Embracing Equifinality with Efficiency: Limits of Acceptability Sampling Using the DREAM$_{(\mathrm{ABC})}$ algorithm

Jasper A. Vrugt [*†‡] and, Keith J. Beven[§¶]

April 6, 2016

[*]Corresponding author. Department of Civil and Environmental Engineering, University of California Irvine, Irvine, CA 92697-1075. Email: jasper@uci.edu

[†]Department of Earth System Science, University of California Irvine, Irvine, CA 92697-1075.

[‡]Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, The Netherlands

[§]Lancaster Environment Centre, Lancaster University, Lancaster, UK

[¶]Department of Earth Sciences, Uppsala University, Uppsala, Sweden

[‖]CATS, London School of Economics, London, UK

Abstract

This essay illustrates some recent developments to the DiffeRential Evolution Adaptive Metropolis (DREAM) MATLAB toolbox of *Vrugt* (2016a) to delineate and sample the behavioral solution space of set-theoretic likelihood functions used within the GLUE (limits of acceptability) framework (*Beven and Binley*, 1992; *Beven and Freer*, 2001; *Beven*, 2006; *Beven and Binley*, 2014a). This work builds on the DREAM$_{(\mathrm{ABC})}$ algorithm of *Sadegh and Vrugt* (2014) and enhances significantly the accuracy and CPU-efficiency of Bayesian inference with GLUE.

Keywords: GLUE, Limits of Acceptability, Markov Chain Monte Carlo, Posterior Sampling, DREAM, DREAM$_{(\mathrm{ABC})}$, Sufficiency, Hydrologic Modeling

1 Introduction and Scope

In any analysis of predictive uncertainty associated with the application of a model a number of decisions have to be made. These may be summarized as:

1. Decide on what parameters and/or input data are to be considered uncertain
2. Decide on prior distributions from which they should be (jointly) sampled
3. Decide on a sampling methodology to generate realizations
4. Decide on a likelihood (or fuzzy membership) measure to express the degree of belief in a model realization
5. Decide on a method for combining likelihood measures if necessary

$\int_{\Theta} f(\theta)\mathrm{d}\theta$ where Θ denotes the feasible search space, $\theta \in \Theta \in \mathbb{R}^1$, which can be equivalent to some upper and lower bound values of the parameter, θ.

None of these choices are simple. All will affect the outcomes and interpretation of an uncertainty analysis. *Beven* (2006) distinguishes between ideal and non-ideal applications. In ideal cases, where uncertainties can be satisfactorily described as aleatory in nature, it will be possible to define prior information as joint statistical distributions, it will be possible to define a likelihood function based on a structural model of the residuals, it will be possible to update likelihoods using Bayes equation, and the outcomes will have a formal probabilistic interpretation. In non-deal cases, where epistemic uncertainties dominate and model residual characteristics may be non-stationary or arbitrary, it may be much more difficult to define prior information, or find a satisfactory structural model of the residuals, and the use of Bayes with a simple statistical likelihood function can lead to nonsensical results (*Beven*, 2012; *Vrugt and Sadegh*, 2013a). Thus, it has been suggested that every uncertainty analysis should be associated with an audit trail of the many simplifying assumptions on which it is based as a way of communicating meaning and limitations to potential users (see *Beven et al.* (2011) for flood inundation modeling case studies).

In this paper we focus on one particular aspect of the uncertainty estimation process, that of the choice of sampling methodology, and its impact on the outcomes of an uncertainty estimation based on the Generalised Likelihood Uncertainty Estimation (GLUE) Limits of Acceptability approach (*Beven*, 2006; *Page et al.*, 2007; *Blazkova and Beven*, 2009; *Liu et al.*, 2009; *Beven*, 2012; *Beven and Binley*, 2014a). Past applications of GLUE have commonly used brute-force random sampling techniques across uniform prior distributions of uncertain parameters lacking better prior information. But when run times for a single model realisation are long, or when there are a large number of uncertain parameters and the dimensionality of the search space is high, then computer limitations can result in a sparse sample of model realisations, many of which may be rejected as non-behavioural (though it is worth noting that the original presentation of GLUE in *Beven and Binley* (1992) was based on a selective sampling algorithm in an attempt to improve efficiency given the computing limitations at that time, see also *Beven and Binley* (2014a)). We should expect that such sparse sampling will result in relatively poor explorations of the model space and consequent uncertainty estimates, regardless of the other decisions in the estimation process.

One advantage of statistical uncertainty estimation is that the formal likelihood assumptions can be closely linked to more efficient search algorithms based on Monte Carlo Markov Chain

techniques. In a series of papers from *Vrugt et al.* (2003) on, efficient search methods have been developed for a variety of problems by combining optimisation and adaptive search algorithms. The latest of these methods, the DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm has been designed to simplify Bayesian inference and speed-up estimation of posterior parameter distributions significantly. DREAM is an adaptation of the Shuffled Complex Evolution Metropolis (*Vrugt et al.*, 2003) algorithm and has the advantage of maintaining detailed balance and ergodicity. Benchmark experiments have shown that DREAM is superior to other adaptive MCMC sampling approaches, and in high-dimensional spaces even provides better solutions than powerful optimization algorithms (*Vrugt et al.*, 2008a, 2009; *Laloy and Vrugt*, 2012a; *Laloy et al.*, 2012b, 2013; *Linde and Vrugt*, 2013; *Lochbühler et al.*, 2014; *Laloy et al.*, 2015) (see also our response in *Vrugt and Laloy* (2014) to the comment of *Chu et al.* (2014)).

In the past few years, DREAM has found widespread application and use in many different fields of study, including (among others) atmospheric chemistry (*Partridge et al.*, 2011, 2012), biogeosciences (*Scharnagl et al.*, 2010; *Braakhekke et al.*, 2013; *Ahrens and Reichstein*, 2014; *Dumont et al.*, 2014; *Starrfelt and Kaste*, 2014), biology (*Coehlo et al.*, 2011; *Zaoli et al.*, 2014), chemistry (*Owejan et al.*, 2012; *Tarasevich et al.*, 2013; *DeCaluwe et al.*, 2014; *Gentsch et al.*, 2014), ecohydrology (*Dekker et al.*, 2011), ecology (*Barthel et al.*, 2011; *Gentsch et al.*, 2014; *Iizumi et al.*, 2014; *Zilliox and Goselin*, 2014), economics and quantitative finance (*Bauwens et al.*, 2011; *Lise et al.*, 2012; *Lise*, 2013), epidemiology (*Mari et al.*, 2011; *Rinaldo et al.*, 2012; *Leventhal et al.*, 2013), geophysics (*Bikowski et al.*, 2012; *Linde and Vrugt*, 2013; *Laloy et al.*, 2012b; *Carbajal et al.*, 2014; *Lochbühler et al.*, 2014), geostatistics (*Minasny et al.*, 2011; *Sun et al.*, 2013), hydrogeophysics (*Hinnell et al.*, 2014), hydrologeology (*Keating et al.*, 2010; *Laloy et al.*, 2013; *Malama et al.*, 2013), hydrology (*Vrugt et al.*, 2008a, 2009; *Shafii et al.*, 2014), physics (*Dura et al.*, 2014; *Horowitz et al.*, 2012; *Toyli et al.*, 2012; *Kirby et al.*, 2013; *Yale et al.*, 2013; *Krayer et al.*, 2014), psychology (*Turner and van Zandt*, 2012), soil hydrology (*Wöhling and Vrugt*, 2011), and transportation engineering (*Kow et al.*, 2012). A recent paper by *Vrugt* (2016a) reviews the basic theory of DREAM and introduces a MATLAB toolbox of this algorithm.

The development of DREAM in *Vrugt et al.* (2008a) and *Vrugt et al.* (2009) was inspired by an urgent need for sampling methods that can search efficiently and reliably for the posterior parameter distribution of dynamic simulation models. An original aim in this and related work was to improve the efficiency of applying Bayes methods using likelihood functions derived from simple statistical assumptions (*Schoups and Vrugt*, 2010). But DREAM can also be used to solve a much wider variety of inference problems, for instance involving discrete/combinatorial search spaces (*Vrugt and ter Braak*, 2011), summary statistics (*Sadegh and Vrugt*, 2014), data assimilation (*Vrugt et al.*, 2013b), informal likelihood functions (*Blasone et al.*, 2008), diagnostic model evaluation (*Vrugt and Sadegh*, 2013a; *Sadegh et al.*, 2015), and model averaging (*Vrugt et al.*, 2008b) and the GLUE limits of acceptability framework of *Beven* (2006).

Within this GLUE framework, behavioural models are defined as those that satisfy limits of acceptability around each observation or summary statistic defined prior to running any model realisations. These limits should reflect the observational error of the variable being compared, together with the effects of input error and commensurability errors resulting from differences in scale (spatial and/or temporal) between observed and predicted values. In a previous paper *Sadegh and Vrugt* (2013) have shown that the limits of acceptability framework of GLUE has many elements in common with approximate Bayesian computation (ABC). In particular, the

approaches are virtually equivalent if each observation of the calibration data record is used as a summary statistic.

This paper illustrates some recent developments to the DiffeRential Evolution Adaptive Metropolis (DREAM) MATLAB toolbox of *Vrugt* (2016a) to delineate and sample the behavioral solution space of set-theoretic likelihood measures used within the limits of acceptability framework (*Beven*, 2006; *Beven and Binley*, 2014a). The work builds on the DREAM$_{(\mathrm{ABC})}$ algorithm of *Sadegh and Vrugt* (2014) and enhances significantly the efficiency of sampling the model space within the GLUE methodology. The DREAM algorithm has important advantages over uniform sampling methods that have commonly been used in GLUE as it will generally provide a more exact estimate of parameter and model predictive uncertainty. In particular, it will be shown herein that the use of inferior sampling methods can lead to erroneous conclusions about model rejection.

The remainder of this paper is organized as follows. Section 2 summarizes the GLUE Limits of Acceptability methodology. In section 3, the connection between the limits of acceptability framework and approximate Bayesian computation is discussed. Section 4 then reviews the DREAM$_{(\mathrm{ABC})}$ algorithm of *Sadegh and Vrugt* (2014) which is used to sample the behavioral parameter space, including the mathematical formulation of the likelihood measure and selection rule used to accept proposals within the limits of acceptability framework. These functions are designed carefully so as not to violate detailed balance and to make sure that the behavioral parameter and simulation space, which satisfy the limits of acceptability, are accurately and efficiently sampled. Section 5 then documents the results of three different case studies involving surface hydrology and vadose zone modeling. In this section we benchmark the sampling efficiency of the DREAM$_{(\mathrm{ABC})}$ algorithm against rejection sampling used within GLUE. Finally, section 6 concludes this paper with a summary of the main findings.

2 The Generalized Likelihood Uncertainty Estimation (GLUE) methodology

GLUE has been used widely in hydrological and other types of modelling (*Beven and Binley*, 1992; *Beven and Freer*, 2001; *Beven*, 2006, 2009; *Beven and Binley*, 2014a). The origins of the method lie in trying to deal with uncertainty estimation problems for which simple theoretical likelihood assumptions do not seem appropriate (although it can include statistical likelihood functions as special cases when the strong assumptions required are justified). The GLUE methodology aims to find a set of representations (model inputs, model structures, model parameter sets, model errors) that are behavioral in the sense of being acceptably consistent with the (non-error-free) observations. This method was inspired by the *Hornberger and Spear* (1981) method of sensitivity analysis and operates within the context of Monte Carlo analysis coupled with Bayesian or fuzzy estimation and propagation of uncertainty.

The GLUE limits of acceptability method proceeds as follows. The index i is used to mean '$for\ all\ i \in \{1, \ldots, N\}$'. For each observation with which the model will be compared, limits of acceptability are defined prior to running the model, to reflect (in so far as is possible) the effects of input and observation error. To allow for the fact that different observations might have quite different scales, the limits of acceptability can expressed as a normalised scale (-1 at the low limit, 0 at the observed value, +1 at the upper limit). Performance weightings within

the limits can also be specified as appropriate.

1. Draw N points from the prior parameter distribution, $P(\boldsymbol{\theta})$ and store the samples in a $N \times d$ matrix $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\}$.
2. Evaluate the model for the ith sample of $\boldsymbol{\Theta}$ and thus $\mathbf{Y}_i \leftarrow \mathcal{F}(\boldsymbol{\theta}_i, \tilde{\mathbf{U}}, \tilde{\mathbf{x}}_0)$ in terms of the minimum absolute normalised score that would be required for each model to be acceptable.
3. Rank the models (parameter vectors) by their minimum scores, and select the top R realisations above some acceptability threshold as behavioral. This threshold would normally be an absolute value of 1, indicating that all observations are reproduced within the specified limits of acceptability. All other realisations are given a likelihood value of zero. Collect these behaviorial solutions in a $R \times d$ matrix $\mathbf{B}$ and the corresponding simulations in a matrix $\mathbf{Z}$ of size $n \times R$.
4. Calculate the likelihood measure, $L(\boldsymbol{\theta}_i|\tilde{\mathbf{Y}})$ of the simulated values, $\mathbf{Y}_i$ based on the performance weightings within the limits of acceptability.
5. Normalize the likelihood values of the $r = \{1, \ldots, R\}$ solutions of $\mathbf{B}$, $\bar{L}(\mathbf{B}_r|\tilde{\mathbf{Y}}) = L(\mathbf{B}_r|\tilde{\mathbf{Y}}) / \sum_{r=1}^{R} L(\mathbf{B}_r|\tilde{\mathbf{Y}})$ so that $\sum_{r=1}^{R} \bar{L}(\mathbf{B}_r|\tilde{\mathbf{Y}}) = 1$.
6. For each time $t = \{1, \ldots, n\}$ calculate the likelihood weighted cumulative density function (cdf) by assigning $\mathbf{Z}_{tr}$ the likelihood $\bar{L}(\mathbf{B}_r|\tilde{\mathbf{Y}})$; $r = \{1, \ldots, R\}$.
7. Derive the 95% simulation uncertainty ranges of $\mathcal{F}_t(\boldsymbol{\Theta}, \tilde{\mathbf{U}}, \tilde{\mathbf{x}}_0)$ from the likelihood weighted cdf.

Note that the limits of acceptability may also be defined with respect to some summary statistics of model performance (see, for example, *Blazkova and Beven* (2009) and *Westerberg and McMillan* (2015)) which is one way of reducing the impact of input errors in model evaluation(*Gupta et al.*, 2008; *Vrugt and Sadegh*, 2013a; *Sadegh and Vrugt*, 2014; *Sadegh et al.*, 2015; *Vrugt*, 2016b).

The GLUE methodology has been applied widely to many different modeling problems in different fields of study where the problems of epistemic uncertainties are significant and formal statistical likelihoods functions difficult to justify when residual characteristics are non-stationary and arbitrary (*Beven*, 2015). These are the non-ideal cases that are difficult to represent using statistical residual models and that were the basis for the concept of equifinality of acceptable models *Beven* (2006). However, GLUE has also been strongly criticised for the use of subjectively chosen likelihood measures that will not provide proper probabilistic estimates of predictive uncertainty (*Mantovan and Todini*, 2006; *Stedinger et al.*, 2008; *Clark et al.*, 2011, 2012) (among others). As the referee comments on an earlier version of this paper show, there is little sign of reconciliation of these two differing viewpoints. This is for good epistemic reasons, because of the lack of theory and practice of how to best treat and deal with model structural error and epistemic uncertainty (*Beven*, 2006; *Beven et al.*, 2008; *Beven*, 2012, 2015). It can be suggested that all estimates of predictive uncertainty will be conditional on the assumptions made, and therefore care should be taken in interpreting the resulting prediction estimates, for example using the condition tree proposal of *Beven et al.* (2014b).

The GLUE approach has mostly used simple randomized sampling of the prior parameter space to create an ensemble of N different parameter combinations for evaluation. This Monte Carlo simulation approach is not particularly efficient and in high parameter spaces (large d)

may only provide a sparse sample of the behavioral solution space even after many millions of simulations (*Iorgulescu et al.*, 2005; *Blasone et al.*, 2008; *Vrugt et al.*, 2009), depending on the degree of equifinality in the model space. Uniform random sampling over the hypercube defined by the parameter axes will not only be very inefficient, it can also provide misleading results where the behavioural parameter space is highly localised. While each behavioural sample is likelihood weighted in representing the posterior distribution in GLUE, the number of samples that fall within the behavioural space will be small. *Blasone et al.* (2008) have demonstrated how the efficiency of GLUE can be enhanced in such cases, sometimes dramatically, by the use of Markov chain Monte Carlo (MCMC) simulation (though again see *Beven and Binley* (1992) for a use of MCMC-like sampling strategy in the original GLUE paper). This paper has received a significant number of citations but the proposed MCMC sampling framework has found little use in the GLUE community, despite the free availability of the source code. In this paper we revisit the use of MCMC simulation for approximate Bayesian inference but consider instead the extended GLUE approach involving the limits of acceptability framework. A slight modification of the DREAM$_{(\mathrm{ABC})}$ algorithm of *Sadegh and Vrugt* (2014) developed in the context of diagnostic model evaluation is ideally suited to solve set-theoretic membership functions such as those used in the limits of acceptability methodology.

