Approximate Bayesian Computation Using Markov Chain Monte Carlo Simulation: DREAM$_{(\mathrm{ABC})}$

Mojtaba Sadegh,[1], and Jasper A. Vrugt[1,2,3]

[1]Department of Civil and Environmental Engineering, University of California, Irvine, CA, USA.

[2]Department of Earth System Science, University of California, Irvine, CA, USA.

[3]IBG-3 Agrosphere, Forschungszentrum Julich, Julich, Germany.

Abstract. The quest for a more powerful method for model evaluation has inspired *Vrugt and Sadegh* [2013] to introduce "likelihood-free" inference as vehicle for diagnostic model evaluation. This class of methods is also referred to as Approximate Bayesian Computation (ABC) and relaxes the need for a residual-based likelihood function in favor of one or multiple different summary statistics that exhibit superior diagnostic power. Here, we propose several methodological improvements over commonly used ABC sampling methods to permit inference of complex system models. Our methodology entitled, DREAM$_{(ABC)}$ uses the DiffeRential Evolution Adaptive Metropolis algorithm [*Vrugt et al.*, 2008, 2009] as its main building block and takes advantage of a continuous fitness function to efficiently explore the behavioral model space. Three case studies demonstrate that DREAM$_{(ABC)}$ is at least 3 - 1,000 times more efficient than commonly used ABC sampling methods.

1. Introduction and Scope

Bayesian methods have become increasingly popular for fitting hydrologic models to data (e.g. streamflow, water chemistry, groundwater table depth, soil moisture, pressure head, snow water equivalent). Bayes' rule updates the prior probability of a certain hypothesis when new data, $\tilde{\mathbf{Y}} = \{\tilde{y}_1, \ldots, \tilde{y}_n\}$ (also referred to as evidence) become available. The hypothesis typically constitutes the parameter values, $\boldsymbol{\theta}$, of a model, F, which simulates the observed data using

$$\tilde{\mathbf{Y}} \leftarrow F(\boldsymbol{\theta}, \tilde{\mathbf{u}}, \tilde{\mathbf{x}}_0) + \mathbf{e}, \tag{1}$$

where $\tilde{\mathbf{u}} = \{\tilde{u}_1, \ldots, \tilde{u}_n\}$ denotes the observed forcing data, $\tilde{\mathbf{x}}_0$ signifies the initial state of the system, and $\mathbf{e} = \{e_1, \ldots, e_n\}$ includes observation error, as well as error due to the fact that the simulator, F, may be systematically different from the real system of interest, $\Im(\boldsymbol{\theta})$, for the parameters $\boldsymbol{\theta}$. If our main interest is in the parameters of the model, Bayes law is given by

$$p(\boldsymbol{\theta}|\tilde{\mathbf{Y}}) = \frac{p(\boldsymbol{\theta})p(\tilde{\mathbf{Y}}|\boldsymbol{\theta})}{p(\tilde{\mathbf{Y}})}, \tag{2}$$

where $p(\boldsymbol{\theta})$ denotes the prior parameter distribution, $p(\tilde{\mathbf{Y}}|\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$ is the likelihood function, and $p(\tilde{\mathbf{Y}})$ represents the evidence. As all statistical inferences of the parameters can be made from the unnormalized density, we conveniently remove $p(\tilde{\mathbf{Y}})$ from the denominator and write $p(\boldsymbol{\theta}|\tilde{\mathbf{Y}}) \propto p(\boldsymbol{\theta})L(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$.

The likelihood function, $L(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$, summarizes, in probabilistic sense, the overall distance between the model simulation and corresponding observations. The mathematical definition of this function has been subject to considerable debate in the hydrologic and

statistical literature [e.g. *Schoups and Vrugt* [2010]; *Smith et al.* [2010]; *Evin et al.* [2013]]. Simple likelihood functions that assume Gaussian error residuals are statistically convenient, but this assumption is often not borne out of the probabilistic properties of the error residuals that show significant variations in bias, variance, and autocorrelation at different parts of the simulated watershed response. Such non-traditional residual distributions are often caused by forcing data and model structural errors, whose probabilistic properties are very difficult, if not impossible, to adequately characterize. This makes it rather difficult, if not impossible, to isolate and detect epistemic errors (model structural deficiencies), a prerequisite to improving our understanding and theory of water flow and storage in watersheds.

