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

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

December 20, 2017

[*]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* (2016) to delineate and sample the behavioural 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*, 2014). 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. In particular it is shown how lack of adequate sampling in the model space might lead to unjustified model rejection.

Keywords: GLUE, Limits of Acceptability, Markov Chain Monte Carlo, Posterior Sampling, DREAM, DREAM$_{(\mathrm{LOA})}$, Sufficiency, Hydrological modelling

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. We have to decide on the model structure or structures to be considered; on the prior distributions for the parameters and/or input data that will be considered uncertain; on how to treat residual errors and a likelihood (or fuzzy membership) to express the degree of belief in a model realization; on a sampling method to generate those realizations; and on a way of combining likelihood measures if necessary.

None of these choices are simple and some have proven to be highly contentious in the hydrological literature. 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-ideal 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*, 2016; *Beven and Smith*, 2015; *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.* (2014) for flood inundation modelling 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, 2016). 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 large, 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* (2016)). 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 improvement over the Shuffled Complex Evolution Metropolis (*Vrugt et al.*, 2003) algorithm and has the advantage of maintaining de-

tailed balance and ergodicity. Benchmark experiments have shown that DREAM is superior to other adaptive MCMC sampling approaches (for instance see *Lu et al.* (2017)), and in high-dimensional spaces even provides better solutions than powerful optimisation 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* (2016) 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 has much wider applicability and can solve inference problems involving the use of 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), model averaging (*Vrugt et al.*, 2008b) and GLUE Limits of Acceptability *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 simulations. 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 simulated values. In a previous paper *Sadegh and Vrugt* (2013) have shown that the Limits of Acceptability framework of GLUE has important elements in common with approximate Bayesian computation (ABC), particularly if each observation of the calibration data record is used as a summary statistic.

This paper illustrates some recent developments to the DREAM toolbox of *Vrugt* (2016) in MATLAB to delineate and sample the behavioural solution space of set-theoretic likelihood measures used within the Limits of Acceptability framework (*Beven*, 2006; *Beven and Binley*, 2014). 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 organised as follows. Section 2 summarises 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 briefly the DREAM$_{(\mathrm{ABC})}$ algorithm of *Sadegh and Vrugt* (2014) and introduces DREAM$_{(\mathrm{LOA})}$ which is designed to sample efficiently the behavioural parameter space within the Limits of Acceptability framework. In this section we are particularly concerned with the definition of the likelihood function and Metropolis acceptance probability so as not to violate detailed balance and to make sure that the behavioural 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 modelling. 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 Model Formulation

Consider a n-vector of measurements, $\widetilde{\mathbf{y}} = \{\widetilde{y}_1, \ldots, \widetilde{y}_n\}$, observed at discrete times, $t = \{1, \ldots, n\}$, which summarizes the response of some (spatially distributed) real-world system, $\Im$, subject to q control inputs, $\mathbf{b} = \{b_1, \ldots, b_q\}$, that may be time and/or space variant. We can use a computer model, $\mathcal{M}(\cdot)$, to emulate $\Im$ and explain the experimental data

$$\widetilde{\mathbf{y}} \leftarrow \mathcal{M}(\mathbf{x}, \boldsymbol{\alpha}, \widetilde{\boldsymbol{\psi}}_0, \widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) + \mathbf{e}, \tag{1}$$

where $\mathbf{x} = \{x_1, \ldots, x_d\}$ is a $1 \times d$-vector of parameters, $\boldsymbol{\alpha}$ signifies a vector with fundamental constants (e.g. gravitational acceleration, light velocity) and/or measurable invariant quantities (surface tension), $\widetilde{\boldsymbol{\psi}}_0$ stores the values of the state variables at the start of simulation, matrix $\widetilde{\mathbf{A}}$ characterizes distributed system properties (e.g. subsurface heterogeneity, topography), $\widetilde{\mathbf{B}}$ is the control matrix with (spatio)temporal measurements of the q forcing variables, and $\mathbf{e} = \{e_1, \ldots, e_n\}$ is a vector of residuals. The residuals may constitute measurement errors on $\widetilde{\mathbf{y}}$, or the effects of model structural errors in $\mathcal{M}(\cdot)$, or the effects of input data errors in $\widetilde{\boldsymbol{\psi}}_0$, $\widetilde{\mathbf{A}}$, and $\widetilde{\mathbf{B}}$ (these are processed through the model to contribute to the residual), or a combination thereof, in which case we can write $e_t = e_{1t} + e_{2t} + e_{3t}$. The index t for time takes on strictly positive integer values in the remainder of this paper, $t \in \mathbb{N}_+$, yet may take on real values, $t \in (0, n] \in \mathbb{R}_+$, in $\mathcal{M}(\cdot)$ to simulate continuous-time processes.

The aim of this paper is to determine our posterior beliefs about the model parameters, $\mathbf{x}$, in light of the computer model, $\mathcal{M}(\cdot)$, transient control variables, $\widetilde{\mathbf{A}}$, input data, $\widetilde{\mathbf{B}}$, experimental data, $\widetilde{\mathbf{y}}$, prior beliefs about the parameters, $P(\mathbf{x})$ and measurement and modelling errors, $\mathbf{e}$. The prevailing Bayesian approach would require a statistical model of the measurement errors of the transient control variables, $\widetilde{\mathbf{A}}$, and other model inputs, $\widetilde{\mathbf{B}}$, and demand assumptions about measurement errors of $\widetilde{\mathbf{y}}$, in pursuit of an adequate likelihood function (*Kavetski et al.*, 2006a,b; *Vrugt et al.*, 2008a, 2009). Instead, we adopt an alternative approach and quantify our posterior

beliefs of $\mathbf{x}$ via Limits of Acceptability on the observed data. These limits are defined a-priori by the modeller and summarize the cumulative impacts of measurement errors of $\widetilde{\mathbf{A}}$, $\widetilde{\mathbf{B}}$ and $\widetilde{\mathbf{y}}$ on the simulated output. The second author of this paper is a strong proponent of this methodology, with philosophy, arguments, justification and methodology well rehearsed in past papers for over a decade and discussed briefly in the next section. Without loss of generality, we restrict the model parameters to a closed space, $\boldsymbol{\chi}$, equivalent to a d-dimensional hypercube, $\mathbf{x} \in \boldsymbol{\chi} \in \mathbb{R}^d$, called the feasible parameter space. Furthermore, as the simulation models used herein exhibit degenerative (negative) feedbacks, we take advantage of a spin-up period of T days to gravitate the moisture status to a stable state and remove the impact of state initialization errors on the model output, $\lim_{t \to T} \left(\mathcal{M}_t(\mathbf{x}, \boldsymbol{\alpha}, \widetilde{\boldsymbol{\psi}}_0, \widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) - \mathcal{M}_t(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\psi}_0, \widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}) \right) \to 0$.

3 The Generalised Likelihood Uncertainty Estimation (GLUE) methodology

The GLUE methodology has been applied widely to many different modelling 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 non-traditional (*Beven and Binley*, 1992; *Beven and Freer*, 2001; *Beven*, 2006, 2009, 2016). These are the non-ideal cases that are difficult to represent using statistical residual models and that require a different philosophical approach to model evaluation to traditional statistical methods (*Beven*, 2016; *Beven and Smith*, 2015).

The GLUE methodology aims to find a set of model representations (model inputs, model structures, model parameter sets, model errors) that are behavioural in the sense of being acceptably consistent with the (non-error-free) observations. Such models are not necessarily limited to a small region of the model space. This is the equifinality thesis (*Beven*, 2006, 2012). Predictions are made using this ensemble of behavioural models, weighted according to some likelihood measure supporting a degree of belief. Given an expectation of complex error structures in hydrological modelling (*Beven*, 2016), the likelihood weight need not be defined by a simple statistical error model. Here it is based on performance relative to Limits of Acceptability defined prior to making any model runs, which allows the residuals to be treated implicitly. This approach was originally 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 inference and propagation of uncertainty.

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. These Limits of Acceptability are defined prior to running the model, and 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*, 2016). To allow for the fact that different observations might have quite different scales, the Limits of Acceptability can be expressed as a normalised scale (-1 at the lower limit, 0 at the observed value, +1 at the upper limit). Performance weightings within the limits can also be specified as appropriate (*Beven*, 2006).

The GLUE Limits of Acceptability method proceeds as follows. The index i is used to mean '$for\ all\ i \in \{1, \ldots, N\}$'.

1. Draw at random N samples from the prior parameter distribution, $P(\mathbf{x})$, and store these realizations in a $N \times d$ matrix $\mathbf{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$.

2. Evaluate the model, $\mathbf{y}^{(i)} \leftarrow \mathcal{M}(\mathbf{x}^{(i)}|\cdot)$, and compute the minimum absolute normalised score for the simulation, $\mathbf{y}^{(i)} = \{y_1^{(i)}, \ldots, y_n^{(i)}\}$, to be acceptable.

3. Rank the N parameter vectors by their minimum scores, and select as behavioural the top R realisations above some acceptability threshold. This threshold would normally be an absolute value of 1 on the normalized scale, indicating that all observations are reproduced within the specified Limits of Acceptability. All other realisations are given a likelihood value of zero.

4. Collect the behaviorial solutions in a $R \times d$ matrix $\mathbf{B}$ and store in a matrix $\mathbf{Y}$ of size $R \times n$ their corresponding simulations.

5. Calculate a likelihood value, $L(\mathbf{x}^{(i)}|\widetilde{\mathbf{y}})$, of the simulated values, $\mathbf{y}^{(i)}$, based on the performance weightings within the Limits of Acceptability. The way in which this is done will depend on the nature of the application (see the suggestions in *Beven* (2006)).

6. Normalize the likelihood of each sample of $\mathbf{B}$

$$\bar{L}(\mathbf{B}_r|\widetilde{\mathbf{y}}) = L(\mathbf{B}_r|\widetilde{\mathbf{y}}) / \sum_{r=1}^{R} L(\mathbf{B}_r|\widetilde{\mathbf{y}}), \tag{2}$$

   where $r = \{1, \ldots, R\}$ so that $\sum_{r=1}^{R} \bar{L}(\mathbf{B}_r|\widetilde{\mathbf{y}}) = 1$.

7. Compute the likelihood-weighted cumulative density function (cdf) by assigning each rth row of $\mathbf{Y}$ the likelihood $\bar{L}(\mathbf{B}_r|\widetilde{\mathbf{y}})$, where $r = \{1, \ldots, R\}$.

8. Derive the 95% simulation uncertainty ranges of $\mathcal{M}(\mathbf{x}|\cdot)$ from the likelihood-weighted cdf.

Past work has applied 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; *Westerberg and McMillan*, 2015; *Gupta et al.*, 2008; *Vrugt and Sadegh*, 2013a; *Sadegh and Vrugt*, 2014; *Sadegh et al.*, 2015). Some earlier publications used similar ideas within GLUE based on fuzzy measures, for which the support also acted as Limits of Acceptability (*Page et al.*, 2003; *Freer et al.*, 2004; *Page et al.*, 2004, 2007; *Pappenberger et al.*, 2005, 2007). The set-theoretic approach used by *Keesman* (1990) and *van Straten and Keesman* (1991) is a similar method of model evaluation. 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 should be defined on the basis of best available hydrological knowledge.