3 Limits of Acceptability

In the manifesto for the equifinality thesis, *Beven* (2006) suggested that a more rigorous approach to model evaluation would involve the use of limits of acceptability for each individual observation against which model simulated values are compared. Behavioural models are defined as those that satisfy the limits of acceptability. The limits of acceptability should reflect the observational error of the variable being compared, together with the effects of input error and commensurability errors resulting from time or space scale differences between observed and predicted values (*Beven and Binley*, 2014a). The limits of acceptability approach applied to both individual observations and summary output statistics has been used by various authors (*Blazkova and Beven*, 2009; *Dean et al.*, 2013; *Krueger et al.*, 2009; *Liu et al.*, 2009; *McMillan et al.*, 2010; *Westerberg et al.*, 2011), although earlier publications used similar ideas within GLUE based on fuzzy measures (*Page et al.*, 2003; *Freer et al.*, 2004; *Page et al.*, 2004, 2007; *Pappenberger et al.*, 2005, 2007) and the set-theoretic model evaluation used by *Keesman* (1990) and *van Straten and Keesman* (1991). The limits of acceptability framework might be considered more objective than the standard GLUE thresholding of a goodness-of-fit measure in defining behavioural models, as the limits are expected to be defined before running the model on the basis of best available hydrologic knowledge. It remains difficult, however, to specify how epistemic input errors should affect limits of acceptability (*Beven and Smith*, 2015).

Consider first the case of a prior distribution, $P(\boldsymbol{\theta}) \sim \mathcal{U}_d(\mathbf{a}, \mathbf{b})$ that is multivariate uniform between some d-vector of values $\mathbf{a}$ and $\mathbf{b}$. For a proposal, $\boldsymbol{\theta}^*$ to be deemed acceptable, $\mathbf{Y}(\boldsymbol{\theta}^*)$ should be contained exclusively within the interval $[\tilde{y}_t - \epsilon_t, \tilde{y}_t + \epsilon_t]$ at each time $t = \{1, \ldots, n\}$. This so called "behavioral simulation space" belongs to the set $\hat{\Omega}_{(\mathbf{Y})}$ and can be defined as (*Keesman*, 1990)

$$\hat{\Omega}_{(\mathbf{Y})} = \left\{ \mathbf{Y} \in \mathbb{R}^n \ : \ y_t = \mathcal{F}(\boldsymbol{\theta}, \tilde{\mathbf{x}}_0, \tilde{\mathbf{U}}) \ ; \ \boldsymbol{\theta} \in \hat{\Omega}_{(\boldsymbol{\theta}|\tilde{\mathbf{Y}})} \ , \ t = 1, \ldots, n \right\}, \tag{1}$$

where $\hat{\Omega}_{(\boldsymbol{\theta}|\tilde{\mathbf{Y}})}$ constitutes the posterior (behavioral) parameter set

$$\hat{\Omega}_{(\boldsymbol{\theta}|\tilde{\mathbf{Y}})} = \Omega_{(\boldsymbol{\theta}|\tilde{\mathbf{Y}})}. \tag{2}$$

The conditional parameter set, $\Omega_{(\boldsymbol{\theta}|\tilde{\mathbf{Y}})}$ is defined as follows

$$\Omega_{(\boldsymbol{\theta}|\tilde{\mathbf{Y}})} = \left\{ \boldsymbol{\theta} \in \boldsymbol{\Theta} \in \mathbb{R}^d \ : \ \tilde{y}_t - \mathcal{F}(\boldsymbol{\theta}, \tilde{\mathbf{x}}_0, \tilde{\mathbf{U}}) = e_t \ ; \ e_t \in [-\epsilon_t, \epsilon_t] \ , \ t = 1, \ldots, n \right\}, \tag{3}$$

and contains solutions that satisfy the limits of acceptability of each observation, and $\boldsymbol{\theta}^* \in \hat{\Omega}_{(\boldsymbol{\theta}|\mathbf{Y})}$. If an informative prior distribution is used then the behavioral (posterior) parameter set is computed as the intersection of the prior parameter set, $\Omega_{(\boldsymbol{\theta})}$ and conditional parameter set

$$\hat{\Omega}_{(\boldsymbol{\theta}|\tilde{\mathbf{Y}})} = \Omega_{(\boldsymbol{\theta})} \cap \Omega_{(\boldsymbol{\theta}|\tilde{\mathbf{Y}})}. \tag{4}$$

Figure 1 summarizes graphically four different outcomes of the limits of acceptability framework. The behavioral solution space exists, if and only if, the conditional parameter set, $\Omega_{(\boldsymbol{\theta}|\tilde{\mathbf{Y}})}$ intersects the prior parameter set, $\Omega_{(\boldsymbol{\theta})}$. If an informative prior distribution is used, then a sufficient condition for the posterior (behavioral) parameter set to exist is that the conditional parameter set, $\Omega_{(\boldsymbol{\theta}|\tilde{\mathbf{Y}})}$ is non-empty.

4 Approximate Bayesian Computation

The limits of acceptability approach has many elements in common with likelihood-free inference (*Sadegh and Vrugt*, 2013). This approach was introduced in the statistical literature about three decades ago (*Diggle and Gratton*, 1984) (actually in different departments in the same University where, independently, the first GLUE experiments were being carried out). It is especially useful in situations where evaluation of the likelihood is computationally prohibitive, or for cases when an explicit likelihood (objective) function cannot be formulated. This class of methods is also referred to as approximate Bayesian computation (ABC) and is currently a "hot" topic in statistics (*Marjoram et al.*, 2003; *Sisson et al.*, 2007; *Joyce and Marjoram*, 2008; *Grelaud et et al.*, 2009; *Ratmann et al.*, 2009; *Del Moral et al.*, 2012).

A schematic overview of the ABC method appears in figure 2 using as example the fitting of a hydrograph. The premise behind ABC is that $\boldsymbol{\theta}^*$ should be a sample from the posterior distribution as long as a distance measure between the observed and simulated data, hereafter referred to as $\rho\big(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta}^*)\big)$ is less than some nominal positive value, ϵ (*Marjoram et al.*, 2003; *Sisson et al.*, 2007). Thus, ABC methods bypass the evaluation of the likelihood function and retain the proposal, $\boldsymbol{\theta}^*$ if

$$\rho\big(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta}^*)\big) \leq \epsilon, \tag{5}$$

where the distance function $\rho(a, b) = |a - b|$ and $|\cdot|$ signifies the modulus (absolute value) operator. The ABC approach converges to the true posterior distribution, $P(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$ when $\epsilon \to 0$ (*Pritchard et al.*, 1999; *Beaumont et al.*, 2002; *Ratmann et al.*, 2009; *Turner and van Zandt*, 2012).

All ABC based methods approximate the likelihood function by simulations, the outcomes of which are compared with the observed data (*Beaumont*, 2010; *Bertorelle et al.*, 2010; *Csilléry et al.*, 2010). In so doing, ABC algorithms attempt to approximate the posterior distribution

by sampling from

$$P(\boldsymbol{\theta}|\tilde{\mathbf{Y}}) \propto \int_{\boldsymbol{\mathcal{Y}}} P(\boldsymbol{\theta}) I(\rho\big(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta})\big) \leq \epsilon) d\boldsymbol{y}, \tag{6}$$

where $\boldsymbol{\mathcal{Y}}$ denotes the support of the simulated data, $\mathbf{Y}(\boldsymbol{\theta})$ is a stochastic model output, and $I(a)$ is an indicator function that returns one if the condition a is satisfied and zero otherwise. The accuracy of the estimated posterior distribution, $P\left(\boldsymbol{\theta}|\rho\big(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta})\big) \leq \epsilon\right)$ depends on the value of ϵ. In the limit of $\epsilon \to 0$ the sampled distribution will converge to the true posterior, $P(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$ (*Pritchard et al.*, 1999; *Beaumont et al.*, 2002; *Ratmann et al.*, 2009; *Turner and van Zandt*, 2012). Yet, this requires the underlying model operator to be stochastic, and hence $\mathbf{Y}(\boldsymbol{\theta})$ must be the output of the deterministic model, plus a n-vector drawn randomly from $P(\mathbf{e})$, a user-defined distribution with probabilistic properties equal to the series of model residuals.

For sufficiently complex models and large data sets the probability of happening upon a simulation run that yields precisely the same simulation as the calibration data set will be very small, often unacceptably so. To resolve this problem, it is often convenient to define $\rho\big(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta}^*)\big)$ as a distance between one or more (sufficient) summary statistics, $S(\mathbf{Y}(\boldsymbol{\theta}^*))$ and $S(\tilde{\mathbf{Y}})$ of the simulated and observed data, respectively. If the distance between the summary statistics, $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big)$ is smaller than ϵ the sample is retained (*Vrugt and Sadegh*, 2013a; *Sadegh and Vrugt*, 2014; *Sadegh et al.*, 2015).

In a previous paper, *Sadegh and Vrugt* (2013) have shown that there is an equivalence between the limits of acceptability framework of *Beven* (2006) and ABC if each observation of the calibration data set is used as a summary metric. This proposition is perhaps more obvious if the following notation is used

$$\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big) = \prod_{t=1}^{n} I(|\tilde{y}_t - y_t(\boldsymbol{\theta})| \leq \epsilon_t), \tag{7}$$

where ϵ_t constitutes the limits of acceptability of the tth observation.

5 The DREAM$_{(\mathrm{ABC})}$ Algorithm

Application of likelihood-free inference with ABC requires the availability of a sampling method that can efficiently search the parameter space in pursuit of the set of behavioral model realisations, $\hat{\Omega}_{(\boldsymbol{\theta}|\tilde{\mathbf{Y}})}$ that satisfies $\rho\big(a, b\big) = 1$ in Equation (7). Commonly used (population Monte Carlo) rejection sampling methods are rather inefficient in locating behavioral solutions. The chance that a random sample from the prior distribution satisfies the limits of acceptability of each observation is disturbingly small, particularly if the prior parameter space is large compared to the posterior (behavioral) solution space and the number of observations, n is large. Fortunately, an efficient MCMC sampling method, the DREAM$_{(\mathrm{ABC})}$ algorithm, has been developed by *Sadegh and Vrugt* (2014) to explore efficiently set-theoretic functions such as Equation (3).

In DREAM$_{(\mathrm{ABC})}$, K ($K > 2$) different Markov chains are run simultaneously in parallel, and multivariate proposals are generated on the fly from the collection of chain states, $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_{i-1}^1, \ldots, \boldsymbol{\theta}_{i-1}^K\}$ (matrix of $K \times d$ with each chain state as row vector) using differential evolution (*Storn and Price*, 1997; *Price et al.*, 2005). If A is a subset of d^*-dimensional space of the original parameter space, $\mathbb{R}^{d^*} \subseteq \mathbb{R}^d$ then a jump in the kth chain, $k = \{1, \ldots, K\}$ at iteration

$i = \{2, \ldots, T\}$ is calculated using

$$\begin{aligned} \mathbf{d}\boldsymbol{\Theta}^k_{\mathrm{A}} &= \boldsymbol{\zeta}_{d^*} + (\mathbf{1}_{d^*} + \boldsymbol{\lambda}_{d^*})\gamma_{(\delta,d^*)} \sum_{j=1}^{\delta} \left(\boldsymbol{\Theta}^{\mathbf{a}_j}_{\mathrm{A}} - \boldsymbol{\Theta}^{\mathbf{b}_j}_{\mathrm{A}}\right) \\ \mathbf{d}\boldsymbol{\Theta}^k_{\neq \mathrm{A}} &= 0, \end{aligned} \tag{8}$$

where $\gamma_{(\delta,d^*)} = 2.38/\sqrt{2\delta d^*}$ is the jump rate, δ denotes the number of chain pairs used to generate the jump, and $\mathbf{a}$ and $\mathbf{b}$ are δ-vectors with integer values drawn without replacement from $\{1, \ldots, k-1, k+1, \ldots, K\}$. The values of $\boldsymbol{\lambda}_{d^*}$ and $\boldsymbol{\zeta}_{d^*}$ are sampled independently from the multivariate uniform distribution, $\mathcal{U}_{d^*}(-c, c)$ and multivariate normal distribution, $\mathcal{N}_{d^*}(0, c^*)$ with, typically, $c = 0.1$ and c^* small compared to the width of the target distribution, $c^* = 10^{-12}$ say. At each 5th generation the value of λ is set to unity to enable direct jumps from one mode of the target distribution to another.

The candidate point of chain k at iteration i then becomes

$$\boldsymbol{\Theta}^k_p = \boldsymbol{\Theta}^k + \mathbf{d}\boldsymbol{\Theta}^k, \tag{9}$$

and a modified selection rule is used to determine whether to accept this proposal or not. This selection rule is defined as

$$P_{\mathrm{acc}}(\boldsymbol{\Theta}^k \rightarrow \boldsymbol{\Theta}^k_p) = \begin{cases} I\big(f(\boldsymbol{\Theta}^k_p) \geq f(\boldsymbol{\Theta}^k)\big) & \text{if } f(\boldsymbol{\Theta}^k_p) < n \\ 1 & \text{if } f(\boldsymbol{\Theta}^k_p) = n \end{cases}, \tag{10}$$

where the fitness function, $f(\cdot)$ is calculated as follows

$$f(\boldsymbol{\theta}) = \sum_{t=1}^{n} I(|\tilde{y}_t - y_t(\boldsymbol{\theta})| \leq \epsilon_t). \tag{11}$$

If the proposal is accepted, then the kth chain moves to this new position, $\boldsymbol{\theta}^k_i = \boldsymbol{\Theta}^k_p$, otherwise it remains at its current location, that is $\boldsymbol{\theta}^k_i = \boldsymbol{\theta}^k_{i-1}$.

The fitness of the proposal $\boldsymbol{\theta}^*$ is equivalent to the number of observations the simulation of $\boldsymbol{\theta}^*$ satisfies within the limits of acceptability. The proposal, $P_{\mathrm{acc}}(\boldsymbol{\Theta}^k \rightarrow \boldsymbol{\Theta}^k_p) = 1$, is accepted if the fitness of $\boldsymbol{\Theta}^k_p$ is larger than that of the current state of the kth chain, $\boldsymbol{\Theta}^k$ or if the simulation of the proposal is consistently within $\boldsymbol{\epsilon} = \{\epsilon_1, \ldots, \epsilon_n\}$ of the observed values, and thus $f(\boldsymbol{\Theta}^k_p) = n$, otherwise the candidate point is rejected. After a burn-in period in which $f(\cdot) < n$, the convergence of DREAM$_{(\mathrm{ABC})}$ can be monitored with the $\hat{R}$ diagnostic of *Gelman and Rubin* (1992). A full description of DREAM$_{(\mathrm{ABC})}$ appears in *Sadegh and Vrugt* (2014) and interested readers are referred to this publication for further details.

A basic code for the DREAM$_{(\mathrm{ABC})}$ algorithm is given in Appendix A of this paper. The results presented herein are derived from the MATLAB toolbox of DREAM, which provides a much wider arsenal of options and capabilities (such as parallel computing). A detailed description of this toolbox appears in *Vrugt* (2016a) and interested readers are referred to this publication for further information.

6 Numerical Examples

Three different numerical examples are considered to illustrate the ability of the DREAM$_{(\text{ABC})}$ algorithm to sample efficiently the behavioral parameter, $\hat{\Omega}_{(\boldsymbol{\theta}|\tilde{\mathbf{Y}})}$ and simulation, $\hat{\Omega}_{(\mathbf{Y})}$ space that satisfy the prior parameter distribution and limits of acceptability of each observation. All the examples assume a noninformative and indepdendent prior distributions, and default values of the algorithmic parameters of DREAM$_{(\text{ABC})}$.

6.1 Unit Hydrograph

The first case study considers the modeling of the instantaneous unit hydrograph using the ordinates of *Nash* (1960) defined as

$$Q_t = \frac{1}{L\Gamma(g)} \left(\frac{t}{L}\right)^{(g-1)} \exp\left(-\frac{t}{L}\right), \tag{12}$$

where Q_t (mm day^{-1}) is the simulated streamflow at time t (days), g (-) denotes the number of reservoirs, L (days) signifies the recession constant, and $\Gamma(\cdot)$ is the gamma function

$$\Gamma(g) = \int_0^\infty x^{g-1} \exp(-x) \mathrm{d}x \qquad \forall g \in \mathbb{R} \tag{13}$$

which satisfies the recursion $\Gamma(g+1) = g\Gamma(g)$.

A $n = 25$ - day period with synthetic daily streamflow data was generated by driving Equation (12) with an artificial precipitation record using $g = 2$ reservoirs, and a recession constant of $L = 4$ days. This artificial data set is subsequently perturbed with a heteroscedastic measurement error (non-constant variance) with standard deviation equal to 10% of the original simulated discharge values. In this case input data and model structure are assumed to be known accurately. The DREAM$_{(\text{ABC})}$ algorithm then uses the observed discharge record, $\tilde{\mathbf{Y}} = \{\tilde{y}_1, \ldots, \tilde{y}_{25}\}$ to estimate the behavioral solution space of g and L using the limits of acceptability, $\epsilon_t = 0.2\tilde{y}_t \; \forall t \in \{1, ..., 25\}$. A bivariate uniform prior distribution, $\mathcal{U}_2[1, 10]$ was used for g and L in the calculations.

Figure 3 summarizes the results of the analysis. The graph at the left-hand-side presents a time series plot of the observed (red dots) and simulated discharge data (gray). These simulated data satisfy the limits of acceptability of each observation and thus belong to the behavioral set, $\hat{\Omega}_{(\mathbf{Y})}$. The two figures at the right-hand-side plot histograms of the behavioral parameter space of g and L respectively. The true parameter values used to generate the synthetic data are separately indicated with the red 'X' symbol. The behavioral simulation space satisfies the limits of acceptability of the entire hydrograph, but fails to bracket the discharge measurements on days 6, 9 and 13. This is not unexpected given that the limits of acceptability were defined a priori to give 95% coverage of the known stochastic variation. The posterior histograms center around their "true" values but appears a little biased (to the left) for parameter g.