The inability of classical likelihood-based fitting methods to detect model malfunctioning is evident if we critically assess the progress that has been made in modeling of the rainfall-runoff transformation. For instance, consider the Sacramento soil moisture accounting (SAC-SMA) model introduced by *Burnash et al.* [1973] in the early 1970s and used by the US National Weather Service for flash-flood forecasting throughout the United States. In about four decades of fitting the SAC-SMA model to (spatially distributed) streamflow data, we have not been able to make any noticeable improvements to the underlying equations of the model. This is even more disturbing given the relative low order complexity of the SAC-SMA model. If for such relatively simple (lumped) hydrologic models our fitting methods are unable to illuminate to what degree a representation of the real world has been adequately achieved and how the model should be improved, the prospects of learning and scientific discovery for the emerging generation of very high order system models are rather poor, because more complex process representations lead

(unavoidably) to greater interaction among model components, and perpetually larger volumes of field and remote sensing data need to be utilized for system characterization and evaluation.

The limitations of classical residual-based fitting methods has stimulated *Gupta et al.* [2008] (amongst others) to propose a signature-based approach to model evaluation. By choosing the signatures so that they each measure different but relevant parts of system behavior, diagnostic evaluation proceeds with analysis of the behavioural similarities (and differences) of the observed data and corresponding model simulations. Ideally, these differences are then related to individual process descriptions, and correction takes place by refining/improving these respective components of the model. What is left is the numerical implementation of diagnostic model evaluation.

In a previous paper, *Vrugt and Sadegh* [2013] advocated the use of "likelihood-free" inference for diagnostic model evaluation. This approach, introduced in the statistical literature about three decades ago [*Diggle and Gratton*, 1984], is especially useful for cases where the likelihood is intractable, too expensive to be evaluated, or impossible to be formulated explicitly. This class of methods is also referred to as Approximate Bayesian Computation (ABC), a term coined by *Beaumont et al.* [2002], and widens the realm of models for which statistical inference can be considered [*Marjoram et al.*, 2003; *Sisson et al.*, 2007; *Del Moral et al.*, 2008; *Joyce and Marjoram*, 2008; *Grelaud et et al.*, 2009; *Ratmann et al.*, 2009]. ABC has rapidly gained popularity in the past few years, in particular for the analysis of complex problems arising in population genetics, ecology, epidemiology, and systems biology. The first application of ABC in hydrology can be found in *Nott et al.* [2012] and establishes a theoretical connection between ABC and

GLUE-based approaches. Other work on this topic can be found in the recent publication by *Sadegh and Vrugt* [2013].

The premise behind ABC is that $\boldsymbol{\theta}^*$ should be a sample from the posterior distribution if the distance between the observed and simulated data, $\rho\big(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta}^*)\big)$, is smaller than some small positive value, ϵ [*Marjoram et al.*, 2003; *Sisson et al.*, 2007]. Figure 1 provides a conceptual overview of the ABC methodology. All ABC based methods approximate the likelihood function by simulations, the outcomes of which are compared with the observed data [*Beaumont*, 2010; *Bertorelle et al.*, 2010; *Csilléry et al.*, 2010]. In so doing, ABC algorithms attempt to approximate the posterior distribution by sampling from

$$p(\boldsymbol{\theta}|\tilde{\mathbf{Y}}) \propto \int_{\boldsymbol{\mathcal{Y}}} p(\boldsymbol{\theta})\text{Model}(\boldsymbol{y}|\boldsymbol{\theta})\text{I}(\rho\big(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta})\big) \leq \epsilon) d\boldsymbol{y}, \tag{3}$$

where $\boldsymbol{\mathcal{Y}}$ denotes the support of the simulated data, $\mathbf{Y} \sim \text{Model}(\boldsymbol{y}|\boldsymbol{\theta})$, and $\text{I}(a)$ is an indicator function that returns one if the condition a is satisfied and zero otherwise. The accuracy of the estimated posterior distribution, $p(\theta|\rho\big(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta})\big) \leq \epsilon)$ depends on the value of ϵ. In the limit of $\epsilon \to 0$ the sampled distribution will converge to the true posterior, $p(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$ [*Pritchard et al.*, 1999; *Beaumont et al.*, 2002; *Ratmann et al.*, 2009; *Turner and van Zandt*, 2012]. Yet, this requires the underlying model operator to be stochastic, and hence Model$(\cdot)$ in Equation (3) is equivalent to the output of the deterministic model F in Equation (1) plus a random error with probabilistic properties equal to those of $\mathbf{e}$.

For sufficiently complex system models and/or large data sets, it will be difficult, if not impossible, to find a model simulation that always fits the data within ϵ. It has therefore become common practice in ABC to use one or more summary statistics of the data rather than the data itself. Ideally, these chosen summary statistics, $S(\cdot)$ are sufficient and thus

provide as much information for the model parameters as the original data set itself. In practice, however, the use of summary statistics usually entails a loss of information and hence results in an approximate likelihood, especially for complex models. Partial least squares [*Wegmann et al.*, 2009] and information-theory [*Barnes et al.*, 2011] can help to determine (approximately) a set of nearly-sufficient marginal statistics. Nonetheless, complex models admitting sufficient statistics are practical exceptions.