Two primary sources of epistemic uncertainty will influence the Limits of Acceptability: uncertainty in the evaluation observations, $\widetilde{\mathbf{y}}$ (e.g. discharge, water table or soil moisture observations), and uncertainty in the model input data, $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}$. The former is generally easiest to handle in that there will be a direct relationship between observed and predicted variables. For example, the measured discharge will be subject to considerable uncertainty due to imperfect knowledge of the rating curve. Fortunately, this curve can be reasonably well estimated via statistical regression (*Krueger et al.*, 2009; *McMillan et al.*, 2012), fuzzy regression (*Blazkova and Beven*, 2009), or other approaches such as the Monte Carlo based voting point method of

McMillan and Westerberg (2015), with or without a water balance constraint (*Holloway et al.*, 2017).

It remains difficult and subjective how errors in the input data, $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}$, should affect the Limits of Acceptability (*Beven and Smith*, 2015). Not only are such errors difficult to estimate a-priori, but their effect will also accumulate in the modeled state variables and produce a complex, non-traditional, time series of residuals. The Limits of Acceptability should be extended to account for input data errors, but at present there is no commonly accepted framework for doing so beyond special synthetic cases with a-priori known error sources and properties. Thus, the degree of extension remains necessarily subjective, and possibly guided by the requirements of an application in evaluating model simulations as fit-for-purpose (*Beven*, 2017). This may purport a methodological flaw, yet is the consequence of the inexact nature of hydrological science.

This paper is not about whether GLUE is a valid choice of methodology, only about the efficiency of applying that methodology. We recognise that all estimates of predictive uncertainty will be conditional on the assumptions made, and therefore care should be exercised when interpreting and communicating the resulting prediction estimates, for example using the condition tree proposal of *Beven et al.* (2014).

The GLUE approach has mostly used simple randomised 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 may only provide a sparse sample of the behavioural solution space in high parameter spaces (large d) 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. This is especially the case when there is little information about the prior distributions of the parameters and only feasible ranges can be specified. Uniform random sampling over the hypercube defined by the parameter ranges 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. In the original GLUE paper, *Beven and Binley* (1992) used a nearest neighbour resampling method to replace samples with low likelihoods, nevertheless, this approach lacks statistical rigor.

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. This paper has received a significant number of citations but the proposed MCMC sampling framework has found little use in the GLUE community, despite source code availability. 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 (GLUE_LoA) framework. This extension demands changes to the sampling approach of *Blasone et al.* (2008) to satisfy efficiently the Limits of Acceptability in pursuit of the behavioral parameter space. A simple adaptation of the DREAM$_{(\mathrm{ABC})}$ algorithm of *Sadegh and Vrugt* (2014) developed in the context of diagnostic model evaluation will suffice to solve set-theoretic membership functions such as those used in the Limits of Acceptability methodology.

4 GLUE_LoA and Approximate Bayesian Computation

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

$$\widehat{\mathbf{\Omega}}_{(\mathbf{y})} = \left\{ \mathbf{y} \in \mathbb{R}^n \ : \ y_t = \mathcal{M}_t(\mathbf{x}|\cdot) \ ; \ \mathbf{x} \in \widehat{\mathbf{\Omega}}_{(\mathbf{x}|\widetilde{\mathbf{y}})} \ , \ t = 1, \ldots, n \right\}, \tag{3}$$

where $\widehat{\mathbf{\Omega}}_{(\mathbf{x}|\widetilde{\mathbf{y}})}$ constitutes the posterior (behavioural) parameter set

$$\widehat{\mathbf{\Omega}}_{(\mathbf{x}|\widetilde{\mathbf{y}})} = \mathbf{\Omega}_{(\mathbf{x}|\widetilde{\mathbf{y}})}. \tag{4}$$

The conditional parameter set, $\mathbf{\Omega}_{(\mathbf{x}|\widetilde{\mathbf{y}})}$, is defined as follows

$$\mathbf{\Omega}_{(\mathbf{x}|\widetilde{\mathbf{y}})} = \left\{ \mathbf{x} \in \boldsymbol{\chi} \in \mathbb{R}^d \ : \ \widetilde{y}_t - \mathcal{M}_t(\mathbf{x}|\cdot) = e_t \ ; \ e_t \in [-\epsilon_t, \epsilon_t] \ , \ t = 1, \ldots, n \right\}, \tag{5}$$

and contains solutions that satisfy the Limits of Acceptability of each observation, and $\mathbf{x}^* \in \widehat{\mathbf{\Omega}}_{(\mathbf{x}|\widetilde{\mathbf{y}})}$. If an informative prior distribution is used then the behavioural (posterior) parameter set, $\widehat{\mathbf{\Omega}}_{(\mathbf{x}|\widetilde{\mathbf{y}})}$, is computed as the intersection of the prior parameter set, $\mathbf{\Omega}_{(\mathbf{x})}$, and conditional parameter set, $\mathbf{\Omega}_{(\mathbf{x}|\widetilde{\mathbf{y}})}$, or

$$\widehat{\mathbf{\Omega}}_{(\mathbf{x}|\widetilde{\mathbf{y}})} = \mathbf{\Omega}_{(\mathbf{x})} \cap \mathbf{\Omega}_{(\mathbf{x}|\widetilde{\mathbf{y}})}. \tag{6}$$

Figure 1 summarises graphically four different outcomes of the Limits of Acceptability framework. The behavioural solution space exists, if and only if, the conditional parameter set, $\mathbf{\Omega}_{(\mathbf{x}|\widetilde{\mathbf{y}})}$, intersects the prior parameter set, $\mathbf{\Omega}_{(\mathbf{x})}$. If an informative prior distribution is used, then a sufficient condition for the posterior (behavioural) parameter set to exist is that the conditional parameter set, $\mathbf{\Omega}_{(\mathbf{x}|\widetilde{\mathbf{y}})}$, is non-empty.

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) (coincidentally in different departments of Lancaster University where, quite 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 receiving a surge of interest in statistics (*Marjoram et al.*, 2003; *Sisson et al.*, 2007; *Joyce and Marjoram*, 2008; *Grelaud et et al.*, 2009; *Del Moral et al.*, 2012) with common applications in genetics (*Pritchard et al.*, 1999; *Tanaka et al.*, 2006; *Ratmann et al.*, 2009; *Beaumont*, 2010), epidemiology (*Blum and Tran*, 2010), population biology (*Bertorelle et al.*, 2010), (evolutionary) ecology (*Beaumont*, 2010; *Csilléry et al.*, 2010), and psychology (*Turner and van Zandt*, 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 $\mathbf{x}^*$ should be a sample from the posterior distribution if its simulated output matches closely and consistently the observed data. Or in

ABC-terminology, the distance, $\rho\big(\widetilde{\mathbf{y}}, \mathbf{y}(\mathbf{x}^*)\big)$, between the simulated and observed data must be less than some nominal value, ϵ (*Marjoram et al.*, 2003; *Sisson et al.*, 2007). Thus, ABC methods do not use a formal likelihood function to infer the posterior parameter distribution, but instead retain a proposal, $\mathbf{x}^*$, if

$$\rho\big(\widetilde{\mathbf{y}}, \mathbf{y}(\mathbf{x}^*)\big) \leq \epsilon, \tag{7}$$

where $\rho(a, b) = |a - b|$ is the distance function and $|\cdot|$ signifies the modulus (absolute value) operator. The most basic ABC algorithm (rejection sampling) then proceeds in the following steps

1. Draw a proposal, $\mathbf{x}^*$, from the prior distribution, $P(\mathbf{x})$.
2. Simulate the model output, $\mathbf{y} \leftarrow \mathcal{M}(\mathbf{x}^*|\cdot)$
3. Accept $\mathbf{x}^*$ if $\rho\big(\widetilde{\mathbf{y}}, \mathbf{y}(\mathbf{x}^*)\big) \leq \epsilon$

The accepted samples will approximate the posterior distribution, $P(\mathbf{x}|\rho\big(\widetilde{\mathbf{y}}, \mathbf{y}(\mathbf{x}^*)\big) \leq \epsilon)$. When $\epsilon \to 0$ the rejection algorithm provides samples from the exact posterior distribution, $P(\mathbf{x}|\widetilde{\mathbf{y}})$, whereas if $\epsilon \to \infty$ the algorithm would generate draws from the prior distribution, $P(\mathbf{x})$. The tolerance, ϵ can therefore be considered a trade-off between computational tractability and accuracy (*Wilkinson*, 2013).

The probability of stumbling upon a simulation that satisfies exactly all n data points of $\widetilde{\mathbf{y}}$ within a small tolerance, ϵ, decreases rapidly with increasing model complexity and length of the data set. To mitigate this problem, it is common practice to replace $\widetilde{\mathbf{y}}$ with one or more summary statistics, $S(\widetilde{\mathbf{y}})$, of the data. These statistics summarize in much lower dimensions the relevant information in $\widetilde{\mathbf{y}}$. Samples are retained if their simulated statistics reside within ϵ of their observed values, or $\rho\left(S\big(\widetilde{\mathbf{y}}\big), S\big(\mathbf{y}(\mathbf{x}^*)\big)\right) \leq \epsilon$ (*Vrugt and Sadegh*, 2013a; *Sadegh and Vrugt*, 2014; *Sadegh et al.*, 2015). If the summary statistics are sufficient and contain as much information as the data, $\widetilde{\mathbf{y}}$, itself then this approach does not introduce errors. However, in practice, models and data of complex systems rarely admit sufficient statistics.

One attractive feature of summary statistics is that they can reduce significantly the impact of poorly known error sources on model and parameter estimation. A textbook example is the runoff index of a catchment. This metric is hardly sensitive to precipitation data measurement errors that otherwise would lead to a complex, non-traditional, time series of discharge residuals (*Vrugt and Sadegh*, 2013a). What is more, summary statistics are useful for hypothesis testing, and temporal analysis of their values can help detect system (catchment) nonstationarity (*Sadegh et al.*, 2015).

In a previous paper, *Sadegh and Vrugt* (2013) have shown an equivalence of 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

$$\rho\left(S\big(\widetilde{\mathbf{y}}\big), S\big(\mathbf{y}(\mathbf{x}^*)\big)\right) = \prod_{t=1}^{n} I\big(|\widetilde{y}_t - y_t(\mathbf{x}^*)| \leq \epsilon_t\big), \tag{8}$$

where $I(a)$ is an indicator function that returns one if the condition a is satisfied and zero otherwise, ϵ_t constitutes the Limits of Acceptability of the tth observation, and $\boldsymbol{\epsilon} = \{\epsilon_1, \ldots, \epsilon_n\}$.

