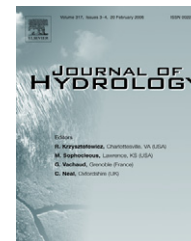




available at [www.sciencedirect.com](http://www.sciencedirect.com)



journal homepage: [www.elsevier.com/locate/jhydrol](http://www.elsevier.com/locate/jhydrol)



# Comparing uncertainty analysis techniques for a SWAT application to the Chaohe Basin in China

Jing Yang <sup>a,\*</sup>, Peter Reichert <sup>a</sup>, K.C. Abbaspour <sup>a</sup>, Jun Xia <sup>b</sup>, Hong Yang <sup>a</sup>

<sup>a</sup> *Eawag: Swiss Federal Institute of Aquatic Science and Technology, Ueberlandstr. 133, P.O. Box 611, 8600 Duebendorf, Switzerland*

<sup>b</sup> *Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographical Sciences and Natural Resources Research, CAS, Datun Road 11A, 100101 Beijing, China*

Received 7 December 2007; accepted 10 May 2008

## KEYWORDS

Uncertainty analysis;  
Watershed modeling;  
Bayesian inference;  
SUFI-2;  
GLUE;  
ParaSol

**Summary** Distributed watershed models are increasingly being used to support decisions about alternative management strategies in the areas of land use change, climate change, water allocation, and pollution control. For this reason it is important that these models pass through a careful calibration and uncertainty analysis. To fulfil this demand, in recent years, scientists have come up with various uncertainty analysis techniques for watershed models. To determine the differences and similarities of these techniques we compared five uncertainty analysis procedures: Generalized Likelihood Uncertainty Estimation (GLUE), Parameter Solution (ParaSol), Sequential Uncertainty Fitting algorithm (SUFI-2), and a Bayesian framework implemented using Markov chain Monte Carlo (MCMC) and Importance Sampling (IS) techniques. As these techniques are different in their philosophies and leave the user some freedom in formulating the generalized likelihood measure, objective function, or likelihood function, a literal comparison between these techniques is not possible. As there is a small spectrum of different applications in hydrology for the first three techniques, we made this choice according to their typical use in hydrology. For Bayesian inference, we used a recently developed likelihood function that does not obviously violate the statistical assumptions, namely a continuous-time autoregressive error model. We implemented all these techniques for the soil and water assessment tool (SWAT) and applied them to the Chaohe Basin in China. We compared the results with respect to the posterior parameter distributions, performances of their best estimates, prediction uncertainty, conceptual bases, computational efficiency, and difficulty of implementation. The comparison results for these categories are listed and the advantages and disadvantages are analyzed. From the point of view of the authors, if computationally feasible, Bayesian-based approaches are most recommendable because of their

\* Corresponding author. Tel.: +41 44 823 5534; fax: +41 44 823 5375.  
E-mail address: [jing.yang@eawag.ch](mailto:jing.yang@eawag.ch) (J. Yang).

solid conceptual basis, but construction and test of the likelihood function requires critical attention.

© 2008 Elsevier B.V. All rights reserved.

## Introduction

Simulation programs implementing models of watershed hydrology and river water quality are important tools for watershed management for both operational and research purposes. In recent years many such simulation programs have been developed such as AGNPS (Agricultural Non Point Source model) (Young et al., 1989), SWAT (Soil and Water Assessment Tool) (Arnold et al., 1998) and HSPF (Hydrologic Simulation Program – Fortran) (Bicknell et al., 2000). Areas of application of watershed models include integrated watershed management (e.g., Zacharias et al., 2005), peak flow forecasting (e.g., Jorgeson and Julien, 2005), test of the effectiveness of measures for the reduction of non-point source pollution (e.g., Bekele and Nicklow, 2005; Santhi et al., 2001), soil loss prediction (e.g., Cochrane and Flanagan, 2005), assessment of the effect of land use change (e.g., Hundecha and Bardossy, 2004; Claessens et al., 2006; Cotler and Ortega-Larrocea, 2006), analysis of causes of nutrient loss (e.g., Abbaspour et al., 2007; Adeuya et al., 2005), and climate change impact assessment (e.g., Claessens et al., 2006; Huang et al., 2005; Pednekar et al., 2005) among many others. This large number of various, and often very specific, applications led to the development of a multitude of watershed models starting in the early 1960s (see Todini, 2007 for a historical review).

As distributed watershed models are increasingly being used to support decisions about alternative management strategies, it is important for these models to pass a careful calibration and uncertainty analysis. Calibration of watershed models, however, is a challenging task because of input, model structure, parameter, and output uncertainty. Sources of model structural uncertainty include processes not accounted for in the model such as unknown activities in the watershed, and model inaccuracy due to over-simplification of the processes considered in the model. Some examples of this type of uncertainty are effects of wetlands and reservoirs on hydrology and chemical transport; interaction between surface and groundwater; occurrence of landslides, and large constructions (e.g., roads, dams, tunnels, bridges) that could produce large amounts of sediment during short time periods affecting water quantity and quality; unknown wastewater discharges into the streams from factories and water treatment plants; imprecisely known application of fertilizers and pesticides, unknown irrigation activities and water diversions, and other activities in the river basin. Input uncertainty is often related to imprecise or spatially interpolated measurements of model input or initial conditions, such as elevation data, land use data, rainfall intensity, temperature and initial groundwater levels. Other uncertainties in distributed models may also arise due to the large number of unknown parameters and the errors in the data used for parameter calibration.

To account for these uncertainties, in the last two decades, many uncertainty-analysis techniques have been

developed and applied to various catchments. The motivation for developing new or modified approaches may stem from the fact that the typical use of frequentist and Bayesian approaches which only consider parameter uncertainty and (independent) measurement error while neglecting input and model structure uncertainty leads to unrealistic prediction uncertainty bounds. The development follows three main categories: (i) Development of new approaches without rigorous statistical assumptions or ad-hoc modifications to existing statistical approaches. These approaches try to represent all uncertainties by an enhanced parameter uncertainty. Examples of such approaches are Generalized Likelihood Uncertainty Estimation (GLUE) (Beven and Binley, 1992) and Sequential Uncertainty Fitting (SUFI-2) (Abbaspour et al., 2004, 2007). (ii) Approaches that account for the effect of input and model structural errors on the output by an additive error model which introduces temporal correlation of the residuals. Representatives of this class are autoregressive error models as used, e.g., by Sorooshian and Dracup (1980), Kuczera (1983), Duan et al. (1988), Bates and Campbell (2001), Yang et al. (2007a), and Schaeffli et al. (2007). Although the approach of autoregressive error models is not a new technique, it is conceptually embedded into the general approach of describing model bias or deficiency by a random process that has recently gained attention in the statistical literature (Kennedy and O'Hagan, 2001; Bayarri et al., 2007). (iii) Development of improved likelihood functions that explicitly represent input errors and/or model structural error of the underlying hydrological model. These approaches include consideration of input uncertainty through rain multipliers (Kavetski et al., 2003, 2006a, b; Kuczera et al., 2006), simultaneous optimization and data assimilation (Vrugt et al., 2005), sequential data assimilation (Moradkhani et al., 2005), integrated Bayesian uncertainty estimation (Ajami et al., 2007), and use of time-dependent parameters for exploring model deficits and considering input and model structure uncertainty (Reichert and Mieleitner, submitted for publication).

Despite the large number of suggested techniques, only rarely more than one technique was applied in the same case study in the literature. To our knowledge only a few papers on comparison of different uncertainty analysis techniques are available and they are limited to applications of simple hydrological models (e.g., Makowski et al., 2002; Vrugt et al., 2003; Mantovan and Todini, 2006). The objective of this paper is to fill this gap. Although the techniques in category (iii) above are the most promising ones, currently these techniques are still computationally too demanding for straightforward application to complex hydrological models. For this reason it is still important to study the advantages and disadvantages of the techniques from categories (i) and (ii) in practical applications of complex hydrological models. We compare the following five techniques: Generalized Likelihood Uncertainty Estimation (GLUE) (Beven and Binley, 1992), Parameter Solution

(ParaSol) (Van Griensven and Meixner, 2006), Sequential Uncertainty Fitting (SUFI-2) (Abbaspour et al., 2004, 2007), Bayesian inference based on Markov chain Monte Carlo (MCMC) (e.g., Kuczera and Parent, 1998; Marshall et al., 2004; Vrugt et al., 2003; Yang et al., 2007a), and Bayesian inference based on importance sampling (IS) (e.g., Kuczera and Parent, 1998). As these uncertainty analysis techniques are different in their philosophies and leave the user some freedom in formulating the generalized likelihood measure, objective function, or likelihood function, a literal comparison between the techniques is not possible. As there is a smaller spectrum of different applications in hydrology for the first three techniques, we could make this choice according to their typical use in hydrology (Nash–Sutcliffe coefficient for GLUE and SUFI-2, sum of squares of the residuals for ParaSol). The choice is difficult for Bayesian inference, as there is much recent development in the formulation of reasonable likelihood functions (see papers cited in categories (ii) and (iii) above). To simplify the comparison, we used a simple likelihood function that does not obviously violate the statistical assumptions, namely a continuous-time autoregressive model applied as an error term additive to the output of the deterministic simulation model. The advantages of a continuous-time autoregressive error model over a discrete-time model are discussed by Yang et al. (2007b). This seems also to be a logical choice as all of these techniques were designed to overcome the problems of Bayesian inference with an independent error model. For the comparison, we used the hydrologic sub-model of the Soil and Water Assessment Tool (SWAT) applied to the Chaohe Basin in China. We compared the results with respect to the posterior parameter distributions, performances of their best estimates (that minimize or maximize the corresponding objective function), prediction uncertainty, conceptual basis, computational efficiency, and difficulty of implementation. Note that, to not unnecessarily complicate our wording, we use the term “posterior distribution” to summarize the inference results for parameters and model output for all techniques despite the fact that it is not a statistically based posterior in the first three cases.

The remainder of this paper is structured as follows: In the second section, we introduce the methodology used for the comparison, give a brief overview of the selected techniques, and then list the criteria for the assessment. In Section “Case study”, we give an overview of the study site, the SWAT hydrological model, and our model application (aggregation of parameters). In Section “Results and discussion” the results are presented and discussed. The last section contains the conclusions.

## Methodology, selected techniques, and criteria for comparison

### Difficulties in comparing estimation methods

There are various difficulties in comparing uncertainty analysis techniques in hydrological modeling. The following list addresses the most important concerns and how we handled them:

- Most techniques are different in their philosophies and subjective choices have to be made in their formulation with respect to prior parameter distribution, likelihood function and/or objective function. We addressed this problem by choosing priors and objective functions for each technique as they would typically be used in hydrological applications. This leads necessarily to different objective functions for different techniques. When discussing the results, we will analyze whether a problem is caused by the conceptual formulation of a particular technique or by the choice of the objective function.
- Different underlying concepts and objective functions from different techniques make the comparison difficult. The values of the objective functions of all techniques will be calculated for the best estimate (minimizing or maximizing the corresponding objective function) for each technique to allow for a fair comparison. In addition, we use measures of computational efficiency and an assessment of the conceptual basis as criteria for the comparison.
- Different techniques obviously lead to different results for different criteria. We will outline the results in all criteria so that the reader can draw his/her own conclusions. Our own conclusions depend to some degree on subjective judgment. As an example, the final conclusions in this paper are based on a subjective trade-off between conceptual and computational advantages.
- The results of the comparison inherently depend on the application. We try to separate the results of specific application from generic results in the discussion.

## Selected techniques

### GLUE

GLUE is an uncertainty analysis technique inspired by importance sampling and regional sensitivity analysis (RSA; Hornberger and Spear, 1981). In GLUE, parameter uncertainty accounts for all sources of uncertainty, i.e., input uncertainty, structural uncertainty, parameter uncertainty and response uncertainty, because “the likelihood measure value is associated with a parameter set and reflects all these sources of error and any effects of the covariation of parameter values on model performance implicitly” (Beven and Freer, 2001). Also, from a practical point of view, “disaggregation of the error into its source components is difficult, particularly in cases common to hydrology where the model is non-linear and different sources of error may interact to produce the measured deviation” (Gupta et al., 2005). In GLUE, parameter uncertainty is described as a set of discrete “behavioral” parameter sets with corresponding “likelihood weights”.

A GLUE analysis consists of the following three steps:

- (1) After the definition of the “generalized likelihood measure”,  $L(\theta)$ , a large number of parameter sets are randomly sampled from the prior distribution and each parameter set is assessed as either “behavioral” or “non-behavioral” through a comparison of the “likelihood measure” with a selected threshold value.

- (2) Each behavioral parameter set is given a ‘‘likelihood weight’’ according to

$$w_i = \frac{L(\theta_i)}{\sum_{k=1}^N L(\theta_k)} \quad (1)$$

where  $N$  is the number of behavioral parameter sets.

- (3) Finally, prediction uncertainty is described by quantiles of the cumulative distribution realized from the weighted behavioral parameter sets. In the literature, the most frequently used likelihood measure for GLUE is the *Nash–Sutcliffe* coefficient (*NS*) (e.g., [Beven and Freer, 2001](#); [Freer et al., 1996](#)), which is also used in this paper:

$$NS = 1 - \frac{\sum_{t_i=1}^n (y_{t_i}^M(\theta) - y_{t_i})^2}{\sum_{t_i=1}^n (y_{t_i} - \bar{y})^2} \quad (2)$$

where  $n$  is the number of the observed data points, and  $y_{t_i}$  and  $y_{t_i}^M(\theta)$  represent the observation and model simulation with parameters  $\theta$  at time  $t_i$ , respectively, and  $\bar{y}$  is the average value of the observations.

### ParaSol and modified ParaSol

ParaSol is based on a modification to the global optimization algorithm SCE-UA ([Duan et al., 1992](#)). The idea is to use the simulations performed during optimization to derive prediction uncertainty because ‘‘the simulations gathered by SCE-UA are very valuable as the algorithm samples over the entire parameter space with a focus on solutions near the optimum/optima’’ ([Van Griensven and Meixner, 2006](#)).

The procedure of ParaSol is as follows:

- (1) After optimization applying the modified SCE-UA (the randomness of the algorithm SCE-UA is increased to improve the coverage of the parameter space), the simulations performed are divided into ‘‘good’’ simulations and ‘‘not good’’ simulations by a threshold value of the objective function as in GLUE. This leads to ‘‘good’’ parameter sets and ‘‘not good’’ parameter sets.
- (2) Prediction uncertainty is constructed by equally weighting all ‘‘good’’ simulations.