The most common ABC algorithm implements simple rejection sampling which relies on satisfying the condition $\rho\big(S((\tilde{\mathbf{Y}})), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big) \leq \epsilon$. This method has the practical advantage of being relatively easy to implement and use, but its efficiency depends critically on the choice of the prior sampling distribution. If this prior distribution is a poor approximation of the actual posterior distribution, then many of the proposed samples will be rejected. This leads to dramatically low acceptance rates, and thus excessive CPU-times. Indeed, *Vrugt and Sadegh* [2013] and *Sadegh and Vrugt* [2013] report acceptance rates of less than 0.1% for hydrologic models with just a handful of parameters. One remedy to this problem is to increase the value of ϵ, but this leads to an inaccurate approximation of the posterior distribution.

A number of methodological advances have been proposed to enhance the sampling efficiency of ABC algorithms. One common approach is to use a set of monotonically decreasing ϵ values. This allows the algorithm to sequentially adapt the prior distribution and converge to a computationally feasible final value of ϵ. Nonetheless, these algorithms still rely on a boxcar kernel (step function) to evaluate the fitness of each sample, and are not particularly efficient in high dimensional search spaces. In this paper we introduce a Markov Chain Monte Carlo (MCMC) simulation method that enhances, some-

times dramatically, the ABC sampling efficiency. This general-purpose method entitled, DREAM$_{(ABC)}$ uses the DiffeRential Evolution Adaptive Metropolis algorithm [*Vrugt et al.*, 2008, 2009] as its main building block, and replaces the indicator function in Equation (3) with a continuous kernel to decide whether to accept candidate points or not. The proposed methodology is benchmarked using synthetic and real-world simulation experiments.

The remainder of this paper is organized as follows. In section 2 we summarize the results of commonly used ABC sampling methods by application to a synthetic benchmark study. Section 3 introduces the main elements of the DREAM$_{(ABC)}$ algorithm and discusses several of its advantages. This is followed in Section 4 with two synthetic and one real-world simulation experiment. In this section we are especially concerned with sampling efficiency and robustness. Finally, section 5 concludes this paper with a discussion and summary of our main findings.

2. Approximate Bayesian Computation

The ABC method provides an excellent vehicle for diagnostic model evaluation [*Vrugt and Sadegh*, 2013] by using one or multiple different summary statistics that, when rooted in the relevant environmental theory, should have a much stronger and compelling diagnostic power than some residual-based likelihood function. Challenges lie in the proper selection of summary metrics that adequately extract all the available information from the calibration data, how to deal with input data uncertainty, how to detect epistemic errors (lack of knowledge), how to determine an appropriate (small) value for ϵ, and how to efficiently sample complex multi-dimensional spaces involving many tens to hundreds of parameters. This paper is focused on the last topic, and proposes several methodological

developments to overcome the shortcomings of standard ABC sampling methods. The other topics will be investigated in subsequent papers.

We first discuss two common ABC sampling methods that have found widespread application and use within the context of likelihood-free inference. We then introduce DREAM$_{(ABC)}$, a Markov chain Monte Carlo (MCMC) implementation of ABC that permits inference of complex system models.

2.1. Rejection Algorithm

Once the summary statistic(s) has(have) been defined we are left with finding all those values of $\boldsymbol{\theta}^*$ for which $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big) \leq \epsilon$. The most basic algorithm to do so uses rejection sampling. This algorithm proceeds as follows

Algorithm 1 ABC-Rejection Sampler

1: **for** $i = 1, \ldots, N$ **do**
2: **while** $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big) > \epsilon$ **do**
3: Sample $\boldsymbol{\theta}^*$ from the prior, $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta})$
4: Simulate data $\mathbf{Y}$ using $\boldsymbol{\theta}^*$, $\mathbf{Y} \sim$ Model$(\boldsymbol{\theta}^*)$
5: Calculate $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big)$
6: **end while**
7: Set $\boldsymbol{\theta}^i \leftarrow \boldsymbol{\theta}^*$
8: Set $w^i \leftarrow \frac{1}{N}$
9: **end for**

In words, the ABC rejection (ABC-REJ) algorithm proceeds as follows. First we sample a candidate point, $\boldsymbol{\theta}^*$, from some prior distribution, $p(\boldsymbol{\theta})$. We then use this proposal to simulate the output of the model, $\mathbf{Y} \sim$ Model$(\boldsymbol{\theta}^*)$. We then compare the simulated data, $\mathbf{Y}$, with the observed data, $\tilde{\mathbf{Y}}$, using a distance function, $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big)$. If this distance function is smaller than some small positive tolerance value, ϵ then the simulation is close enough to the observations that the candidate point, $\boldsymbol{\theta}^*$ has some nonzero proba-

bility of being in the approximate posterior distribution, $\hat{p}(\boldsymbol{\theta}|\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}))\big) \leq \epsilon)$. By repeating this process N times, ABC-REJ provides an estimate of the actual posterior distribution.

