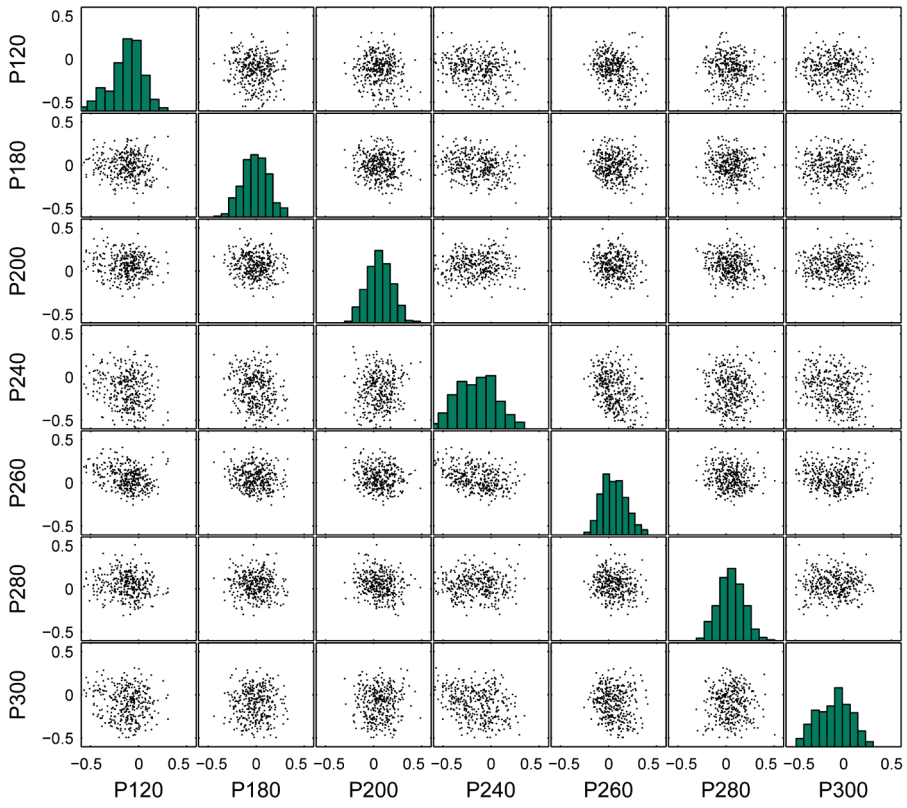

Figure 8. Scatter-plots and histograms of the estimated input errors for different rainfall gauges during the flood on September 22, 1981.