Nevertheless, a fundamental difference between ABC and the Limits of Acceptability framework is that ABC assumes use of a stochastic model so that repeated runs with the exact same

parameter values will produce a range of possible simulations. Otherwise, the posterior distribution, $P(\mathbf{x}|\rho(\widetilde{\mathbf{y}}, \mathbf{y}(\mathbf{x}^*)) \leq \epsilon)$ cannot converge to a stable distribution in the limit of $\epsilon \to 0$, and instead would converge to a single solution (if it existed). For ABC to work with a deterministic model, we must corrupt the simulated output with a draw from $P_{\mathbf{e}}(\cdot)$, a n-variate distribution with probabilistic properties (e.g. mean, variance, correlation structure, bias) equivalent to the residual time series, $\mathbf{e}$, in Equation (1). Thus, while traditional Bayesian approaches draw inferences on the posterior beliefs of $\mathbf{x}$ via a prior distribution, $P(\mathbf{x}$, and likelihood function, $L(\mathbf{x}|\widetilde{\mathbf{y}})$, which summarises the expected statistical properties of the residuals, ABC methods approximate the likelihood function by repeated numerical simulation, the outcomes of which are compared with the observed data (*Beaumont*, 2010; *Bertorelle et al.*, 2010; *Csilléry et al.*, 2010). If, the Limits of Acceptability are rather large in comparison to the residuals as in *Sadegh and Vrugt* (2013) then perturbation of the simulated values will have only a minimal effect on the posterior parameter distribution.

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 behavioural model realisations, $\widehat{\Omega}_{(\mathbf{x}|\widetilde{\mathbf{y}})}$ that satisfies $\rho(a, b) = 1$ in Equation (8). Commonly used (population Monte Carlo) rejection sampling methods are rather inefficient in locating behavioural 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 (behavioural) 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 (5).

In DREAM$_{(\mathrm{ABC})}$, N ($N > 2$) different Markov chains are run simultaneously in parallel, and multivariate proposals are generated on the fly from the collection of chain states, $\mathbf{X} = \{\mathbf{x}_{t-1}^{(1)}, \ldots, \mathbf{x}_{t-1}^{(N)}\}$ (matrix of $N \times d$ 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 ith chain, $i = \{1, \ldots, N\}$, at iteration $t = \{2, \ldots, T\}$ is calculated using

$$\begin{aligned} \Delta \mathbf{x}_A^{(i)} &= \boldsymbol{\zeta}_{d^*} + (\mathbf{1}_{d^*} + \boldsymbol{\lambda}_{d^*}) \gamma_{(\delta, d^*)} \sum_{j=1}^{\delta} \left(\mathbf{X}_A^{\mathbf{r}_{1j}} - \mathbf{X}_A^{\mathbf{r}_{2j}} \right) \\ \Delta \mathbf{x}_{\neq A}^{(i)} &= 0, \end{aligned} \tag{9}$$

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 $\mathbf{r}_1$ and $\mathbf{r}_2$ are δ-vectors with integer values drawn without replacement from $\{1, \ldots, i-1, i+1, \ldots, N\}$. 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. Every fifth 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 the ith chain at iteration t then becomes

$$\mathbf{x}_{\mathrm{p}}^{(i)} = \mathbf{x}^{(i)} + \Delta\mathbf{x}^{(i)}, \tag{10}$$

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

$$P_{\mathrm{acc}}(\mathbf{x}_{t-1}^{(i)} \rightarrow \mathbf{x}_{\mathrm{p}}^{(i)}) = \begin{cases} I\big(f(\mathbf{x}_{\mathrm{p}}^{(i)}) \geq f(\mathbf{x}_{t-1}^{(i)})\big) & \text{if } f(\mathbf{x}_{\mathrm{p}}^{(i)}) < n \\ 1 & \text{if } f(\mathbf{x}_{\mathrm{p}}^{(i)}) = n \end{cases}, \tag{11}$$

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

$$f(\mathbf{x}) = \sum_{t=1}^{n} I(|\widetilde{y}_t - y_t(\mathbf{x})| \leq \epsilon_t). \tag{12}$$

If the proposal is accepted, then the ith chain moves to this new position, $\mathbf{x}_t^{(i)} = \mathbf{x}_{\mathrm{p}}^{(i)}$, otherwise it remains at its current location, that is $\mathbf{x}_t^{(i)} = \mathbf{x}_{t-1}^{(i)}$.

The fitness of a parameter vector thus equates to the number of observations its simulation satisfies within the Limits of Acceptability. A proposal, $\mathbf{x}_{\mathrm{p}}^{(i)}$, in chain i is accepted, $P_{\mathrm{acc}}(\mathbf{x}_{t-1}^{(i)} \rightarrow \mathbf{x}_{\mathrm{p}}^{(i)}) = 1$, if its fitness is higher than, or equal to, that of the current state of the ith chain, $\mathbf{x}_{t-1}^{(i)}$, or, if its fitness is equal to n, and thus, $\mathbf{y}(\mathbf{x}_{\mathrm{p}}^{(i)})$ is contained in the interval $[\widetilde{y}_t - \epsilon_t, \widetilde{y}_t + \epsilon_t]$ for all $t \in \{1, \ldots, n\}$, otherwise $\mathbf{x}_{\mathrm{p}}^{(i)}$ is rejected. After a burn-in period in which $f(\mathbf{x}) < n$, the convergence of DREAM$_{(\mathrm{ABC})}$ can be monitored with the univariate, $\widehat{R}$, and multivariate, $\widehat{R}^d$, scale reduction factors of *Gelman and Rubin* (1992) and *Brooks and Gelman* (1998), respectively. The weight of each simulation in the behavioral space, $\widehat{\mathbf{\Omega}}_{(\mathbf{y})}$, is thus proportional to the number of times its parameter vector appears in the posterior distribution sampled by the joint chains. Interested readers are referred to *Sadegh and Vrugt* (2014) for a full description of the DREAM$_{(\mathrm{ABC})}$ algorithm.

The DREAM$_{(\mathrm{ABC})}$ algorithm was originally designed to speed up ABC inference. In our present application, we use the algorithm to sample efficiently the behavioural parameter space conditional on the Limits of Acceptability. To be comparable to GLUE (section 3) this necessitates use of a deterministic model in DREAM$_{(\mathrm{ABC})}$. To make this distinction obvious, we therefore introduce a new member of the DREAM family, coined DREAM$_{(\mathrm{LOA})}$, which is equivalent to DREAM$_{(\mathrm{ABC})}$ but with the use of a deterministic model.

Appendix A presents a basic implementation of the DREAM$_{(\mathrm{LOA})}$ algorithm in MATLAB. The results presented herein are derived from the MATLAB toolbox of DREAM, which includes a much wider arsenal of options and capabilities (such as postprocessing and multi-core computation). A detailed description of this toolbox appears in *Vrugt* (2016).

6 Numerical Examples

Three different numerical examples are considered to illustrate the ability of the DREAM$_{(\mathrm{LOA})}$ algorithm to sample efficiently the behavioural parameter, $\widehat{\mathbf{\Omega}}_{(\mathbf{x}|\widetilde{\mathbf{y}})}$, and simulation, $\widehat{\mathbf{\Omega}}_{(\mathbf{y})}$, space that satisfy the prior parameter distribution and Limits of Acceptability of each observation. All the examples assume a non-informative and independent prior distributions, and default values of the algorithmic parameters of DREAM$_{(\mathrm{ABC})}$ listed by *Sadegh and Vrugt* (2014).

6.1 Unit Hydrograph

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

$$Q_t = \frac{1}{k\Gamma(m)} \left(\frac{t}{k}\right)^{(m-1)} \exp\left(-\frac{t}{k}\right), \tag{13}$$

where Q_t (mm day^{-1}) is the simulated streamflow at time t (days), $m \in [1, \infty)$ (-) denotes the number of reservoirs, $k > 0$ (days) signifies the recession constant, and $\Gamma(\cdot)$ is the gamma function

$$\Gamma(m) = \int_0^\infty x^{m-1} \exp(-x) \mathrm{d}x \qquad \forall m \in \mathbb{Z}_+ \tag{14}$$

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

A $n = 25$ - day period with synthetic daily streamflow data was generated by driving Equation (13) with an artificial precipitation record using $m = 2$ reservoirs, and a recession constant of $k = 4$ 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 study forcing data and model structure are assumed to be known accurately. The DREAM$_{(\mathrm{LOA})}$ algorithm then uses the observed discharge record, $\widetilde{\mathbf{y}} = \{\widetilde{y}_1, \ldots, \widetilde{y}_{25}\}$ to estimate the behavioural solution space of m and k using the Limits of Acceptability, $\epsilon_t = 0.2\widetilde{y}_t \; \forall t \in \{1, ..., 25\}$. A bivariate uniform prior distribution, $\mathcal{U}_2[1, 10]$ was used for m and k in the calculations. Appendix B presents a MATLAB implementation of Equation (13) and lists an input file of the DREAM$_{(\mathrm{LOA})}$ algorithm with the setup and data used in this case study.

Figure 3 summarises 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 (grey). These simulated data satisfy the Limits of Acceptability of each observation and thus belong to the behavioural set, $\widehat{\mathbf{\Omega}}_{(\mathbf{y})}$. The two Figures at the right-hand-side plot histograms of the behavioural parameter space of m and k respectively. The true parameter values used to generate the synthetic data are separately indicated with the red 'X' symbol. The behavioural 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 centre around their "true" values but appears a little biased (to the left) for parameter m.

To provide insights into the convergence rate of DREAM$_{(\mathrm{LOA})}$ to the posterior set, $\widehat{\mathbf{\Omega}}_{(\mathbf{x}|\widetilde{\mathbf{y}})}$, Figure 4 plots trace plots of the $\widehat{R}$-convergence diagnostic of *Gelman and Rubin* (1992) computed using the samples in the second half of the $N = 8$ different Markov chains. About $2,000$ function evaluations are required to satisfy the convergence threshold of $\widehat{R} \leq 1.2$. The acceptance rate of proposals is equivalent to about 33% (not shown herein), which means that, on average, every third proposal of DREAM$_{(\mathrm{LOA})}$ 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 k and m (see Figure 5). This conjecture is confirmed by numerical simulation. Only 28 samples (indicated with blue dots) were deemed behavioural 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$_{(\mathrm{LOA})}$. This difference in sampling efficiency between DREAM$_{(\mathrm{LOA})}$ and uniform random (rejection) sampling is clearly evident in Figure 5. Not only does DREAM$_{(\mathrm{LOA})}$ produce many diverse samples of $\widehat{\mathbf{\Omega}}_{(\mathbf{x}|\tilde{\mathbf{y}})}$, the posterior parameter set, the algorithm also sharply delineates the behavioural solution space.

In this trivial example it is quite easy to do generate many millions of samples from the uniform prior distribution to compensate for a poor sampling efficiency, nevertheless, 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$_{(\mathrm{LOA})}$ can help to avoid the rejection of models in high-dimensional parameter spaces as a result of sparse and inadequate sampling. Past work has also shown how the DREAM$_{(\mathrm{LOA})}$ methodology can be successful in identifying multiple regions of behavioural models (*Sadegh et al.*, 2015).

6.2 Rainfall-Runoff Modelling

The second case study involves the modelling of the rainfall-runoff transformation of the Leaf River watershed in Mississippi. This temperate 1944 km^2 watershed has been studied extensively in the hydrological 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 initialisation.