Unfortunately, standard rejection sampling method typically requires massive computational resources to generate a sufficient number of samples from the posterior distribution. Failure to maintain an adequate sampling density may result in under sampling probable regions of the parameter space. This inefficiency can provide misleading results, particularly if $p(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$ is high dimensional and occupies only a small region interior to the prior distribution. Only if $p(\boldsymbol{\theta})$ is a good representation of the actual posterior parameter distribution then ABC-REJ can achieve an adequate sampling efficiency.

2.2. Population Monte Carlo Simulation

To guarantee convergence to the appropriate limiting distribution, the value of ϵ in Algorithm 1 (ABC-REJ) needs to be taken very small. Values of $0.01 \leq \epsilon \leq 0.05$ are often deemed appropriate. Unfortunately, this will produce very low acceptance rates, particularly if the prior distribution is poorly chosen and extends far beyond the posterior distribution. To increase sampling efficiency, it would seem logical to stepwise reduce the value of ϵ and to use the accepted samples to iteratively adapt the prior distribution. This is the principal idea behind population Monte Carlo (PMC) algorithms. These methods are used extensively in physics and statistics for many-body problems, lattice spin systems and Bayesian inference, and also referred to as "quantum Monte Carlo", "transfer-matrix Monte Carlo", "Monte Carlo filter", "particle filter" and "sequential Monte Carlo". The PMC sampler of *Beaumont et al.* [2009]; *Turner and van Zandt* [2012] is specifically designed for ABC-inference and works as follows

Algorithm 2 ABC-Population Monte Carlo sampler

At iteration $j = 1$,
for $i = 1, \ldots, N$ **do**
  while $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big) > \epsilon_1$ **do**
    Sample $\boldsymbol{\theta}^*$ from the prior, $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta})$
    Simulate data $\mathbf{Y}$ using $\boldsymbol{\theta}^*$, $\mathbf{Y} \sim \text{Model}(\boldsymbol{\theta}^*)$
    Calculate $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big)$
  end while
  Set $\boldsymbol{\Theta}_1^i \leftarrow \boldsymbol{\theta}^*$
  Set $\mathbf{w}_1^i \leftarrow \frac{1}{N}$
end for
Set $\boldsymbol{\Sigma}_1 \leftarrow 2\text{Cov}(\boldsymbol{\Theta}_1)$,
At iteration $j > 1$,
for $j = 2, \ldots, J$ **do**
  for $i = 1, \ldots, N$ **do**
    while $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^{**}))\big) > \epsilon_j$ **do**
      Sample $\boldsymbol{\theta}^*$ from the previous population, $\boldsymbol{\theta}^* \sim \boldsymbol{\Theta}_{j-1}$ with probability $\mathbf{w}_{j-1}$
      Perturb $\boldsymbol{\theta}^*$ by sampling $\boldsymbol{\theta}^{**} \sim \mathcal{N}_d(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_{j-1})$
      Simulate data $\mathbf{Y}$ using $\boldsymbol{\theta}^{**}$, $\mathbf{Y} \sim \text{Model}(\boldsymbol{\theta}^{**})$
      Calculate $\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^{**}))\big)$
    end while
    Set $\boldsymbol{\Theta}_j^i \leftarrow \boldsymbol{\theta}^{**}$
    Set $w_j^i \leftarrow \dfrac{p(\boldsymbol{\theta}_j^i)}{\sum_{u=1}^{N} \mathbf{w}_{j-1}^u q_d(\boldsymbol{\theta}_{j-1}^u | \boldsymbol{\theta}_j^i, \boldsymbol{\Sigma}_{j-1})}$ (4)
  end for
  Set $\boldsymbol{\Sigma}_j \leftarrow 2\text{Cov}(\boldsymbol{\Theta}_j)$
end for