To provide insights into the convergence rate of DREAM$_{(\text{ABC})}$ to the posterior set, $\hat{\Omega}_{(\boldsymbol{\theta}|\tilde{\mathbf{Y}})}$, figure 4 plots trace plots of the $\hat{R}$-convergence diagnostic of *Gelman and Rubin* (1992) computed using the samples in the second half of the $K = 8$ different Markov chains. About $2,000$ function evaluations are required to satisfy the convergence threshold of $\hat{R} \leq 1.2$. The acceptance rate of proposals is equivalent to about 33% (not shown herein), which means that, on average, every

third proposal of DREAM$_{(\mathrm{ABC})}$ satisfies the limits of acceptability. This acceptance rate would be orders of magnitude lower if uniform random sampling were used, particularly since there is a nearly linear correlation of -0.93 between the posterior parameter samples of L and g (see figure 3). This conjecture is confirmed by numerical simulation. Only 28 samples (indicated with blue dots) were deemed behavioral out of 20,000 draws from the prior distribution. The resulting acceptance rate of approximately 0.14% is more than two orders of magnitude lower than its counterpart derived from MCMC simulation with DREAM$_{(\mathrm{ABC})}$. This difference in sampling efficiency between DREAM$_{(\mathrm{ABC})}$ and uniform random (rejection) sampling is clearly evident in figure 5. Not only does DREAM$_{(\mathrm{ABC})}$ produce many diverse samples of $\hat{\Omega}_{(\boldsymbol{\theta}|\tilde{\mathbf{Y}})}$, the posterior parameter set, the algorithm also sharply delineates the behavioral solution space.

In this trivial model case it is quite possible to do many millions of uniform random samples to compensate for this lack of sampling efficiency but the prospects for much higher dimensional search spaces with much more complex parameter interactions are not very encouraging. The use of a proper sampling method is of crucial importance for correct GLUE inference and the DREAM$_{(\mathrm{ABC})}$ can help to avoid the rejection of models as a result of sparse sampling in higher dimensional parameter spaces with more complex parameter interactions. Past work has also shown how the $_{(\mathrm{ABC})}$ methodology can be successful in identifying multiple regions of behavioral models (*Sadegh et al.*, 2015).

6.2 Rainfall-Runoff Modeling

The second case study involves the modeling of the rainfall-runoff transformation of the Leaf River watershed in Mississippi. This temperate 1944 km^2 watershed has been studied extensively in the hydrologic literature which simplifies comparative analysis of the results. A 10-year historical record (1/10/1952 - 30/9/1962) with daily data of discharge (mm day^{-1}), mean areal precipitation (mm day^{-1}), and mean areal potential evapotranspiration (mm day^{-1}) is used herein for model calibration and evaluation. A 65-day spin-up period is used to reduce sensitivity of the model to state-value initialization.

The rainfall-discharge relationship of the Leaf River basin is simulated using the Sacramento soil moisture accounting (SAC-SMA) model of *Burnash et al.* (1973). This lumped conceptual watershed model is used by the National Weather Service for flood forecasting through the United States. The SAC-SMA model uses six reservoirs (state variables) to represent the rainfall-runoff transformation. These reservoirs represent the upper and lower part of the soil and are filled with "tension" and "free" water, respectively. The upper zone simulates processes such as direct runoff, surface runoff, and interflow, whereas the lower zone is used to mimic groundwater storage and the baseflow component of the hydrograph.

Figure 6 provides a schematic overview of the SAC-SMA model. Nonlinear equations are used to relate the absolute and relative storages of water within each reservoir and their states control the main watershed hydrologic processes such as the partitioning of precipitation into overland flow, surface runoff, direct runoff, infiltration to the upper zone, interflow, percolation to the lower zone, and primary and supplemental baseflow. Saturation excess overland flow occurs when the upper zone is fully saturated and the rainfall rate exceeds interflow and percolation capacities. Percolation from the upper to the lower layer is controlled by a nonlinear process that depends on the storage in both soil zones.

The SAC-SMA model has thirteen user-specifiable (and three fixed) parameters and an

evapotranspiration demand curve (or adjustment curve). Inputs to the model include mean areal precipitation (MAP) and potential evapotranspiration (PET) while the outputs are estimated evapotranspiration and channel inflow. A Nash-Cascade series of three linear reservoirs is used to route the upper zone channel inflow while the baseflow generated by the lower zone recession is passed directly to the gauging point. This configuration adds one parameter and three state variables to the SAC-SMA model. The use of the three reservoirs improves considerably the CPU-efficiency as it avoids the need for computationally expensive convolution (though see the data-based modeling of *Young* (2013) that suggests a longer routing kernel might be appropriate for the Leaf River data set). Our formulation of the model therefore has fourteen time-invariant parameters which are subject to inference using observed discharge data. Table 1 summarizes the fourteen SAC-SMA parameters and six main state variables, and their ranges.

In this case study there is no information about the uncertainties associated with either the forcing rainfall data of each discharge observation. To define the limits of acceptability we follow the approach of *Sadegh and Vrugt* (2013) and use a multiple of an estimated discharge measurement error, hereafter referred to as $\hat{\boldsymbol{\sigma}}_{\tilde{\mathbf{Y}}} = \{\hat{\sigma}_{\tilde{y}_1}, \ldots, \hat{\sigma}_{\tilde{y}_n}\}$. This was estimated by *Vrugt et al.* (2005) using a nonparametric estimator to be of the order of $0.1\tilde{y}_t$. The limits of acceptability in Equation (7) are now computed as multiple of $\hat{\sigma}_{\tilde{\mathbf{Y}}}$ or $\boldsymbol{\epsilon} = \phi\hat{\sigma}_{\tilde{\mathbf{Y}}}$ using $\phi = 4$. This leads to effective observation errors on the order of $\epsilon_t = 0.4\tilde{y}_t$.

Figure 7 plots traces of the sampled fitness values in a selected set of ten Markov chains simulated with DREAM$_{(\mathrm{ABC})}$. The different chains are coded with a different color and/or symbol. The chains converge to a stable fitness value of Equation (11) of around 2,800 after about 80,000 function evaluations. That is about 76% of the discharge observations are fitted within their limits of acceptability. In the philosophy of GLUE the SAC-SMA model should be rejected as it does not satisfy all the prior estimates of the limits of acceptability, even though the model describes accurately a significant portion of the discharge data (see figure 8).

To benchmark the results of the DREAM$_{(\mathrm{ABC})}$ algorithm, a total of $100,000$ samples were drawn randomly from the ranges listed in Table 1. The maximum value of the fitness of this sample is equivalent to $2,401$, much lower than its counterpart of $2,800$ derived from the DREAM$_{(\mathrm{ABC})}$ algorithm. This gives further weight to the argument that adequate sampling is essential to inference using a GLUE limits of acceptability approach but does not alter the conclusion that the SAC-SMA model should be rejected based on these limits.

Further detailed inspection of the complete time series demonstrates that the SAC-SMA model fits most of the recession periods adequately well and that the limits are being exceeded predominantly during a substantial number of storm events. The misfit during these events cannot be contributed solely to model structural error but suggests that there are important epistemic errors associated with the rainfall inputs such that some events may be disinformative for model evaluation (see *Beven and Smith* (2015)). Such errors not only propagate nonlinearly through the SAC-SMA model but also accumulate in the resolved state-variables, hence their impact might be seen in consequent events. What is more, rainfall data errors exhibit non-stationarity. These effects (nonlinearity, non-stationarity and memory) are difficult to encapsulate in limits of acceptability unless detailed prior knowledge is available about the error characteristics of individual storm events. For instance, consider the model-data mismatch observed between days 2,180 - 2,200 and days 2,350 - 2,375 of the calibration data record. This discrepancy is likely due to errors in the precipitation data (too much and too little recorded rainfall, respectively). No conceptual watershed model will be able to describe these events us-

ing reasonable limits of acceptability. Instead what is needed is a careful analysis of the errors of each individual storm event. In addition, such errors can have an important effect in prediction since it is not known a priori whether the next prediction event has well estimated forcing data or not (as demonstrated in *Beven* (2015), for example).

This also demonstrates, however, why it is important that the limits of acceptability should be set prior to running the model. Otherwise it would be rather too easy to exclude those events for which the model does not satisfy those limits as subject to epistemic input errors. In that case no model would be rejected. As *Beven* (2012) points out, the science will not progress if we are not prepared to reject models and explore the reasons for such failures. In this case it could be either a failure of the model structure, or of epistemic uncertainty in the forcing data. It poses the question as to just how good do we expect our models to be, in both calibration and prediction, when we suspect that there are non-stationary input errors. An advantage of the use of summary statistics within the GLUE or DREAM$_{(\mathrm{ABC})}$ framework is that the summary statistics are not so readily affected by outliers as the residuals associated with individual observations. Indeed, *Sadegh et al.* (2015) show how such metrics can help to diagnose and detect catchment non-stationarity. The equivalent disadvantage is that summary statistics may conceal some of the prediction problems revealed in this case study with the possibility of making both Type I and Type II errors in testing models as hypotheses.

6.3 Vadose Zone Modeling

The third and last case study considers the modeling of the soil moisture regime of an agricultural field near Jülich, Germany. Soil moisture content was measured with Time Domain Reflectometry (TDR) probes at 6 cm deep at 61 locations in a 50 × 50 m plot. The TDR data were analysed using the algorithm described in *Heimovaara and Bouten* (1990) and the measured apparent dielectric permittivities were converted to soil moisture contents using the empirical relationship of *Topp et al.* (1980). Measurements were taken on 29 days between 19 March and 14 October 2009, comprising a measurement campaign of 210 days. For the purpose of the present study, the observed soil moisture data at the 61 locations were averaged to obtain a single time series of water content. Precipitation and other meteorological variables were recorded at a meteorological station located 100 m west of the measurement site. Details of the site, soil properties, experimental design and measurements are given by *Scharnagl et al.* (2011) and interested readers are referred to this publication for further details.

The HYDRUS-1D model of *Šimůnek et al.* (2008) was used to simulate variably saturated water flow in the agricultural field (see figure 9). This model solves Richards' equation for given (measured) initial and boundary conditions

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial z}\left[K(h)\left(\frac{\partial h}{\partial z} + 1\right)\right] \tag{14}$$

where θ (cm^3 cm^{-3}) here denotes soil moisture content (not to be confused with parameter values!), t (days) denotes time, z (cm) is the vertical (depth) coordinate, h (cm) signifies the pressure head, and $K(h)$ (cm day^{-1}) the unsaturated soil hydraulic conductivity.

To solve Equation (14) numerically the soil hydraulic properties need to be defined. Here

the van Genuchten-Mualem (VGM) model (*van Genuchten*, 1980) was used:

$$\begin{aligned} \theta(h) &= \theta_r + (\theta_s - \theta_r)[1 + (\alpha|h|)^n]^{-m} \\ K(h) &= K_s S_e(h)^\lambda [1 - (1 - S_e(h)^{1/m})^m]^2, \end{aligned} \tag{15}$$

where θ_s and θ_r ($\mathrm{cm}^3\ \mathrm{cm}^{-3}$) signify the saturated and residual soil water content, respectively, α (cm^{-1}), n (-) and $m = 1 - 1/n$ (-) are shape parameters, K_s ($\mathrm{cm\ day}^{-1}$) denotes the saturated hydraulic conductivity, and $\lambda = 1/2$ (-) represents a pore-connectivity parameter. The effective saturation, S_e (-) is defined as

$$S_e(h) = \frac{\theta(h) - \theta_r}{\theta_s - \theta_r}. \tag{16}$$

Observations of daily precipitation and potential evapotranspiration were used to define the upper boundary condition. In the absence of direct measurements, a constant head lower boundary condition was assumed, h_{bot} (cm), whose value is subject to inference within the GLUE LOA framework using $\mathrm{DREAM}_{(\mathrm{ABC})}$. The aim here is to obtain a simulation of the mean behavior of the field soil moisture, as constrained by the observed soil moisture contents.

Table 2 lists the parameters of the HYDRUS-1D model and their prior uncertainty ranges which are subject to inference using the 210-day period of the averaged observed soil moisture measurements. In this study the limits of acceptability $\epsilon = \{\epsilon_1, \ldots, \epsilon_n\}$ are based on the observed spatial variability of the soil moisture data in the 2,500 m^2 field plot. *Scharnagl et al.* (2011) depict in their figure 8 (p. 3055), the 95% ranges of the observed soil moisture data at each measurement time. From these, the 95% confidence in the mean soil moisture content could also be derived, but given the nonlinearity inherent in the soil water flux process and the expected heterogeneity of the boundary conditions, this would be expected to underestimate the potential uncertainty in modeling the mean field water content and soil water fluxes. Thus, for the purpose of this sampling case study, the limits of acceptability have been set to half the width of the 95% interval of the distributed moisture content observations. This equates to an average value of the limits of acceptability of 0.047 ($\mathrm{cm}^3\mathrm{cm}^{-3}$). To speed-up posterior exploration, the $N = 8$ different chains are ran on different processors using the MATLAB parallel computing toolbox.

Figure 10 presents histograms of the marginal posterior distribution of the six HYDRUS-1D model parameters considered in this study. The bottom panel presents a time series plot of the behavioral simulation set, $\hat{\Omega}_{(\mathbf{Y})}$. The observed soil moisture data are indicated separately with red dots. The behavioral HYDRUS-1D model nicely tracks the observed average soil moisture measurements within behavioral simulation space defined in this way. The root mean square error (RMSE) of the behavioral (posterior) mean simulation equates to about 0.0149 $\mathrm{cm}^3/\mathrm{cm}^{-3}$, a value somewhat larger than derived separately using Bayesian inference with a Gaussian likelihood function (*Vrugt*, 2016a). The behavioral parameter space of most parameters extend across a large part of their respective prior ranges with marginal distributions that deviate markedly from normality. The prior ranges are rather narrow and derived from Monte Carlo simulation with the ROSETTA pedotransfer toolbox using textural data (percentages of sand, silt, and clay) as main input variables. Pedotransfer functions are, however, derived from small volume sample measurements and may not always be appropriate in simulating field scale behavior (*Beven and Germann*, 2013).

The acceptance rate of $\mathrm{DREAM}_{(\mathrm{ABC})}$ averages about 15.1%. Thus every sixth proposal generated with $\mathrm{DREAM}_{(\mathrm{ABC})}$ satisfies the limits of acceptability of the soil moisture observations.

This efficiency is considerably higher than derived from rejection sampling. Out of $10,000$ samples drawn from the prior distribution in Table 2 only 47 were deemed behavioral. This equates to an acceptance rate of approximately 0.47%. This efficiency, is about 35 times lower than DREAM$_{(\mathrm{ABC})}$, and expected to deteriorate further with increasing dimensionality and size of the parameter space.

To provide further insights into the convergence speed of DREAM$_{(\mathrm{ABC})}$, Figure 11 plots the evolution of the $\hat{R}$-diagnostic for the six HYDRUS-1D model parameters in the top panel and traces of the sampled fitness values of the $K = 8$ different chains simulated with DREAM$_{(\mathrm{ABC})}$ in the bottom panel. The $\hat{R}$-diagnostic of *Gelman and Rubin* (1992) satisfies the convergence threshold (black line) after about $4,800$ function evaluations. This means that the last 50% of the chains, between function evaluations $2,400$ - $4,800$ and their corresponding sample numbers 300 - 600 satisfy convergence. This conclusion is confirmed in the bottom panel which demonstrates that about 300 samples are needed in each chain to satisfy the limits of acceptability of each observation (fitness score of 29). The subsequent 300 samples are used for the chains to explore fully the behavioral parameter space. It is interesting to observe that the two diagnostics, albeit quite different proxies for convergence, provide remarkably similar results.

One should however be particularly careful to judge convergence based on the $\hat{R}$-statistic. This convergence diagnostic is only meaningful if all the chains satisfy reversibility. This condition is however not satisfied in the present case with the use of the acceptance probability in Equation (10). This selection rule of proposals directs the DREAM$_{(\mathrm{ABC})}$ algorithm to the posterior parameter set, $\hat{\Omega}_{(\boldsymbol{\theta}|\tilde{\mathbf{Y}})}$ but violates detailed balance to do so in the first part of the chain until the target distribution is reached. Of course, in the case that the behavioral solution set is empty and the model is rejected (as with the SAC-SMA model in previous study), the DREAM$_{(\mathrm{ABC})}$ algorithm cannot converge formally.

Finally, Figure 12 shows how the posterior parameter set translates into uncertainty of the soil water retention (left) and unsaturated soil hydraulic conductivity (right) functions. The light grey region corresponds to the range of the prior parameter set whereas the dark grey is used to denote the behavioral (posterior) solution set. The posterior mean soil hydraulic functions are indicated with the solid black line. The posterior uncertainty of the soil hydraulic functions appears rather large in response to the limited constraints provided by a single depth of measurement, with uncertain upper and lower boundary conditions (see also *Binley and Beven* (2003))

7 Summary and Conclusions

In the manifesto for the equifinality thesis, *Beven* (2006) suggested that a more rigorous approach to hydrologic model evaluation would involve the use of limits of acceptability for each individual observation against which model simulated values are compared. Within this framework, behavioural models are defined as those that satisfy the limits of acceptability for each observation. Ideally, the limits of acceptability should reflect the observational error of the variable being compared, together with the effects of input error and commensurability errors resulting from time or space scale differences between observed and predicted values (*Beven et al.*, 2014b). In the GLUE: 20 years on paper, *Beven and Binley* (2014a) argue that the limits of acceptability framework might be considered more objective than the standard GLUE approach advocated in *Beven and Binley* (1992) as the limits are defined before running the model on

the basis of best available hydrologic knowledge.