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 throughout 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 hydrological 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 modelling 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 summarises 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 $\widehat{\boldsymbol{\sigma}}_{\widetilde{\mathbf{y}}} = \{\widehat{\sigma}_{\widetilde{y}_1}, \ldots, \widehat{\sigma}_{\widetilde{y}_n}\}$. The n-values of $\widehat{\boldsymbol{\sigma}}_{\widetilde{\mathbf{y}}}$ were derived using the nonparametric estimator by *Vrugt et al.* (2005) and shown to be on the order of $0.1\widetilde{y}_t$. The Limits of Acceptability in Equation (8) are now computed as multiple of $\widehat{\boldsymbol{\sigma}}_{\widetilde{\mathbf{y}}}$ or $\boldsymbol{\epsilon} = \phi\widehat{\boldsymbol{\sigma}}_{\widetilde{\mathbf{y}}}$ using $\phi = 4$. This leads to effective observation errors on the order of $\epsilon_t = 0.4\widetilde{y}_t$.

Figure 7 plots traces of the sampled fitness values in a selected set of ten Markov chains simulated with DREAM$_{(\mathrm{LOA})}$. The different chains are coded with a different colour and/or symbol. The chains converge to a stable fitness value of Equation (12) 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$_{(\mathrm{LOA})}$ algorithm, a total of $100,000$ samples were drawn randomly from the ranges listed in Table 1. While this is a small number for $d = 14$ and some past GLUE applications of hydrological models, it is still a large number of model runs and we use it here to make a point. 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$_{(\mathrm{LOA})}$ algorithm. This therefore gives further weight to the argument that adequate sampling is essential to inference using a GLUE_LoA 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 using reasonable Limits of Acceptability or with a simple statistical error model since the same issues apply. One of the advantages of the Limits of Acceptability approach is that it highlights events with problems (rather than just allowing the error variance or event rainfall multiplier distribution to expand to cover such event). Indeed, what is ideally 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 and Smith* (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$_{(\mathrm{LOA})}$ 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 Modelling

The third and last case study considers the modelling 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

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial z}\left[K(h)\left(\frac{\partial h}{\partial z} + 1\right)\right] \tag{15}$$

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 (15) numerically the soil hydraulic properties need to be defined. Here

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

$$\begin{aligned} \theta(h) &= \theta_{\mathrm{r}} + (\theta_{\mathrm{s}} - \theta_{\mathrm{r}})[1 + (\alpha|h|)^n]^{-m} \\ K(h) &= K_{\mathrm{s}} S_{\mathrm{e}}(h)^{\lambda}[1 - (1 - S_{\mathrm{e}}(h)^{1/m})^m]^2, \end{aligned} \tag{16}$$

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

$$S_{\mathrm{e}}(h) = \frac{\theta(h) - \theta_{\mathrm{r}}}{\theta_{\mathrm{s}} - \theta_{\mathrm{r}}}. \tag{17}$$

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 $\mathrm{DREAM}_{(\mathrm{LOA})}$. The aim here is to obtain a simulation of the mean behaviour 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. The prior ranges are taken deliberately large so as not to constrain too much our field scale soil moisture simulations. In this study the Limits of Acceptability, $\boldsymbol{\epsilon} = \{\epsilon_1, \ldots, \epsilon_n\}$, 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 modelling the mean field water content and soil water fluxes. Thus, for the purpose of this study, the Limits of Acceptability, $\boldsymbol{\epsilon} = \{\epsilon_1, \ldots, \epsilon_n\}$, are set equal to half the width of the 95% interval ($= 2\epsilon$) of the distributed moisture content observations. This equates to an average value of epsilon, $\overline{\boldsymbol{\epsilon}} = 0.047$ ($\mathrm{cm}^3\mathrm{cm}^{-3}$). Thus, a HYDRUS-1D run is classified as behavioural if the simulated moisture contents lie within $[\widetilde{y}_t - \epsilon_t, \widetilde{y}_t + \epsilon_t]$ for $t = \{1, ..., n\}$. The behavioral range thus matches exactly the 95% ranges of the distributed moisture content observations at each time. 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 behavioural simulation set, $\widehat{\boldsymbol{\Omega}}_{(\mathbf{y})}$. The observed soil moisture data are indicated separately with red dots. The behavioural HYDRUS-1D model nicely tracks the observed average soil moisture measurements within behavioural simulation space defined in this way. The root mean square error (RMSE) of the behavioural (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*, 2016). The behavioural parameter space of most parameters extend across a large part of their respective prior ranges with marginal distributions that deviate markedly from normality. The VGM parameters θ_{r} and K_{s} and the lower boundary condition, h_{bot}, are not well defined. The poor sensitivity of the simulated moisture contents to θ_{r} is well understood in the absence of a sustained dry period with low soil moisture contents.

Indeed, the information for θ_r appears outside the range of measured soil moisture contents (*Vrugt et al.*, 2001). The large posterior uncertainties of θ_s and K_s are explained by the imposed Limits of Acceptability which promote considerable variation in the hydraulic functions. Furthermore, inference of h_{bot} suffers from a lack of soil moisture observations in the deeper parts of the profile. Interestingly, the behavioral values of α and n are in excellent agreement with their values derived separately from ROSETTA (*Schaap et al.*, 2001) using soil textural data as main input variables (*Scharnagl et al.*, 2011). Pedotransfer functions are, however, derived from small volume sample measurements and may not always be appropriate in simulating field scale behaviour (*Beven and Germann*, 2013).

The acceptance rate of DREAM$_{(LOA)}$ averages about 15.1%. Thus every sixth proposal generated with DREAM$_{(LOA)}$ 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 behavioural. This equates to an acceptance rate of approximately 0.47%. This efficiency, is about 35 times lower than DREAM$_{(LOA)}$, and expected to deteriorate further with increasing dimensionality and size of the parameter space.

To provide further insights into the convergence speed of DREAM$_{(LOA)}$, Figure 11 plots the evolution of the $\widehat{R}$-diagnostic for the six HYDRUS-1D model parameters in the top panel and traces of the sampled fitness values of the $N = 8$ different chains simulated with DREAM$_{(LOA)}$ in the bottom panel. The $\widehat{R}$-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 behavioural 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 $\widehat{R}$-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 (11). This selection rule of proposals directs the DREAM$_{(LOA)}$ algorithm to the posterior parameter set, $\widehat{\mathbf{\Omega}}_{(\mathbf{x}|\tilde{\mathbf{y}})}$ 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 behavioural solution set is empty and the model is rejected (as with the SAC-SMA model in previous study), the DREAM$_{(LOA)}$ 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 behavioural (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 hydrological 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*, 2016). In the GLUE: 20 years on paper, *Beven and Binley* (2014) 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 hydrological 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 behavioural 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 behavioural simulations are found because of inadequate sampling, models might be wrongly rejected. Thus inadequate sampling alone can increase the possibility rejecting a model that would be useful in prediction. In this paper the reversible chain MCMC simulation with the DREAM$_{(\mathrm{LOA})}$ 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$_{(\mathrm{LOA})}$ within the GLUE_LoA framework. The most important results are as follows

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

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

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

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

(5) The use of inferior sampling methods can lead to inexact inference about the behavioural 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$_{(\mathrm{LOA})}$ algorithm. Depending on the initial set of chains, this may still not identify all areas of behavioural models in complex model spaces, but will still be expected to identify regions of behavioural 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 five 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 core implementation of the DREAM$_{(\mathrm{LOA})}$ algorithm in MATLAB. This code can serve as template for users to delineate the behavioural parameter space for set-theoretic likelihood functions and Limits of Acceptability sampling. Symbols and notation match, in so far possible, variables used in the main text. The variable `x` stores the parameter vector, and `X` signifies the $N \times d$ matrix with the state of the chains, $\{\mathbf{x}_{t-1}^{(i)}, \ldots, \mathbf{x}_{t-1}^{(N)}\}$ at iteration $t-1$. Built-in functions are highlighted with a low dash.

```
function [chains] = dream_LOA(prior,N,T,d,problem)
% DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm for Limits of Acceptability sampling

[delta,c,c_star,nCR,p_unit] = deal(3,0.1,1e-12,3,.2);     % Default values DREAM algorithmic parameters
for i = 1:N, ind(i,1:N-1) = setdiff(1:N,i); end           % Matrix with for each chain index other chains
CR = [1:nCR]/nCR; pCR = ones(1,nCR)/nCR;                  % Crossover values and selection probabilities
chains = nan(T,d+1,N);                                    % Preallocate memory chain trajectories with fitness

X = prior(N,d);                                           % Draw initial state of each chain
for i = 1:N, f_X(i,1) = fitness(X(i,1:d),problem); end    % Compute fitness of initial state each chain
chains(1,1:d+1,1:N) = reshape([X f_X]',1,d+1,N);          % Store in chains the initial states and their fitness

for t = 2:T                                               % Dynamic part: Evolve the N chains T-1 steps
    [~,draw] = sort(rand(N-1,N));                             % Randomly permute [1,...,N-1] N times
    dX = zeros(N,d);                                          % Set to zero jump vector of each chain
    lambda = unifrnd(-c,c,N,1);                               % Draw N different lambda values
    for i = 1:N                                               % Each chain: Create proposal and accept/reject
        D = randsample([1:delta],1,'true');                       % Select delta (equal selection probability)
        r1 = ind(i,draw(1:D,i)); r2 = ind(i,draw(D+1:2*D,i));     % Unpack r1 and r2; "r1" n.e. "r2" n.e. "i"
        cr = randsample(CR,1,'true',pCR);                         % Draw at random one crossover value
        A = find(rand(1,d) < cr);                                 % Derive subset A with dimensions to sample
        d_star = numel(A);                                        % Cardinality of A: Number dimensions to sample
        g_RWM = 2.38/sqrt(2*D*d_star);                            % Jump rate RWM for D chains + d_star dimensions
        gamma = randsample([g_RWM 1],1,'true',[1-p_unit p_unit]); % Select gamma: 80/20 mix of default and unity
        dX(i,A) = (1+lambda(i))*gamma*sum(X(r1,A)-X(r2,A),1) ...  % Compute ith jump with differential evolution
            + c_star*randn(1,d_star);
        Xp(i,1:d) = X(i,1:d) + dX(i,1:d);                         % Compute ith proposal
                                                                  % ---> Enforce parameter ranges via folding <---
        f_Xp(i,1) = fitness(Xp(i,1:d),problem);                   % Calculate fitness of ith proposal
        P_acc = f_Xp(i,1) >= f_X(i,1);                            % Compute acceptance probability (0 or 1)
        if P_acc, X(i,1:d) = Xp(i,1:d); f_X(i,1) = f_Xp(i,1); end % True: Accept proposal
    end                                                       % End each chain: Create proposal and accept/reject
    chains(t,1:d+1,1:N) = reshape([X f_X]',1,d+1,N);          % Add to chains current position and fitness
                                                              % ---> Monitor convergence of sampled chains <---
    [X,f_X] = outlier(X,f_X);                                 % Outlier detection and correction
end                                                       % End of dynamic part: Evolve the N chains T-1 steps
```

The `dream_LOA` function has five input arguments, including `prior`, an anonymous function handle of the prior distribution, `N`, the number of chains, `T`, the number of generations, `d`, the dimensionality of the target distribution, and `problem`, a structure array with data containers called fields which are required to compute the fitness of each proposal (more of which later). Based on these input arguments the code creates a three-dimensional matrix, `chains` of size `T` by `d+1` by `N` with T parameter vectors and corresponding fitness values in the `N` different Markov chains. `randsample` draws with replacement ('true') the value of the jump rate, `gamma` from the vector `[g_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 `r1` and `r2`. The function `outlier()` is a patch for outlier chains (*Vrugt*, 2016). The jump vector, `dx(i,1:d)` of the `i`th chain contains the desired information about the scale and orientation of the proposal distribution and is derived from the remaining `N-1` chains. The remaining functions `nan()`, `reshape()`, `setdiff()`, `sort()`, `zeros`), `find()`, `numel()`, `sqrt()`, and `ceil()` are explained in introductory textbooks and/or the MATLAB "help" utility. Note that this basic code of DREAM$_{(\mathrm{LOA})}$ does not monitor convergence of the sampled chains, does not enforce parameter constraints (to honor prior ranges), and does not

adapt the selection probabilities of the crossover values.