In short, the ABC-PMC sampler starts out as ABC-REJ during the first iteration, $j = 1$, but using a much larger initial value for ϵ. This will significantly enhance the initial acceptance rate. During each successive iteration, $j = \{2, \ldots, J\}$, the value of ϵ is decreased and the multi-normal proposal distribution, $q = \mathcal{N}_d(\boldsymbol{\theta}_{j-1}^k, \boldsymbol{\Sigma}_{j-1})$, adapted using $\boldsymbol{\Sigma}_{j-1} = 2\text{Cov}(\boldsymbol{\theta}_{j-1}^1, \ldots, \boldsymbol{\theta}_{j-1}^N)$ with $\boldsymbol{\theta}_{j-1}^k$ drawn from a multinomial distribution, $\mathfrak{F}(\boldsymbol{\Theta}_{j-1}|\mathbf{w}_{j-1})$, where $\boldsymbol{\Theta}_{j-1} = \{\boldsymbol{\theta}_{j-1}^1, \ldots, \boldsymbol{\theta}_{j-1}^N\}$ is a $N \times d$ matrix and $\mathbf{w}_{j-1} = \{w_{j-1}^1, \ldots, w_{j-1}^N\}$ is a N-vector of normalized weights, $\sum_{i=1}^{N} w_{j-1}^i = 1$ and $w_{j-1}^i \geq 0$. In summary, a sequence of (multi)normal proposal distributions is used to iteratively refine the samples and explore the posterior distribution. This approach, similar in spirit as the adaptive Metropolis sampler of [*Haario et al.*, 1999, 2001], achieves a much higher sampling efficiency than ABC-REJ, particularly for cases where the prior distribution, $p(\boldsymbol{\theta})$, is a poor approximation of the actual target distribution.

The PMC sampler of *Turner and van Zandt* [2012] assumes that the sequence of $\boldsymbol{\epsilon}$ values is specified by the user. Practical experience suggests that a poor selection of $\boldsymbol{\epsilon} = \{\epsilon_1, \ldots, \epsilon_J\}$ can lead to very low acceptance rates or even premature convergence. *Sadegh and Vrugt* [2013] have therefore introduced an alternative variant of ABC-PMC with adaptive selection of $\epsilon_{j(j>1)}$. This method requires the user to specify only the initial kernel bandwidth, ϵ_1, and subsequent values of $\boldsymbol{\epsilon}$ are determined from the $\rho(\cdot)$ values of the N most recently accepted samples. This approach is not only more practical, but also enhances convergence speed to the posterior distribution. We therefore prescribe the sequence of $\boldsymbol{\epsilon}$ values in ABC-PMC using the outcome of several adaptive runs.

It is interesting to note that the PMC sampler has elements in common with genetic algorithms (GA) [*Higuchi*, 1997] in that a population of individuals is used to search the

parameter space. The main difference between both approaches is that PMC is specifically designed for statistical inference of the marginal and joint parameter distributions, whereas GAs are specialized in optimization. Yet, it is not difficult to modify the PMC sampler so that it converges to a single "best" solution. Nonetheless, one should be particularly careful using common GA operators such as crossover and mutation. Such genetic operators can significantly improve the search capabilities of ABC-PMC in high dimensional search spaces, but can harm convergence properties.

To benchmark the efficiency of ABC-REJ and ABC-PMC, we start by fitting a relatively simple mixture of two Gaussian distributions, which has become a classical problem in the ABC literature [*Sisson et al.*, 2007; *Beaumont et al.*, 2009; *Toni et al.*, 2009; *Turner and Sederberg*, 2012]

$$p(\theta) = \frac{1}{2}\mathcal{N}(0, \frac{1}{100}) + \frac{1}{2}\mathcal{N}(0, 1), \tag{5}$$

where $\mathcal{N}(a, b)$ is a normal distribution with mean, a and standard deviation, b. The solid black line in Figure 2 plots the actual target distribution.

We now test the efficiency of ABC-REJ and ABC-PMC using the following distance function,

$$\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big) = \begin{cases} |\overline{\mathbf{Y}}| & \text{with probability of } 50\% \\ |y_1| & \text{with probability of } 50\% \end{cases}$$

where $\mathbf{Y} = \{y_1, \ldots, y_{100}\}$, $y_i \sim \mathcal{N}(\theta, 1)$, and the operator $|\cdot|$ signifies the modulus (absolute value). In keeping with the statistical literature, we assume a uniform prior distribution, $p(\theta) \sim \mathcal{U}[-10, 10]$, with $\epsilon = 0.025$ (ABC-REJ) and $\boldsymbol{\epsilon} = \{1, 0.75, 0.5, 0.25, 0.1, 0.05, 0.025\}$ (ABC-PMC). Fig. 2 plots histograms of the ABC-REJ (A: left plot) and ABC-PMC (B: middle plot) derived posterior distribution of θ using $N = 1,000$ samples.

The marginal distribution derived with ABC-REJ and ABC-PMC are in good agreement with the known target distribution (black line). Table 1 summarizes the performance of the ABC-REJ and ABC-PMC sampler. The rejection sampler (ABC-REJ) requires about $386,742$ function evaluations to find $N = 1,000$ behavioral solutions. This corresponds to an acceptance rate (AR, %) of approximate 0.26%, which can be considered highly inefficient. The ABC-PMC sampler on the other hand requires fewer function evaluations (170,848) to explore the target distribution, with an acceptance rate of about 0.59%. Note however that ABC-PMC underestimates the sampling density of points in the tails of the posterior suggesting that the algorithm has not been able to fully explore $p(\theta)$. The results of DREAM$_{(ABC)}$ that are listed at the bottom of Table 1 will be discussed in section 4.1 of this paper.