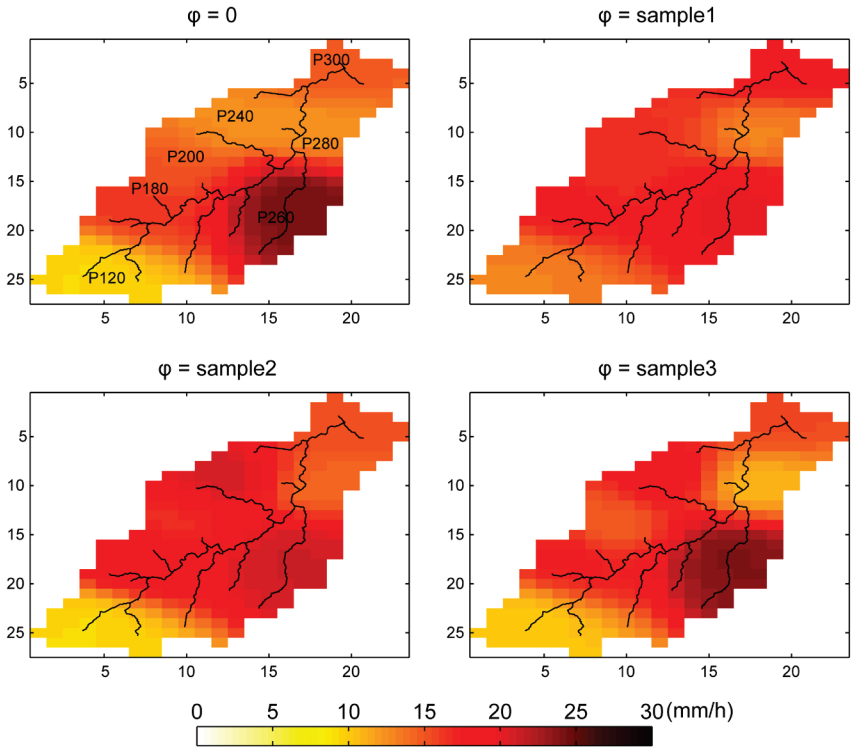

Figure 9. Observed and estimated spatial distributions of precipitation over the Chuzhou catchment for a time step during the flood on September 22, 1981. The angular distance-weighted method is used for the spatial interpolation.