`prior()` is an anonymous function that draws `N` samples from a `d`-variate prior distribution. For the instantaneous unit hydrograph (first case study), the prior distribution would equate to

```
prior = @(N,d) unifrnd(1,10,N,d)
```
(18)

where the `@` operator creates the function handle. The prior distribution determines the initial state of the N Markov chains.

`fitness()` is a function which evaluates the fitness of Equation (12) for each proposal, `x`. We provide a template of this function for the first case study with a call to the `Nash_Cascade()` function of Equation (13).

```
function f = fitness(x,problem)
% This function computes the fitness of a parameter vector
% ------------------------------------------------------------------------
% Input:    x           1 x d — vector with (proposal) parameter values
%           problem     structure with additional input arguments
% Output:   f           fitness of proposal
% ------------------------------------------------------------------------

y = Nash_Cascade(x,problem);                        % Run forward model, Equation (1), for x(1:d)
f = sum(abs(problem.y_obs-y) <= problem.epsilon);   % Equation (12), # observations within limits?
```

This function demands as second input argument the structure `problem` with fields `y_obs` and `epsilon`, which store the observed data and their Limits of Acceptability, respectively. This structure also allows the user to pass additional arguments (in fields) to their forward model.

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

10 Appendix B

This Appendix presents a MATLAB implementation of the `Nash_Cascade()` forward model used in section 6.1 of this paper.

```
function [y] = Nash_Cascade(x,problem)
% Nash-Cascade unit hydrograph -- series of linear reservoirs

k = x(1); m = x(2);                                          % Unpack recession constant and # reservoirs)

if k < 1
    warning('Recession constant < 1 day --> numerical errors');
end

W = zeros(problem.maxt,problem.maxt);                        % Initialize matrix W
IUH = 1/(k*gamma(m)) * (problem.t/k).^(m-1) .* exp(-problem.t/k);   % Calculate instantaneous unit hydrograph
for t = 1:problem.maxt                                       % Loop over time
    W(t:problem.maxt,t) = problem.P(t) * IUH(1:problem.maxt-(t-1)); % Calculate discharge
end                                                          % End of time loop
y = sum(W,2);                                                % Total daily discharge
```

The problem setup is defined in the following MATLAB script and used to execute the `dream_LOA` algorithm.

```
%% ------------------------------------------------------------------------ %%
%% EXAMPLE IMPLEMENTATION OF CASE STUDY 1: NASH-CASCADE UNIT HYDROGRAPH  %%
%% ------------------------------------------------------------------------ %%

problem.maxt = 25;                              % Define maximum simulation time (days)
problem.P = [ 10 25 8 2 zeros(1,21) ]';         % Define the 25-days precipitation record
problem.t = 1:problem.maxt;                     % Define times (days) of simulated output
k = 4;                                          % Synthetic data: recession constant (days)
m = 2;                                          % Synthetic data: number of reservoirs (-)
y = Nash_Cascade([k m],problem);                % Create artificial streamflow data (mm/day)
problem.y_obs = normrnd(y,0.1*y);               % Perturb the data and store as observations
problem.epsilon = 0.2 * problem.y_obs;          % Define epilon as multiple of observed data
prior = @(N,d) unifrnd(1,10,N,d);               % Create handle for prior distribution
[chains] = dream_LOA(prior,8,1000,2,problem);   % Run dream_LOA to sample behavioural space
```

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.

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.

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

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, 2014.

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

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.

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, "On hypothesis testing in hydrology: why falsification of models is still a really good idea," *WIRES Water*, In Press.

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.

M.G.B. Blum, and V.C. Tran, "HIV with contact tracing: a case study in approximate Bayesian computation," *Biostatistics*, vol. 11(4), pp. 644-660, 2010.

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.

S.P. Brooks, and A. Gelman, "General methods for monitoring convergence of iterative simulations," *Journal of Computational and Graphical Statistics*, vol. 7, pp. 434-455, 1998.

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.

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.

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.

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.

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, 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.

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.

M. Hollaway, K.J. Beven, C.Mc.W.H. Benskin, M.C.Ockenden, and P.M. Haygarth, "A method for uncertainty constraint of catchment discharge and load estimates," *Journal of Hydrology*, Submitted.

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.

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*, vol. 42 (3), W03407, doi:10.1029/2005WR004368, 2006a.

D. Kavetski, G. Kuczera, and S.W. Franks, "Bayesian analysis of input uncertainty in hydrological modeling: 2. Application", *Water Resources Research*, vol. 42 (3), W03408, doi:10.1029/2005WR004376, 2006b.

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.

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.

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.

D. Lu, D. Ricciuto, A. Walker, C. Safta, and W. Munger, "Bayesian calibration of terrestrial ecosystem models: A study of advanced Markov chain Monte Carlo methods," *Biogeosciences Discussions*, https://doi.org/10.5194/bg-2017-41, Accepted.

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.

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. 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.

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.

H. McMillan, T. Krueger, and J. Freer, "Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality," *Hydrological Processes*, vol. 26 (26), pp. 4078-4111, 2012, doi:10.1002/hyp.9384.

H.K. McMillan, and I.K. Westerberg, "Rating curve estimation under epistemic uncertainty," *Hydrological Processes*, vol. 29 (7), pp. 1873-1882, 2015, doi:10.1002/hyp.10419.

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.

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.

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), "Bridging the gap between GLUE and formal statistical approaches: approximate Bayesian computation," *Hydrology and Earth System Sciences*, vol. 17, pp. 4831-4850, doi:10.5194/hess-17-4831-2013, 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$_{(ABC)}$," *Water Resources Research*, vol. 51, pp. 9207-9231, doi:10.1002/2014WR016805, 2015.

M.G. Schaap, F.J. Leij, and M.T. van Genuchten, "ROSETTA: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions," *Journal of Hydrology*, vol. 251, pp. 163- 176, doi:10.1016/S0022-1694(01)00466-8, 2001.

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.

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.

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.

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.

M., Sunnåker, A.G. Busetto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz, "Approximate Bayesian Computation," *PLOS Computational Biology*, vol. 9 (1), 2013.

M.M. Tanaka, A.R. Francis, F. Luciani, S.A. Sisson, "Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data," *Genetics*, vol. 173, pp. 1511-1520, 2006.

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, W. Bouten, and A.H. Weerts, "Information content of data for identifying soil hydraulic properties from outflow experiments," *Soil Science Society of America Journal*, vol. 65 (1), pp. 19-27, 2001.

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, 2016.

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.

R.D. Wilkinson, "Approximate Bayesian computation (ABC) gives exact results under the assumption of model error", *Statistical Applications in Genetics and Molecular Biology*, 12(2), 129-141, 2013.

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. 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.

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}	0.00	0.10	$\mathrm{cm}^3\ \mathrm{cm}^{-3}$
Saturated soil moisture content	θ_{s}	0.30	0.55	$\mathrm{cm}^3\ \mathrm{cm}^{-3}$
Reciprocal of air-entry value	α	0.02	0.50	cm^{-1}
Curve shape parameter	n	1.05	2.50	-
Saturated hydraulic conductivity	K_{s}	0.24	100.00	$\mathrm{cm\ day}^{-1}$
Pressure head at the lower boundary	h_{bot}	-500	-10	cm

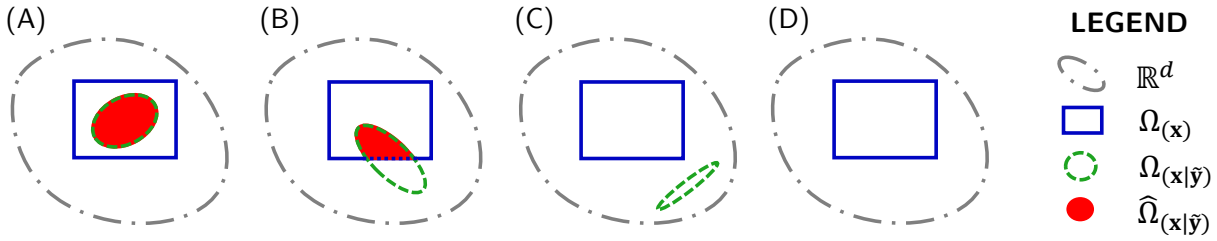

Figure 1 Set-theoretic approach to quantification of parameter uncertainty. The blue, green, and red colours delineate the prior, $\mathbf{\Omega}_{(\mathbf{x})}$, conditional, $\mathbf{\Omega}_{(\mathbf{x}|\tilde{\mathbf{y}})}$, and posterior, $\widehat{\mathbf{\Omega}}_{(\mathbf{x}|\tilde{\mathbf{y}})}$ parameter set respectively, whereas the grey ellipsoidal defines the feasible parameter space, $\mathbf{x} \in \boldsymbol{\chi} \in \mathbb{R}^d$. 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 behavioural solution space.