This then raises the issue of how to identify efficiently the behavioural parameter sets that satisfy the limits of acceptability. In most GLUE applications, random sampling from the prior distribution has been used to delineate the behavioral parameter space. This method, known as rejection sampling when combined with a membership-set likelihood function, is not particularly efficient and if applied with an inadequate sampling density might result in a misrepresentation of the posterior parameter distribution. It is also possible that when no behavioral simulations are found because of inadequate sampling, models might be wrongly rejected. Thus inadequate sampling can increase the possibility of Type II errors of rejecting a model that would be useful in prediction. In this paper the reversible chain MCMC simulation with the DREAM$_{(\mathrm{ABC})}$ algorithm has been used to enhance, sometimes dramatically, the accuracy and efficiency of limits of acceptability sampling.

Three different case studies have been used to demonstrate the usefulness and practical application of MCMC simulation with DREAM$_{(\mathrm{ABC})}$ within the GLUE limits of acceptability framework. The most important results are as follows

(1) The DREAM$_{(\mathrm{ABC})}$ algorithm achieves equivalent results to the limits of acceptability approach of GLUE if all observations are used as summary statistics and the values of ϵ are set equal to the effective observation error.

(2) Reversible MCMC simulation with DREAM$_{(\mathrm{ABC})}$ is orders of magnitude more efficient than rejection sampling used within the GLUE limits of acceptability framework.

(3) The DREAM$_{(\mathrm{ABC})}$ algorithm provides a diverse and dense sample of the behavioral parameter set.

(4) The DREAM$_{(\mathrm{ABC})}$ algorithm delineates sharply the behavioral parameter space.

(5) The use of inferior sampling methods can lead to inexact inference about the behavioral parameter set and erroneous conclusions about model rejection.

We should expect that the problems with any sampling method become increasingly problematic with increasing dimensionality of the parameter space, increasing numbers of local regions of behavioural models, and increasing model run times. The only way around these issues is to use efficient sampling methods such as the DREAM$_{(\mathrm{ABC})}$ algorithm. Depending on the initial set of chains, this may still not identify all areas of behavioral models in complex model spaces, but will still be expected to identify regions of behavioral models with much greater reliability and efficiency. This should therefore lead to more reliable and robust inference based on the GLUE methodology.

8 Acknowledgements

This version of the paper reflects the useful comments of two anonymous reviewers. The DREAM toolbox used herein is available from the first author, `jasper@uci.edu` upon request.

9 Appendix A

This Appendix presents a basic implementation of the DREAM$_{(\text{ABC})}$ algorithm in MATLAB. The core of DREAM$_{(\text{ABC})}$ can be written in MATLAB in about 30 lines of code. This code can be used as starting point for users to implement MCMC simulation to sample the behavioral parameter space for limits of acceptability sampling. Notation matches for large part variable names used in main text. For convenience the variable `x` is used for $\boldsymbol{\theta}$, the parameters of the model, and `X` signifies $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_{i-1}^1, \ldots, \boldsymbol{\theta}_{i-1}^K\}$, the $K \times d$ matrix with the collection of chains at generation $i-1$.

```
function [x,p_x] = dream_ABC(prior,fitness,K,T,d)
% DiffeRential Evolution Adaptive Metropolis (DREAM) ABC algorithm

[delta,c,c_star,nCR,p_g] = deal(3,0.1,1e-12,3,0.2);  % Default values DREAM algorithmic parameters
x = nan(T,d,K); f_x = nan(T,K);                      % Preallocate memory for chains and density
X = prior(K,d);                                      % Create initial population
for k = 1:K, f_X = fitness(X(k,1:d)); end            % Compute fitness initial states of chains
x(1,1:d,1:K) = reshape(X',1,d,K); f_x(1,1:K) = f_X'; % Store initial position of chain and density
for k = 1:K, R(k,1:K-1) = setdiff(1:K,k); end        % R-matrix: ith chain, the index of chains for DE
CR = [1:nCR]/nCR; pCR = ones(1,nCR)/nCR;             % Crossover values and their selection probability

for i = 2:T,                                         % Dynamic part: Evolution of K chains
    [~,draw] = sort(rand(K-1,K));                        % Randomly permute [1,...,K-1] K times
    dX = zeros(K,d);                                     % Set K jump vectors equal to zero
    lambda = unifrnd(-c,c,K,1);                          % Draw K lambda values
    for k = 1:K,                                         % Create proposal each chain and accept/reject
        D = randsample([1:delta],1,'true');              % Select delta (equal selection probability)
        a = R(k,draw(1:D,k)); b = R(k,draw(D+1:2*D,k));  % Unpack vectors a and b; "a" n.e. "b" n.e. "k"
        cr = randsample(CR,1,'true',pCR);                % Select crossover value
        A = find(rand(1,d) < cr);                        % Derive subset A with dimensions to sample
        d_star = numel(A);                               % How many dimensions are sampled?
        gamma_d = 2.38/sqrt(2*D*d_star);                 % Calculate jump rate
        g = randsample([gamma_d 1],1,'true',[1-p_g p_g]);% Select gamma: 80/20 mix (default)
        dX(k,A) = (1+lambda(k))*g*sum(X(a,A)-...
            X(b,A),1) + c_star*randn(1,d_star);          % Compute kth jump with differential evolution
        Xp(k,1:d) = X(k,1:d) + dX(k,1:d);                % Compute kth proposal
        f_Xp(k,1) = fitness(Xp(k,1:d));                  % Calculate fitness of ith proposal
        p_acc = f_Xp(k,1) >= f_X(k,1);                   % Compute selection rule
        if p_acc,
            X(k,1:d) = Xp(k,1:d); f_X(k,1) = f_Xp(k,1);  % True: Accept proposal
        end
    end
    x(t,1:d,1:K) = reshape(X',1,d,K); f_x(t,1:K) = f_X'; % Add current position and density to chain
    [X,f_X,f_x(1:t,1:K)] = outlier(X,f_X,f_x(1:t,1:K));  % Outlier detection and correction
end
```

Built-in functions are highlighted with a low dash. The `persistent` declaration helps retain variables `Y_obs` and `epsilon` in local memory after the first function call has been completed. This is computationally appealing, as it avoids having to reload these variables each time a proposal is being evaluated.

The `dream_ABC` function has five input arguments, including `prior` an anonymous function of the prior distribution, `f` a handle of the fitness function of Equation (11), `K` the number of chains, `T` the number of generations, and `d` the dimensionality of the target distribution. Based on these input arguments the code creates `K` different Markov chains, `x` with their corresponding fitness values, `f_x`, which measures the number of observations whose limits of acceptability are satisfied. `randsample` draws with replacement (true) the value of the jump rate, `gamma` from the vector `[gamma_RWM 1]` using selection probabilities `[0.8 0.2]`. `ones()` returns a unit vector of size `nCR`, and `randn()` draws `d_star` values from a standard normal distribution. `deal()` assigns default values to the algorithmic variables of DREAM. `sum()` computes the sum of the columns `A` of the chain pairs `a` and `b`. The function `outlier()` detects for potential outlier chains and corrects their states. The jump vector, `dx(k,1:d)` of the `k`th chain contains the desired information about the scale and orientation of the proposal distribution and is derived from the remaining `K-1` chains. Please refer to introductory textbooks and/or the MATLAB "help"

utility for the remaining functions `nan()`, `reshape()`, `setdiff()`, `sort()`, `zeros`), `find()`, `numel()`, `sqrt()`, and `ceil()`. Note, the basic code of DREAM$_{(\text{ABC})}$ given above does not monitor convergence

`prior()` is an anonymous function that draws `K` samples from a `d`-variate prior distribution, for example

```
prior = @(K,d) unifrnd(-10,10,K,d)
```
(A1)

where the @ operator creates the function handle. This anonymous function accepts as input the value of K and d and returns as output, a $K \times d$ of draws from the uniform distribution between -10 and 10, or $\mathtt{X} \sim \mathcal{U}_d(-10, 10)$. These values define the initial state of each of the K Markov chains. `fitness()` is another anonymous function which calculates the fitness of each proposal using Equation (11)

```
fitness = @(x) model(x).
```
(A2)

The function `model` is written by the user. A template of this function is given below.

```
function f = model(X_p);
% This function computes the fitness of the proposal x_p

persistent Y_obs epsilon            % Retain in memory after first call

if isempty(Y_obs),
    %% Load observations and corresponding limits of acceptability, epsilon
    Y_obs = load('data_file.txt'); epsilon = load('limits.txt');
end

%% Run forward model, Equation (6), for parameter values X(k,1:d)
Y_sim = own_model_script(X_p);

%% Now solve Equation (27), how many observations are within their limits?
f = sum( abs(Y_obs - Y_sim) <= epsilon );
```

The reader is referred to *Vrugt* (2016a) for a detailed introduction to the MATLAB toolbox of DREAM and related algorithms.

References

B. Ahrens, and M. Reichstein, "Reconciling ^{14}C and minirhizotron-based estimates of fine-root turnover with functions," *Journal of Plant Nutrition and Soil Science*, vol. 177, pp. 287-296, doi:10.1002/jpln.201300110, 2014.

M. Barthel, A. Hammerle, P. Sturm, T. Baur, L. Gentsch, and A. Knohl, "The diel imprint of leaf metabolism on the δ^{13}C signal of soil respiration under control and drought conditions," *New Phytologist*, vol. 192, pp. 925-938, doi: 10.1111/j.1469-8137.2011.03848.x, 2011.

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

L. Bauwens, B. de Backer, and A. Dufays, "Estimating and forecasting structural breaks in financial time series," *Economics, Finance, Operations Research, Econometrics, and Statistics*, Discussion paper, pp. 1-23, 2011.

T. Bayes, and R. Price, "An essay towards solving a problem in the doctrine of chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, A.M.F.R.S.", *Philosophical Transactions of the Royal Society of London*, vol. 53 (0), pp. 370 - 418, doi:10.1098/rstl.1763.0053.

M.A. Beaumont, W. Zhang, and D.J. Balding, "Approximate Bayesian computation in population genetics," *Genetics*, vol. 162 (4), pp. 2025-2035, 2002.

M.A. Beaumont, "Approximate Bayesian computation in evolution and ecology," *Annual Review of Ecology, Evolution, and Systematics*, vol. 41, pp. 379-406, 2010.

G. Bertorelle, A. Benazzo, and S. Mona, "ABC as a flexible framework to estimate demography over space and time: some cons, many pros," *Molecular Ecology*, vol. 19, pp. 2609-2625, 2010.

K.J. Beven, and A.M. Binley, "The future of distributed models: model calibration and uncertainty prediction," *Hydrological Processes*, vol. 6, pp. 279-98, 1992.

K.J. Beven, and J. Freer, "Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology," *Journal of Hydrology*, vol. 249 (1-4), pp. 11-29, doi:10.1016/S0022-1694(01)00421-8, 2001.

K.J. Beven, "A manifesto for the equifinality thesis," *Journal of Hydrology*, vol. 320 (1), pp. 18-36, 2006.

K.J. Beven, P.J. Smith, and J. Freer, "So just why would a modeler choose to be incoherent?," *Journal of Hydrology*, vol. 354, pp. 15-32, 2008.

K.J. Beven, "*Environmental Modelling: An Uncertain Future?*," Routledge, London.

K.J. Beven, P.J. Smith, and A. Wood, "On the color and spin of epistemic error (and what we might do about it)," *Hydrologic and Earth System Sciences*, 15, 3123-3133, doi:10.5194/hess-15-3123-2011, 2011.

K.J. Beven, "Causal models as multiple working hypotheses about environmental processes," *Comptes Rendus Geoscience*, Académie de Sciences, Paris, 344: 77-88, doi:10.1016/j.crte.2012.01.005, 2012.

K.J. Beven and P.F. Germann, "Macropores and water flow in soils revisited", *Water Resources Research*, 49 (6), pp. 3071-3092, doi:10.1002/wrcr.20156, 2013.

K.J. Beven, and A.M. Binley, "GLUE: 20 years on," *Hydrological Processes*, vol. 28, pp. 5879-5918, 2014, doi:10.1002/hyp.10082, 2014a.

K.J. Beven, D.T. Leedal, S. McCarthy, "Framework for assessing uncertainty in fluvial flood risk mapping," CIRIA report C721, available at `http://www.ciria.org/Resources/Free_publications/fluvial_flood_risk_mapping.aspx`, 2014b.

K.J. Beven, "EGU Leonardo Lecture: Facets of Hydrology - epistemic error, non-stationarity, likelihood, hypothesis testing, and communication," *Hydrologic Sciences Journal*, 10.1080/02626667.2015.1031761, 2015.

K.J. Beven, and P.J. Smith, "Concepts of information content and likelihood in parameter calibration for hydrologic simulation models," *ASCE Journal of Hydrologic Engineering*, doi: 10.1061/(ASCE)HE.1943-5584.0000991, 2015.

J. Bikowski, J.A. Huisman, J.A. Vrugt, H. Vereecken, and J. van der Kruk, "Inversion and sensitivity analysis of ground penetrating radar data with waveguide dispersion using deterministic and Markov chain Monte Carlo methods," *Near Surface Geophysics*, vol. 10 (6), pp. 641-652, doi:10.3997/1873-0604.2012041, 2012.

A. Binley, and K.J. Beven, "Vadose zone model uncertainty as conditioned on geophysical data", *Ground Water*, 41(2), 119-127, 2003.

R.S. Blasone, J.A. Vrugt, H. Madsen, D. Rosbjerg, G.A. Zyvoloski, and B.A. Robinson, "Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov chain Monte Carlo sampling," *Advances in Water Resources*, vol. 31, pp. 630-648, doi:10.1016/j.advwatres.2007.12.003, 2008.

S. Blazkova, and K.J Beven, "A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic," *Water Resources Research*, vol. 45, W00B16, doi:10.1029/2007WR006726, 2009.

K. Bogner, F. Pappenberger, and H.L. Cloke, "Technical Note: The normal quantile transformation and its application in a flood forecasting system," *Hydrology and Earth System Sciences*, vol. 16, pp. 1085-1094, 2012.

G.E.P. Box, and D.R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society*, Series B, vol. 26 (2), pp. 211-252, 1964.

M.C. Braakhekke, T. Wutzler, C. Beer, J. Kattge, M. Schrumpf, B. Ahrens, I. Schöning, M.R. Hoosbeek, B. Kruijt, P. Kabat, and M. Reichstein, "Modeling the vertical soil organic matter profile using Bayesian parameter estimation", *Biogeosciences*, vol. 10, pp. 399-420, doi:10.5194/bg-10-399-2013, 2013.

R.J.C. Burnash, R.L. Ferral, and R.A. McGuire, "A generalized streamflow simulation system: conceptual models for digital computers," Joint Federal-State River Forecast Center, Sacramento, CA, 1973.

W. Chu, T. Yang, and X. Gao, "Comment on 'High-dimensional posterior exploration of hydrologic models using multiple-try DREAM$_{(ZS)}$ and high-performance computing' by E. Laloy and J.A. Vrugt, *Water Resources Research*, vol. 50, doi:10.1002/2012WR013341, 2014.

M. Clark, D. Kavetski, and F. Fenicia, "Pursuing the method of multipleworking hypotheses for hydrological modeling," *Water Resources Research*, vol. 47 (9), doi:10.1029/2010WR009827, 2011.

M. Clark, D. Kavetski and F. Fenicia, Reply to comment by K. Beven et al. on "Pursuing the method of multiple working hypotheses for hydrological modeling," *Water Resources Research*, vol. 48, W11802, doi:10.1029/2012WR012547, 2012.

F.C. Coelho, C.T. Codeço, and M.G.M. Gomes, "A Bayesian framework for parameter estimation in dynamical models," *PLoS ONE*, vol. 6 (5), e19616, doi:10.1371/journal.pone.0019616, 2011.

G. Coxon, J. Freer, I.K. Westerberg, T. Wagener, R. Woods, and P.J. Smith. "A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations." *Water Resources Research.* 51(7), 5531-5546, 2015.

K. Csilléry, M.G.B. Blum, O.E. Gaggiotti, and O. François, "Approximate Bayesian computation (ABC) in practice," *Trends in Ecology & Evolution*, vol. 25, pp. 410-418, 2010.

K. Csilléry, M.G.B. Blum, O.E. Gaggiotti, and O. François, "Approximate Bayesian computation (ABC) in practice," *Trends in Ecology & Evolution*, vol. 25, pp. 410-418, 2010.

S. Dean, J.E. Freer, K.J. Beven, A.J. Wade, and D. Butterfield, "Uncertainty assessment of a process-based integrated catchment model of phosphorus (INCA-P)," *Stochastic Environmental Research and Risk Assessment*, vol. 23, pp. 991-1010, doi:10.1007/s00477-008-0273-z, 2009.

S.C. DeCaluwe, P.A. Kienzle, P. Bhargava, A.M. Baker, and J.A. Dura, "Phase segregation of sulfonate groups in Nafion interface lamellae, quantified via neutron reflectometry fitting techniques for multi-layered structures," *Soft Matter*, vol. 10, pp. 5763-577, doi:10.1039/C4SM00850B, 2014.