3. Markov Chain Monte Carlo Simulation

The adaptive capabilities of the ABC-PMC sampler offer significant computational advantages over ABC-REJ. However, further methodological improvements are warranted to enable inference of complex simulation models involving high dimensional parameter spaces. The use of a boxcar fitness kernel (zero probability everywhere except for a small interval where it is a constant) is theoretically convenient, but makes it very difficult for any sampling algorithm to determine the preferred search direction. All rejected simulations receive a similar score, irrespective of whether their $\rho(\cdot)$ values are in close proximity of the threshold, ϵ or far removed. This is certainly not desirable and unnecessarily complicates posterior exploration. Furthermore, ABC-REJ and ABC-PMC update all entries of the parameter vector simultaneously. This is equivalent to a crossover of 100%, and adequate for low-dimensional problems involving just a handful of parameters, but not

necessarily efficient in high dimensional search spaces. For such problems, conditional (or subspace) sampling has desirable properties and can enhance, sometimes dramatically, the speed of convergence. The method we propose herein is more commonly known as Metropolis-within-Gibbs, and samples individual dimensions (or groups of parameters) in turn.

3.1. Continuous fitness kernel

A boxcar kernel has the disadvantage that all samples with $\rho(\cdot)$ value larger than ϵ are considered equal and discarded. This is the basis of rejection sampling, and can be very inefficient particularly if the prior sampling distribution is poorly chosen. To solve this problem, *Turner and Sederberg* [2012] recently introduced an alternative ABC sampling approach using MCMC simulation with a continuous fitness kernel. This method is based on the concept of noisy-ABC [*Beaumont et al.*, 2002; *Blum and François*, 2010] and perturbs the model simulation with a random error, $\boldsymbol{\xi}$

$$\mathbf{Y} \leftarrow \text{Model}(\boldsymbol{\theta}^*) + \boldsymbol{\xi} \tag{6}$$

If we assume that $\boldsymbol{\xi}$ follows a multivariate normal distribution, $\mathcal{N}(\mathbf{0}_n, \alpha)$, then we can evaluate the probability density, $p\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big)$, of $\boldsymbol{\theta}^*$ using

$$p(\boldsymbol{\theta}^*|\alpha) = \frac{1}{\alpha\sqrt{2\pi}} \exp\left(-\frac{1}{2}\alpha^{-2}\Big(\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big)\Big)^2\right), \tag{7}$$

where α is an algorithmic (or free) parameter that defines the width of the kernel. As a result, the approximation in Equation (3) becomes [*Turner and Sederberg*, 2012]

$$p(\boldsymbol{\theta}|\tilde{\mathbf{Y}}) \propto \int_{\boldsymbol{\mathcal{Y}}} p(\boldsymbol{\theta})\text{Model}(\boldsymbol{y}|\boldsymbol{\theta})p\Big(\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big)\Big)d\boldsymbol{y} \tag{8}$$

This approach has recently been coined kernel-based ABC (KABC) in the statistical literature, and opens up an arsenal of advanced Monte Carlo based sampling methods to explore the posterior distribution. Three preliminary case studies of different complexity (not shown herein) demonstrate that KABC with DREAM is at least $3 - 1,000$ times more efficient than ABC-PMC. Unfortunately, KABC can run into a fatal problem. The MCMC algorithm may produce a nice bell shaped posterior distribution but with simulated summary statistics that are far removed from their observed values. For example, let's assume an extreme case in which the model is unable to fit any of the observed summary statistics within the required tolerance ϵ. Sampling would still produce a limiting distribution, but with probability densities of Equation (7) that are practically zero. These results are undesirable, and have nothing to do with the actual MCMC method used, replacing this with another sampling approach would give similar findings. The culprit is the continuous kernel of Equation (7), which does not a-priori bound the feasible space of the posterior solutions. Changing Equation (7) to a boxcar function with $p(\cdot) = 1$ in the interval $[-\epsilon, \epsilon]$ around the measured summary statistic(s), and exponential decline of the density outside this interval runs into the same problems and is thus also futile.

We here propose an alternative method for fitness assignment that helps convergence of a MCMC simulator to the posterior distribution. We define the fitness of $\boldsymbol{\theta}^*$ as follows

$$f(\boldsymbol{\theta}^*, \phi) = \phi - \rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big), \tag{9}$$

were $\phi > 0$ is a coefficient that bounds the fitness values between $(-\infty, \phi]$. The smaller the distance of the simulated summary metrics to their observed values, the higher the fitness. By setting $\phi = \epsilon$, then $f(\boldsymbol{\theta}^*) \in [0, \epsilon]$ is a necessary condition for a sample, $\boldsymbol{\theta}^*$, to be

called a posterior solution, otherwise the sample is non behavioral and can be discarded. This condition is easily verified a-posteriori from the sampled fitness values of the Markov chains.