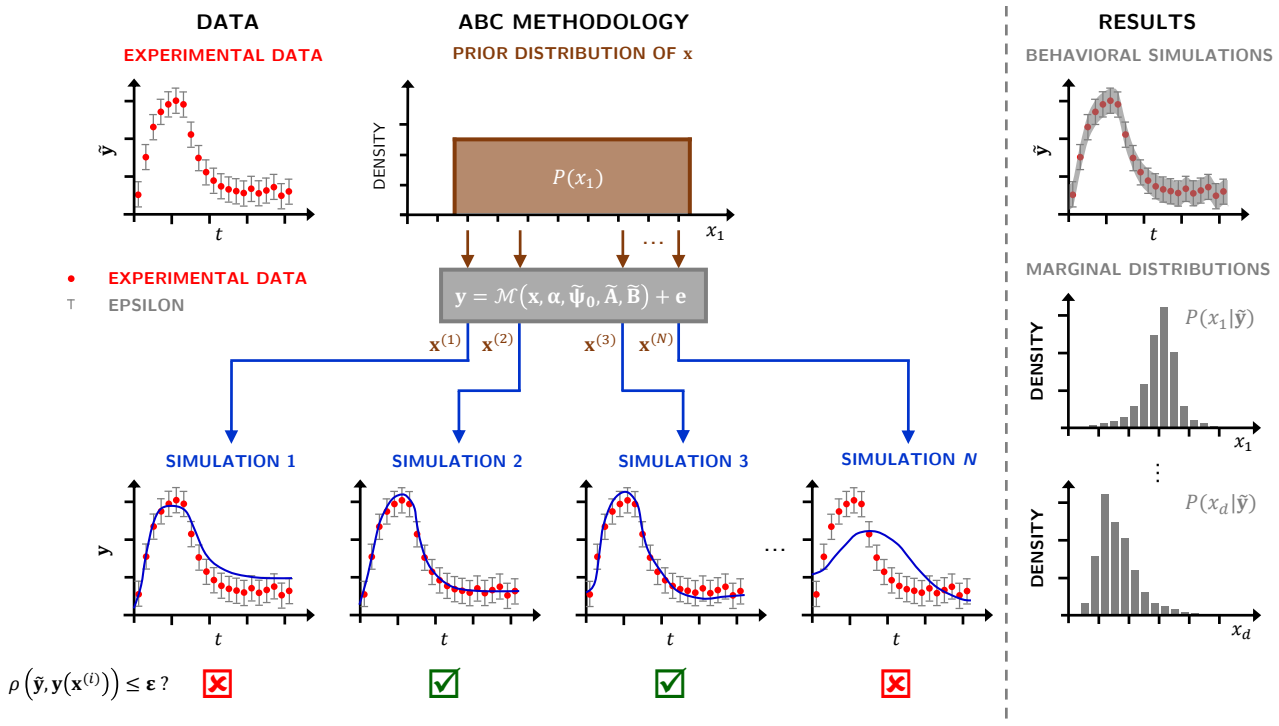

Figure 2 Conceptual overview of approximate Bayesian computation (ABC) for a hypothetical one-dimensional parameter estimation problem (inspired by *Sunnåker et al.* (2013)). First, N samples are drawn from a user-defined prior distribution, $\mathbf{x}^{(i)} \sim P(\mathbf{x})$, where $i = \{1, \ldots, N\}$. Then, each parameter vector is evaluated with the model and corrupted with a residual time series, $\mathbf{e}$, drawn randomly from $P_{\mathbf{e}}(\cdot)$. This creates an ensemble of N different simulations. If the distance between the observed and simulated data, $\rho\left(\tilde{\mathbf{y}}, \mathbf{y}(\mathbf{x}^{(i)})\right)$ is smaller than or equal to some nominal positive value, ϵ then $\mathbf{x}^{(i)}$ is retained, otherwise the simulation is discarded. The accepted samples are then used to approximate the posterior parameter distribution, $P(\mathbf{x}|\tilde{\mathbf{y}})$. For complex models and large data sets the probability of happening upon a simulation run that describes exactly the observations will be very small. Therefore, $\rho\left(\tilde{\mathbf{y}}, \mathbf{y}(\mathbf{x}^{(i)})\right)$ is usually defined as a distance between summary statistics of the simulated, $S\left(\mathbf{y}(\mathbf{x}^{(i)})\right)$, and observed, $S\left(\tilde{\mathbf{y}}\right)$, data, respectively.

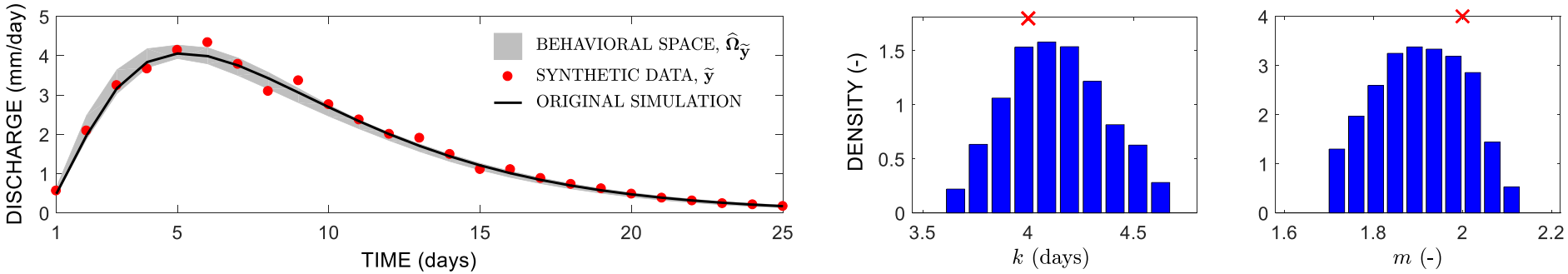

Figure 3 Results of case study I: Nash-Cascade series of reservoirs. (A) Comparison of the observed and simulated hydrograph. The solid black line and red dots denote the original and corrupted data record, respectively, and the gray region is made up of behavioural simulations that satisfy the Limits of Acceptability at each discharge observation. (B),(C) histograms of the marginal posterior distribution of the model parameters k and m in Equation (13). The parameter values of the (uncorrupted) synthetic data record are separately indicated with the red cross ('X') symbols.

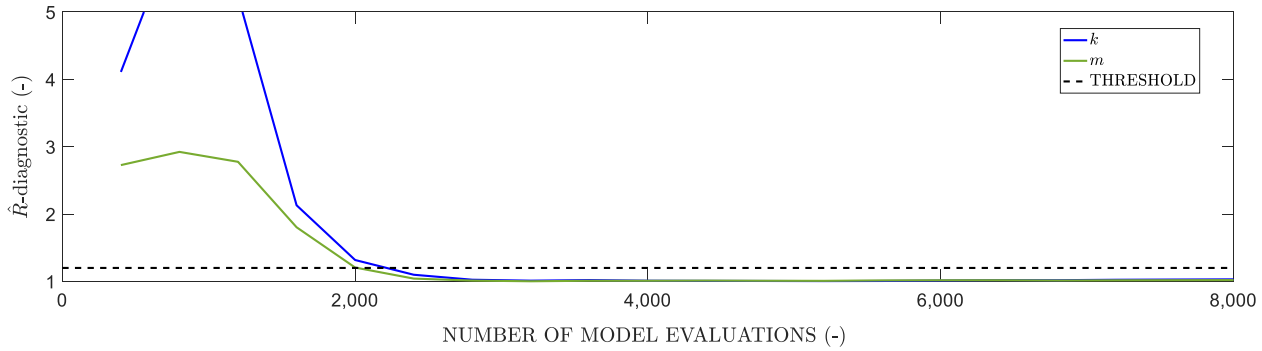

Figure 4 Results of case study I: Nash-Cascade series of reservoirs. Evolution of the $\widehat{R}$-diagnostic of *Gelman and Rubin* (1992) used to judge when convergence of the $N = 8$ Markov chains to a limiting distribution has been achieved. The two parameters are coded with a different colour. About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

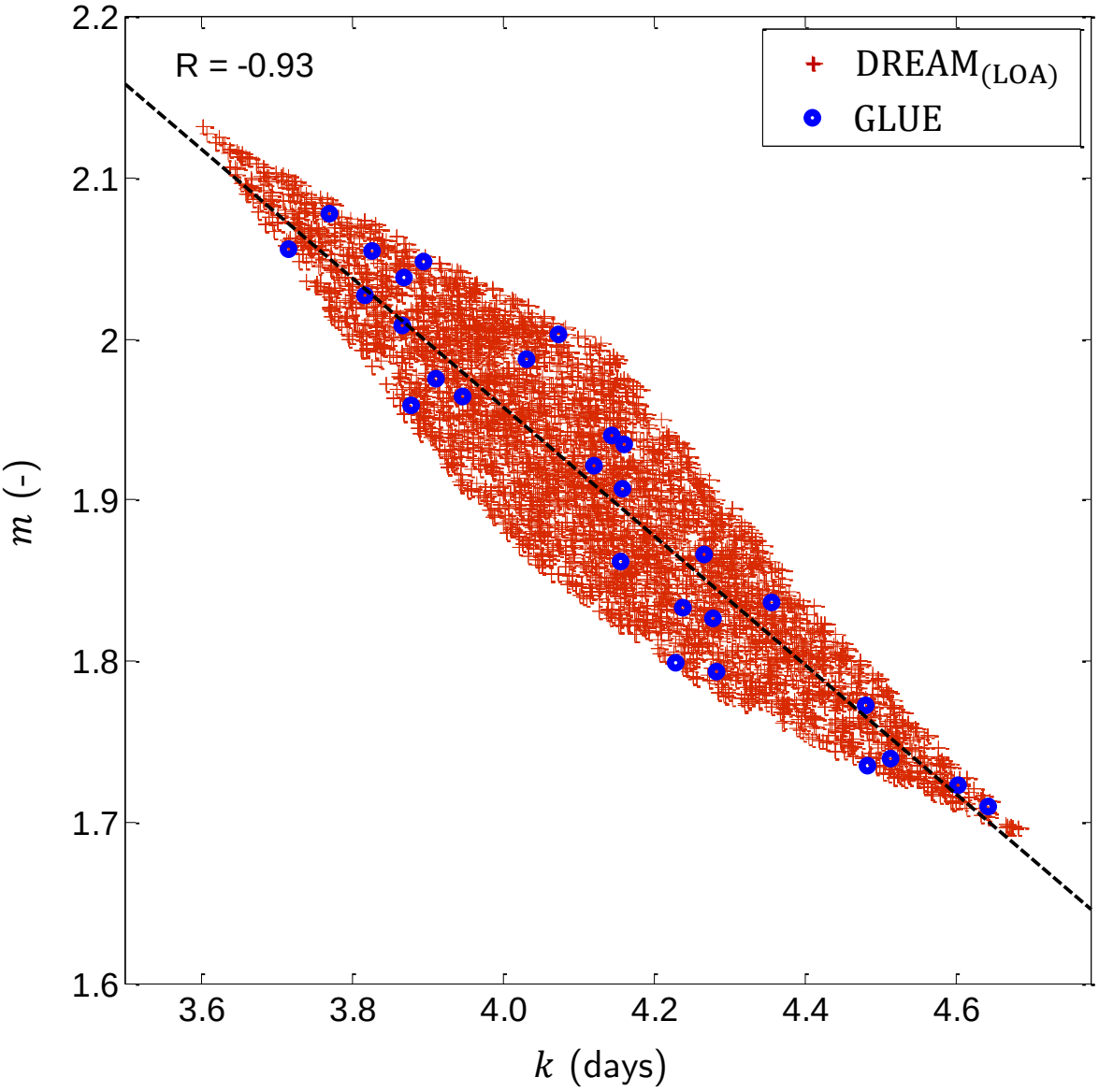

Figure 5 Results of case study I: Nash-Cascade series of reservoirs. Bivariate scatter plot of the behavioural (posterior) samples of k and m derived from MCMC simulation with DREAM$_{(\mathrm{LOA})}$ (dark red) and uniform random sampling (blue dots). The dashed black line plots the least-squares fit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