S.C. Dekker, J.A. Vrugt, and R.J. Elkington, "Significant variation in vegetation characteristics and dynamics from ecohydrologic optimality of net carbon profit," *Ecohydrology*, vol. 5, pp. 1-18, doi:10.1002/eco.177, 2010.

P. Del Moral, A. Doucet, and A. Jasra, "An adaptive sequential Monte Carlo method for approximate Bayesian computation," *Statistics & Computing*, vol. 22, pp. 1009-1020, doi:10.1007/s11222-011-9271-y, 2012.

P.J. Diggle, and R.J. Gratton, "Monte Carlo methods of inference for implicit statistical models," *Journal of the Royal Statistical Society Series B*, vol. 46, pp. 193-227, 1984.

Q. Duan, J.J. Schaake, V. Andréassian, S. Franks, G. Goteti, H.V. Gupta, Y.M. Gusev, F. Habets, A. Hall, L. Hay, T. Hogue, M. Huang, G. Leavesley, X. Liang, O.N. Nasonova, J. Noilhan, L. Oudin, S. Sorooshian, T. Wagener, and E.F. Wood, "Model parameter estimation experiment (MOPEX): An overview of science strategy and major results from the second and third workshops," *Journal of Hydrology*, vol. 320, pp. 3-17, 2006.

B. Dumont, V. Leemans, M. Mansouri, B. Bodson, J.-P. Destain, M.-F Destain, "Parameter identification of the STICS crop model, using an accelerated formal MCMC approach," *Environmental Modeling & Software*, vol. 52, pp. 121-135, 2014.

J.A. Dura, S.T. Kelly, P.A. Kienzle, J.-H. Her, T.J. Udovic, C.F. Majkrzak, C.-J. Chung, and B.M. Clemens, "Porous Mg formation upon dehydrogenation of MgH_2 thin films," *Journal of Applied Physics*, vol. 109, 093501, 2011.

G. Evin, D. Kavetski, M. Thyer, and G. Kuczera, "Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration," *Water Resources Research*, vol. 49, 4518âĂŞ4524, doi:10.1002/wrcr.20284, 2013.

J. Freer, K.J. Beven, B. Ambroise, "Bayesian estimation of uncertainty in runoff prediction and the value of data: an application of the GLUE approach," *Water Resources Research*, vol. 32, pp. 2161-2173, 1996.

J. Freer, H. McMillan, J.J. McDonnell, and K.J. Beven, "Constraining dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures," *Journal of Hydrology*, vol. 291, pp. 254-277, 2004.

A. Gelman, and D.B. Rubin, "Inference from iterative simulation using multiple sequences," *Statistical Science*, vol. 7, pp. 457-472, 1992.

L. Gentsch, A. Hammerle, P. Sturm, J. Ogée, L. Wingate, R. Siegwolf, P. Plüss, T. Baur, N. Buchmann, and A. Knohl, "Carbon isotope discrimination during branch photosynthesis of Fagus sylvatica: a Bayesian modeling approach," *Plant, Cell & Environment*, vol. 37, pp. 1516-1535, doi: 10.1111/pce.12262, 2014.

A. Grelaud, C. Robert, J. Marin, F. Rodolphe, and J. Taly, "ABC likelihood-free methods for model choice in Gibbs random fields," *Bayesian Analysis*, vol. 4 (2), pp. 317-336, 2009.

H.V. Gupta, S. Sorooshian, and P.O. Yapo, "Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information," *Water Resources Research*, vol. 34 (4), pp. 751-763, 1998.

H.V. Gupta, T. Wagener, and Y. Liu, "Reconciling theory with observations: elements of a diagnostic approach to model evaluation," *Hydrological Processes*, vol. 22 (18), pp. 3802-3813, 2008.

H.V. Gupta, H. Kling, K.K. Yilmaz, and G.F. Martinez, "Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling," *Journal of Hydrology*, vol. 377 (1-2), pp. 80-91, 2009.

K.W. Harrison, S.V. Kumar, C.D. Peters-Lidard, and J.A. Santanello, "Quantifying the change in soil moisture modeling uncertainty from remote sensing observations using Bayesian inference techniques," *Water Resources Research*, vol. 48, W11514, doi:10.1029/2012WR012337, 2012.

T.J. Heimovaara, and W. Bouten, "A computer-controlled 36-channel time domain reflectometry system for monitoring soil water contents," *Water Resources Research*, vol. 26, pp. 2311-2316, doi:10.1029/WR026i010p02311, 1990.

A.W. Hinnell, T.P.A. Ferré, J.A. Vrugt, S. Moysey, J.A. Huisman, and M.B. Kowalsky, "Improved extraction of hydrologic information from geophysical data through coupled hydrogeophysical inversion," *Water Resources Research*, vol. 46, W00D40, doi:10.1029/2008WR007060, 2010.

R.M. Hornberger, and R.C. Spear, "An approach to the preliminary analysis of environmental systems," *Journal of Environmental Management*, vol. 12, pp. 7-18, 1981.

V.R. Horowitz, B.J. Aleman, D.J. Christle, A.N. Cleland, and D.D. Awschalom, "Electron spin resonance of nitrogen-vacancy centers in optically trapped nanodiamonds," *Proceedings of the National Academy of the United States of America*, vol. 109 (34), pp. 13493-13497, doi:10.1073/pnas.1211311109, 2012.

N.E. Huang, S.R. Long, and Z. Shen, "The mechanism for frequency downshift in nonlinear wave evolution," *Advances in Applied Mechanics*, vol. 32, pp. 59-111, doi:10.1016/S0065-2156(08)70076-0, 1996.

N.E. Huang, Z. Shen, and R.S. Long, "A new view of nonlinear water waves - The Hilbert spectrum," *Annual Review of Fluid Mechanics*, vol. 31, pp. 417-457, doi:10.1146/annurev.fluid.31.1.417, 1999.

N.E. Huang, and Z. Wu, "A review on Hilbert-Huang transform: Method and its applications to geophysical studies," *Reviews of Geophysics*, vol. 46, doi:10.1029/2007RG000228, 2008.

R.P. Ibbitt, and T. O'Donnell, "Designing conceptual catchment models for automatic fitting methods," In: Mathematical Models in Hydrology Symposium, IAHS-AISH, Publication No. 2, pp. 461-475, 1974.

T. Iizumi, Y. Tanaka, G. Sakurai, Y. Ishigooka, and M. Yokozawa, "Dependency of parameter values of a crop model on the spatial scale of simulation," *Journal of Advances in Modeling Earth Systems*, vol. 06, doi:10.1002/2014MS000311, 2014.

I. Iorgulescu, K. Beven, and A. Musy, "Data-based modelling of runoff and chemical tracer concentrations in the Haute-Mentue research catchment (Switzerland)," *Hydrological Processes*, vol. 19, pp. 2557-2573, doi:10.1002/hyp.5731, 2005.

P. Joyce, and P. Marjoram, "Approximately sufficient statistics and Bayesian computation," *Statistical Applications in Genetics and Molecular Biology*, vol. 7 (1), 2008.

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

E.H. Keating, J. Doherty, J.A. Vrugt, and Q. Kang, "Optimization and uncertainty assessment of strongly nonlinear groundwater models with high parameter dimensionality," *Water Resources Research*, vol. 46, W10517, doi:10.1029/2009WR008584, 2010.

K. Keesman, "Membership-set estimation using random scanning and principal component analysis," *Mathematics and Computers in Simulation*, vol. 32, pp. 535-543, 1990.

K. Kelly, and R. Krzysztofowicz, "A bivariate meta-Gaussian density for use in hydrology," *Stochastic Hydrology and Hydraulics*, vol. 11, pp. 17-31, 1997.

B.J. Kirby, M.T. Rahman, R.K. Dumas, J.E. Davies, C.H. Lai, and K. Liu, "Depth-resolved magnetization reversal in nanoporous perpendicular anisotropy multilayers," *Journal of Applied Physics*, vol. 113, 033909, doi:10.1063/1.4775819, 2013.

W.Y. Kow, W.L. Khong, Y.K. Chin, I. Saad, K.T.K. Teo, "Enhancement of Markov chain monte Carlo convergence speed in vehicle tracking using genetic operator," 2012 Fourth International Conference on Computational Intelligence, Modeling and Simulation (CIMSiM), pp. 270-275, doi:10.1109/CIMSim.2012.61, 2012.

L. Krayer, J.W. Lau, and B.J. Kirby, "Structural and magnetic etch damage in CoFeB," *Journal of Applied Physics*, vol. 115, 17B751, 2014.

T. Krueger, J.N. Quinton, J. Freer, C.J. Macleod, G.S. Bilotta, R.E. Brazier, and P.M. Haygarth," Uncertainties in data and models to describe event dynamics of agricultural sediment and phosphorus transfer," *Journal of Environmental Quality*, vol. 38 (3), pp. 1137-1148, 2009.

R. Krzysztofowicz, "Transformation and normalization of variates with specified distributions," *Journal of Hydrology*, vol. 197, pp. 286-292, 1997.

R. Krzysztofowicz, and K. Kelly, "Hydrologic uncertainty processor for probabilistic river stage forecasting," *Water Resources Research*, vol. 36, pp. 3265-3277, 2000.

R. Krzysztofowicz, and C.J. Maranzano, "Hydrologic uncertainty processor for probabilistic stage transition forecasting," *Journal of Hydrology*, vol. 293, pp. 57-73, 2004.

G. Kuczera, "Improved parameter inference in catchment models, 1. Evaluating parameter uncertainty," *Water Resources Research*, vol. 19 (5), pp. 1151-1162, doi:10.1029/WR019i005p01151, 1983.

E. Laloy, and J.A. Vrugt, "High-dimensional posterior exploration of hydrologic models using multiple-try DREAM$_{(ZS)}$ and high-performance computing," *Water Resources Research*, vol. 48, W01526, doi:10.1029/2011WR010608, 2012a.

E. Laloy, N. Linde, and J.A. Vrugt, "Mass conservative three-dimensional water tracer distribution from Markov chain Monte Carlo inversion of time-lapse ground-penetrating radar data," *Water Resources Research*, vol. 48, W07510, doi:10.1029/2011WR011238, 2012b.

E. Laloy, B. Rogiers, J.A. Vrugt, D. Jacques, and D. Mallants, "Efficient posterior exploration of a high-dimensional groundwater model from two-stage Markov chain Monte Carlo simulation and polynomial chaos expansion," *Water Resources Research*, vol. 49 (5), pp. 2664-2682, doi:10.1002/wrcr.20226, 2013.

E. Laloy, N. Linde, D. Jacques, and J.A. Vrugt, "Probabilistic inference of multi-Gaussian fields from indirect hydrological data using circulant embedding and dimensionality reduction," *Water Resources Research*, vol. 51, 4224-4243, doi:10.1002/2014WR016395, 2015.

G.E. Leventhal, H.F. Günthard, S. Bonhoeffer, and T. Stadler, "Using an epidemiological model for phylogenetic inference reveals density dependence in HIV transmission," *Molecular Biology and Evolution*, vol. 31 (1), pp. 6-17, doi:10.1093/molbev/mst172, 2013.

N. Linde, and J.A. Vrugt, "Distributed soil moisture from crosshole ground-penetrating radar travel times using stochastic inversion," *Vadose Zone Journal*, vol. 12 (1), doi:10.2136/vzj2012.0101, 2013.

J. Lise, C. Meghir, and J-M. Robin, "Mismatch, sorting and wage dynamics," National Bureau of Economic Research, Working paper, 18719, pp. 1-43, http://www.nber.org/papers/w18719, 2012.

J. Lise, "On the job search and precautionary savings," *Review of economic studies*, vol. 80 (3), pp. 1086-1113, doi:10.1093/restud/rds042, 2013.

Y. Liu, J.E. Freer, K.J. Beven, and P. Matgen, "Towards a limits of acceptability approach to the calibration of hydrological models: extending observation error," *Journal of Hydrology*, 367, pp. 93-103, doi:10.1016/j.jhydrol.2009.01.016, 2009.

T. Lochbühler, S.J. Breen, R.L. Detwiler, J.A. Vrugt, and N. Linde, "Probabilistic electrical resistivity tomography for a CO_2 sequestration analog," *Journal of Applied Geophysics*, vol. 107, pp. 80-92, doi:10.1016/j.jappgeo.2014.05.013, 2014.

T. Lochbühler, J.A. Vrugt, M. Sadegh, and N. Linde, "Summary statistics from training images as prior information in probabilistic inversion," *Geophysical Journal International*, vol. 201, pp. 157-171, doi:10.1093/gji/ggv008.

B. Malama, K.L. Kuhlman, and S.C. James, "Core-scale solute transport model selection using Monte Carlo analysis," *Water Resources Research*, vol. 49, pp. 3133-3147, doi:10.1002/wrcr.20273, 2013.

P. Mantovan, and E. Todini, "Hydrological forecasting uncertainty assessment: incoherence of the GLUE methodology," *Journal of Hydrology*, vol. 330, pp. 368âĂŞ381, 2006.

P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré, "Markov chain Monte Carlo without likelihoods," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100 (26), pp. 15324-15328, 2003.

L. Mari, E. Bertuzzo, L. Righetto, R. Casagrandi, M. Gatto, I. Rodriguez-Iturbe, and A. Rinaldo, "Modeling cholera epidemics: the role of waterways, human mobility and sanitation," *Journal of the Royal Society Interface*, vol. 9 (67), pp. 376-388, 2011.

L.A. Marshall, A. Sharma, and D.J. Nott, "Towards dynamic catchment modelling: A Bayesian hierarchical mixtures of experts framework," *Hydrological Processes*, vol. 21, pp. 847-861, 2007.

H. McMillan, J. Freer, F. Pappenberger, T. Krueger, and M. Clark, "Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions," *Hydrological Processes*, vol. 24 (10), pp. 1270-1284, 2010.

Y. Meyer, "Wavelets and operators," *Cambridge: Cambridge University Press*, ISBN 0-521-42000-8, 1992.

B. Minasny, J.A. Vrugt, and A.B. McBratney, "Confronting uncertainty in model-based geostatistics using Markov chain Monte Carlo simulation," *Geoderma*, vol. 163, pp. 150-622, doi:10.1016/j.geoderma.2011.03.011, 2011.

J.E. Nash, "A unit hydrograph study with particular reference to British catchments," *Proceedings - Institution of Civil Engineers*, vol. 17, pp. 249-282, 1960.

D.J. Nott, L. Marshall, and J. Brown, "Generalized likelihood uncertainty estimation (GLUE) and approximate Bayesian computation: What's the connection?," *Water Resources Research*, vol. 48 (12), doi:10.1029/2011WR011128, 2012.

J.E. Owejan, J.P. Owejan, S.C. DeCaluwe, and J.A Dura, "Solid electrolyte interphase in Li-ion batteries: Evolving structures measured in situ by neutron reflectometry", *Chemistry of Materials*, vol. 24, pp. 2133-2140, 2012.

T. Page, K.J. Beven, J. Freer, A. Jenkins, "Investigating the uncertaintyin predicting responses to atmospheric deposition using the model of acidification of groundwater in catchments (MAGIC) within a generalised likelihood uncertainty estimation (GLUE) framework," *Water Soil and Air Pollution*, vol. 142, pp. 71-94, 2003.

T. Page, K.J. Beven, D. Whyatt, "Predictive capability in estimating changes in water quality: long-term responses to atmospheric deposition," *Water Soil and Air Pollution*, vol. 151, pp. 215-244, 2004.

T. Page, K.J. Beven, J. Freer, "Modelling the chloride signal at the Plynlimon catchments, Wales using a modified dynamic TOPMODEL," *Hydrological Processes*, vol. 21, pp. 292-307, 2007.

F. Pappenberger, K. Beven, M. Horritt, S. Blazkova, "Uncertainty in the calibration of effective roughness parameters in HEC-RAS using inundation and downstream level observations," *Journal of Hydrology*, vol. 302, pp. 46-69, 2005.

F. Pappenberger, K. Frodsham, K.J. Beven, R. Romanovicz, and P. Matgen, "Fuzzy set approach to calibrating distributed flood inundation models using remote sensing observations," *Hydrology and Earth System Sciences*, vol. 11 (2), pp. 739-752, 2007.

D.G. Partridge, J.A. Vrugt, P. Tunved, A.M.L. Ekman, D. Gorea, and A. Sorooshian, "Inverse modeling of cloud-aerosol interactions - Part I: Detailed response surface analysis," *Atmospheric Chemistry and Physics*, vol. 11, pp. 4749-4806, doi:10.5194/acpd-11-4749-2011, 2011.

D.G. Partridge, J.A. Vrugt, P. Tunved, A.M.L. Ekman, H. Struthers, and A. Sorooshian, "Inverse modeling of cloud-aerosol interactions - Part II: Sensitivity tests on liquid phase clouds using Markov chain Monte Carlo simulation approach," *Atmospheric Chemistry and Physics*, vol. 12, pp. 2823-2847, doi:10.5194/acp-12-2823-2012, 2012.

K.V. Price, R.M. Storn, and J.A. Lampinen, "Differential evolution, A practical approach to global optimization," Springer, Berlin, 2005.

J.K. Pritchard, M.T. Seielstad, A. Perez-Lezaun, and M.T. Feldman, "Population growth of human Y chromosomes: A study of Y chromosome microsatellites," *Molecular Biology and Evolution*, vol. 16 (12), pp. 1791-1798, 1999.

O. Ratmann, C. Andrieu, C. Wiuf, and S. Richardson, "Model criticism based on likelihood-free inference, with an application to protein network evolution," *Proceedings of the National Academy of Sciences of the United States of Ameria*, vol. 106, pp. 1-6, 2009.