We are now left with a definition of the selection rule to help determine whether to accept trial moves or not. The original scheme proposed by *Metropolis et al.* [1953] was constructed using the condition of detailed balance. If $p(u)$ ($p(i)$) denotes the probability to find the system in state u (i) and $q(u \rightarrow i)$ ($q(i \rightarrow u)$) is the conditional probability to perform a trial move from u to i (i to u), then the probability $p_{\rm acc}(u \rightarrow i)$ to accept the trial move from u to i is related to $p_{\rm acc}(i \rightarrow u)$ according to:

$$p(u)q(u \rightarrow i)p_{\rm acc}(u \rightarrow i) = p(i)q(i \rightarrow u)p_{\rm acc}(i \rightarrow u) \tag{10}$$

If we assume a symmetric jumping distribution, that is $q(u \rightarrow i) = q(i \rightarrow u)$, then it follows that

$$\frac{p_{\rm acc}(u \rightarrow i)}{p_{\rm acc}(i \rightarrow u)} = \frac{p(i)}{p(u)} \tag{11}$$

This Equation does not yet determine the acceptance probability. *Metropolis et al.* [1953] made the following choice:

$$p_{\rm acc}(u \rightarrow i) = \min\left[1, \frac{p(i)}{p(u)}\right], \tag{12}$$

to determine whether to accept a trial move or not. This selection rule has become the basic building block of many existing MCMC algorithms. *Hastings* [1970] extended Equation (12) to non-symmetrical jumping distributions in which $q(u \rightarrow i) \neq q(i \rightarrow u)$.

Unfortunately, Equation (9) is not a proper probability density function, and hence application of (12) will lead to a spurious approximation of the posterior distribution. The

same problem arises with other fitness functions derived from Equation (7), for example, the Nash-Sutcliffe efficiency, pseudolikelihoods and a composite of multiple objective functions. We therefore calculate the acceptance probability using

$$p_{\rm acc}(u \rightarrow i) = \max\Big(\mathrm{I}\big(f(i) \geq f(u)\big), \mathrm{I}\big(f(i) \geq (\phi - \epsilon)\big)\Big) \tag{13}$$

where $\mathrm{I}(\cdot)$ is an indicator function, and the operator $\max\big(\mathrm{I}(a), \mathrm{I}(b)\big)$ returns one if a and/or b is true, and zero otherwise. Thus, $p_{\rm acc}(u \rightarrow i) = 1$ (we accept) if the fitness of the proposal i is higher than or equal to that of the current state of the chain, u. On the contrary, if the fitness of i is smaller than that of u then $p_{\rm acc}(u \rightarrow i) = 0$ and the proposal is rejected, unless $f(i) \geq 0$, then we still accept.

The binary acceptance probability of Equation (13) differs fundamentally from a regular MCMC selection rule, but has important practical advantages that promulgate converge to the correct limiting distribution. Initially, when the samples are rather inferior, Equation (13) enforces the MCMC algorithm to act as an optimizer and only accept proposals with a higher fitness and thus smaller distance to the observed values of the summary statistics. The transitions of the Markov chain during this time are irreversible with a zero acceptance probability of the previous state (backward jump). This changes the moment a candidate point has been sampled whose fitness, $f(\cdot) \geq 0$. From this point forward, the acceptance probability of Equation (13) leads to a reversible Markov chain and the successive samples can be used to approximate the posterior target distribution.

An important limitation of Equation (13) is that it cannot incorporate non-symmetric jump distributions, such as the snooker updater used in DREAM$_{(ZS)}$ and MT-DREAM$_{(ZS)}$

[*Laloy and Vrugt*, 2012]. This would require a Hastings-type correction, but cannot be readily incorporated within the current framework.

3.2. Pseudo-code of DREAM$_{(ABC)}$

We can solve for the posterior distribution of Equation (9) using MCMC simulation with DREAM [*Vrugt et al.*, 2008, 2009]. This method uses subspace sampling to increase search efficiency and overcome some of the main limitations of ABC-REJ and ABC-PMC. In DREAM, K ($K > 2$) different Markov chains are run simultaneously in parallel, and multivariate proposals are generated on the fly from the collection of chains, $\boldsymbol{\Theta}^{t-1}$ (matrix of $K \times d$ with each chain state as row vector), using differential evolution [*Storn and Price*, 1997; *Price et al.*, 2005]. If A is a subset of δ-dimensional space of the original parameter space, $\mathbb{R}^{\delta} \subseteq \mathbb{R}^{d}$, then a jump in the kth chain, $k = \{1, \ldots, K\}$ at iteration $t = \{2, \ldots, T\}$ is calculated using