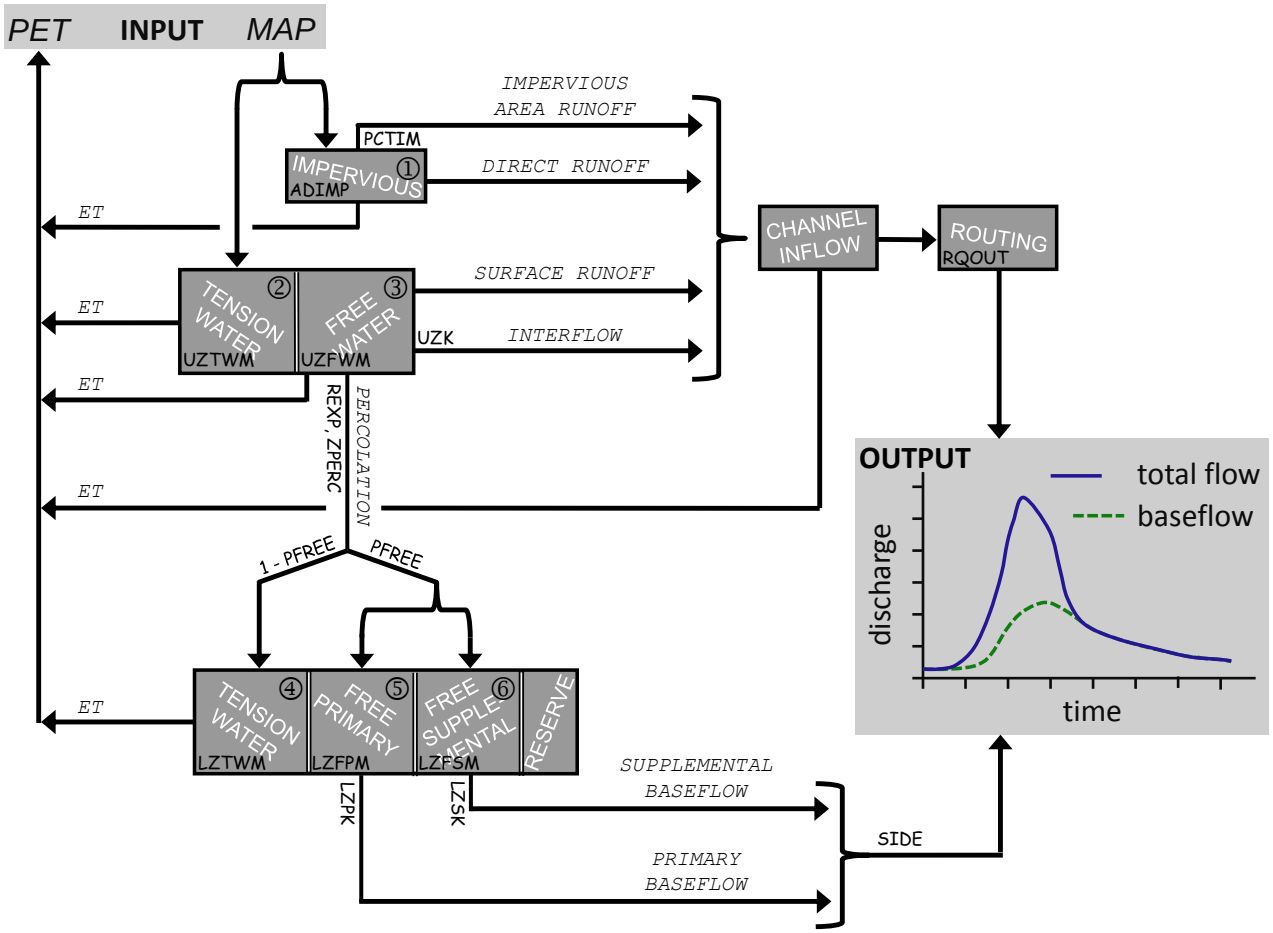

Figure 6 Schematic representation of the Sacramento soil moisture accounting (SAC-SMA) conceptual watershed model. The parameters of the SAC-SMA model appear in Comic Sans font type (black), whereas Courier font type is used to denote the individual fluxes computed by the model. Numbers are used to denote 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.

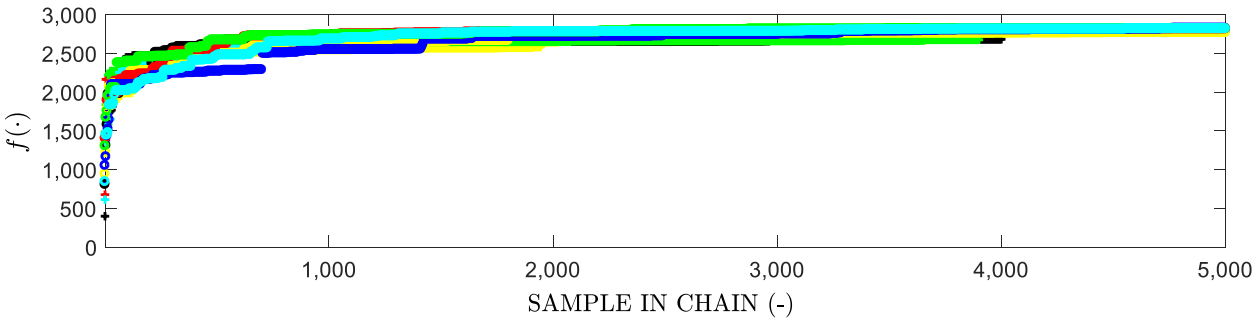

Figure 7 Results of case study II: The SAC-SMA conceptual watershed model. Trace plot of the sampled fitness values of Equation (12) in a randomly selected set of the $N = 20$ different Markov chains of the DREAM$_{(\mathrm{LOA})}$ algorithm. Each chain is coded with a different colour and/or symbol. The computed fitness is equivalent to the number of times the simulated value honors the Limits of Acceptability, $\boldsymbol{\epsilon} = 0.4\widetilde{\mathbf{y}}$ 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.

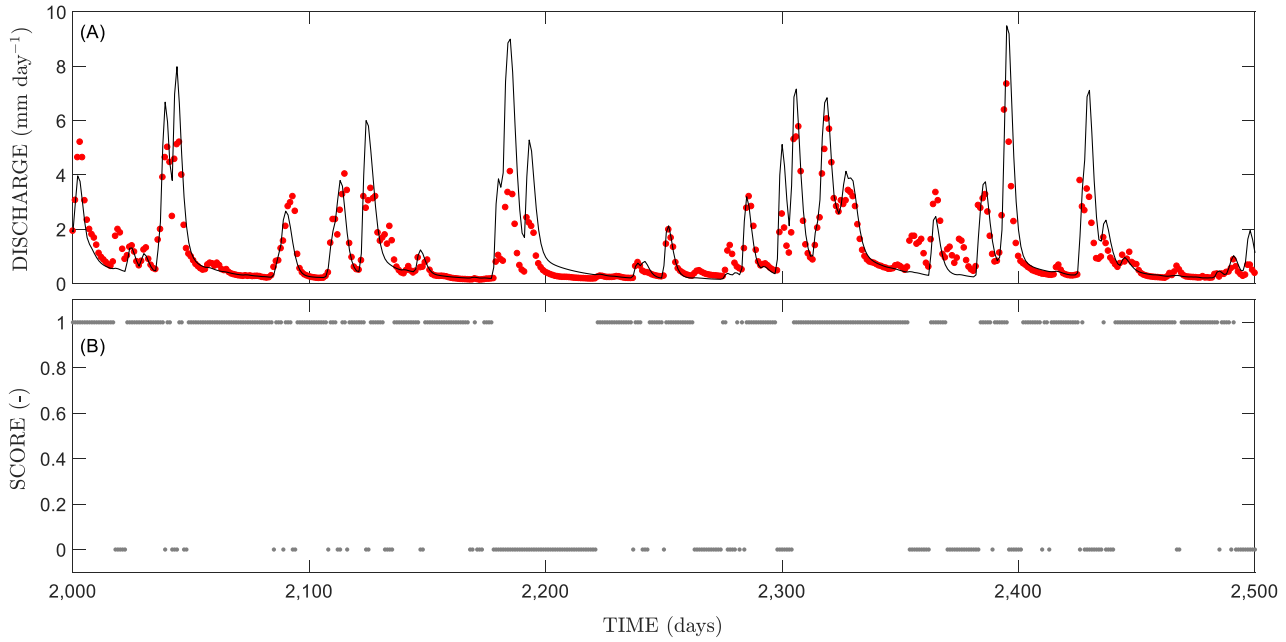

Figure 8 Results of case study II: The SAC-SMA conceptual watershed model. (A) Comparison of the observed (red dots) and simulated (black line) discharge data for a selected 365-day portion of the calibration data period. The simulated values correspond to the DREAM$_{(\mathrm{LOA})}$ sample with highest 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 non-behavioural solution.

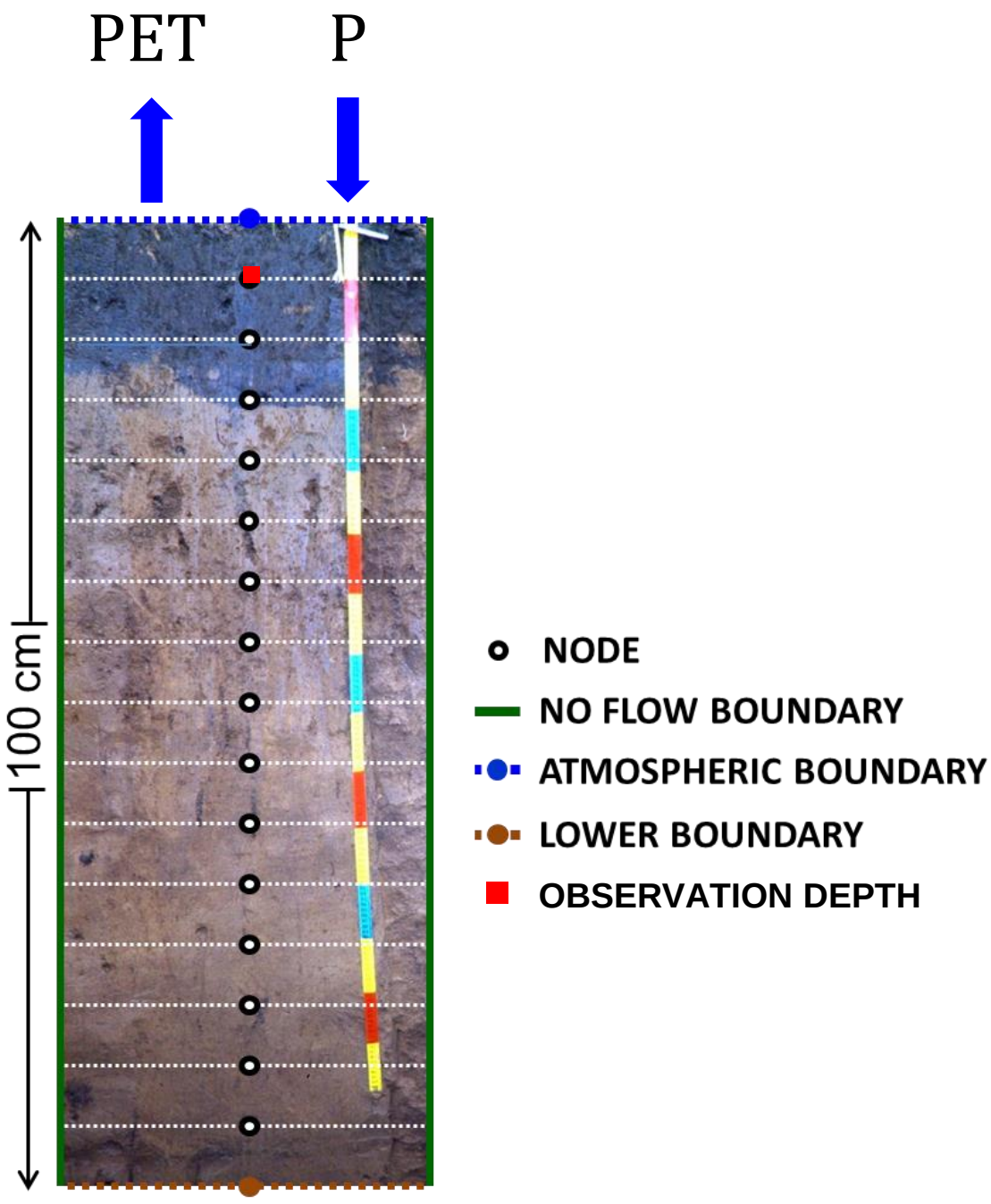

Figure 9 Schematic representation of the HYDRUS-1D model setup for the experimental field plot near Jülich, Germany.