P. Reichert, and J. Mieleitner, "Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters," *Water Resources Research*, vol. 45, W10402, doi:10.1029/2009WR007814, 2009.

P. Reichert, and N. Schuwirth, "Linking statistical bias description to multiobjective model calibration, *Water Resources Research*, vol. 48, W09543, doi:10.1029/2011WR011391, 2012.

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

A. Rinaldo, E. Bertuzzo, L. Mari, L. Righetto, M. Blokesch, M. Gatto, R. Casagrandi, M. Murray, S.M. Vesenbeckh, and I. Rodriguez-Iturbe, "Reassessment of the 2010-2011 Haiti cholera outbreak and rainfall-driven multiseason projections," *Proceedings of the National Academy of the United States of America*, vol. 109 (17), pp. 6602-6607, 2012.

M. Rosas-Carbajal, N. Linde, T. Kalscheuer, and J.A. Vrugt, "Two-dimensional probabilistic inversion of plane-wave electromagnetic data: Methodology, model constraints and joint inversion with electrical resistivity data," *Geophysical Journal International*, vol. 196 (3), pp. 1508-1524, doi: 10.1093/gji/ggt482, 2014.

M. Sadegh, and J.A. Vrugt (2013), Approximate Bayesian Computation in hydrologic modeling: equifinality of formal and informal approaches, *Hydrology and Earth System Sciences - Discussions*, 10, 4739-4797, doi:10.5194/hessd-10-4739-2013.

M. Sadegh, and J.A. Vrugt, "Approximate Bayesian computation using Markov chain monte Carlo simulation: DREAM$_{(\mathrm{ABC})}$," *Water Resources Research*, vol. 50, doi:10.1002/2014WR015386, 2014.

M. Sadegh, J.A. Vrugt, C. Xu, and E. Volpi, "The stationarity paradigm revisited: Hypothesis testing using diagnostics, summary metrics, and DREAM$_{(\mathrm{ABC})}$," *Water Resources Research*, vol. 51, pp. 9207-9231, doi:10.1002/2014WR016805, 2015.

B. Scharnagl, J.A. Vrugt, H. Vereecken, and M. Herbst, "Information content of incubation experiments for inverse estimation of pools in the Rothamsted carbon model: a Bayesian perspective," *Biogeosciences*, vol. 7, pp. 763-776, 2010.

B. Scharnagl, J.A. Vrugt, H. Vereecken, and M. Herbst, "Bayesian inverse modeling of soil water dynamics at the field scale: using prior information about the soil hydraulic properties," *Hydrology and Earth System Sciences*, vol. 15, pp. 3043-3059, doi:10.5194/hess-15-3043-2011, 2011.

B. Scharnagl, S.C. Iden, W. Durner, H. Vereecken, and M. Herbst, "Inverse modelling of in situ soil water dynamics: accounting for heteroscedastic, autocorrelated, and non-Gaussian distributed residuals," *Hydrology and Earth System Sciences Discussions*, vol. 12, pp. 2155-2199, 2015.

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

M. Shafii, B. Tolson, and. L.S. Matott, "Uncertainty-based multi-criteria calibration of rainfall-runoff models: a comparative study," *Stochastic Environmental Research and Risk Assessment*, vol. 28 (6), pp. 1493-1510, 2014.

J. Šimůnek, M.Šejna, H. Saito, M. Sakai, and M.T. van Genuchten, "The HYDRUS-1D software package for simulating the one-dimensional movement of water, heat and multiple solutes in variably-saturated media (Version 4.0)," Department of Environmental Sciences, University of California Riverside, Riverside, CA, USA, 2008.

S.A. Sisson, Y. Fan, and M.M. Tanaka, "Sequential Monte Carlo without likelihoods," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104 (6), pp. 1760-1765, 2007.

T.A. Smith, A. Sharma, L. Marshall, R. Mehrotra, and S. Sisson, "Development of a formal likelihood function for improved Bayesian inference of ephemeral catchments," *Water Resources Research*, vol. 46, W12551, doi:10.1029/2010WR009514, 2010.

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

J. Starrfelt, and Ø. Kaste, "Bayesian uncertainty assessment of a semi-distributed integrated catchment model of phosphorus transport," *Environmental Science: Processes & Impacts*, vol. 16, pp. 1578-1587, doi:10.1039/C3EM00619K, 2014.

J.R. Stedinger, R.M. Vogel, S.U. Lee, R. Batchelder, "Appraisal ofthe Generalized Likelihood Uncertainty Estimation (GLUE) method," *Water Resources Research*, vol. 44, W00B06, doi:10.1029/2008WR006822, 2008.

S.M. Stigler, "Who discovered Bayes's theorem?," *The American Statistician*, vol. 37 (4), Part 1, pp. 290-296, 1983.

R. Storn, and K. Price, "Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization*, vol. 11, pp. 341-359, 1997.

X-L. Sun, S-C. Wu, H-L. Wang, Yu-G. Zhao, G-L. Zhang, Y.B. Man, M.H. Wong, "Dealing with spatial outliers and mapping uncertainty for evaluating the effects of urbanization on soil: A case study of soil pH and particle fractions in Hong Kong," *Geoderma*, vol. 195-196, pp. 220-233, 2013.

B.J. Tarasevich, U. Perez-Salas, D.L. Masic, J. Philo, P. Kienzle, S. Krueger, C.F. Majkrzak, J.L. Gray, and W.J. Shaw, "Neutron reflectometry studies of the adsorbed structure of the Amelogenin, LRAP", *The Journal of Physical Chemistry B*, vol. 117 (11), pp. 3098-3109, doi:10.1021/jp311936j, 2013.

G.C. Topp, J.L. Davis, and A.P. Annan, "Electromagnetic determination of soil water content: measurements in coaxial transmission lines," *Water Resources Research*, vol. 16, pp. 574-582, doi:10.1029/WR016i003p00574, 1980.

D.M. Toyli, D.J. Christle, A. Alkauskas, B.B. Buckley, C.G. van de Walle, and D.D. Awschalom, "Measurement and control of single nitrogen-vacancy center spins above 600 K," *Physical Review X*, vol. 2, 031001, doi:10.1103/PhysRevX.2.031001, 2012.

B.M. Turner, and T. van Zandt, "A tutorial on approximate Bayesian computation," *Journal of Mathematical Psychology*, vol. 56, pp. 69-85, 2012.

M.T. van Genuchten, "A closed-form equation for predicting the hydraulic conductivity of unsaturated soils," *Soil Science Society of America Journal*, vol. 44 (5), pp. 892-898, 1980.

G. T. van Straten and K.J. Keesman, ĂIJUncertainty propagation and speculation in projective forecasts of environmental change: A lake-eutrophication example, *J. Environmental Forecasting*. vol. 10 (1-2) pp. 163âĂŞ190, 1991.

J.A. Vrugt, H.V. Gupta, W. Bouten, and S. Sorooshian, "A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters," *Water Resources Research*, vol. 39 (8), 1201, doi:10.1029/2002WR001642, 2003.

J.A. Vrugt, C.G.H. Diks, W. Bouten, H.V. Gupta, and J.M. Verstraten, "Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation," *Water Resources Research*, vol. 41 (1), W01017, doi:10.1029/2004WR003059, 2005.

J.A. Vrugt, C.J.F. ter Braak, M.P. Clark, J.M. Hyman, and B.A. Robinson, "Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation," *Water Resources Research*, vol. 44, W00B09, doi:10.1029/2007WR006720, 2008a.

J.A. Vrugt, C.G.H. Diks, and M.P. Clark, "Ensemble Bayesian model averaging using Markov chain Monte Carlo sampling," *Environmental Fluid Mechanics*, vol. 8 (5-6), pp. 579-595, doi:10.1007/s10652-008-9106-3, 2008b.

J.A. Vrugt, C.J.F. ter Braak, H.V. Gupta, and B.A. Robinson, "Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling?," *Stochastic Environmental Research and Risk Assessment*, vol. 23 (7), pp. 1011-1026, 2009.

J.A. Vrugt, and C.J.F. ter Braak, "DREAM$_{(D)}$: an adaptive Markov chain Monte Carlo simulation algorithm to solve discrete, noncontinuous, and combinatorial posterior parameter estimation problems," *Hydrology and Earth System Sciences*, vol. 15, pp. 3701-3713, doi:10.5194/hess-15-3701-2011, 2011.

J.A. Vrugt, and M. Sadegh, "Toward diagnostic model calibration and evaluation: Approximate Bayesian computation," *Water Resources Research*, vol. 49, doi:10.1002/wrcr.20354, 2013a.

J.A. Vrugt, C.J.F. ter Braak, C.G.H. Diks, and G. Schoups, "Advancing hydrologic data assimilation using particle Markov chain Monte Carlo simulation: theory, concepts and applications," *Advances in Water Resources*, Anniversary Issue - 35 Years, 51, 457-478, doi:10.1016/j.advwatres.2012.04.002, 2013b.

J.A. Vrugt, and E. Laloy, "Reply to comment by Chu et al. on "High-dimensional posterior exploration of hydrologic models using multiple-try DREAM$_{(ZS)}$ and high-performance computing," *Water Resources Research*, vol. 50, pp. 2781-2786, doi:10.1002/2013WR014425, 2014.

J.A. Vrugt, "Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB Implementation," *Environmental Modeling & Software*, vol. 75, pp. 273-316, doi:10.1016/j.envsoft.2015.08.013, 2016a.

J.A. Vrugt, "The scientific method, Bayes theorem, diagnostic model evaluation, and summary metrics as prior distribution," *Water Resources Research*, vol. XX, no. XX, pp. XX-XX, In Prep, 2016b.

I.K. Westerberg, J-L. Guerrero, P.M. Younger, K.J. Beven, J. Seibert, S. Halldin, J.E. Freer, C-Y. Xu, "Calibration of hydrological models using flow-duration curves," *Hydrology and Earth System Sciences*, vol. 15, pp. 2205-2227, doi:10.5194/hess-15-2205-2011, 2011.

I.K. Westerberg, and H.K. McMillan, "Uncertainty in hydrological signatures", *Hydrology and Earth System Sciences*, 19(9), 3951-3968, 2015.

T. Wöhling, and J.A. Vrugt, "Multi-response multi-layer vadose zone model calibration using Markov chain Monte Carlo simulation and field water retention data," *Water Resources Research*, vol. 47, W04510, doi:10.1029/2010WR009265, 2011.

C.G. Yale, B.B. Buckley, D.J. Christle, G. Burkard, F.J. Heremans, L.C. Bassett, and D.D. Awschalom, "All-optical control of a solid-state spin using coherent dark states," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110 (19), pp. 7595-7600, doi:10.1073/pnas.1305920110, 2013.

P.C. Young, "Hypothetico-inductive data-based mechanistic modeling of hydrologic systems," *Water Resources Research*, 49(2), 915-935, 2013.

S.L. Zabell, "The rule of succession," *Erkenntnis*, vol. 31 (2-3), pp. 283-321, 1989.

S. Zaoli, A. Giometto, M. Formentin, S. Azaele, A. Rinaldo, and A. Maritan, "Phenomenological modeling of the motility of self-propelled microorganisms," *arXiv*, 1407.1762, 2014.

C. Zilliox, and Frédéric Gosselin, "Tree species diversity and abundance as indicators of understory diversity in French mountain forests: Variations of the relationship in geographical and ecological space," *Forest Ecology and Management*, vol. 321 (1), pp. 105-116, 2014.

D. Zhang, K.J. Beven, and A. Mermoud, A comparison of nonlinear least square and GLUE for model calibration and uncertainty estimation for pesticide transport in soils. Advances in Water Resources, 29, 1924-1933, 2006.

Table 1 Parameters and state variables of the SAC-SMA model and their ranges.

Parameter	Symbol	Lower	Upper	Units
Upper zone tension water maximum storage	UZTWM	1.0	150.0	mm
Upper zone free water maximum storage	UZFWM	1.0	150.0	mm
Lower zone tension water maximum storage	LZTWM	1.0	500.0	mm
Lower zone free water primary maximum storage	LZFPM	1.0	1000.0	mm
Lower zone free water supplemental maximum storage	LZFSM	1.0	1000.0	mm
Additional impervious area	ADIMP	0.0	0.40	-
Upper zone free water lateral depletion rate	UZK	0.1	0.5	day^{-1}
Lower zone primary free water depletion rate	LZPK	0.0001	0.025	day^{-1}
Lower zone supplemental free water depletion rate	LZSK	0.01	0.25	day^{-1}
Maximum percolation rate	ZPERC	1.0	250.0	-
Exponent of the percolation equation	REXP	1.0	5.0	-
Impervious fraction of the watershed area	PCTIM	0.0	0.1	-
Fraction from upper to lower zone free water storage	PFREE	0.0	0.6	-
Recession constant three linear routing reservoirs	RQOUT	0.0	1.0	day^{-1}
State variables				
Upper-zone tension water storage content	UZTWC	0.0	150.0	mm
Upper-zone free water storage content	UZFWC	0.0	150.0	mm
Lower-zone tension water storage content	LZTWC	0.0	500.0	mm
Lower-zone free primary water storage content	LZFPC	0.0	1000.0	mm
Lower-zone free secondary water storage content	LZFSC	0.0	1000.0	mm
Additional impervious area content	ADIMC	0.0	650.0	mm

Table 2 Parameters of the HYDRUS-1D model and their prior uncertainty ranges.

Parameter	Symbol	Lower	Upper	Units
Residual soil moisture content	θ_r	0.043	0.091	$cm^3\ cm^{-3}$
Saturated soil moisture content	θ_s	0.409	0.481	$cm^3\ cm^{-3}$
Reciprocal of air-entry value	α	0.078	0.126	cm^{-1}
Curve shape parameter	n	1.196	1.306	-
Saturated hydraulic conductivity	K_s	0.107	0.923	$cm\ day^{-1}$
Pressure head at the lower boundary	h_{bot}	-500	-10	cm

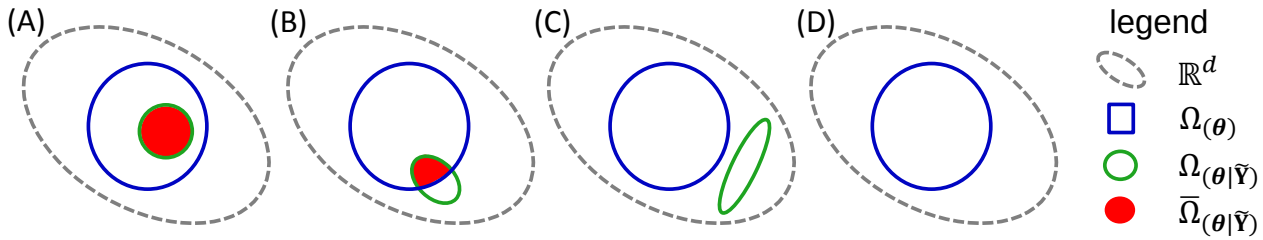

Figure 1 Set-theoretic approach to quantification of parameter uncertainty. The blue, green, and red colors delineate the prior, $\Omega_{(\boldsymbol{\theta})}$, conditional, $\Omega_{(\boldsymbol{\theta}|\tilde{\mathbf{Y}})}$, and posterior, $\hat{\Omega}_{(\boldsymbol{\theta}|\tilde{\mathbf{Y}})}$ parameter set respectively, whereas the grey ellipsoidal defines the feasible parameter space, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \in \mathbb{R}^d$. The four examples each portray a different outcome, (A) the conditional parameter set intersects fully the prior parameter set, (B) the conditional parameter set intersects only partially the prior parameter set, (C) the conditional and prior parameter set are disjoint (have no elements in common), and (D) the conditional parameter set is empty (no solutions exist that satisfy the limits of acceptability). For the last two examples there does not exist a behavioral solution space.