The objective function used in ParaSol is the sum of the squares of the residuals (*SSQ*):

$$SSQ = \sum_{t_i=1}^n (y_{t_i}^M(\theta) - y_{t_i})^2 \quad (3)$$

The relationship between *NS* and *SSQ* is

$$NS = 1 - \frac{1}{\sum_{t_i=1}^n (y_{t_i} - \bar{y})^2} \cdot SSQ \quad (4)$$

where  $\sum_{t_i=1}^n (y_{t_i} - \bar{y})^2$  is a fixed value for given observations. To improve the comparability with GLUE, all objective function values of ParaSol were converted to *NS*.

As the choice of the threshold of the objective function in ParaSol is based on the  $\chi^2$ -statistics it mainly accounts for parameter uncertainty under the assumption of independent measurement errors. For the purpose of comparison with GLUE, as an alternative, we choose the same threshold

as used by GLUE and we call this method ‘‘modified ParaSol’’.

### SUFI-2 procedure

In SUFI-2, parameter uncertainty is described by a multivariate uniform distribution in a parameter hypercube, while the output uncertainty is quantified by the 95% prediction uncertainty band (95PPU) calculated at the 2.5% and 97.5% levels of the cumulative distribution function of the output variables ([Abbaspour et al., 2007](#)). Latin hypercube sampling is used to draw independent parameter sets ([Abbaspour et al., 2007](#)). Similarly to GLUE, SUFI-2 represents uncertainties of all sources through parameter uncertainty in the hydrological model.

The procedure of SUFI-2 is as follows:

- (1) In the first step, the objective function  $g(\theta)$  and meaningful parameter ranges  $[\theta_{\text{abs min}}, \theta_{\text{abs max}}]$  are defined.
- (2) Then a Latin Hypercube sampling is carried out in the hypercube  $[\theta_{\text{min}}, \theta_{\text{max}}]$  (initially set to  $[\theta_{\text{abs min}}, \theta_{\text{abs max}}]$ ), the corresponding objective functions are evaluated, and the sensitivity matrix  $J$  and the parameter covariance matrix  $C$  are calculated according to

$$J_{ij} = \frac{\Delta g_i}{\Delta \theta_j}, \quad i = 1, \dots, C_2^m, \quad j = 1, \dots, n, \quad (5)$$

$$C = s_g^2 (J^T J)^{-1} \quad (6)$$

where  $s_g^2$  is the variance of the objective function values resulting from the  $m$  model runs.

- (3) A 95% predictive interval of a parameter  $\theta_j$  is computed as follows:

$$\theta_{j,\text{lower}} = \theta_j^* - t_{v,0.025} \sqrt{C_{jj}}, \quad \theta_{j,\text{upper}} = \theta_j^* + t_{v,0.025} \sqrt{C_{jj}} \quad (7)$$

where  $\theta_j^*$  is the parameter  $\theta_j$  for the best estimates (i.e., parameters which produce the optimal objective function), and  $v$  is the degrees of freedom ( $m - n$ ).

- (4) The 95PPU is calculated. And then two indices, i.e., the  $p$ -factor (the percent of observations bracketed by the 95PPU) and the  $r$ -factor, are calculated:

$$r\text{-factor} = \frac{\frac{1}{n} \sum_{t_i=1}^n (y_{t_i}^{M,97.5\%} - y_{t_i}^{M,2.5\%})}{\sigma_{\text{obs}}} \quad (8)$$

where  $y_{t_i}^{M,97.5\%}$  and  $y_{t_i}^{M,2.5\%}$  represent the upper and lower boundary of the 95PPU, and  $\sigma_{\text{obs}}$  stands for the standard deviation of the measured data.

The goodness of calibration and prediction uncertainty is judged on the basis of the closeness of the  $p$ -factor to 100% (i.e., all observations bracketed by the prediction uncertainty) and the  $r$ -factor to 1 (i.e., achievement of rather small uncertainty band). As all uncertainties in the conceptual model and inputs are reflected in the measurements (e.g., discharge), bracketing most of the measured data in the prediction 95PPU ensures that all uncertainties are depicted by the parameter uncertainties. If the two factors have satisfactory values, then a uniform distribution in the parameter hypercube  $[\theta_{\text{min}}, \theta_{\text{max}}]$  is interpreted as the

posterior parameter distribution. Otherwise,  $[\theta_{\min}, \theta_{\max}]$  is updated according to

$$\begin{aligned}\theta_{j,\min,\text{new}} &= \theta_{j,\text{lower}} - \max\left(\frac{\theta_{j,\text{lower}} - \theta_{j,\min}}{2}, \frac{\theta_{j,\max} - \theta_{j,\text{upper}}}{2}\right) \\ \theta_{j,\max,\text{new}} &= \theta_{j,\text{upper}} + \max\left(\frac{\theta_{j,\text{lower}} - \theta_{j,\min}}{2}, \frac{\theta_{j,\max} - \theta_{j,\text{upper}}}{2}\right)\end{aligned}\quad (9)$$

and another iteration needs to be performed.

SUFI-2 allows its users several choices of the objective function (for instance the NS coefficient). In the literature, the weighted root mean square error (RMSE) (Abbaspour et al., 2004) and the weighted sum of squares SSQ (Abbaspour et al., 2007) were used. In this study we chose the NS coefficient for the sake of comparison with other techniques.

### Bayesian inference

According to Bayes' theorem, the probability density of the posterior parameter distribution  $f_{\theta_{\text{post}}|\mathbf{Y}}(\theta|\mathbf{y}_{\text{meas}})$  is derived from the prior density  $f_{\theta_{\text{pri}}}(\theta)$  and measured data  $\mathbf{y}_{\text{meas}}$  as

$$f_{\theta_{\text{post}}|\mathbf{Y}}(\theta|\mathbf{y}_{\text{meas}}) = \frac{f_{\mathbf{Y}^M|\theta}(\mathbf{y}_{\text{meas}}|\theta) \cdot f_{\theta_{\text{pri}}}(\theta)}{\int f_{\mathbf{Y}^M|\theta}(\mathbf{y}_{\text{meas}}|\theta') f_{\theta_{\text{pri}}}(\theta') d\theta'} \quad (10)$$

where  $f_{\mathbf{Y}^M|\theta}(\mathbf{y}_{\text{meas}}|\theta)$  is the likelihood function of the model, i.e., the probability density for model results for given parameters with the measurements substituted for the model results, and  $\mathbf{Y}^M$  is the vector of random variables that characterizes the hydrologic model including all uncertainties. Posterior prediction uncertainty is usually represented by quantiles of the posterior distribution. The crucial point of applying this technique is the formulation of the likelihood function. If the statistical assumptions for formulating the likelihood function are violated, the results of Bayesian inference are unreliable. Unfortunately, when formulating likelihood functions in hydrological applications, it is often assumed that the residuals between measurements and simulations are independently and identically (usually normally) distributed (iid). However, this assumption is often violated. To avoid this problem in our case study, we constructed the likelihood function by combining a Box–Cox transformation (Box and Cox, 1964, 1982) with a continuous-time autoregressive error model (Brockwell and Davis, 1996; Brockwell, 2001) as follows:

$$\begin{aligned}f_{\mathbf{Y}^M|\theta}(\mathbf{y}|\theta) &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{1}{2} \frac{[g(\mathbf{y}_{t_0}) - g(\mathbf{y}_{t_0}^M(\theta))]^2}{\sigma^2}\right) \cdot \left|\frac{dg}{dy}\right|_{y=\mathbf{y}_{t_0}} \\ &\cdot \prod_{i=1}^n \left[ \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma \sqrt{1 - \exp(-2 \frac{t_i - t_{i-1}}{\tau})}} \exp\left(-\frac{1}{2} \frac{[g(\mathbf{y}_{t_i}) - g(\mathbf{y}_{t_i}^M(\theta)) - [g(\mathbf{y}_{t_{i-1}}) - g(\mathbf{y}_{t_{i-1}}^M(\theta))] \exp(-\frac{t_i - t_{i-1}}{\tau})]^2}{\sigma^2 (1 - \exp(-2 \frac{t_i - t_{i-1}}{\tau}))}\right) \cdot \left|\frac{dg}{dy}\right|_{y=\mathbf{y}_{t_i}} \right]\end{aligned}\quad (11)$$

where  $\sigma$  is the asymptotic standard deviation of the errors,  $\tau$  is the characteristic correlation time,  $\theta$  is the vector of model parameters,  $\mathbf{y}_{t_i}$  and  $\mathbf{y}_{t_i}^M(\theta)$  are the observation and model simulation, respectively, at time  $t_i$ , and  $g(\cdot)$

represents the Box–Cox transformation (Box and Cox, 1964, 1982):

$$\begin{aligned}g(\mathbf{y}) &= \begin{cases} \frac{(\mathbf{y} + \lambda_2)^{\lambda_1} - 1}{\lambda_1} & \lambda_1 \neq 0 \\ \ln(\mathbf{y} + \lambda_2) & \lambda_1 = 0 \end{cases}, \\ g^{-1}(\mathbf{z}) &= \begin{cases} (\lambda_1 \mathbf{z} + 1)^{1/\lambda_1} - \lambda_2 & \lambda_1 \neq 0 \\ \exp(\mathbf{z}) - \lambda_2 & \lambda_1 = 0 \end{cases}, \quad \frac{dg}{dy} = (\mathbf{y} + \lambda_2)^{\lambda_1 - 1}\end{aligned}\quad (12)$$

This model extends earlier works with discrete-time autoregressive error models in hydrological applications (e.g., Kuczera, 1983; Duan et al., 1988; Bates and Campbell, 2001). More details are given by Yang et al. (2007a).

Two generic Monte Carlo approaches to sample from the posterior distribution are Markov chain Monte Carlo and Importance Sampling (Gelman et al., 1995; Kuczera and Parent, 1998). Both techniques are used as implemented in the systems analysis tool UNCSIM (Reichert, 2005; <http://www.uncsim.eawag.ch>).

### Markov chain Monte Carlo (MCMC)

MCMC methods are a class of algorithms for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. The simplest technique from this class is the Metropolis algorithm (Metropolis et al., 1953; Gelman et al., 1995), which is applied in this study. A sequence (Markov chain) of parameter sets representing the posterior distribution is constructed as follows:

- (1) An initial starting point in the parameter space is chosen.
- (2) A candidate for the next point is proposed by adding a random realization from a symmetrical jump distribution,  $f_{\text{jump}}$ , to the coordinates of the previous point of the sequence:

$$\theta_{k+1}^* = \theta_k + \text{rand}(f_{\text{jump}}) \quad (13)$$

- (3) The acceptance of the candidate points depends on the ratio  $r$ :

$$r = \frac{f_{\theta_{\text{post}}|\mathbf{Y}}(\theta_{k+1}^*|\mathbf{y}_{\text{meas}})}{f_{\theta_{\text{post}}|\mathbf{Y}}(\theta_k|\mathbf{y}_{\text{meas}})} \quad (14)$$

If  $r \geq 1$ , then the candidate point is accepted as a new point, else it is accepted with probability  $r$ . If the candidate point is rejected, the previous point is used as the next point of the sequence.

In order to avoid long burn-in periods (or even lack of convergence to the posterior distribution) the chain is started at a numerical approximation to the maximum of the posterior distribution calculated with the aid of the shuffled complex global optimization algorithm (Duan et al., 1992).

### Importance sampling (IS)

Importance sampling is a well established technique for randomly sampling from a probability distribution (Gelman et al., 1995; Kuczera and Parent, 1998). The idea is to draw randomly from a sampling distribution  $f_{\text{sample}}$  and calculate weights for the sampling points to make the weighted sample a sample from the posterior distribution. The procedure consists of the following steps:

- (1) Choose a sampling distribution and draw a random sample from this sampling distribution.
- (2) For each parameter set,  $\theta_i$ , of the sample, calculate a weight according to

$$w_i = \frac{f_{\theta_{\text{post}}|Y}(\theta_i|y_{\text{meas}})/f_{\text{sample}}(\theta_i)}{\sum_{k=1}^N f_{\theta_{\text{post}}|Y}(\theta_k|y_{\text{meas}})/f_{\text{sample}}(\theta_k)} \quad (15)$$

- (3) Use the weighted sample to derive properties of the posterior distribution, for example, by calculating the expected value of a function  $h$  according to

$$E_{f_{\text{post}}}(h) \approx \sum_{k=1}^N w_k h(\theta_k) \quad (16)$$

The computational efficiency of this procedure depends strongly on how close the sampling distribution is to the posterior distribution, and hence, the choice of the sampling distribution is crucial (Geweke, 1989; Gelman et al., 1995). Three practical choices for the sampling distribution are sampling from the prior distribution (often uniform sampling over a hypercube referred to in the following as primitive IS or naive IS), use of an over-dispersed multi-normal distribution as a sampling distribution (e.g., Kuczera and Parent, 1998), and the method of iteratively adapting the sampling distribution and using efficient sampling techniques (Reichert et al., 2002). Each of the above methods has some disadvantages. Primitive IS is very inefficient if the posterior is significantly different from the prior, particularly for high dimensional parameter spaces. It is also worth nothing that primitive IS is a special case of GLUE, in which no generalizations are made to the likelihood function and all parameter sets are accepted as behavioral (although some will get a very small weight). For the method with over-dispersed multi-normal distribution, it is difficult to determine a priori for the dispersion coefficients (Kuczera and Parent, 1998). The method of iteratively adapting the sampling distribution becomes more and more difficult to implement as the dimensionality of the parameter space increases (Reichert et al., 2002). This is because larger samples are required to get sufficient information on the shape of the posterior and it becomes more and more difficult to find a reasonable parameterized sampling distribution to approximate the posterior. In this study,

only the primitive IS is implemented, as this also allows us to study the behavior of GLUE with different likelihood measures.

### Criteria for the comparison

We use the following five categories of criteria to compare the performances of the uncertainty analysis techniques:

1. The best parameter estimate at the mode of the posterior distribution, parameter uncertainty, and correlation coefficients between parameters.
2. Performance of the simulation at the mode of the posterior distribution, evaluated for all criteria (i.e., Nash–Sutcliffe coefficient,  $R^2$ , and values of the objective functions).
3. Model prediction uncertainty.

Three indices are used to compare the derived 95% probability band (95PPU). Those indices are the width of 95PPU (i.e.,  $r$ -factor as used in SUFI-2), percentage of the measurements bracketed by this band (i.e.,  $p$ -factor in SUFI-2), and the Continuous Rank Probability Score (CRPS). An ideal uncertainty analysis technique would lead to a 95% probability band that is as narrow as possible while still being a correct estimate under the statistical assumptions of the technique. The percentage of measurements bracketed by the band provides empirical evidence that the estimate is realistic (for the prediction of new measurements, not the mean).