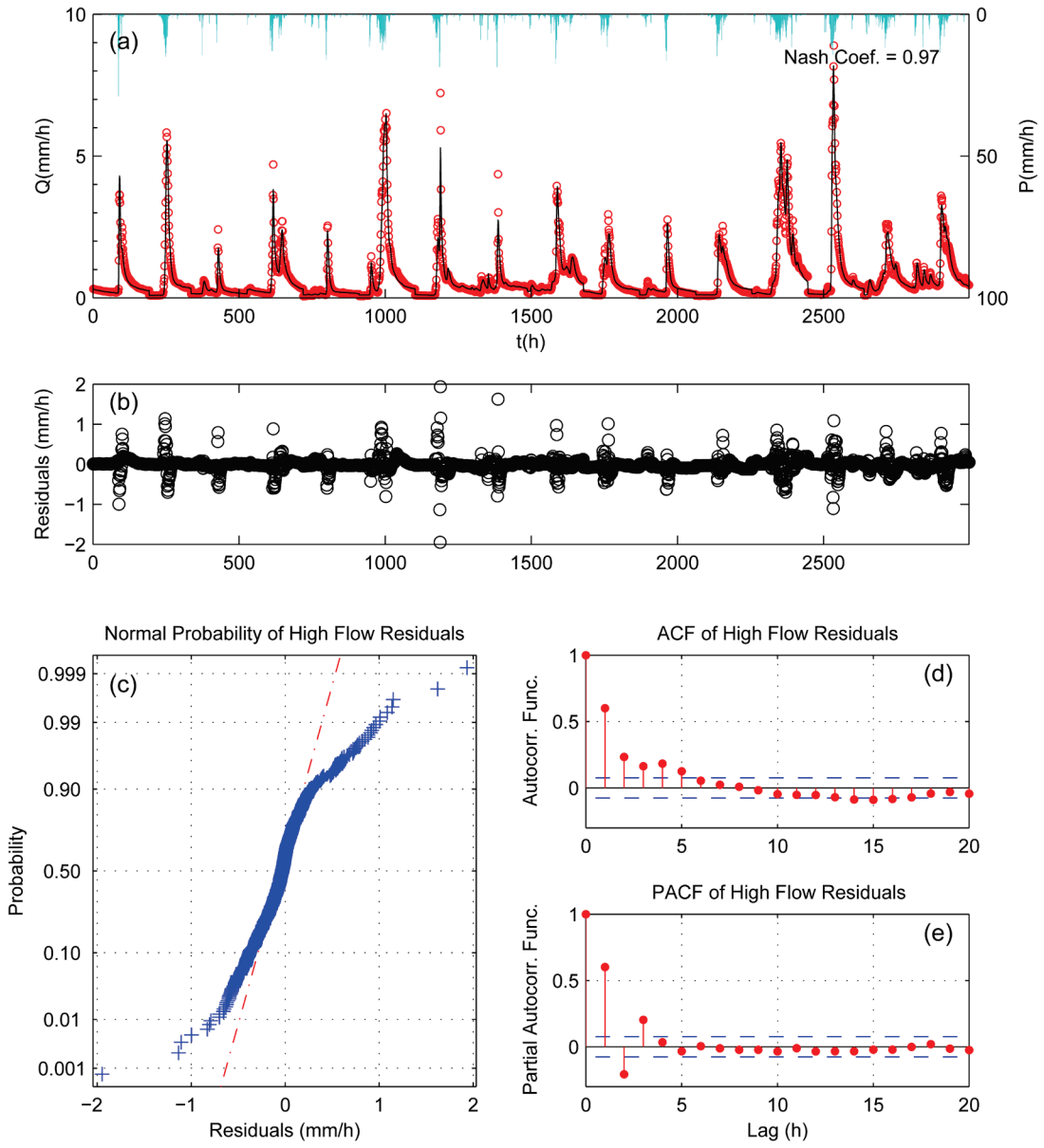

Figure 10. Evaluation of the estimated discharge medians for the calibration period: (a) medians of the estimated discharges (black curves) and observed discharges (red circles); (b) plots of residuals between the simulated and observed discharges; (c) normal probability plot of residuals for the high flow data; (d) autocorrelation functions (ACF) of high flow residuals; and (e) partial autocorrelation functions (PACF) of high flow residuals. For readability, 16 isolated floods are plotted together in (a) and (b).

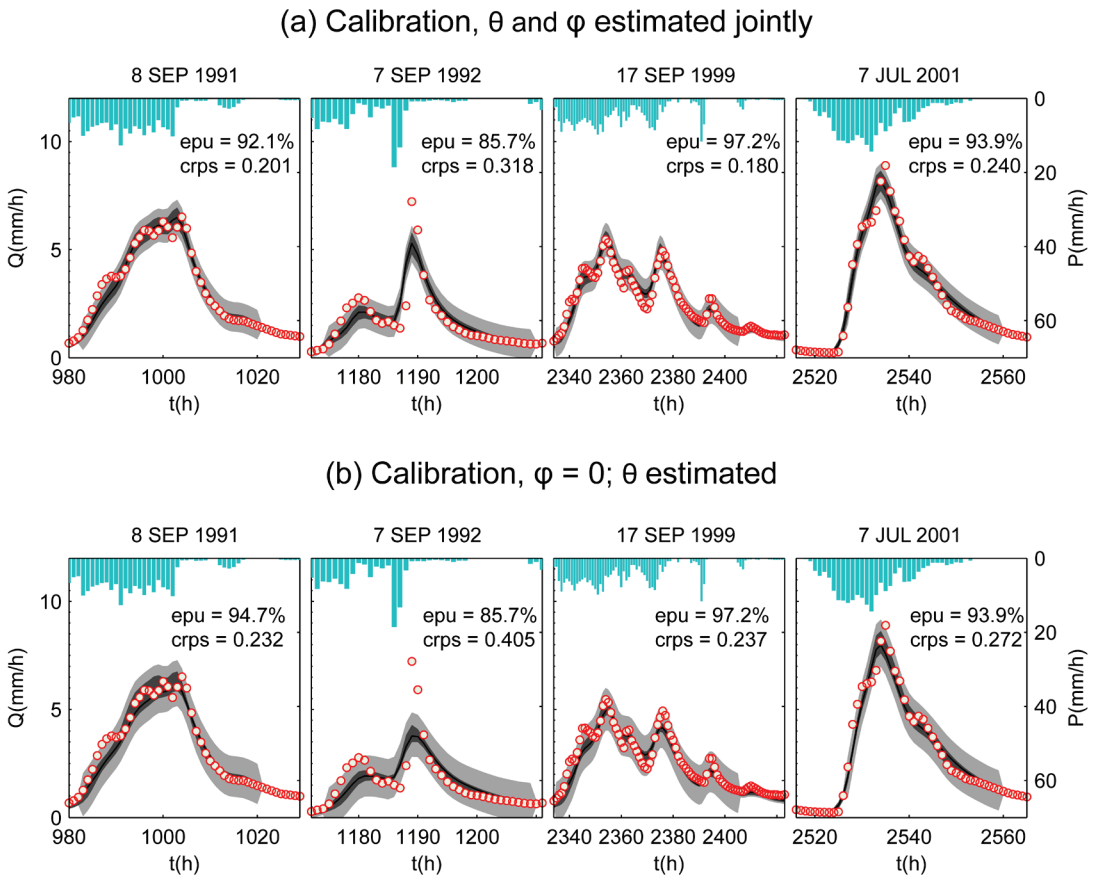

Figure 11. Simulated and observed discharges for the calibration period: (a) Bayesian model that estimates **φ** together with GBHM parameters in calibration. The darker shadow areas show the 95% model simulation uncertainty related to the uncertainty of GBHM parameters and input errors. The lighter shadow areas show the total uncertainty with the additional influence of model residuals. Within a flood event, discharges with relatively wide uncertainty ranges are those classified into high flow data; others are classified into low flow data. (b) Fixing **φ** = 0 in calibration. Thus, the dark shadow areas are related exclusively to parameter uncertainty.