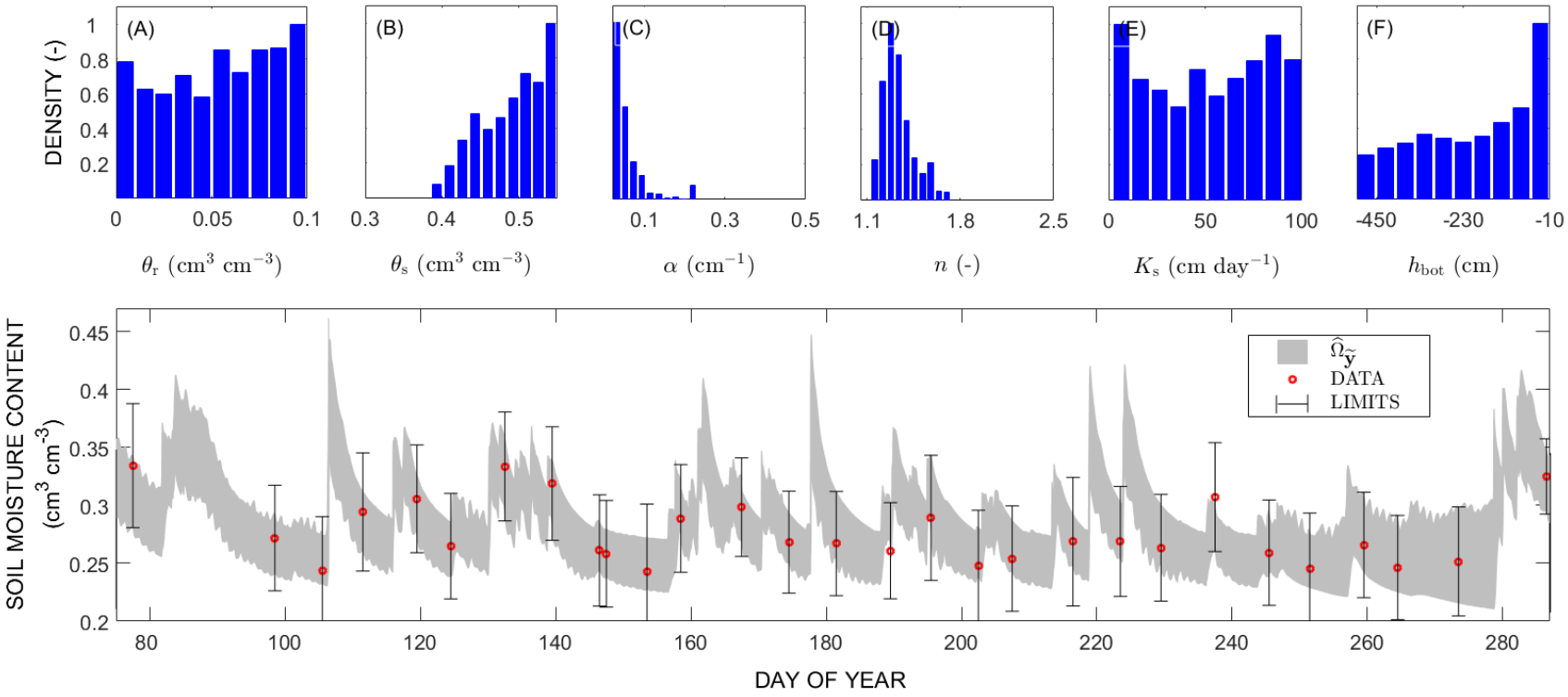

Figure 10 Results of case study III: The HYDRUS-1D variably saturated flow model. (A) Histograms of the behavioural parameter set, $\widehat{\boldsymbol{\Omega}}_{(\mathbf{x}|\tilde{\mathbf{y}})}$ of the soil hydraulic parameters, (A) θ_r, (B) θ_s, (C) α, (D) n, (E) K_s, and (F) h_{bot}. Each x-axis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

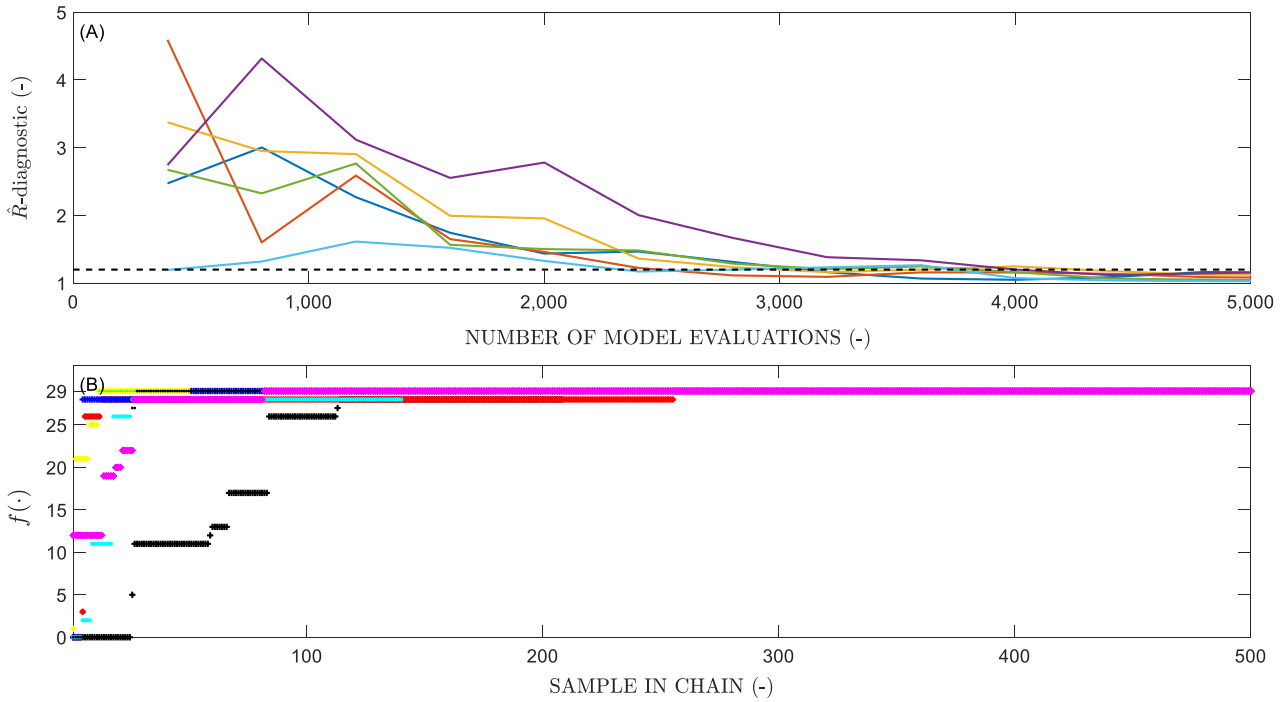

Figure 11 Results of case study III: The HYDRUS-1D variably saturated flow model. Trace plots of the (A) $\widehat{R}$-convergence diagnostic of *Gelman and Rubin* (1992), and (B) sampled fitness values in each of the different Markov chains simulated with DREAM$_{(\mathrm{LOA})}$. The parameters and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

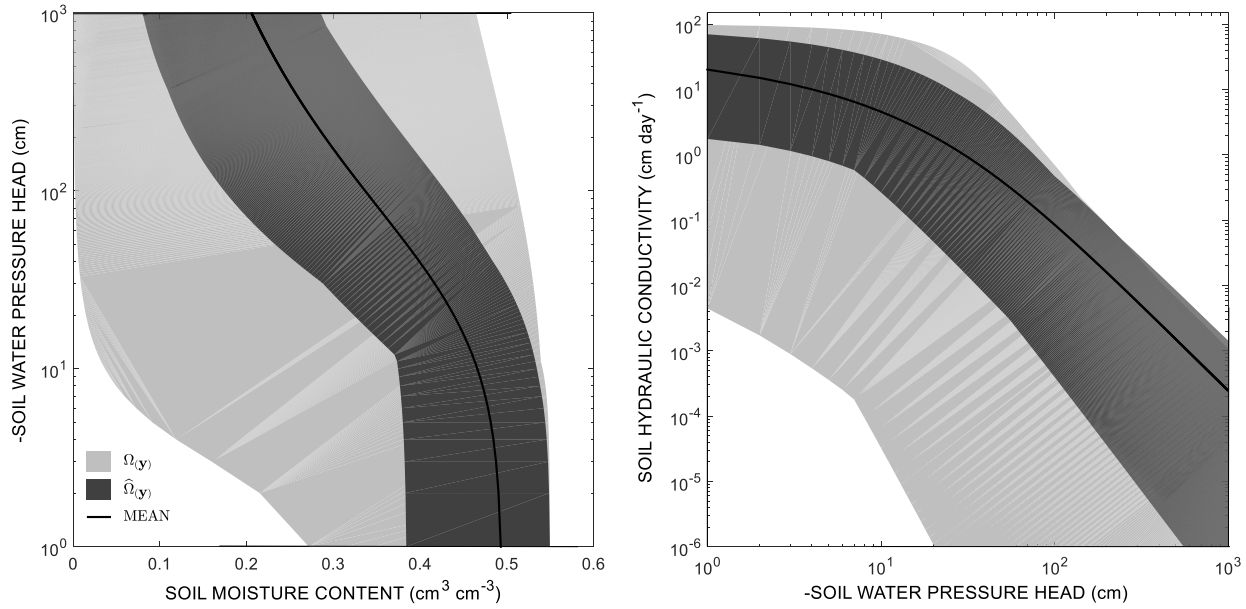

Figure 12 Results of case study III: The HYDRUS-1D variably saturated flow model. Comparison of the prior (dark grey) and posterior (light grey) ranges of the (A) soil water retention, and (B) unsaturated soil hydraulic conductivity function. The black line plots the posterior (behavioural) mean hydraulic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