CRPS is widely used in weather forecast as a measure of the closeness of the predicted and occurred cumulative distributions and sharpness of the predicted probability density function (PDF) (e.g., Hersbach, 2000). For a time series, the CRPS at time  $t$  can be defined as

$$\text{CRPS}_t = \int_{-\infty}^{\infty} (F_t(y) - H(y - y_t))^2 dy \quad (17)$$

where  $F_t(y)$  stands for the predicted cumulative density function (CDF) at time  $t$ ,  $H$  is the Heaviside function (returning zero for negative and unity for non-negative arguments), and  $y_t$  is the observed at time  $t$ . The minimal value zero of  $\text{CRPS}_t$  is only achieved when the empirical distribution is identical to the predicted distribution, that is, in the case of a perfect deterministic forecast (Hersbach, 2000). In practice the CRPS is averaged over a time series:

$$\text{CRPS} = \sum_t w_t \cdot \text{CRPS}_t \quad (18)$$

where  $w_t$  is the weight for corresponding  $\text{CRPS}_t$  at time  $t$  and we take equal weights in our study. Therefore, the smaller the CRPS the narrow would be the prediction uncertainty.

4. The conceptual basis of the technique (theoretical basis, testability and fulfillment of statistical assumptions, capability of exploring the parameter space, coverage of regions with high objective function values).
5. Difficulty of implementation and computational efficiency of the technique (programming effort and number of simulations required to get reasonable results).

## Case study

### The Chaohe Basin and data

The Chaohe Basin is situated in North China with a drainage area of 5300 km<sup>2</sup> upstream of the Xiahui station (see Fig. 1 in Yang et al., 2007a). The climate is temperate continental, semi-humid and semi-arid. From 1980 to 1990 the average daily maximum temperature was 6.2 °C, the average daily minimum temperature 0.9 °C, and the yearly rainfall varied between 350 and 690 mm. The elevation varies from 200 m at the basin outlet to 2400 m at the highest point in the catchment. The topography is characterized by high mountain ranges, steep slopes and deep valleys. The average channel slope is 1.87% which leads to fast water flow in the river. Average daily flow at the catchment outlet (Xiahui station) is 9.3 m<sup>3</sup> s<sup>-1</sup> and varies irregularly from around 798 m<sup>3</sup> s<sup>-1</sup> during the flood season to lower than 1 m<sup>3</sup> s<sup>-1</sup> in the dry season. The runoff coefficient (the ratio of runoff to precipitation) at the Xiahui station to the rainfall in this basin decreased from 0.24 in 1980 to 0.09 in 1990. It is believed that the decline is mainly due to the intensified human activities, including increasing water use and building of more (small scale) water retention structures.

### The watershed model

The soil and water assessment tool (SWAT) (Arnold et al., 1998) is a continuous-time, spatially distributed simulator of water, sediment, nutrients and pesticides transport at a catchment scale. It runs on a daily time step. In SWAT, a watershed is divided into a number of sub-basins based on a given digital elevation model (DEM) map. Within each sub-basin, soil and landuse maps are overlaid to create a number of unique hydrologic response units (HRUs). SWAT simulates surface and subsurface processes, accounting for snow fall and snow melt, vadose zone processes (i.e., infiltration, evaporation, plant uptake, lateral flows, and percolation into aquifer). Runoff volume is calculated using the Curve Number method (USDA Soil Conservation Service, 1972). Sediment yield from each sub-basin is generated using the Modified Universal Soil Loss Equation (MUSLE) (Williams, 1995). The model updates the C factor of the MUSLE on a daily basis using information from the crop growth module. The routing phase controls the movement of water using the variable storage method or the Muskingum method (Cunge, 1969; Chow et al., 1988).

### Model application

Parameterization of spatially-distributed hydrologic models can potentially lead to a large number of parameters. To effectively limit the number of parameters, we developed an aggregating scheme based on hydrologic group (A, B, C, or D), soil texture, land use, sub-basin, and the spatial distribution of default values. This scheme was implemented in an interface, iSWAT, that allows systems analysis programs to access SWAT parameters that are distributed over many input files (Yang et al., 2005; <http://www.uncsim.eawag.ch/interfaces/swat>). The names of aggregate

parameters specified in the interface iSWAT have the following format:

$$x_{-}(\text{parname}) \cdot \langle \text{ext} \rangle_{-}(\text{hydrogrp})_{-}(\text{soltext})_{-}(\text{landuse})_{-}(\text{subbsn}) \quad (19)$$

where  $x$  represents the type of change to be applied to the parameter ( $v$ : replacement;  $a$ : absolute change; or  $r$ : relative change),  $\langle \text{parname} \rangle$  is the SWAT parameter name;  $\langle \text{ext} \rangle$  represents the extension of the SWAT input file which contains the parameter;  $\langle \text{hydrogrp} \rangle$  is the identifier for the hydrologic group;  $\langle \text{soltext} \rangle$  is the soil texture;  $\langle \text{landuse} \rangle$  is the landuse; and  $\langle \text{subbsn} \rangle$  is the sub-basin number, the crop index, or the fertilizer index. The interface exchanges parameter values with the systems analysis tool based on a simple text file-based interface (Reichert, 2006).

Following our previous work (Yang et al., 2007a), 10 aggregate SWAT parameters related to discharge at the watershed outlet were selected. These parameters, listed in Table 1, represent single global values, global multipliers, or global additive terms to the distributed default values of the corresponding SWAT parameters (compare parameter names in Table 1 with the explanations of expression (19)). The likelihood function for the Bayesian approach requires the additional parameters  $\sigma$  and  $\tau$  characterizing the standard deviation and characteristic correlation time of the continuous-time autoregressive error model (see Eq. (11)). These parameters were considered to be dependent on the seasons, i.e.,  $\sigma_{\text{dry}}$  and  $\tau_{\text{dry}}$  were used for dry season (October–May), and  $\sigma_{\text{wet}}$  and  $\tau_{\text{wet}}$  were used for wet season (July–August), and we assumed a linear transition from one value to the other in June and September (Yang et al., 2007a).

The priors of all the parameters above were assumed to be independent of each other. For the SWAT parameters, uniform priors within reasonable ranges were assumed for all the techniques. For the parameters  $\sigma$  and  $\tau$ , densities proportional to  $1/\sigma$  and  $1/\tau$  were chosen (see, e.g., Arnold, 1996), which is equivalent to assuming that the logarithms of these parameters are uniformly distributed. Table 1 gives an overview of the parameters used for calibration and their prior distributions.

As we cannot specify reasonable initial values for all storage volumes considered in the model, SWAT is operated for a “warm-up” period of 5 years without comparison of model results with data. We found that such a “warm-up period” was sufficient to minimize the effects of the initial state of SWAT variables on river flow. Furthermore, in order to verify the calibrated model parameters, the model was calibrated and tested based on the observed discharges at the watershed outlet (Xiahui station) using a split sample procedure. The data from the years 1985 to 1988 with omission of a single outlier in 1985 was used for calibration, and the data from 1989 to 1990 was used to test the model. This strategy was applied for all the techniques.

## Results and discussion

We start with a description of the results for each technique and then compare and discuss the results according to the categories of criteria given in the Section “Criteria for the comparison”.

**Table 1** Selected parameters for uncertainty analysis and their prior distributions

Aggregate parameters <sup>a</sup>	Name and meaning of underlying SWAT parameter	Range of initial underlying SWAT parameters	Prior distribution of aggregate parameters <sup>b</sup>
$a\_CN2.mgt$	CN2: curve number	72–92 <sup>b</sup>	$U[-30, 5]$
$v\_ESCO.hru$	ESCO: soil evaporation compensation factor	0.95	$U[0, 1]$
$v\_EPCO.hru$	EPCO: plant uptake compensation factor	1.00	$U[0, 1]$
$r\_SOL\_K.sol$	SOL_K: soil hydraulic conductivity (mm/h)	1.6–328.27	$U[-0.8, 0.8]$
$a\_SOL\_AWC.sol$	SOL_AWC: soil available water capacity (mm H <sub>2</sub> O/mm soil)	0–0.13	$U[0, 0.15]$
$v\_ALPHA\_BF.gw$	ALPHA_BF: base flow alpha factor (1/day)	0.048	$U[0, 1]$
$r\_SLSUBBSN.hru^c$	SLSUBBSN: average slope length (m)	9.461–91.463	$U[-0.6, 0.6]$
$a\_CH\_K2.rte$	CH_K2: effective hydraulic conductivity in main channel alluvium (mm/h)	0	$U[0, 150]$
$a\_OV\_N.hru$	OV_N: overland manning roughness	0.06–0.15	$U[0, 0.2]$
$v\_GW\_DELAY.gw$	GW_DELAY: groundwater delay time (days)	31	$U[0, 300]$
$\lambda_1$	Parameter of Box–Cox transformation		$U[0, 1]$
$\sigma$	$\sigma_{dry}$ Standard deviation during dry season in Eqs. (9) and (10)		Inv
	$\sigma_{wet}$ Standard deviation during wet season in Eqs. (9) and (10)		Inv
$\tau$	$\tau_{dry}$ Characteristic correlation time of autoregressive process during dry season (days) in Eq. (10)		Inv
	$\tau_{wet}$ characteristic correlation time of autoregressive process during wet season (days) in Eq. (10)		Inv

The last five parameters are only used for MCMC and primitive IS.

<sup>a</sup> Aggregate parameters are constructed based on Eq. (1). ‘‘a\_’’, ‘‘v\_’’ and ‘‘r\_’’ means an absolute increase, a replacement, and a relative change to the initial parameter values, respectively.

<sup>b</sup> This range is based on the initial parameter estimates of the project.

<sup>c</sup> Prior distributions of aggregate parameters are based on Neitsch et al. (2001);  $U[a, b]$  in this column means the distribution of this parameter/aggregate parameters is uniform over the interval  $[a, b]$ ; ‘‘Inv’’ means that the probability density at the value  $x$  is proportional to  $1/x$  (the logarithm of the parameter is uniform distributed).

## Results of GLUE implementation with likelihood measure NS

GLUE is convenient and easy to implement and widely used in hydrology (e.g., Freer et al., 1996; Cameron et al., 2000a, b; Blazkova et al., 2002). The drawback of this approach is its prohibitive computational burden imposed by its random sampling strategy (Hossain et al., 2004).

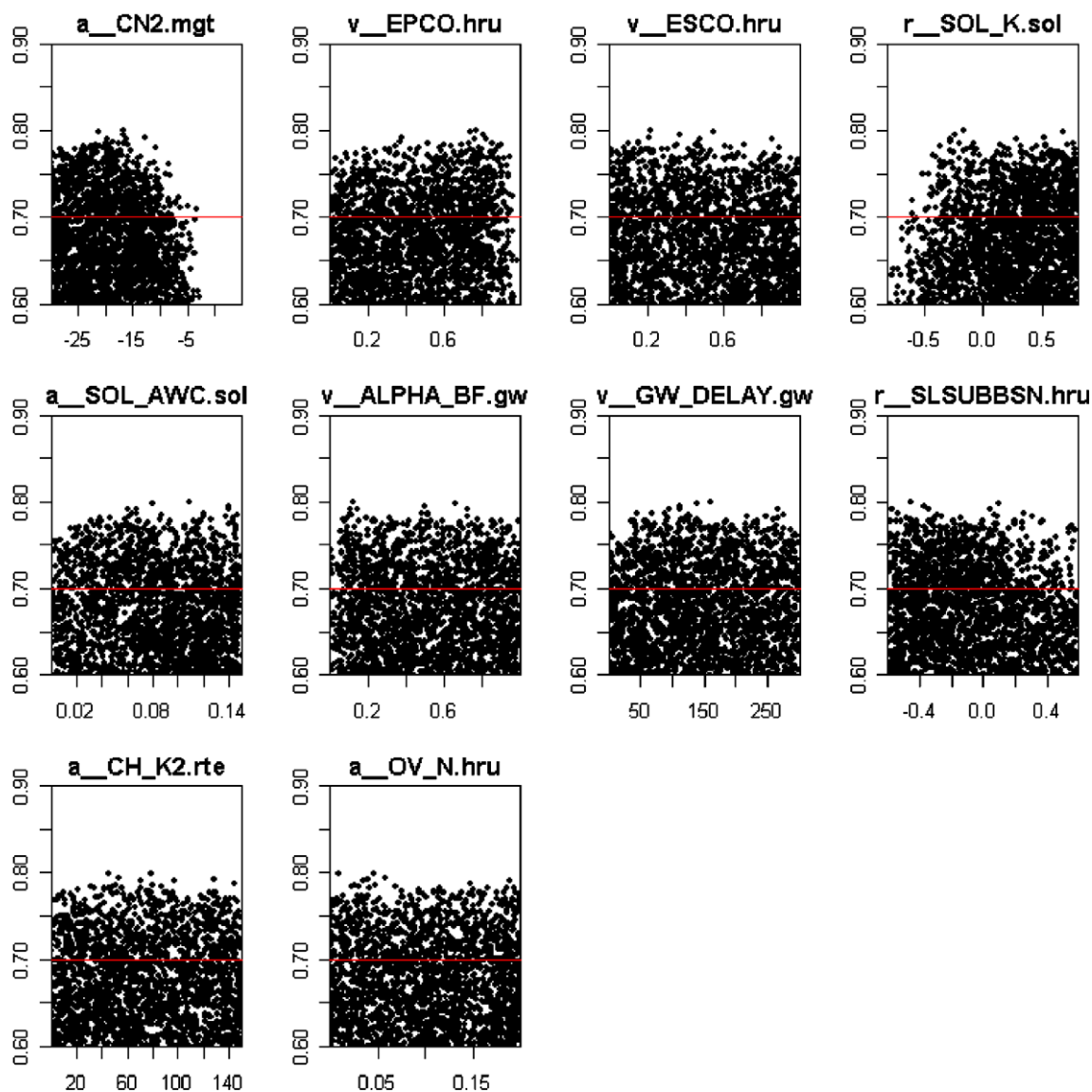
In this study, the threshold value of GLUE application is chosen to be 0.70, i.e., the simulations with NS values larger than 0.70 are behavioral otherwise non-behavioral. Four GLUE simulations were performed with sample sizes of 1000, 5000, 10,000, and 20,000. For each simulation, the dot plot, cumulative posterior distribution and 95PPU are analyzed. The comparison shows that there are some differences in the results between 1000, 5000 and 10,000 while there is no significant difference between 10,000 and 20,000. The following analysis of results and comparison are based on a sample size of 10,000. The dot plot shown in Fig. 1 demonstrates that for each parameter solutions with similarly good values of the NS coefficient can be found within the complete prior range. The posteriors of most aggregate parameters follow closely the uniform prior distribution. Table 2 shows the mean, standard deviation and correlation matrix of the posterior parameter distribution. The correlations between most parameters are small except between  $a\_CN2.mgt$  and  $a\_SOL\_AWC.sol$ ,  $v\_ESCO.hru$  and  $a\_SOL\_AWC.sol$ , and  $r\_SOL\_K.sol$  and  $r\_SLSUBBSN.hru$ , with values of 0.44, 0.56 and 0.67, respectively. The third

column in Table 2 shows the standard deviations of the parameters. Fig. 2 shows the 95PPU of the model results for both calibration and validation periods. Most of the observations are bracketed by the 95PPUs (79% during the calibration period and 69% during the validation period, see  $p$ -factor in Table 5).