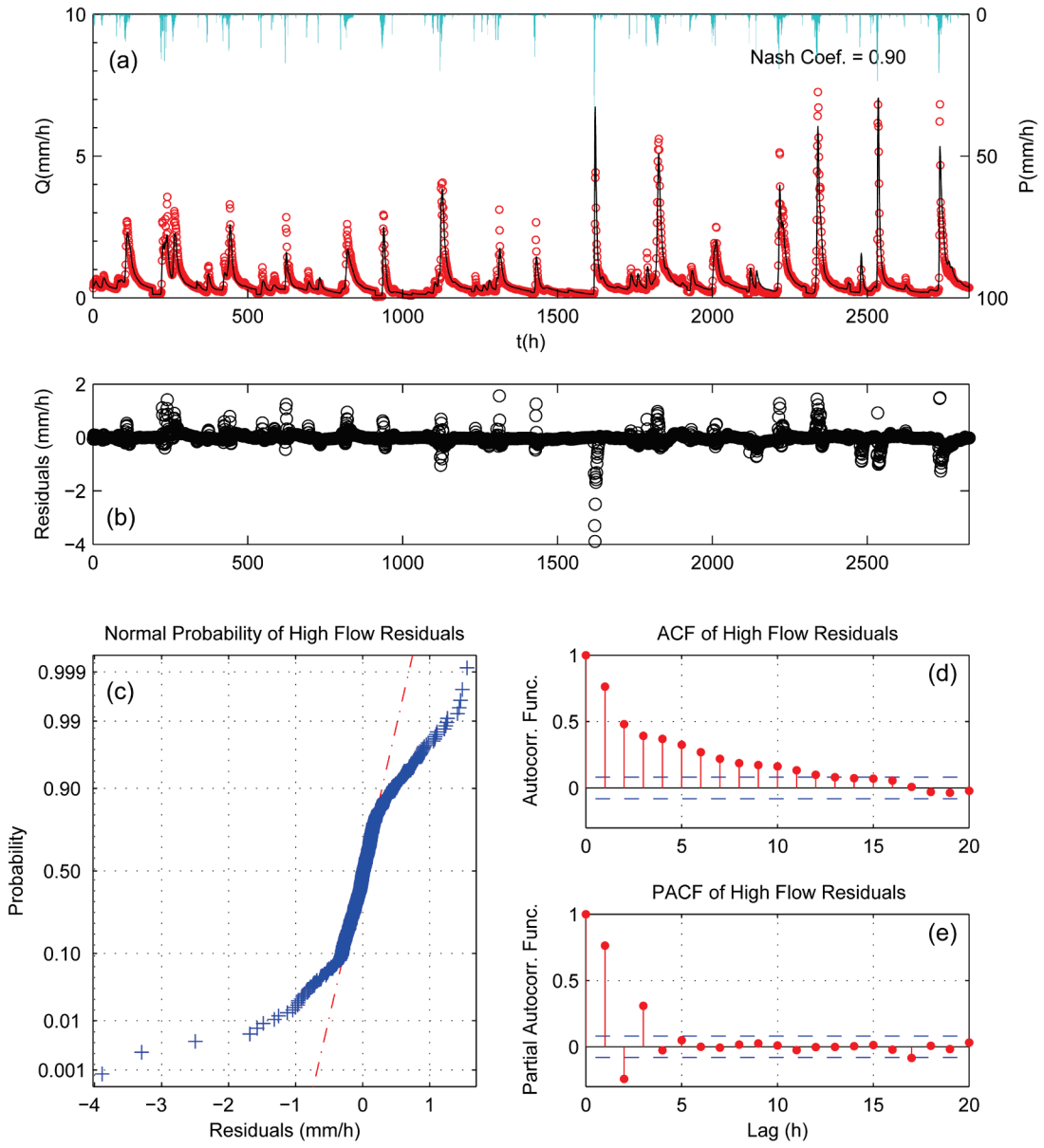

Figure 12. Evaluation of the estimated discharge medians for the validation period: (a) medians of the estimated discharges (black curves) and observed discharges (red circles); (b) plots of residuals between the simulated and observed discharges; (c) normal probability plot of residuals for the high flow data; (d) autocorrelation functions (ACF) of high flow residuals; and (e) partial autocorrelation functions (PACF) of high flow residuals. For readability, 16 isolated floods are plotted together in (a) and (b).