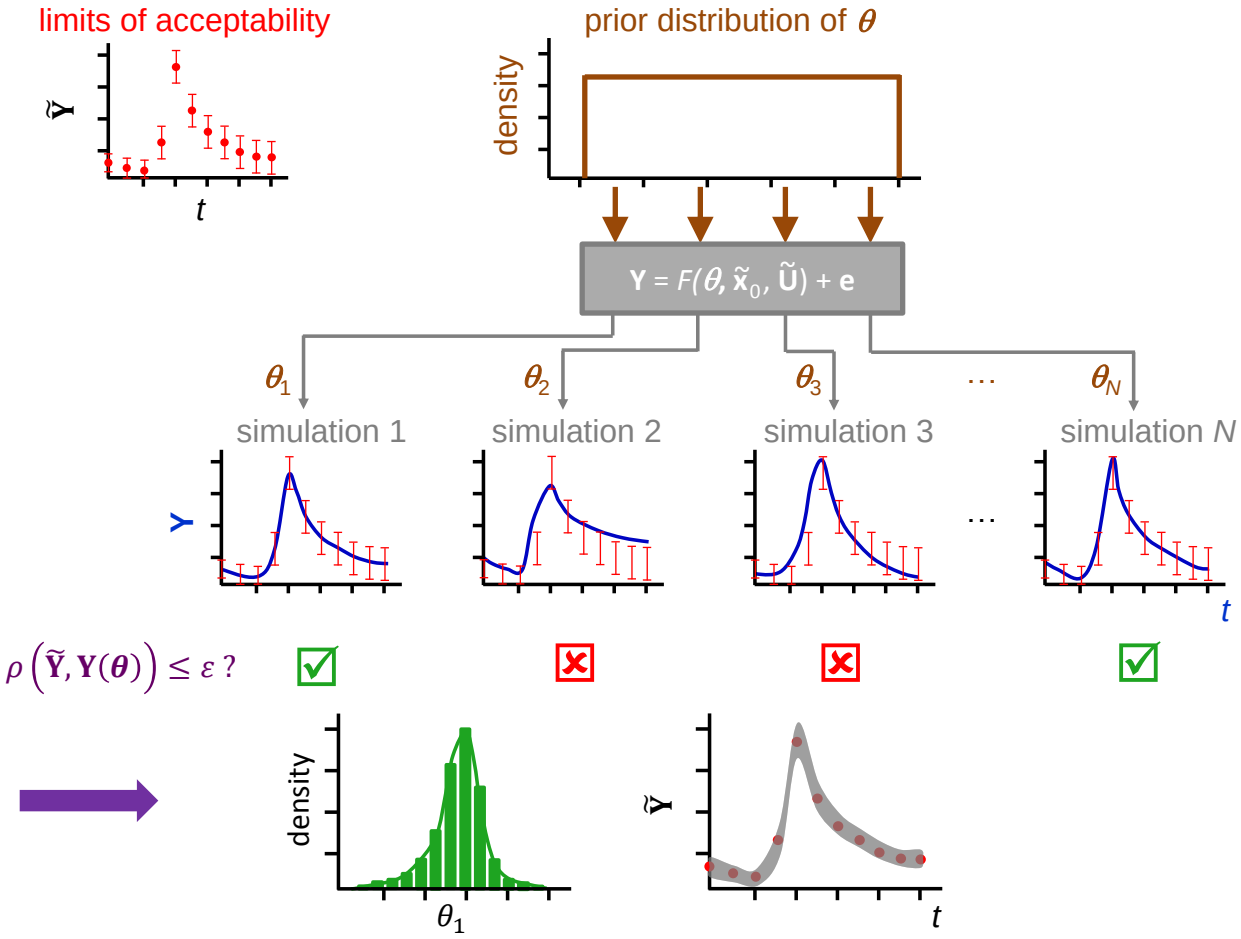

Figure 2 Conceptual overview of approximate Bayesian computation (ABC) for a hypothetical one-dimensional parameter estimation problem. First, N samples are drawn from a user-defined prior distribution, $\boldsymbol{\theta}^* \sim P(\boldsymbol{\theta})$. Then, this ensemble is evaluated with the model (and perturbed with a stochastic error representing exactly the probabilistic properties of the residuals, $\mathbf{e}$) and creates N model simulations. If the distance between the observed and simulated data, $\rho(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta}^*))$ is smaller than or equal to some nominal value, ϵ then $\boldsymbol{\theta}^*$ is retained, otherwise the simulation is discarded. The accepted samples are then used to approximate the posterior parameter distribution, $P(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$. Note that for sufficiently complex models and large data sets the probability of happening upon a simulation run that yields precisely the same simulated values as the observations will be very small, often unacceptably so. Therefore, $\rho(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta}^*))$ is usually defined as a distance between summary statistics of the simulated, $S(\mathbf{Y}(\boldsymbol{\theta}^*))$ and observed, $S(\tilde{\mathbf{Y}})$ data, respectively.

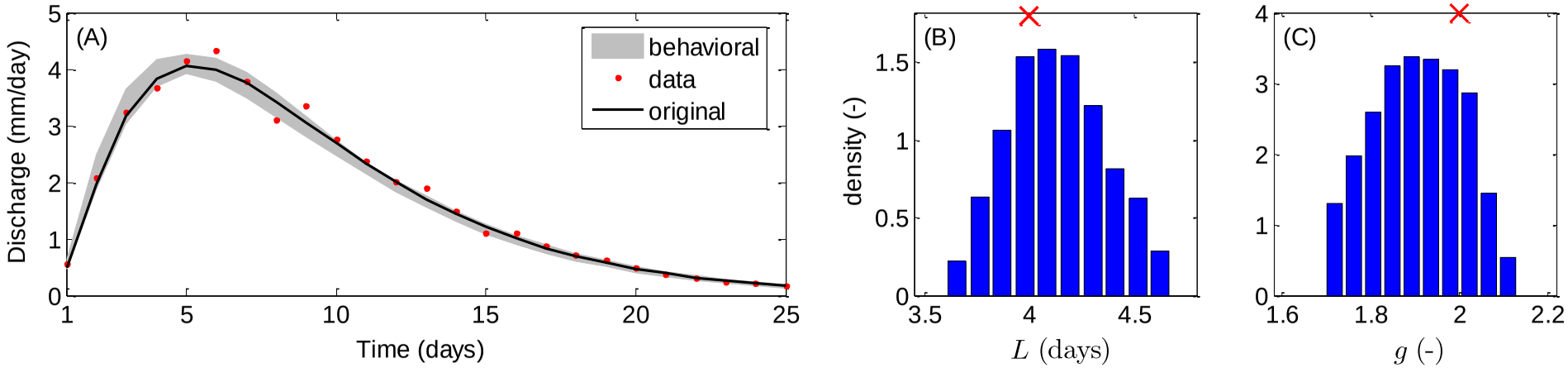

Figure 3 Results of case study I: Nash-Cascade series of reservoirs. (A) Comparison of the observed and simulated hydrograph. The solid black line and red dots denote the original and corrupted data record, respectively, and the gray region is made up of behavioral simulations that satisfy the limits of acceptability at each discharge observation. (B),(C) histograms of the marginal posterior distribution of the model parameters L and g in Equation (12). The parameter values of the (uncorrupted) synthetic data record are separately indicated with the red cross ('X') symbols.

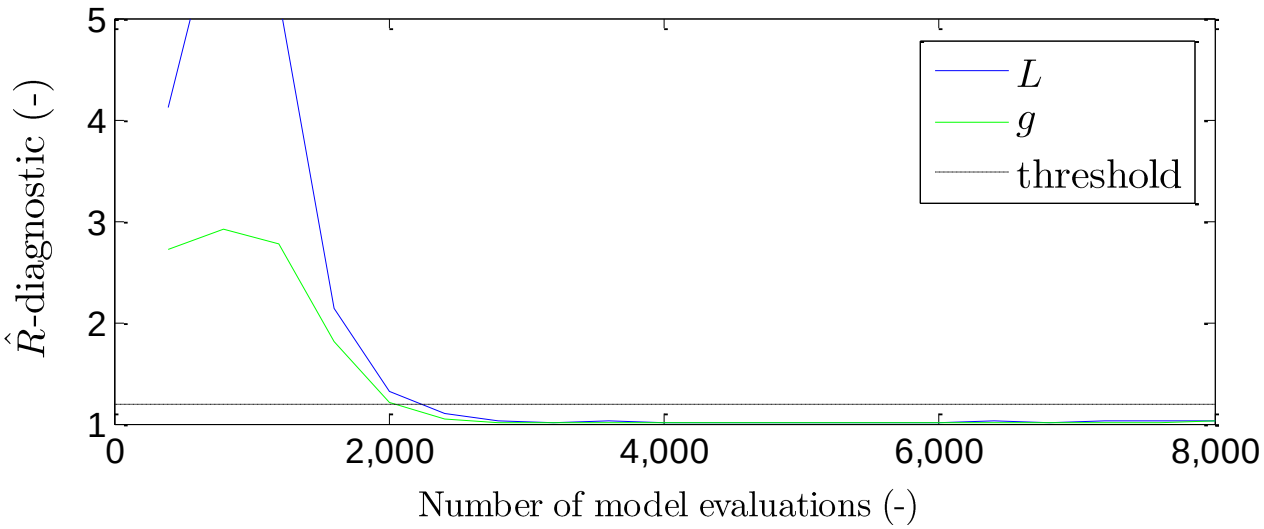

Figure 4 Results of case study I: Nash-Cascade series of reservoirs. Evolution of the $\hat{R}$-diagnostic of *Gelman and Rubin* (1992) used to judge when convergence of the $N = 8$ Markov chains to a limiting distribution has been achieved. The two parameters are coded with a different color. About $2,000$ function evaluations are required to satisfy the convergence threshold of $\hat{R}_j \leq 1.2; j \in \{1, 2\}$.

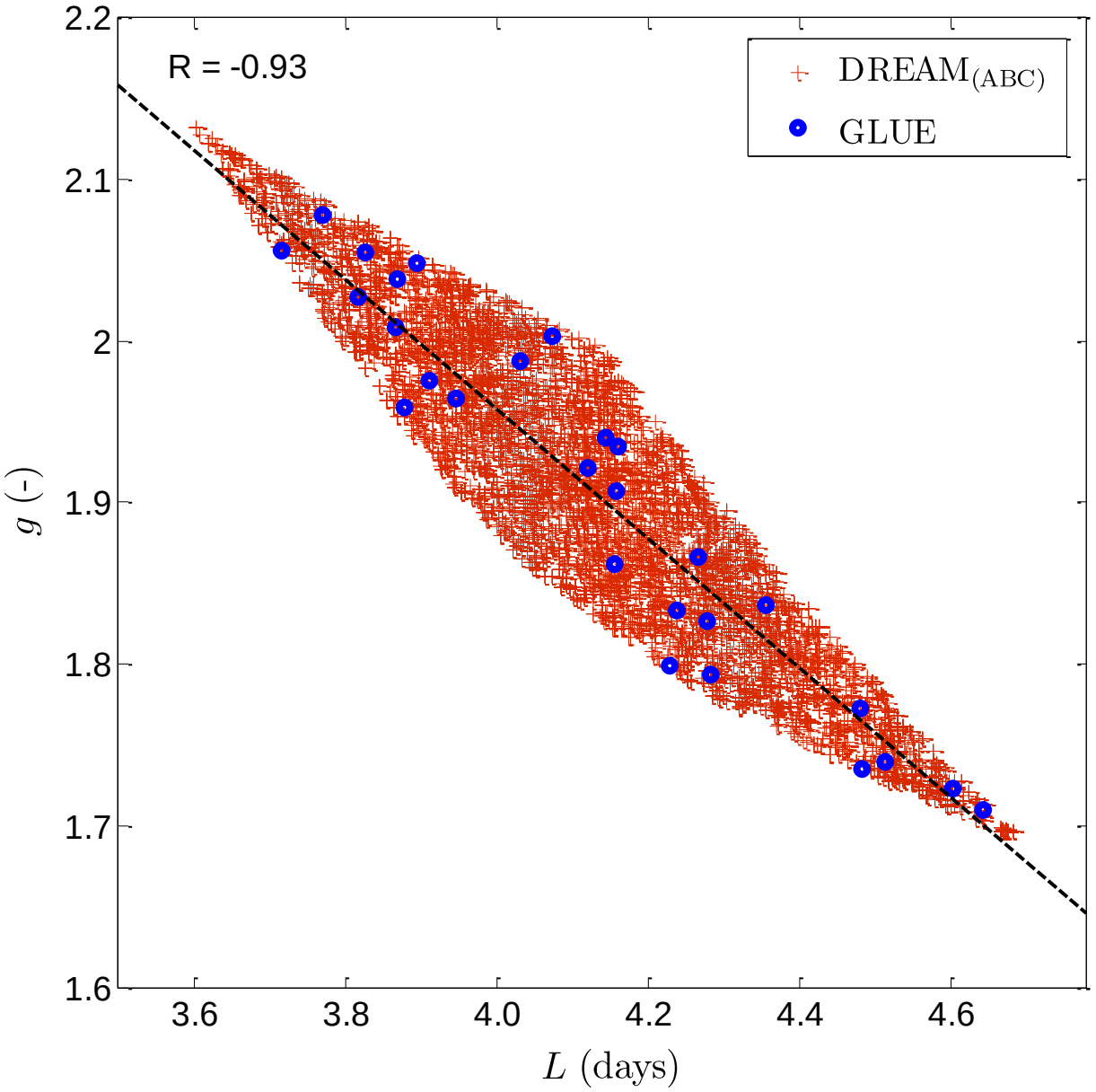

Figure 5 Results of case study I: Nash-Cascade series of reservoirs. Bivariate scatter plot of the behavioral (posterior) samples of L and g derived from MCMC simulation with DREAM$_{(ABC)}$ (dark red) and uniform random sampling (blue dots). The dashed black line plots the least-squares fit to the DREAM$_{(ABC)}$ sample of points. The correlation coefficient equals -0.93.

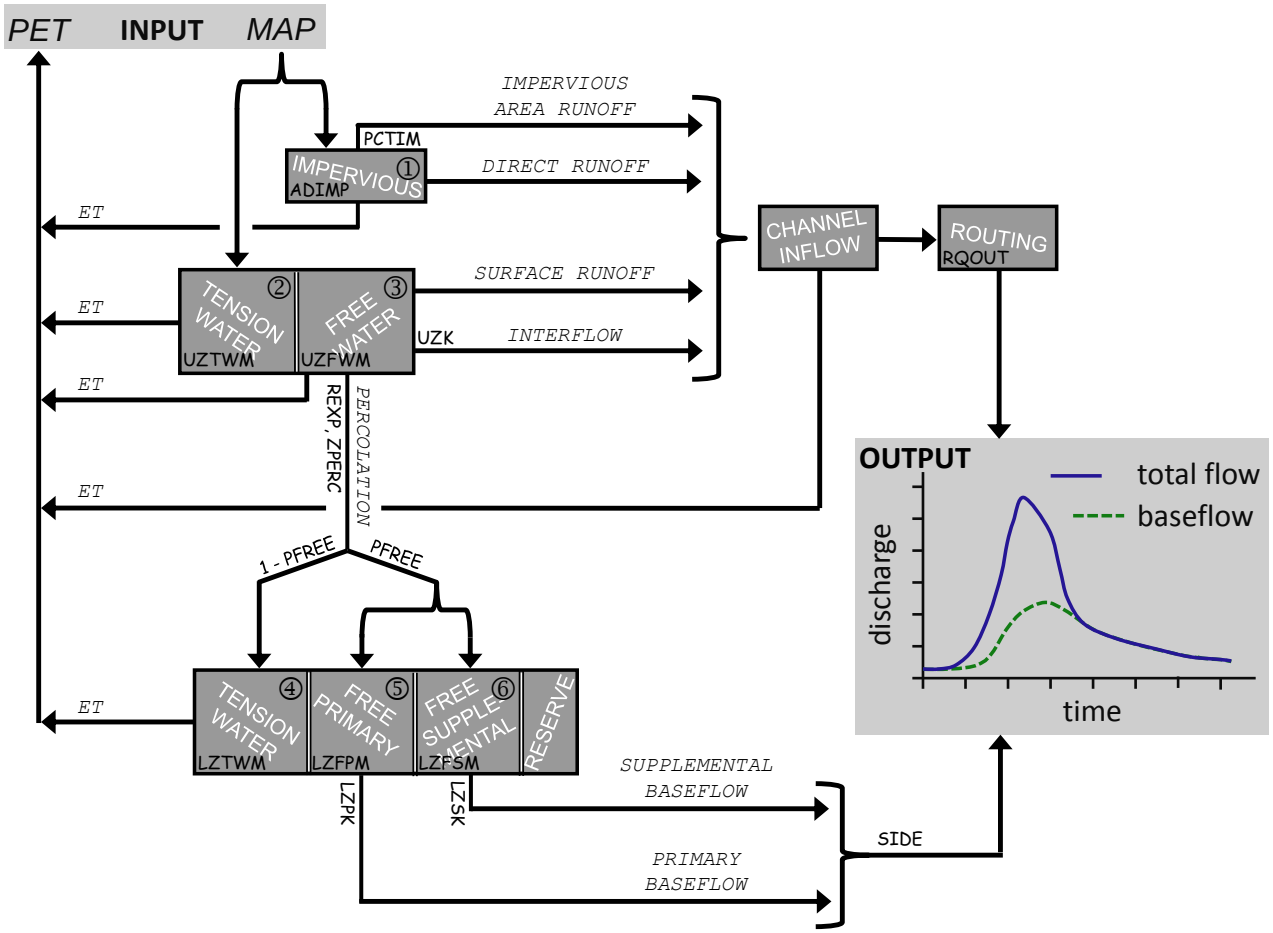

Figure 6 Schematic representation of the Sacramento soil moisture accounting (SAC-SMA) conceptual watershed model. The parameters of the SAC-SMA model appear in Comic Sans font type (black), whereas Courier font type is used to denote the individual fluxes computed by the model. Numbers are used to denote the different SAC-SMA state variables, (1) ADIMC, (2) UZTWC, (3) UZFWC, (4) LZTWC, (5) LZFPC, and (6) LZFSC. The ratio of deep recharge to channel base flow (SIDE) and other remaining SAC-SMA parameters RIVA and RSERV are set to their default values of 0.0, 0.0 and 0.3, respectively.