## Results of ParaSol and modified ParaSol implementations with objective function SSQ

Implementation of ParaSol is relatively easy and the computation depends only on the convergence of the optimization process (SCE-UA algorithm). Once the optimization is done, ParaSol will determine the behavioral and non-behavioral parameter sets and produce prediction uncertainty.

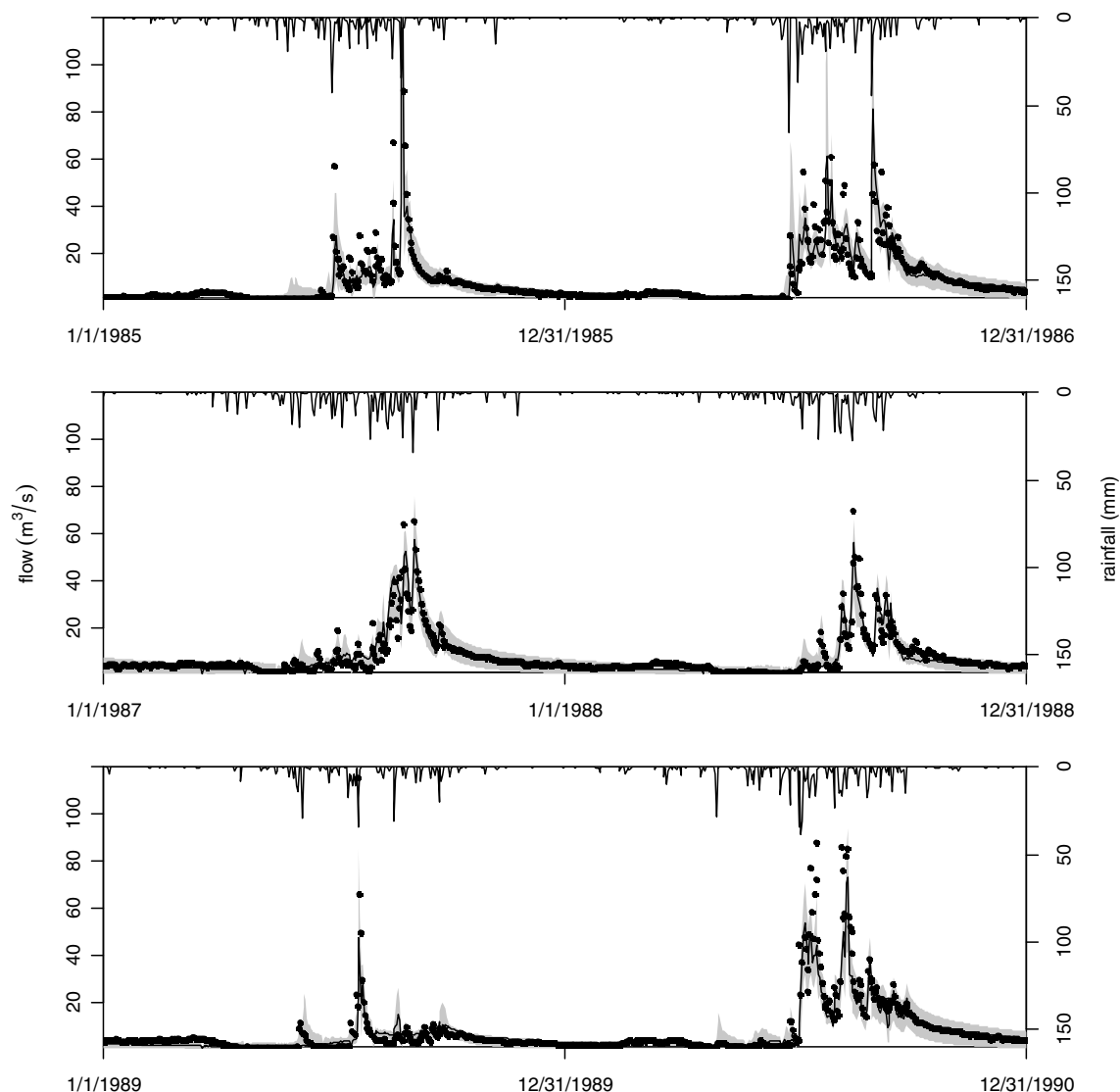
The application of ParaSol resulted in 851 behavioral parameter sets out of a total of 7550 samples (the threshold value based on the  $\chi^2$ -statistics is equivalent to NS = 0.819). Fig. 3 shows the dot plot of the NS coefficient against each parameter. Clearly, the parameter samples are very dense around the maximum. This is confirmed by very steep cumulative distribution functions (not shown) and small standard deviations of the estimated model parameters (third column in Table 3). ParaSol based on the SCE-UA is very efficient in detecting the area with high objective-function values in the response surface. The threshold line (blue line) in Fig. 3 separates the parameters sets into behavioral parameter sets (above the blue line) and non-behavioral



**Figure 1** Dotty plot of NS coefficient against each aggregate SWAT parameters conditioning with GLUE based on 10,000 samples with threshold 0.70 (red line), above which the parameter sets are behavioral. (For the interpretation of color in this figure legend, the reader is referred to the Web version of this article.)

**Table 2** Mean, standard deviation (SD) and correlation matrix of the posterior distribution resulting from application of the GLUE technique

Aggregate parameters	Mean	SD											
$a\_CN2.mgt$	-20.27	5.56	1										
$v\_EPCO.hru$	0.47	0.29	-0.02	1									
$v\_ESCO.hru$	0.51	0.25	0.04	0.18	1								
$r\_SOL\_K.sol$	0.30	0.33	-0.04	0.03	-0.10	1							
$a\_SOL\_AWC.sol$	0.08	0.04	0.44	-0.09	0.56	-0.09	1						
$v\_ALPHA\_BF.gw$	0.51	0.28	-0.19	0.03	0.02	-0.07	-0.06	1					
$v\_GW\_DELAY.gw$	149.46	81.96	0.03	0.01	0.15	0.06	-0.04	-0.14	1				
$r\_SLSUBBSN.hru$	-0.13	0.28	0.09	-0.03	0.24	0.67	0.19	0.07	0.01	1			
$a\_CH\_K2.rte$	74.99	42.42	0.16	-0.02	-0.12	0.00	-0.08	-0.01	0.04	-0.13	1		
$a\_OV\_N.hru$	0.10	0.06	-0.02	0.03	0.01	0.05	-0.02	-0.03	0.03	0.02	0.08	1	



**Figure 2** 95PPU (shaded area) derived by GLUE during the calibration period (top and middle) and validation period (bottom). The dots correspond to the observed discharge at the basin outlet, while the solid line represents the best simulation obtained by GLUE.

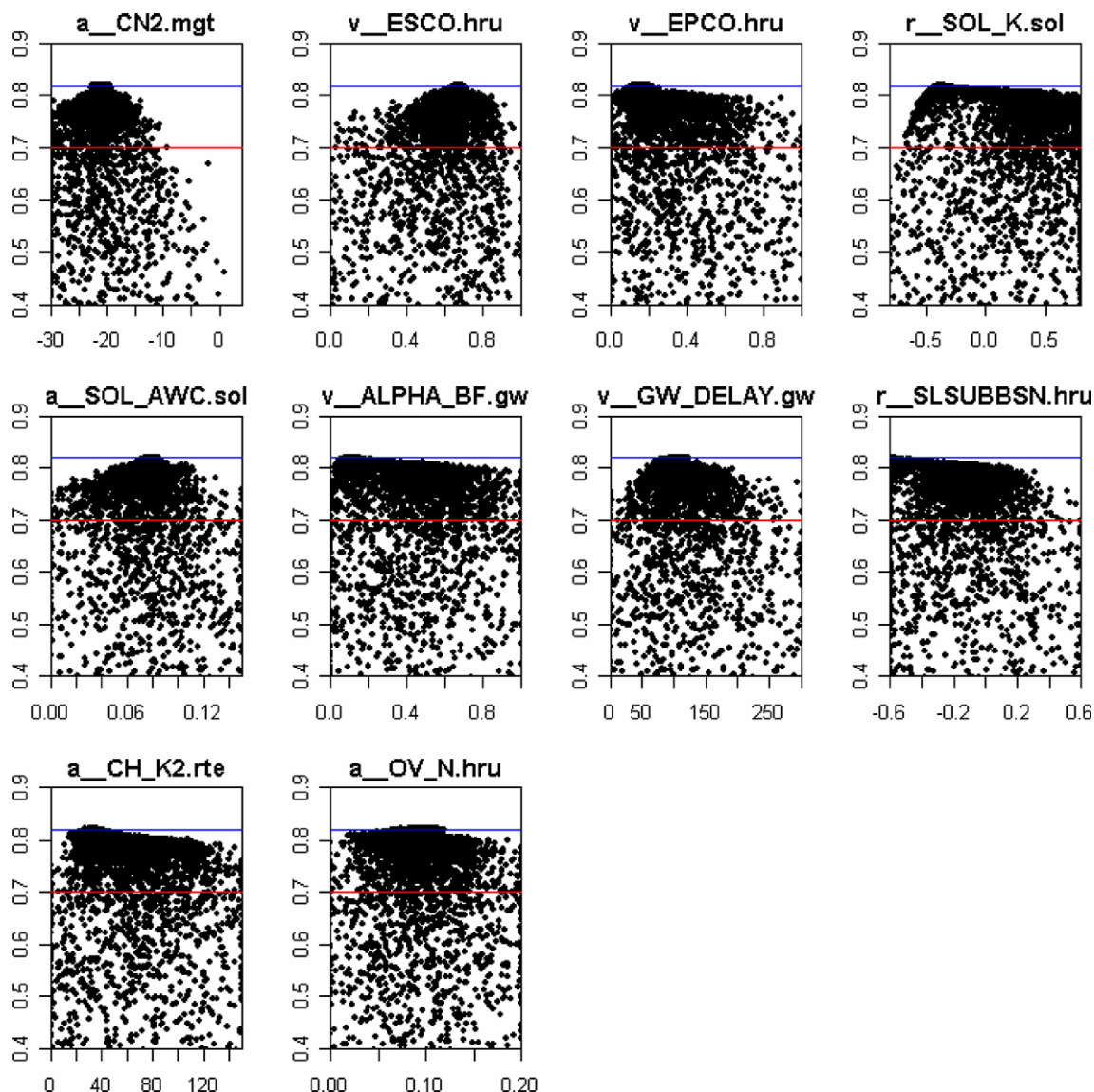
parameter sets (below the blue line). However, as can be seen, both the number and area of the behavioral parameter sets are extremely small, and the corresponding parameter ranges are very narrow. This also leads to a very narrow 95PPU for model predictions shown in Fig. 4 (dark gray area). ParaSol failed to derive the reasonable prediction uncertainty (only 18% of measurements were bracketed by 95PPU during the calibration period) though the best simulation matches the observation quite well with NS equals 0.82 during the calibration period. This is because ParaSol does not consider the error in the model structure, measured input and measured response, which results in an underestimation of the prediction uncertainty. The developer of ParaSol solved this problem by reducing the threshold to include the correct number of data points (technique "SUNGLASSES"). SUNGLASSES is not applied here because it needs to take into account the observations during the validation period, which will complicate the comparison.

As to the modified ParaSol with threshold value 0.70, Fig. 3 shows the behavioral and non-behavioral parameter sets separated by threshold line with value 0.70 (red line<sup>1</sup>), and light grey area in Fig. 4 describes the 95PPU. There are 60% of measurements bracketed by 95PPU during calibration period and 52% during validation period (see  $p$ -factor in Table 5).

### Result of SUFI-2 implementation with objective function NS

SUFI-2 is also convenient to use. The drawback of this approach is that it is semi-automated and requires the interaction of the modeler for checking a set of suggested posterior

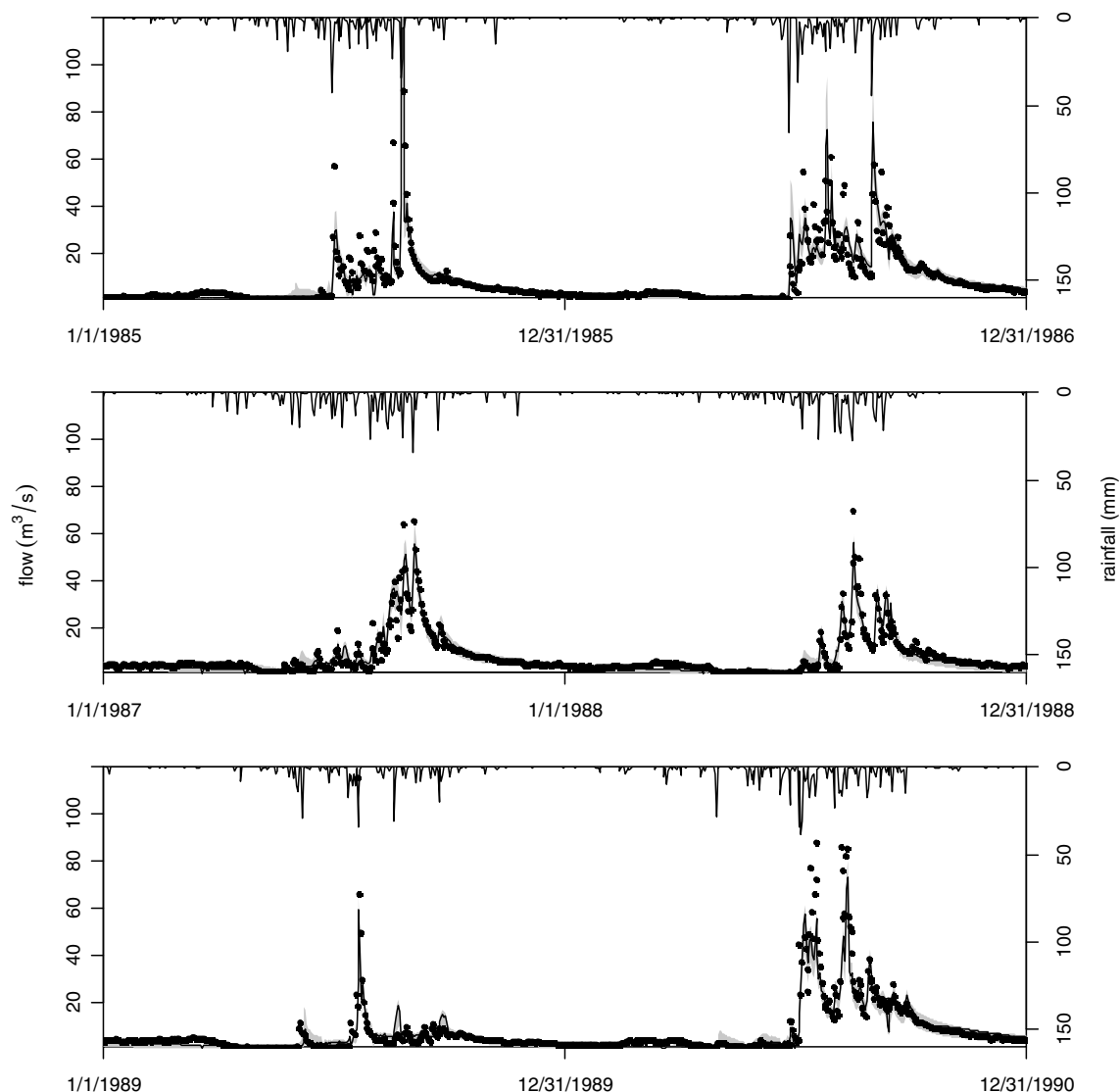
<sup>1</sup> For the interpretation of color in Fig. 3, the reader is referred to the Web version of this article.