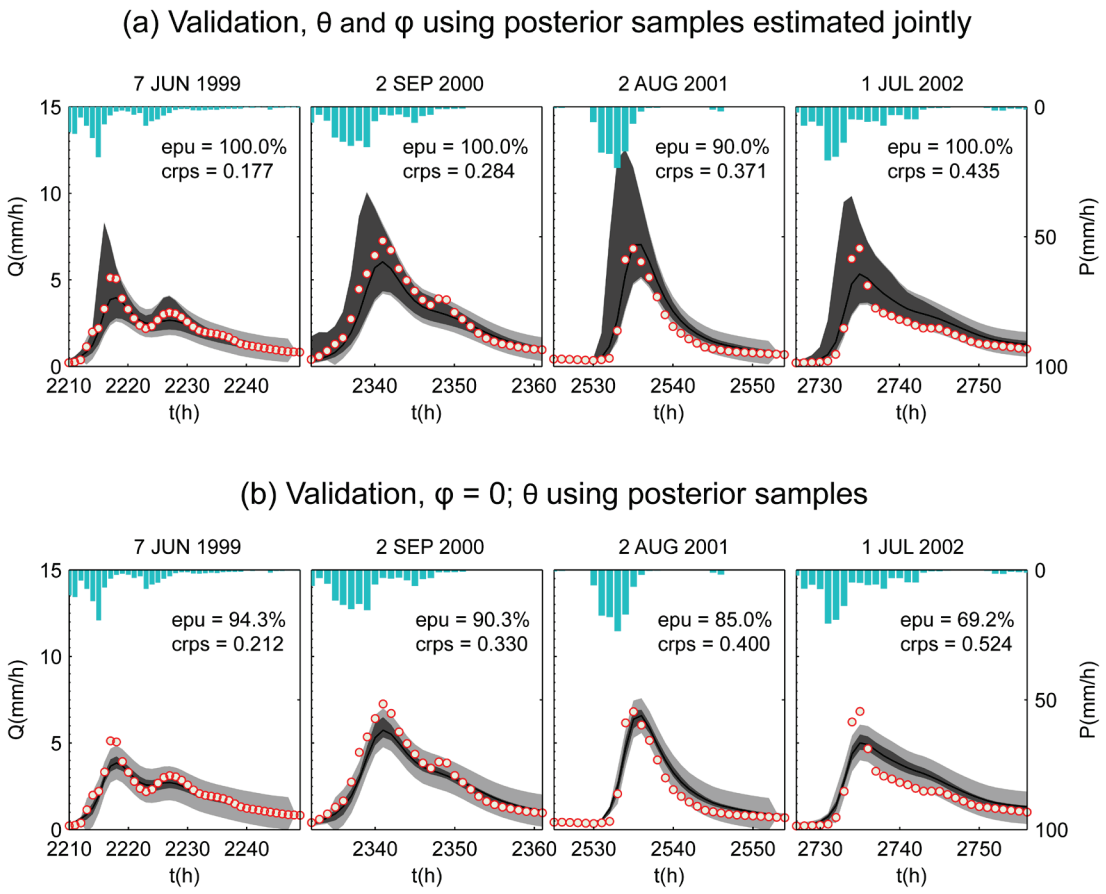

Figure 13. Predicted and observed discharges for the validation period: (a) Bayesian model that estimates **φ** together with GBHM parameters in prediction. The darker shadow areas show the 95% model prediction uncertainty related to the uncertainty of GBHM parameters and input errors. The lighter shadow areas show the total uncertainty with additional influence of model residuals; and (b) Fixing **φ** = 0 in prediction. Thus, the dark shadow areas are related exclusively to parameter uncertainty.