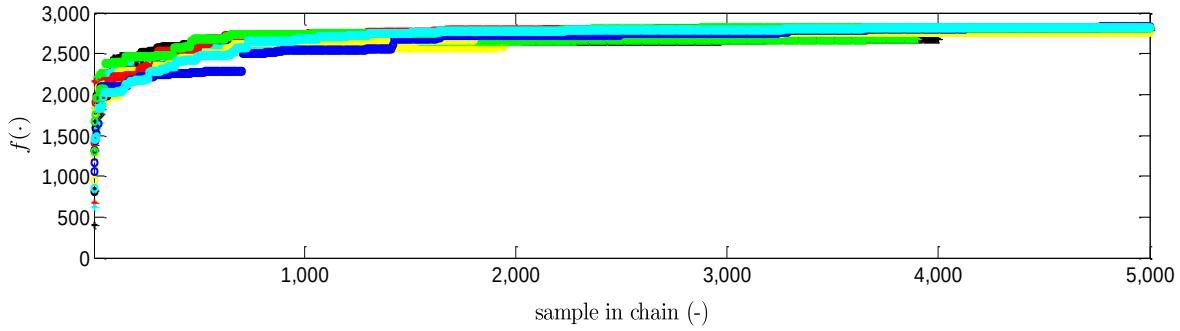

Figure 7 Results of case study II: The SAC-SMA conceptual watershed model. Trace plot of the sampled fitness values of Equation (11) in a randomly selected set of the $K = 20$ different Markov chains of the DREAM$_{(\text{ABC})}$ algorithm. Each chain is coded with a different color and/or symbol. The computed fitness is equivalent to the number of times the simulated value honors the limits of acceptability, $\boldsymbol{\epsilon} = 0.4\tilde{\mathbf{Y}}$ of the observed discharge data. The SAC-SMA model can only fit a portion of the $n = 3,652$ discharge observations of the calibration data set, and is thus rejected as not fit-for-purpose.

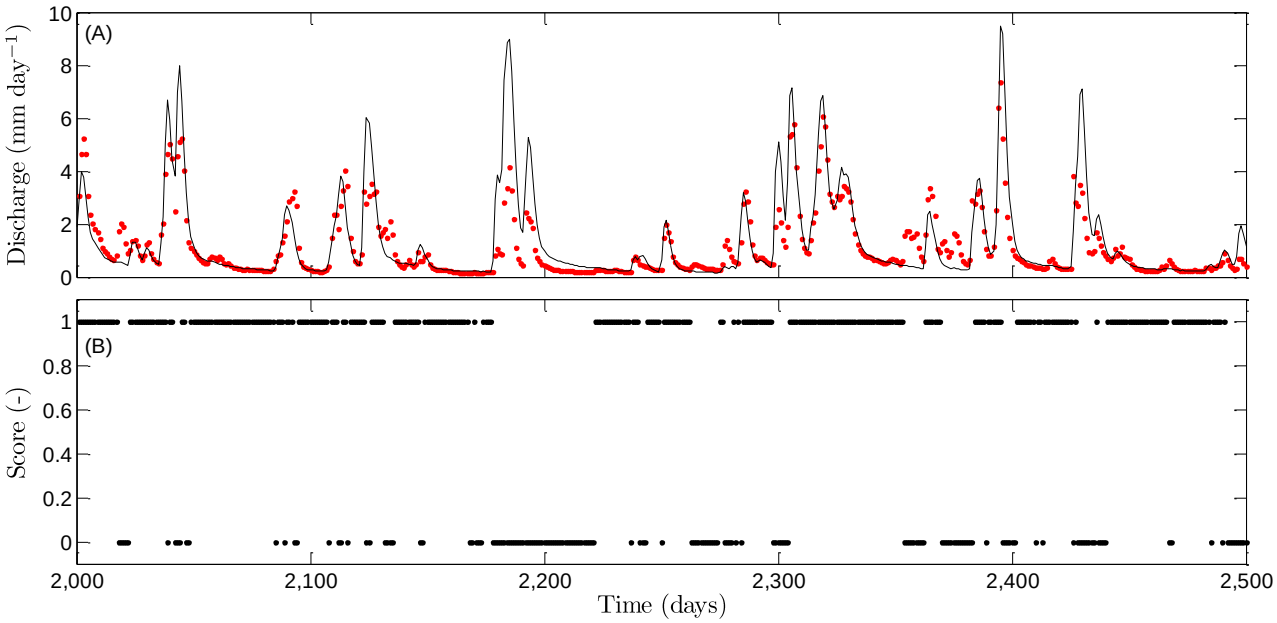

Figure 8 Results of case study II: The SAC-SMA conceptual watershed model. (A) Comparison of the observed (red dots) and simulated (black line) discharge data for a selected 365-day portion of the calibration data period. The simulated values correspond to the DREAM$_{(\mathrm{ABC})}$ sample with highest fitness. (B) score plot of the limits of acceptability. A daily score of unity signifies that the simulated value satisfies the limits of acceptability of the corresponding observation, whereas a daily score of zero denotes a nonbehavioral solution.

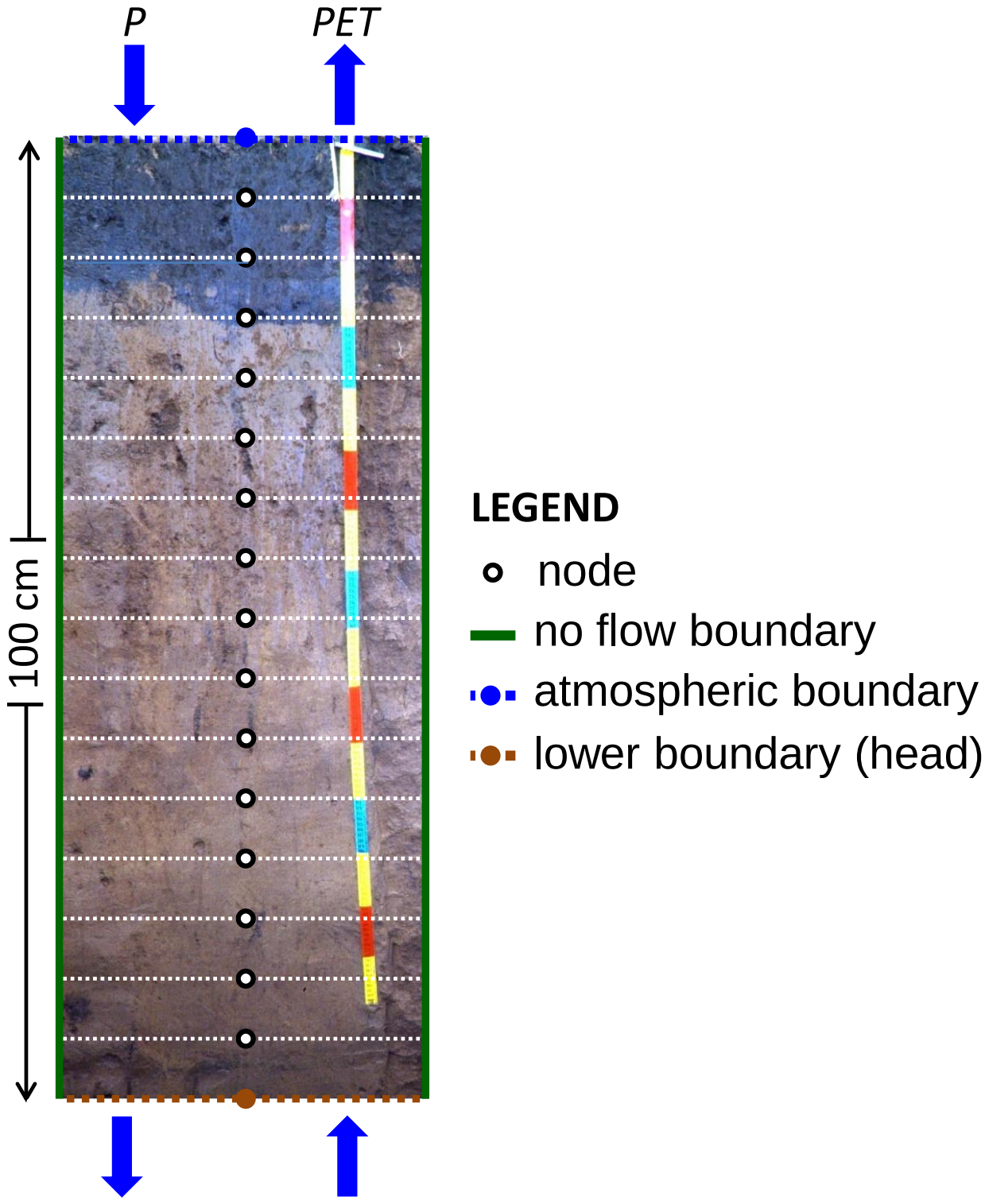

Figure 9 Schematic representation of the HYDRUS-1D model setup for the experimental field plot near Jülich, Germany.

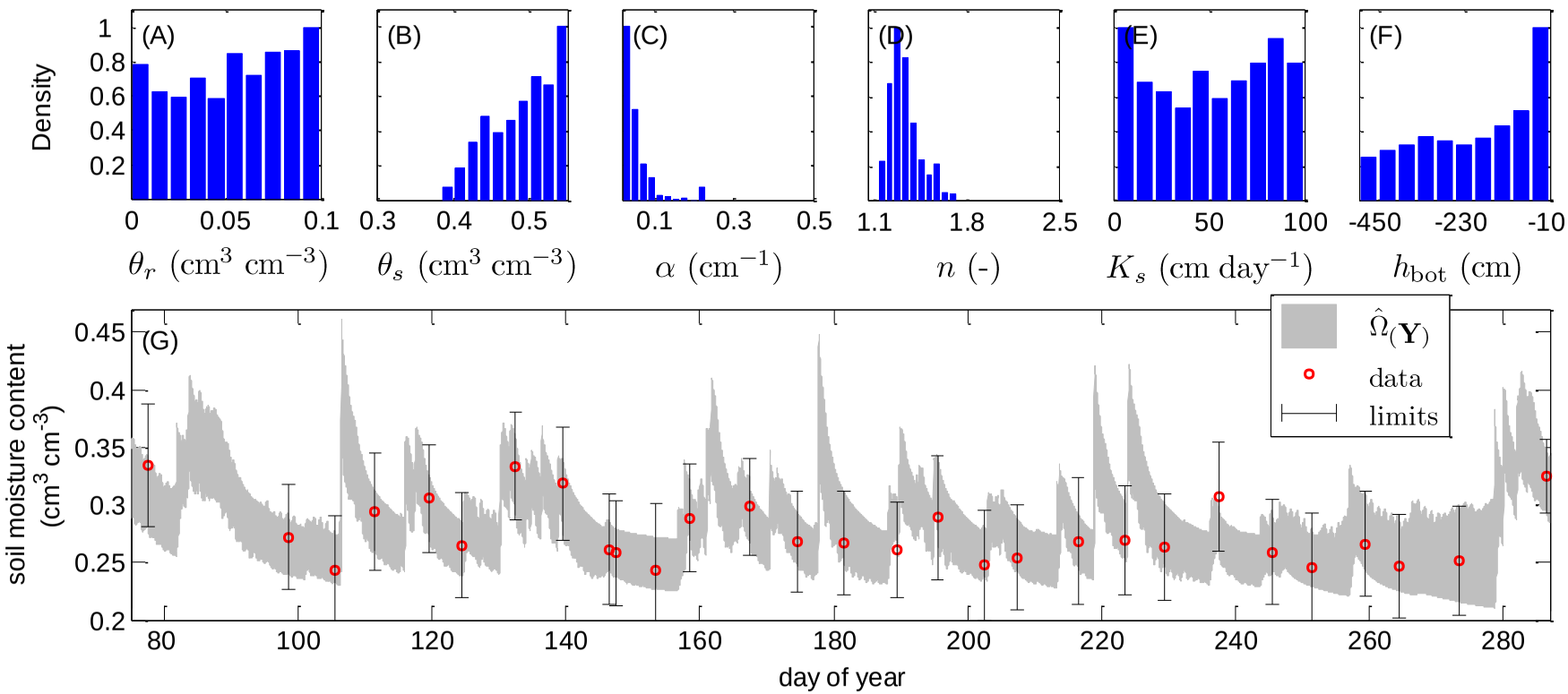

Figure 10 Results of case study III: The HYDRUS-1D variably saturated flow model. (A) Histograms of the behavioral parameter set, $\hat{\Omega}_{(\boldsymbol{\theta}|\tilde{\mathbf{Y}})}$ of the soil hydraulic parameters, (A) θ_r, (B) θ_s, (C) α, (D) n, (E) K_s, and (F) h_{bot}. Each x-axis matches exactly the (uniform) prior distribution. (G) Comparison of observed (red dots) and posterior simulated, $\hat{\Omega}_{(\mathbf{Y})}$ (grey region) soil moisture content.

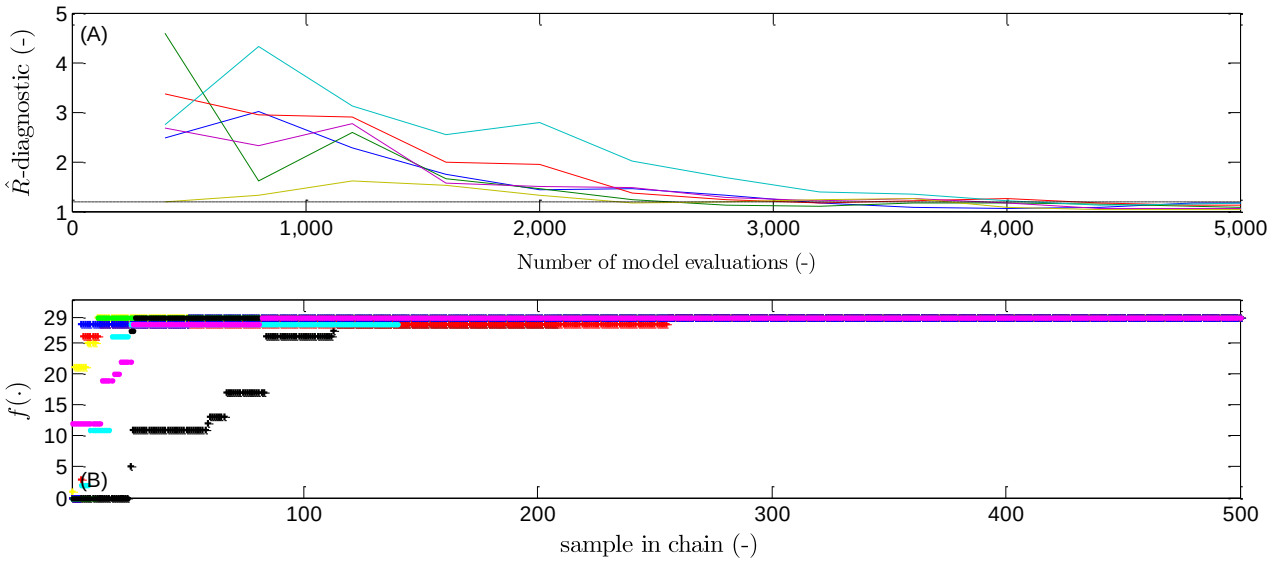

Figure 11 Results of case study III: The HYDRUS-1D variably saturated flow model. Trace plots of the (A) $\hat{R}$-convergence diagnostic of *Gelman and Rubin* (1992), and (B) sampled fitness values in each of the different Markov chains simulated with DREAM$_{(\mathrm{ABC})}$. The parameters and chains are coded with a different symbol and color.

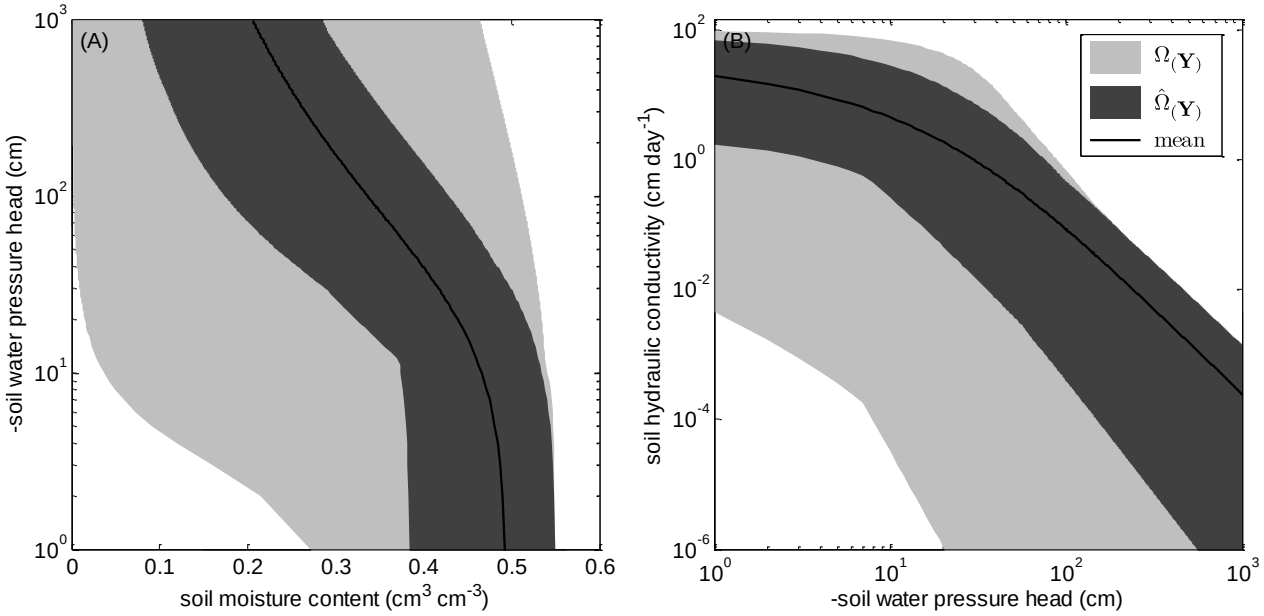

Figure 12 Results of case study III: The HYDRUS-1D variably saturated flow model. Comparison of the prior (dark grey) and posterior (light grey) ranges of the (A) soil water retention, and (B) unsaturated soil hydraulic conductivity function. The black line plots the posterior (behavioral) mean hydraulic functions.