**Figure 3** Dotty plot of NS coefficient against aggregate SWAT parameters conditioning with ParaSol. The blue line is the threshold determined by ParaSol, and red line is the threshold with value 0.70 for modified ParaSol. (For the interpretation of color in this figure legend, the reader is referred to the Web version of this article.)

**Table 3** Mean, standard deviation (SD) and correlation matrix of the posterior distribution resulting from application of the ParaSol technique

Aggregate parameters	Mean	SD											
<i>a</i> _CN2.mgt	-21.08	1.81	1										
<i>v</i> _ESCO.hru	0.65	0.07	0.18	1									
<i>v</i> _EPCO.hru	0.22	0.13	-0.06	0.04	1								
<i>r</i> _SOL_K.sol	0.00	0.38	-0.03	-0.14	0.50	1							
<i>a</i> _SOL_AWC.sol	0.08	0.01	0.42	0.54	-0.18	-0.20	1						
<i>v</i> _ALPHA_BF.gw	0.29	0.21	-0.15	-0.08	0.57	0.84	-0.14	1					
<i>v</i> _GW_DELAY.gw	106.62	24.91	-0.08	-0.02	0.35	0.36	-0.27	0.38	1				
<i>r</i> _SLSUBBSN.hru	-0.35	0.24	0.03	-0.07	0.48	0.96	-0.11	0.85	0.33	1			
<i>a</i> _CH_K2.rte	49.58	23.41	-0.07	-0.19	0.54	0.76	-0.36	0.73	0.45	0.72	1		
<i>a</i> _OV_N.hru	0.09	0.02	-0.09	-0.11	0.30	0.31	-0.18	0.26	0.16	0.28	0.36	1	



**Figure 4** 95PPUs derived by ParaSol (dark gray area) and modified ParaSol (light gray area) during the calibration period (top and middle) and validation period (bottom). The dots correspond to the observed discharge at the basin outlet, while the solid line represents the best simulation obtained by ParaSol.

parameters, hence, requiring a good knowledge of the parameters and their effects on the output. This may add an additional error, i.e., “modeler’s uncertainty” to the list of other uncertainties.

For the SUFI-2 approach, we did two iterations with 1500 model runs in each iteration. In the second iteration, the 95PPU brackets 84% of the observations and  $r$ -factor equals 1.03 which is very close to a suggested value of 1. The following analysis is based on the second iteration. Posterior distributions in SUFI-2 are always independent and uniformly distributed, and expressed as narrowed parameter ranges (see the interval bracketed by parentheses in category 1 in Table 5). Fig. 5 shows the dot plot conditioned on SUFI-2, and all these sampled parameter sets are taken as behavioral samples and contributing to the 95PPU. Obviously, there are some parameter sets with low NS values (e.g.,  $-1.5$ ) in Fig. 5. Fig. 6 shows 95PPU for model results derived by SUFI-2 for the second iteration. As can be seen, most of the observations are bracketed by the 95PPU (84%

during calibration period and 82% during validation period), indicating SUFI-2 is capable of capturing the observations during both calibration and validation periods. The 95PPU is quite suitable to bracket the observations in 1985, 1988 and 1989, while it is somehow slightly overestimated in 1986, 1987 and 1990 especially in the recession part. This indicates there is a lot of uncertainty in the recession calculation of SWAT. However, as SUFI-2 is a sequential procedure, i.e., one more iteration can always be made leading to a smaller 95PPU at the expense of more observation points falling out of the prediction band.

#### Result of MCMC implementation of Bayesian analysis with autoregressive error model

Implementation of Bayesian inference is not so easy especially for complex models because it requires formulating and testing of a likelihood function that characterizes the

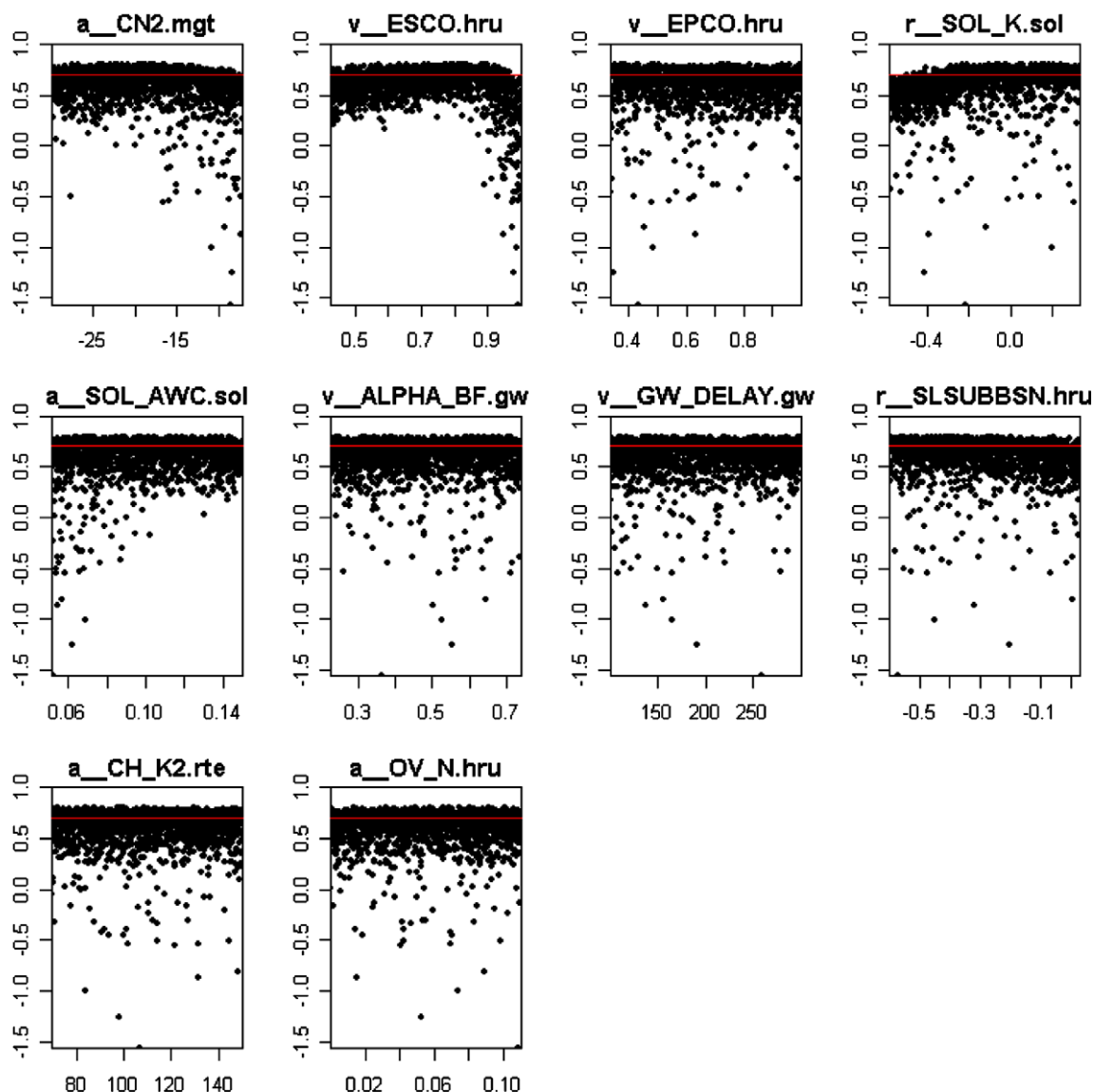


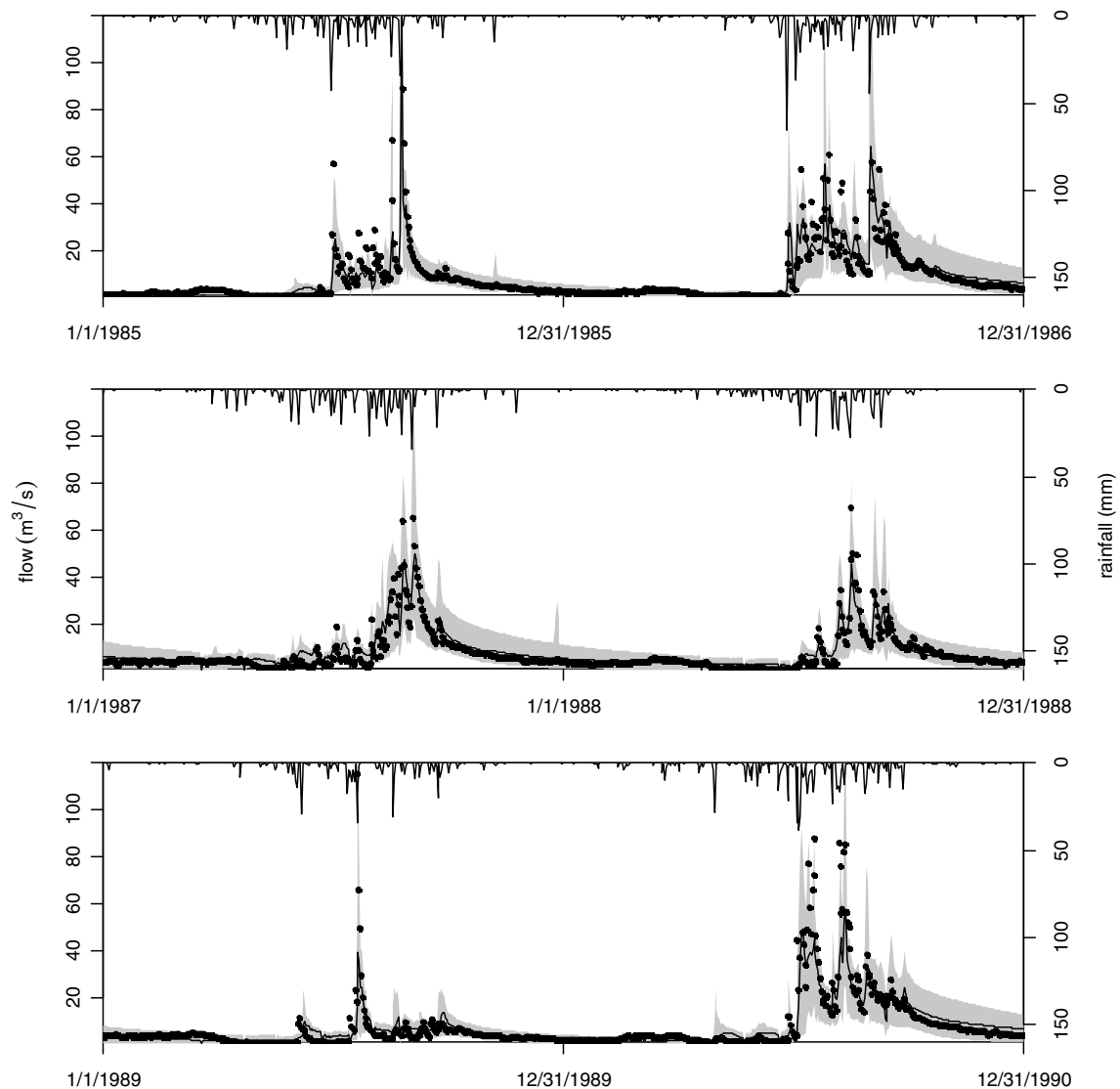
Figure 5 Dotty plot of NS coefficient against each aggregate SWAT parameter conditioning with SUFI-2.

stochasticity of the observations. This usually requires several iterations of the inference procedure for different likelihood functions as statistical tests of residuals can only be performed after the analysis is completed. Once the constructed likelihood function is validated (i.e., the statistical assumptions for the likelihood function are validated), MCMC must be conducted and the resulting chain must be analyzed for the burn-in and stationary periods. Only points from the stationary period should be used for inference.

In this study, the Markov chain was started at a numerical approximation to the maximum of the posterior distribution calculated with the aid of the SCE-UA (Duan et al., 1992) to keep the burn-in period short. The Markov chain was run until 20,000 simulations were reached after the convergence of the chain to the stationary distribution monitored by the Heidelberger and Welch method (Heidelberger and Welch, 1983; Cowles and Carlin, 1996). The ‘‘CODA’’ package (Best et al., 1995) as implemented in the statistical

software package R (<http://www.r-project.org>) was used to perform this test. As shown in Yang et al. (2007a), the statistical assumptions of the likelihood function (Eq. (11)) were not significantly violated, so that we can be confident about the derived prediction uncertainties.