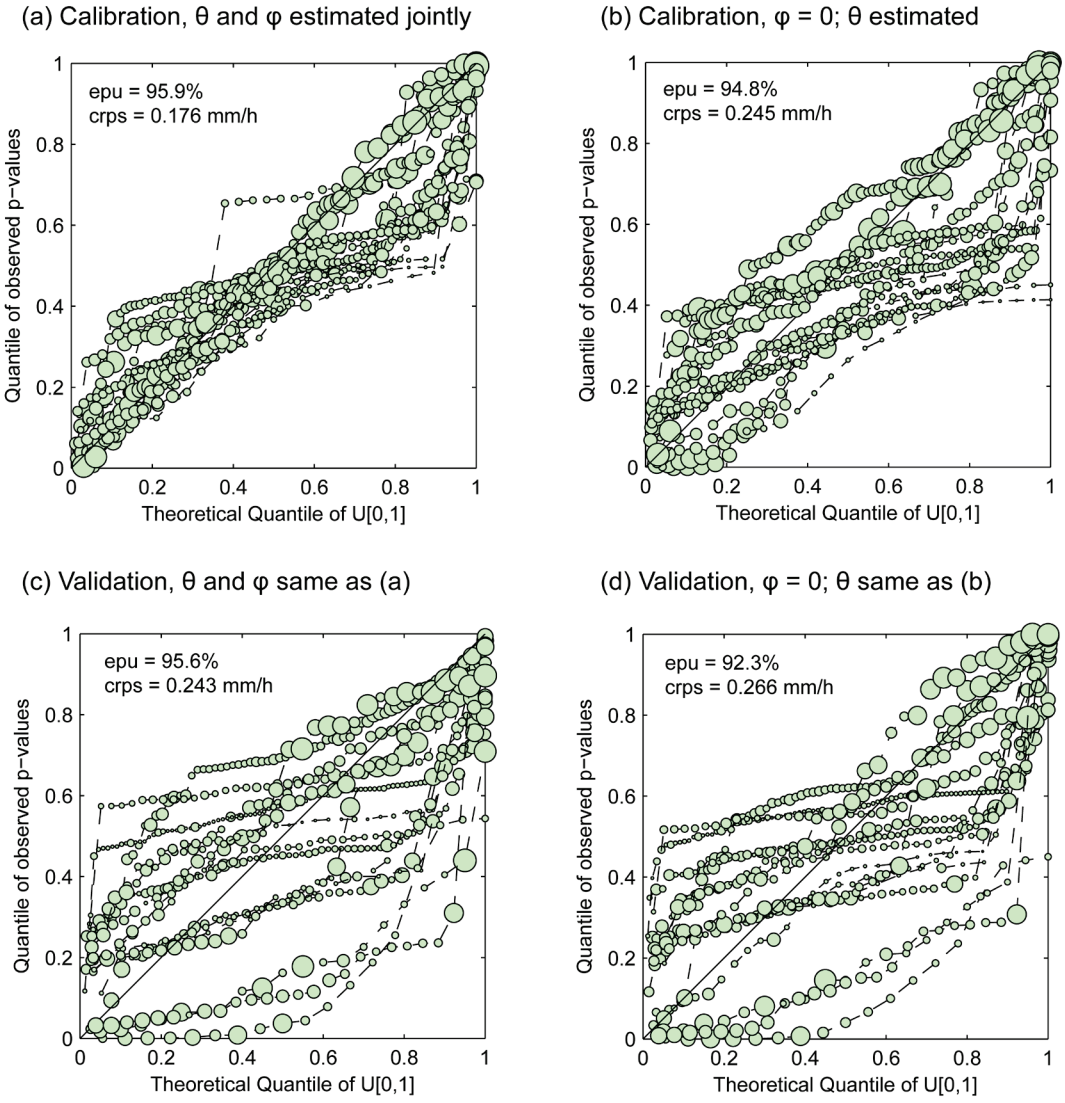

Figure 14. Predictive quantile-quantile plot (PQQ) of the estimated predictive uncertainty for both calibration and validation periods. The PQQ curves of 16 flood events are plotted together. The symbol size is proportional to the magnitude of the observed data.