Fig. 7 shows histograms which approximate the marginal posterior distributions of parameters conditioned with Bayesian MCMC. Except for the parameter  $a_{OV\_N.hru}$  which has the approximate uniform distribution of its prior, all other parameters exhibit different posterior distributions than their priors in both parameter range and shape of the distributions. Table 4 lists the means, standard deviations, and correlation matrix of the posterior parameter distribution. As can be seen from Table 4, with the exception of the high correlation between the parameters  $r_{SOL\_K.sol}$  and  $r_{SLSUBBSN.hru}$ , correlations between aggregate parameters are not very high. The high correlations between the parameters of the autoregressive error model ( $\sigma_{dry}$ ,  $\sigma_{wet}$ ,  $\tau_{dry}$ , and  $\tau_{wet}$ ) indicate strong interactions among

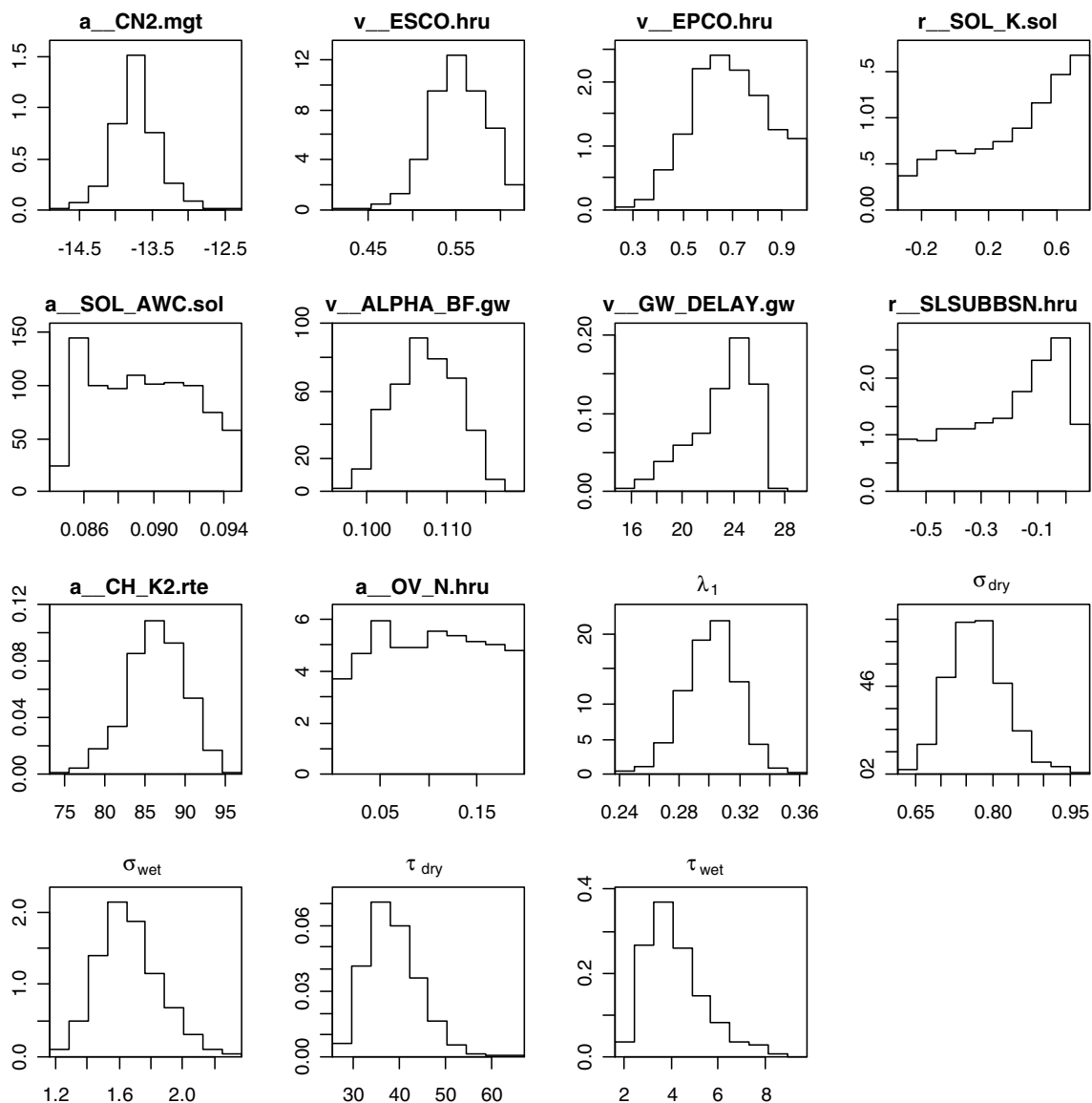


**Figure 6** 95PPU (shaded area) derived by SUFI-2 during the calibration period (top and middle) and validation period (bottom). The dots correspond to the observed discharge at the basin outlet, while the solid line represents the best simulation obtained by SUFI-2.

those parameters. Fig. 8 shows the 95PPU of the model results arising from parameter uncertainty only (dark shaded area) and from total uncertainty (light shaded area) due to parameters, input, model structure and output represented by parameter uncertainty and the autoregressive error model. As can be seen, although the prediction uncertainty based on the parameter uncertainty alone in MCMC is quite narrow, that from parameter uncertainty and uncertainty sources represented by the autoregressive error model brackets over 80% of the observed points for both calibration and validation periods. This indicates that there is a large uncertainty in input, output and model structure in addition to parameter uncertainty. As can also be seen, there is a slight overestimation of prediction uncertainty during the wet season, and this suggests more attention should be paid to the wet season when constructing the likelihood function.

### Result of primitive IS implementation of Bayesian analysis with autoregressive error model

The application with primitive IS is extremely inefficient. In this study, within 100,000 model runs only one parameter set got a weight significantly different from zero. This shows that IS based on the prior as a sampling distribution is computationally too inefficient to be applied to such hydrological problems. As importance sampling is only an alternative numerical implementation of Bayesian inference, it should lead to the same results as those using MCMC in the previous subsection. However, the results of primitive importance sampling using the prior as a sampling distribution (the present application) demonstrate that the technique is too inefficient to produce meaningful results. An iterative narrowing of the sampling distribution that already starts with a good guess (e.g., close to the maximum of the pos-



**Figure 7** Histograms approximating the marginal posterior distributions of aggregate SWAT parameters conditioning with Bayesian MCMC.

terior) would be required to make IS computationally more feasible. However this is still very difficult for high dimensional parameter spaces.

**Comparison**

Table 5 summarizes the results of the comparison in the categories of criteria introduced in Section ‘‘Criteria for the comparison’’. We will exclude primitive IS from further discussion as obviously the numerical technique of primitive IS from the prior fails to give a reasonable approximation to the posterior at the sample sizes we can afford.

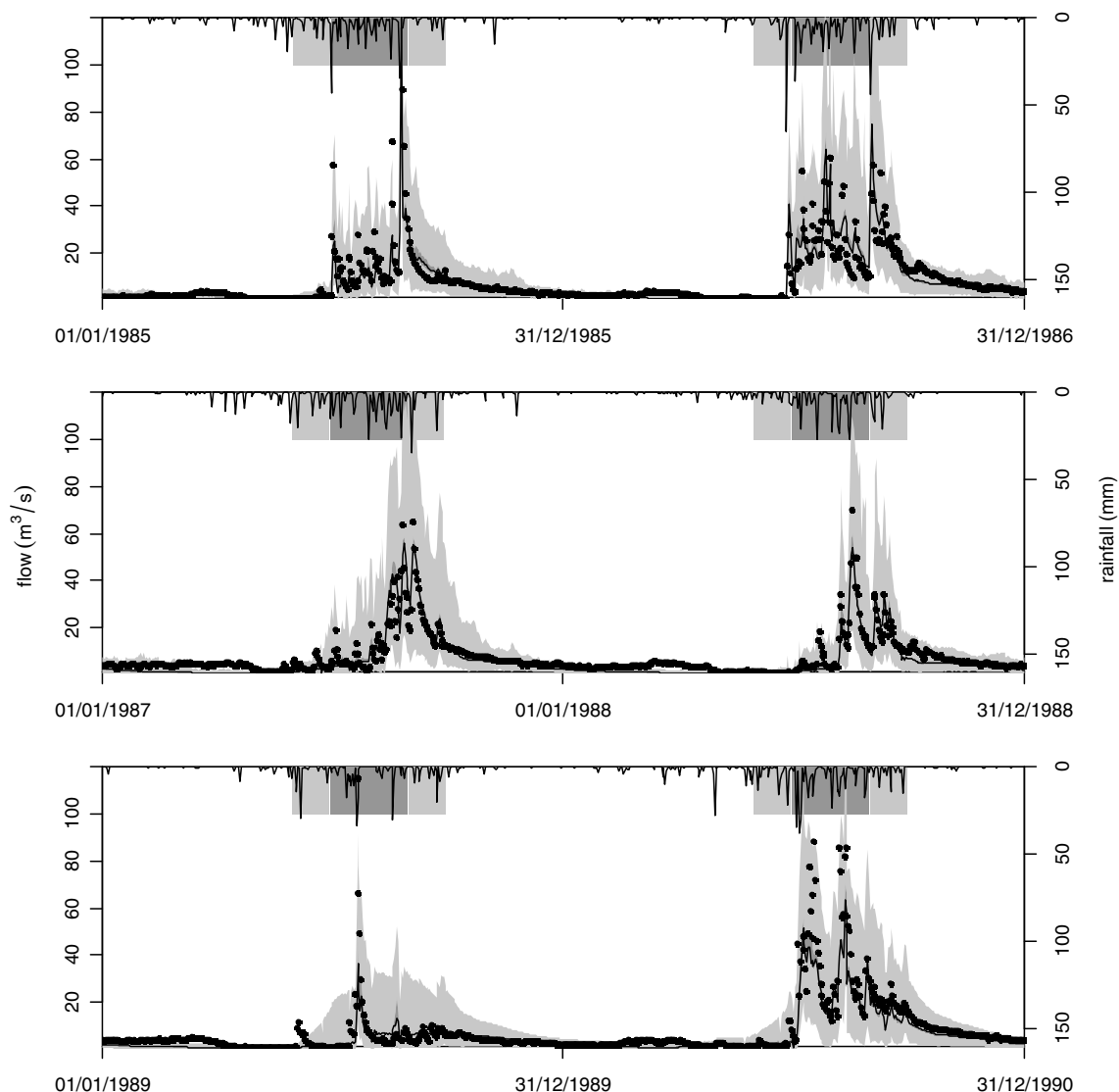
**Parameter estimates and parameter uncertainty**

Results of the marginal posterior parameter distributions are shown as dotted plots in Figs. 1, 3, and 5 or marginal distributions in Fig. 7. In addition, posterior means, standard

deviations and correlation matrices of the techniques that provide these estimates are given in Tables 2–4. Finally, best estimates and 95% parameter uncertainty ranges are summarized in Table 5 (category 1). In general, different techniques result in different posterior parameter distributions, which are represented by different 95% parameter uncertainty ranges, dotted plots and correlation matrices.

Category 1 in Table 5 shows the 95% uncertainty ranges of the marginals of all parameter distributions resulting from GLUE, ParaSol and MCMC, and the posterior parameter intervals resulting from SUFI-2. As can be seen, GLUE provided the widest 95% parameter uncertainty ranges, followed by SUFI-2, MCMC and ParaSol. Most of the uncertainty intervals derived by GLUE contain the corresponding intervals from SUFI-2, MCMC and ParaSol. However, not all the parameter intervals derived by SUFI-2 contain the corresponding intervals of MCMC (for example, a\_OV\_N.hru). Some uncertainty intervals from SUFI-2 do





**Figure 8** 95PPUs associated with parameter uncertainty (dark shaded area) and with total uncertainty (light shaded area) due to parameters, input, model structure and output represented by parameter uncertainty and the autoregressive error model during the calibration period (top and middle) and validation period (bottom). The dots correspond to the observed discharge at the basin outlet, while the line stands for the simulated discharge at the maximum of the posterior distribution.

also indicate the phenomenon that the real response surface is flattened by GLUE. This is in correspondence to other analyses of the GLUE methodology (Mantovan and Todini, 2006). In SUFI-2, parameter correlations are neglected.

#### Performance of the simulation at the mode of the posterior distribution

The performances of the simulation at the mode of posterior distribution are listed in category 2 of Table 5. It is not astonishing that ParaSol (for NS) and MCMC (for log posterior) find the best fit of their respective objective functions because these techniques are based on global optimization algorithms (at least as a first step for MCMC). Such algorithms are much more efficient for finding the maximum of the objective function than random or Latin hypercube sampling. It should be noted that the optimized MCMC parameters do not correspond to the maximum of the

Nash–Sutcliffe as a different objective function was optimized. In despite of this, the simulation with optimized MCMC parameters has similar good NS values as others. The reader can compare other measures of performance in category 2 of Table 5.

#### Model prediction uncertainty

Category 3 in Table 5 lists the relative coverages of measurements ( $p$ -factor), the relative width ( $r$ -factor) and the CRPSs of the 95PPUs for model predictions for all techniques. For the reasons mentioned in Section “Parameter estimates and parameter uncertainty”, ParaSol gave too narrow prediction uncertainty bands which are hardly distinguishable from its best prediction (i.e., the one with the best value of the objective function). GLUE and SUFI-2 led to similar  $p$ -factors (one is 79% and the other is 84%) but different  $r$ -factors (0.65 for GLUE and 1.03 for SUFI-2)

**Table 5** Comparison of criteria of the different inference and uncertainty analysis techniques (see Section “Criteria for the comparison” for the interpretation of the differences)

Category	Criterion	GLUE [NS (Eq. (2))] <sup>a</sup>	ParaSol [SSQ (Eq. (3))]	SUF1-2 [NS (Eq. (2))]	Bayesian inference with cont. autoregr. error model [Log. unnorm. post. prob. density (Eqs. (10)–(12))] <sup>h</sup>	
					MCMC	Primitive IS
1	Best estimate and uncertainty range <sup>b</sup>					
	$a_{CN2.mgt}$	−16.78 (−29.58, −9.84)	−20.97 (−21.93, −20.08)	−26.9 (−30.00, −7.23)	−13.75 (−14.35, −13.04)	−11.57
	$v_{ESCO.hru}$	0.76 (0.02, 0.97)	0.67 (0.65, 0.69)	0.82 (0.43, 1.00)	0.55 (0.49, 0.61)	0.16
	$v_{EPCO.hru}$	0.22 (0.04, 0.90)	0.16 (0.13, 0.20)	1 (0.34, 1.00)	0.62 (0.40, 0.98)	0.73
	$r_{SOL_K.sol}$	−0.16 (−0.36, 0.78)	−0.37 (−0.41, −0.34)	−0.1 (−0.58, 0.34)	0.01 (−0.26, 0.78)	−0.29
	$a_{SOL_AWC.sol}$	0.11 (0.01, 0.15)	0.07 (0.08, 0.08) <sup>c</sup>	0.07 (0.05, 0.15)	0.09 (0.09, 0.09)	0.11
	$v_{ALPHA_BF.gw}$	0.12 (0.06, 0.97)	0.12 (0.08, 0.13)	0.51 (0.23, 0.74)	0.10 (0.10, 0.11)	0.41
	$v_{GW_DELAY.gw}$	159.58 (9.72, 289.29)	107.70 (91.23, 115.20)	190.07 (100.24, 300.00)	24.00 (17.42, 26.11)	43.07
	$r_{SLSUBBSN.hru}$	−0.45 (−0.56, 0.46)	−0.59 (−0.60, −0.58)	−0.52 (−0.60, 0.03)	−0.41 (−0.57, 0.04)	−0.39
	$a_{CH_K2.rte}$	78.19 (6.01, 144.82)	35.70 (27.72, 37.67)	83.95 (69.42, 150.00)	90.18 (78.87, 93.26)	83.85
	$a_{OV_N.hru}$	0.05 (0.00, 0.20)	0.11 (0.07, 0.10)	0.06 (0.00, 0.11)	0.19 (0.01, 0.20)	0.17
	$\lambda$	—	—	—	0.31 (0.26, 0.34)	0.34
	$\sigma_{dry}$	—	—	—	0.73 (0.67, 0.90)	1.11
	$\sigma_{wet}$	—	—	—	1.38 (1.32, 2.08)	6.05
	$\tau_{dry}$	—	—	—	31.12 (29.63, 51.75)	93.47
	$\tau_{wet}$	—	—	—	2.48 (2.39, 7.48)	27.64
	Parameter correlations	Yes	Yes	No	Yes	Yes
2	NS for calibration	0.80	0.82	0.80	0.77	0.60
	NS for validation	0.78	0.81	0.75	0.73	0.51
	$R^2$ for calibration	0.80	0.82	0.81	0.78	0.73
	$R^2$ for validation	0.84	0.85	0.81	0.81	0.80
	Log PDF for calibration <sup>d</sup>	−2124	−2293	−2620	−1460	−1662
	Log PDF for validation	−994	−1237	−1232	−815	−884

3 <sup>g</sup>	<i>p</i> -factor for calibration <sup>e</sup>	79%	18%	84%	85%(10%) <sup>g</sup>	—
	<i>p</i> -Factor for validation	69%	20%	82%	85%(7%)	—
	<i>r</i> -Factor for calibration <sup>f</sup>	0.65	0.08	1.03	1.48 (0.08)	—
	<i>r</i> -Factor for validation	0.51	0.07	0.82	1.16 (0.06)	—
	CRPS for calibration	1.64	0.58	1.62	1.90 (0.54)	—
	CRPS for validation	1.87	0.56	2.03	1.95 (0.57)	—
4	Uncertainty described by parameter uncertainty	All sources of uncertainty	Parameter uncertainty only	All sources of uncertainty	Parameter uncertainty only	Parameter uncertainty only
	Source of prediction uncertainty	Parameter uncertainty	Parameter uncertainty	Parameter uncertainty	Parameter uncertainty + all other uncertainties described by the autoregressive error model	Parameter uncertainty + all other uncertainties described by the autoregressive error model
	Theoretical basis	a. Normalization of generalized likelihood measure b. Primitive random sampling strategy	a. Least squares (probability theory) b. SCE-UA based sampling strategy	a. Generalized objective function b. Latin hypercube sampling; restriction of sampling intervals	a. Likelihood function (Probability theory) b. MCMC starting from optimal parameter set based on SCE-UA	a. Likelihood function (Probability theory) b. Primitive random sampling strategy
	Testability of stat. assum.	No	Yes	No	Yes	Yes
	Result of test		Violated		No contradiction	
5	Difficulty of implement	Very easy	Easy	Easy	More complicated	More complicated
	Number of runs	10,000	7500	1500 + 1500	5000 + 20,000 + 20,000	100,000

<sup>a</sup> The bracketed is the objective function used by the corresponding uncertainty analysis technique.

<sup>b</sup>  $c(a, b)$  for each parameter means:  $c$  is the best parameter estimate,  $(a, b)$  is the 95% parameter uncertainty range except SUFI-2 (in SUFI-2, this interval denotes the posterior parameter distribution).

<sup>c</sup> In ParaSol for parameter  $a_{SOL\_AWC.sol}$ , the optimal value and sampled range are 0.07416 and (0.07242, 0.085), respectively, while 95% parameter uncertainty range is (0.075765, 0.084416). Therefore, the optimal value is outside of the 95% parameter uncertainty.

<sup>d</sup> The  $\sigma_{dry}$ ,  $\sigma_{wet}$ ,  $\tau_{dry}$ , and  $\tau_{wet}$  are used to calculate the logarithm of the posterior probability density function (PDF) which are from the best of MCMC.

<sup>e</sup> *p*-Factor means the percentage of observations covered by the 95PPU (see Section "Criteria for the comparison").

<sup>f</sup> *r*-Factor means relative width of 95PPU (see Section "Criteria for the comparison", defined by Eq. (8)).

<sup>g</sup> In category 3 column "MCMC", the values in the bracket mean the corresponding values from the prediction uncertainty from the parameter uncertainty only.

<sup>h</sup> "Log. unnorm. post. prob. density" means the logarithm of the unnormalized posterior probability density.

during the calibration period, and both different  $p$ -factors (69% for GLUE and 82% for SUFI-2) and  $r$ -factors (0.51 for GLUE and 0.82 for SUFI-2) in the validation period. The reason for this may be that the uncertainty width ( $r$ -factor) of the 95PPU based on GLUE is determined not only by the threshold but also its capability of exploring the parameter space while that of SUFI-2 is determined by the inclusion of some parameter sets with poor objective function in the posterior hypercube. In MCMC, the  $p$ -factors are similar to those of GLUE and SUFI-2, however, the  $r$ -factor is a bit higher. This may be because of the overestimation of errors in the input, output and model structure. It is worth nothing that the coverage ( $p$ -factor) of GLUE and modified ParaSol can be increased at the expense of increasing the  $r$ -factor by decreasing the threshold. And in SUFI-2 this can be done by performing one more iteration. This is not true for MCMC as the coverage does not depend on an arbitrary threshold of the technique.

An examination on the dynamics of these 95PPUs in Figs. 2, 4, 6 and 8 reveals that the uncertainty analysis techniques based on NS show a better coverage in the recession part of the hydrographs than, for example, the peaks and there is also a clear yearly variation (overestimated in 1986, 1987 and 1990) for GLUE and SUFI-2, while MCMC has a better balance between years, but there seems to be a slight overestimation of prediction uncertainty during the wet season. The reason is that in the application to the Chaohe Basin the autoregressive error model explicitly specifies the seasonally dependent values of the  $\sigma$ 's and  $\tau$ 's which reflect the seasonal impacts of input uncertainty, model structural uncertainty and measured response uncertainty. In GLUE and SUFI-2 total uncertainty is expressed as parameter uncertainty, which leads to an equally weighted impact on wet season and dry season.

The CRPS values in Table 5 illustrate the relative width of the uncertainties in the predictions of these applications. While the smaller values for ParaSol (0.58 and 0.56 for calibration and validation periods) and MCMC with parameter uncertainty only (0.54 and 0.57 for calibration and validation periods) indicate narrow uncertainty bands in their predictions, the large values for GLUE (1.64 and 1.87 for calibration and validation periods), SUFI-2 (1.62 and 2.03 for calibration and validation periods), and MCMC (1.90 and 1.95 for calibration and validation periods) reflect wide uncertainty bands in their predictions. These results corroborate our analyses above: that the uncertainties from ParaSol and MCMC with parameter uncertainty only are underestimated as they only consider parameter uncertainty; the similar CRPS values for GLUE, SUFI-2 and MCMC indicate similarly good results as far as the prediction uncertainty band is concerned though visually the prediction uncertainty band of Fig. 8 is much wider than those of Figs. 4 and 6. In MCMC, we can easily see that the parameter uncertainty only contributes to around 30% (0.54/1.90) of the total uncertainty.

### Conceptual basis of the technique

The crucial criteria with respect of the conceptual basis of the techniques are summarized in category 4 of Table 5.

The first two criteria describe how different sources of uncertainty are dealt with. In GLUE and SUFI-2, all sources

of uncertainty are mapped to (enlarged) parameter uncertainty, which will result in wider parameter marginals than ParaSol, MCMC and primitive IS. ParaSol ignores other sources of uncertainty except parameter uncertainty. Finally, the autoregressive error model maps the effect of input, output and model structure uncertainty to a continuous-time autoregressive error model, hence, producing smaller parameter uncertainty ranges.

The conceptual basis of ParaSol, MCMC and primitive IS is probability theory. This has the advantage that the statistical assumptions must be clearly stated and are testable. The statistical assumptions underlying ParaSol (independent and normally distributed residuals) are clearly violated whereas there is no significant violation of the assumptions made by the autoregressive error model (see Yang et al., 2007a). The conceptual bases of GLUE and SUFI-2 are different and their statistical bases are weak (Mantovan and Todini, 2006). GLUE and SUFI-2 allow the users to formulate different likelihood measures (or objective functions) which certainly could have the form of the likelihood function used in the Bayesian framework (e.g., Eq. (11)). However, when using generalized rather than ordinary likelihood functions, GLUE and SUFI-2 lose the probabilistic interpretation of the results. In the last step of the GLUE application, weights are normalized and again interpreted as probabilities. This procedure lacks a consistent and testable statistical formulation. Also SUFI-2 lacks a rigorous probabilistic formulation. Parameter uncertainty formulated by a uniform distribution in a hypercube is propagated through the hydrologic model correctly, but the lack of consideration of parameter correlation and the inclusion of some simulations with poor objective function values are the problems with this methodology.

### Difficulty of implementation and computational efficiency

The final category of comparison criteria (category 5) in Table 5 is difficulty of implementation and computational efficiency.

Implementation of GLUE is straightforward and very easy. Due to the calculation of sensitivity measures and global optimization, implementation of SUFI-2 and ParaSol is somewhat more complicated than GLUE but still easy compared to the Bayesian techniques (see below). Due to the most complicated likelihood function and processing technique, the Bayesian techniques (i.e., MCMC and primitive IS) need more effort to be implemented (such as the construction of likelihood function, test of the statistical assumptions, etc.).

Due to an efficient optimization procedure SCE-UA, ParaSol does not require intensive computations (7000 model runs). Concerning SUFI-2, taking into account a relatively sparse coverage of the parameter space, it is also not very computationally expensive to run (3000 model runs). Depending on the required coverage, GLUE can be run with smaller or bigger sample sizes (10,000 model runs in this study). The computationally most expensive technique is Bayesian inference in this study: MCMC takes 45,000 model runs while the primitive IS is too inefficient to obtain any reasonable result even after 100,000 model runs. This is certainly the major disadvantage of this technique.

## Conclusions

After comparing the applications of different uncertainty analysis techniques to a distributed watershed model (SWAT) for the Chaohe Basin in North China, we come to the following conclusions:

- (1) *Application of GLUE based on the Nash–Sutcliffe coefficient.* This application led to the widest marginal parameter uncertainty intervals of the model parameters and to a good prediction uncertainty in the sense of coverage of measurements by the uncertainty bands. On the other hand, the inefficiency of the global sampling procedure leads to problems in locating the maximum or maxima of the objective function. When using the likelihood measure NS, this technique tends to flatten the response surface of the objective function. The wide parameter uncertainty ranges are a result of this flat response surface and the chosen threshold value for behavioral solutions.
- (2) *Application of ParaSol based on the Nash–Sutcliffe coefficient.* ParaSol was able to find a good approximation to the global maximum of NS, however, it led to too narrow prediction uncertainty bands due to a violation of the statistical assumption of independently and normally distributed errors. Decreasing the threshold value in modified ParaSol increases its prediction uncertainty but the choice of the threshold value may be hard to justify.
- (3) *Application of SUFI-2 based on the Nash–Sutcliffe coefficient.* This technique could be run with the smallest number of model runs to achieve good prediction uncertainty ranges in the sense of a reasonable coverage of data points by the prediction uncertainty bands. This characteristic is very important for computationally demanding models. However, the choice of a small sample size obviously decreases the exploration of the parameter space and the poorly defined convergence criterion and not considering parameter correlations decreases the ability of finding a unique posterior.
- (4) *Application of MCMC based on a continuous-time autoregressive error model.* Due to the global optimization performed before starting the Markov chain, MCMC achieved a good approximation to the maximum of the posterior. The statistical assumptions of the error model are testable and in reasonable agreement with empirical evidence. The additional parameters of the error model give the user some freedom in the description of the effect of input and model structure error (such as seasonal dependence of the magnitude of these effects). The main disadvantages of this technique are the difficulty of constructing the likelihood function, the large number of simulations required to get a good approximation to the posterior, and the difficulty of covering multi-modal distributions caused by the numerical implementation of MCMC.
- (5) *Application of primitive IS based on a continuous-time autoregressive error model.* The implementa-

tion of primitive importance sampling is much too inefficient to get a reasonable approximation to the posterior. This is a similar problem as for GLUE, but it is much more extreme here as GLUE uses a flatter (generalized) likelihood function. Importance sampling could only be an alternative to MCMC if a careful and adaptive process is applied for choosing the likelihood function.

- (6) *About choosing the objective functions.* GLUE and SUFI-2 are very flexible by allowing for arbitrary likelihood measures/objective functions. On the other hand, GLUE and SUFI-2 lose their statistical basis when using this additional freedom. The real capability of exploring the parameter space is also seriously affected by the choice of the objective function. In ParaSol, though the objective function and the way to split the parameter set are statistically based, the underlying statistical assumptions are seriously violated. This makes the results unreliable. The likelihood function used for MCMC has a testable statistical basis and the test of our result did not indicate a severe violation of the assumptions. This makes the Bayesian inference which is based on this likelihood function conceptually the most satisfying technique.

Despite these big differences in concepts and performance, GLUE, SUFI-2 and MCMC led to similar prediction uncertainty bands. Our preference is for MCMC because Bayesian inference has a sound theoretical foundation and the statistical assumptions underlying the likelihood function based on the autoregressive error model is testable and did not indicate significant violations of the assumptions. However, further efforts are required to improve the formulation of likelihood functions used in hydrological applications. In particular, it would be interesting to formulate a likelihood function that not only describes the effect of input, model structure and output uncertainty on model output (e.g., our autoregressive error model), but also resolves the different sources of uncertainty. As mentioned in the introduction, such techniques are under development (see category (iii) in the introduction), but still are computationally too expensive for straightforward use with complex hydrological models.

## Acknowledgements

We would like to thank Dr. Ann van Griensven from UNESCO-IHE for providing the ParaSol program and suggestions, and Dr. Jasper Vrugt from the Los Alamos National Laboratory for constructive comments.

## References

- Abbaspour, K.C., Johnson, C.A., van Genuchten, M.T., 2004. Estimating uncertain flow and transport parameters using a sequential uncertainty fitting procedure. *Vadose Zone Journal* 3 (4), 1340–1352.
- Abbaspour, K.C., Yang, J., Maximov, I., Siber, R., Bogner, K., Mieleitner, J., Zobrist, J., Srinivasan, R., 2007. Spatially-distributed modelling of hydrology and water quality in the

- pre-alpine/alpine Thur watershed using SWAT. *Journal of Hydrology* 333, 413–430.
- Adeuya, R.K., Lim, K.J., Engel, B.A., Thomas, M.A., 2005. Modeling the average annual nutrient losses of two watersheds in Indiana using GLEAMS-NAPRA. *Transactions of the Asae* 48 (5), 1739–1749.
- Ajami, N.K., Duan, Q.Y., Sorooshian, S., 2007. An integrated hydrologic Bayesian multimodal combination framework: confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resources Research* 43 (1), Art. No. W01403.
- Arnold, J.G., Srinivasan, R., Muttiah, R.S., Williams, J.R., 1998. Large area hydrologic modeling and assessment – Part 1: Model development. *Journal of the American Water Resources Association* 34 (1), 73–89.
- Arnold, Z., 1996. *An Introduction to Bayesian Inference in Econometrics* Wiley Classics Library Edition. John Wiley, New York, p. 66.
- Bates, B.C., Campbell, E.P., 2001. A Markov chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall–runoff modeling. *Water Resources Research* 37 (4), 937–947.
- Bayarri, M.J., Berger, J.O., Paulo, R., Sacks, J., Cafeo, J.A., Cavendish, J., Lin, C.-H., Tu, J., 2007. A framework for validation of computer models. *Technometrics* 49 (2), 138–154.
- Bekele, E.G., Nicklow, J.W., 2005. Multiobjective management of ecosystem services by integrative watershed modeling and evolutionary algorithms. *Water Resources Research* 41, W10406. doi:10.1029/2005WR004090.
- Best, N.G., Cowles, M.K., Vines, S.K., 1995. *Convergence Diagnosis and Output Analysis software for Gibbs Sampler output: Version 0.3*. Medical Research Council Biostatistics Unit, Cambridge.
- Beven, K., Binley, A., 1992. The future of distributed models – model calibration and uncertainty prediction. *Hydrological Processes* 6 (3), 279–298.
- Beven, K., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of Hydrology* 249 (1–4), 11–29.
- Bicknell, B.R., Imhoff, J., Kittle, J., Jobes, T., Donigian, A.S., 2000. *Hydrological Simulation Program – Fortran User's Manual*. Release 12, US EPA.
- Blazkova, S., Beven, K., Tacheci, P., Kulasova, A., 2002. Testing the distributed water table predictions of TOPMODEL (allowing for uncertainty in model calibration): the death of TOPMODEL? *Water Resources Research* 38 (11), 1257. doi:10.1029/2001WR000912.
- Box, G.E.P., Cox, D.R., 1964. An analysis of transformations. *Journal of the Royal Statistical Society Series B*, 211–252.
- Box, G.E.P., Cox, D.R., 1982. An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association* 77 (377), 209–210.
- Brockwell, P.J., Davis, R.A., 1996. *Introduction to Time Series and Forecasting*. Springer, New York.
- Brockwell, P.J., 2001. Continuous-time ARMA processes. In: Shanbhag, D.N., Rao, C.R. (Eds.), *Stochastic Processes: Theory and Methods*, Handbook of Statistics, vol. 19. Elsevier, Amsterdam, pp. 249–276.
- Cameron, D., Beven, K., Naden, P., 2000a. Flood frequency estimation by continuous simulation under climate change (with uncertainty). *Hydrology and Earth System Sciences* 4 (3), 393–405.
- Cameron, D., Beven, K., Tawn, J., Naden, P., 2000 b. Flood frequency estimation by continuous simulation (with likelihood based uncertainty estimation). *Hydrology and Earth System Sciences* 4 (1), 23–34.
- Chow, V.T., Maidment, D.R., Mays, L.W., 1988. *Applied Hydrology*. McGraw-Hill, New York.
- Claessens, L., Hopkinson, C., Rastetter, E., Vallino, J., 2006. Effect of historical changes in land use and climate on the water budget of an urbanizing watershed. *Water Resources Research* 42, W03426. doi:10.1029/2005WR004131.
- Cochrane, T.A., Flanagan, D.C., 2005. Effect of DEM resolutions in the runoff and soil loss predictions of the WEPP watershed model. *Transactions of the Asae* 48 (1), 109–120.
- Cotler, H., Ortega-Larrocea, M.P., 2006. Effects of land use on soil erosion in a tropical dry forest ecosystem, Chamela watershed, Mexico. *Catena* 65 (2), 107–117.
- Cowles, M.K., Carlin, B.P., 1996. Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association* 91 (434), 883–904.
- Cunge, J.A., 1969. On the subject of a flood propagation method (Muskingum method). *Journal of Hydraulics Research* 7 (2), 205–230.
- Duan, Q.Y., Sorooshian, S., Ibbitt, R.P., 1988. A maximum likelihood criterion for use with data collected at unequal time intervals. *Water Resources Research* 24 (7), 1163–1173.
- Duan, Q.Y., Sorooshian, S., Gupta, V., 1992. Effective and efficient global optimization for conceptual rainfall–runoff models. *Water Resources Research* 28 (4), 1015–1031.
- Freer, J., Beven, K., Ambrose, B., 1996. Bayesian estimation of uncertainty in runoff prediction and the value of data: an application of the GLUE approach. *Water Resources Research* 32 (7), 2161–2173.
- Gelman, S., Carlin, J.B., Stren, H.S., Rubin, D.B., 1995. *Bayesian Data Analysis*. Chapman and Hall, New York.
- Geweke, J., 1989. Bayesian-Inference in Econometric-Models Using Monte-Carlo Integration. *Econometrica* 57 (6), 1317–1339.
- Gupta, H.V., Beven, K.J., Wagener, T., 2005. Model calibration and uncertainty estimation. In: Anderson, M.G. (Ed.), *Encyclopedia of Hydrological Sciences*. John Wiley, New York, pp. 2015–2031.
- Heidelberger, P., Welch, P.D., 1983. Simulation run length control in the presence of an initial transient. *Operations Research* 31 (6), 1109–1144.
- Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting* 15 (5), 559–570.
- Hornberger, G.M., Spear, R.C., 1981. An approach to the preliminary-analysis of environmental systems. *Journal of Environmental Management* 12 (1), 7–18.
- Hossain, F., Anagnostou, E.N., Lee, K.H., 2004. A non-linear and stochastic response surface method for Bayesian estimation of uncertainty in soil moisture simulation from a land surface model. *Nonlinear Processes in Geophysics* 11 (4), 427–440.
- Huang, Y.F., Zou, Y., Huang, G.H., Maqsood, I., Chakma, A., 2005. Flood vulnerability to climate change through hydrological modeling – a case study of the swift current creek watershed in western Canada. *Water International* 30 (1), 31–39.
- Hundecha, Y., Bardossy, A., 2004. Modeling of the effect of land use changes on the runoff generation of a river basin through parameter regionalization of a watershed model. *Journal of Hydrology* 292 (1–4), 281–295.
- Jorgeson, J., Julien, P., 2005. Peak flow forecasting with radar precipitation and the distributed model CASC2D. *Water International* 30 (1), 40–49.
- Kavetski, D., Franks, S.W., Kuczera, G., 2003. Confronting input uncertainty in environmental modelling. In: Duan, Q., Gupta, H.V., Sorooshian, S., Rousseau, A.N., Turcotte, R. (Eds.), *Calibration of Watershed Models*. American Geophysical Union, Washington, DC, pp. 49–68.
- Kavetski, D., Kuczera, G., Franks, S.W., 2006a. Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resources Research* 42 (3), Art. No. W03407.
- Kavetski, D., Kuczera, G., Franks, S.W., 2006 b. Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water Resources Research* 42 (3), Art. No. W03408.

- Kennedy, M.C., O'Hagan, A., 2001. Bayesian calibration of computer models. *Journal of the Royal Statistical Society Series B* 63 (3), 425–464.
- Kuczera, G., 1983. Improved parameter inference in catchment models. 1. Evaluating parameter uncertainty. *Water Resources Research* 19 (5), 1151–1162.
- Kuczera, G., Parent, E., 1998. Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm. *Journal of Hydrology* 211 (1–4), 69–85.
- Kuczera, G., Kavetski, D., Franks, S., Thyer, M., 2006. Towards a Bayesian total error analysis of conceptual rainfall–runoff models: characterising model error using storm-dependent parameters. *Journal of Hydrology* 331, 161–177.
- Makowski, D., Wallach, D., Tremblay, M., 2002. Using a Bayesian approach to parameter estimation: comparison of the GLUE and MCMC methods. *Agronomie* 22 (2), 191–203.
- Mantovan, P., Todini, E., 2006. Hydrological forecasting uncertainty assessment: incoherence of the GLUE methodology. *Journal of Hydrology* 330, 368–381.
- Marshall, L., Nott, D., Sharma, A., 2004. A comparative study of Markov chain Monte Carlo methods for conceptual rainfall–runoff modeling. *Water Resources Research* 40, W02501. doi:10.1029/2003WR002378.
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. *Journal of Chemical Physics* 21 (6), 1087–1092.
- Moradkhani, H., Hsu, K.-L., Gupta, H., Sorooshian, S., 2005. Uncertainty assessment of hydrologic model states and parameters: sequential data assimilation using the particle filter. *Water Resources Research* 41, W05012. doi:10.1029/2004WR003604.
- Neitsch, S.L., Arnold, J.G., Kiniry, J.R., Williams, J.R., 2001. Soil and Water Assessment Tool User's Manual, version 2000. Grassland, Soil and Water Research Laboratory, Agricultural Research Service, Blackland Research Center, Texas Agricultural Experiment Station.
- Pednekar, A.M., Grant, S.B., Jeong, Y., Poon, Y., Oancea, C., 2005. Influence of climate change, tidal mixing, and watershed urbanization on historical water quality in Newport Bay, a saltwater wetland and tidal embayment in southern California. *Environmental Science and Technology* 39 (23), 9071–9082.
- Reichert, P., Schervish, M., Small, M.J., 2002. An efficient sampling technique for Bayesian inference with computationally demanding models. *Technometrics* 44 (4), 318–327.
- Reichert, P., 2005. UNCSIM – A computer programme for statistical inference and sensitivity, identifiability, and uncertainty analysis. In: Teixeira, J.M.F., Carvalho-Brito, A.E. (Eds.), *Proceedings of the 2005 European Simulation and Modelling Conference (ESM 2005)*, October 24–26, Porto, Portugal, EUROSIS-ETI. pp. 51–55.
- Reichert, P., 2006. A standard interface between simulation programs and systems analysis software. *Water Science and Technology* 53 (1), 267–275.
- Reichert, P., Mieleitner, J., submitted for publication. Analyzing input and structural uncertainty of a hydrological model with stochastic, time-dependent parameters. *Water Resources Research*, submitted for publication.
- Santhi, C., Arnold, J.G., Williams, J.R., Hauck, L.M., Dugas, W.A., 2001. Application of a watershed model to evaluate management effects on point and nonpoint source pollution. *Transactions of the Asae* 44 (6), 1559–1570.
- Schaefli, B., Talamba, D.B., Musy, A., 2007. Quantifying hydrological modeling errors through a mixture of normal distributions. *Journal of Hydrology* 332, 303–315.
- Sorooshian, S., Dracup, J.A., 1980. Stochastic parameter estimation procedures for hydrologic rainfall–runoff models – correlated and heteroscedastic error cases. *Water Resources Research* 16 (2), 430–442.
- Todini, E., 2007. Hydrological catchment modelling: past, present and future. *Hydrology and Earth System Sciences* 11 (1), 468–482.
- USDA Soil Conservation Service, 1972. Section 4: Hydrology. In: *National Engineering Handbook*. SCS.
- Van Griensven, A., Meixner, T., 2006. Methods to quantify and identify the sources of uncertainty for river basin water quality models. *Water Science and Technology* 53 (1), 51–59.
- Vrugt, J.A., Gupta, H.V., Bouten, W., Sorooshian, S., 2003. A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resources Research* 39 (8), Art. No.-1201.
- Vrugt, J.A., Diks, C.G.H., Bouten, W., Gupta, H.V., Verstraten, J.M., 2005. Towards a complete treatment of uncertainty in hydrologic modelling: combining the strengths of global optimization and data assimilation. *Water Resources Research* 41 (1), W01017. doi:10.1029/2004WR003059.
- Williams, J.R., 1995. The EPIC model. In: Singh, V.P. (Ed.), *Computer Models of Watershed Hydrology*. Water Resources Publication, Colorado, USA, pp. 909–1000.
- Young, R.A., Onstad, C.A., Bosch, D.D., Anderson, W.P., 1989. AGNPS – A nonpoint-source pollution model for evaluating agricultural watersheds. *Journal of Soil and Water Conservation* 44 (2), 168–173.
- Yang, J., Abbaspour, K.C., Reichert, P., 2005. Interfacing SWAT with systems analysis tools: a generic platform. In: Srinivasan, R., Jacobs, J., Day, D., Abbaspour, K. (Eds.), *Third International SWAT Conference Proceedings*, July 11–15, Zurich, Switzerland. pp. 169–178.
- Yang, J., Reichert, P., Abbaspour, K.C., Yang, H., 2007a. Hydrological modelling of the Chaohe Basin in China: statistical model formulation and Bayesian inference. *Journal of Hydrology* 340, 167–182.
- Yang, J., Reichert, P., Abbaspour, K.C., 2007b. Bayesian uncertainty analysis in distributed hydrologic modeling: a case study in the Thur River basin (Switzerland). *Water Resources Research* 43, W10401. doi:10.1029/2006WR005497.
- Zacharias, I., Dimitriou, E., Koussouris, T., 2005. Integrated water management scenarios for wetland protection: application in Trichonis Lake. *Environmental Modelling and Software* 20 (2), 177–185.