$$\begin{aligned} \boldsymbol{\Delta}^{*}_{k,A} &= (\mathbf{1}_{\delta} + \boldsymbol{\lambda})\gamma(\delta)\left[\sum_{j=1}^{\tau} \boldsymbol{\theta}^{t-1}_{\mathbf{g}_j,A} - \sum_{j=1}^{\tau} \boldsymbol{\theta}^{t-1}_{\mathbf{r}_j,A}\right] + \boldsymbol{\zeta} \\ \boldsymbol{\Delta}^{*}_{k,\neq A} &= 0, \end{aligned} \tag{14}$$

where $\gamma = 2.38/\sqrt{2\tau D}$ is the jump rate, τ denotes the number of chain pairs used to generate the jump, and $\mathbf{g}$ and $\mathbf{r}$ are τ-vectors with integer values drawn without replacement from $\{1, \ldots, k-1, k+1, \ldots, K\}$. The values of $\boldsymbol{\lambda}$ and $\boldsymbol{\zeta}$ are sampled independently from $\mathcal{U}_{\delta}(-c, c)$ and $\mathcal{N}_{\delta}(0, c^{*})$ with, typically, $c = 0.1$ and c^{*} small compared to the width of the target distribution, $c^{*} = 10^{-12}$ say.

The candidate point of chain k at iteration t then becomes

$$\boldsymbol{\theta}^{*}_{k} = \boldsymbol{\theta}^{t-1}_{k} + \boldsymbol{\Delta}^{*}_{k}, \tag{15}$$

and the Metropolis ratio is used to determine whether to accept this proposal or not. The DREAM algorithm solves an important practical problem in MCMC simulation, namely that of choosing an appropriate scale and orientation of the proposal distribution. Section 3.3 will detail the procedure for selecting the subset A of dimensions of the parameter space that will be updated each time a proposal is created.

We now proceed with a pseudo-code of DREAM$_{(ABC)}$.

Algorithm 3 DREAM$_{(ABC)}$-Markov chain Monte Carlo sampler

At iteration $t = 1$,
for $k = 1, \ldots, K$ **do**
  Sample $\boldsymbol{\theta}_k^1$ from the prior, $\boldsymbol{\theta}_k^1 \sim p(\boldsymbol{\theta})$
  Simulate data $\mathbf{Y}$ using $\boldsymbol{\theta}_k^1$, $\mathbf{Y} \sim \text{Model}(\boldsymbol{\theta}_k^1)$
  Calculate the fitness, $f(\boldsymbol{\theta}_k^1) = \epsilon - \rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}_k^1))\big)$
end for
At iteration $t > 1$,
for $t = 2, \ldots, T$ **do**
  for $k = 1, \ldots, K$ **do**
    Determine subset A, the dimensions of the parameter space to be updated.
    Calculate the jump vector, $\boldsymbol{\Delta}_k^*$ using Equation (14)
    Compute the proposal, $\boldsymbol{\theta}_k^* = \boldsymbol{\theta}_k^{t-1} + \boldsymbol{\Delta}_k^*$
    Simulate data $\mathbf{Y}$ using $\boldsymbol{\theta}_k^*$, $\mathbf{Y} \sim \text{Model}(\boldsymbol{\theta}_k^*)$
    Calculate the fitness, $f(\boldsymbol{\theta}_k^*) = \epsilon - \rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}_k^*))\big)$
    Calculate the acceptance probability using Equation (13),

$$p_{\text{acc}}(\boldsymbol{\theta}_k^*) = \max\Big(\text{I}\big(f(\boldsymbol{\theta}_k^*) \geq f(\boldsymbol{\theta}_k^{t-1})\big), \text{I}\big(f(\boldsymbol{\theta}_k^*) \geq 0\big)\Big)$$

    If $p_{\text{acc}}(\boldsymbol{\theta}_k^*) = 1$, set $\boldsymbol{\theta}_k^t = \boldsymbol{\theta}_k^*$ otherwise remain at current state, $\boldsymbol{\theta}_k^t = \boldsymbol{\theta}_k^{t-1}$
  end for
  Compute $\hat{R}$-statistic for each entry of $\boldsymbol{\theta}$ using last 50% of samples in each chain.
  If $\hat{R}_j \leq 1.2$ for all $j = \{1, \ldots, d\}$ stop, otherwise continue.
end for

In summary, DREAM$_{(ABC)}$, runs K different Markov chains in parallel. Multivariate proposals in each chain $k = \{1, \ldots, K\}$ are generated by taking a fixed multiple of the

difference of two or more randomly chosen members (chains) of $\boldsymbol{\Theta}$ (without replacement). By accepting each jump with binary probability of Equation (13) a Markov chain is obtained, the stationary or limiting distribution of which is the posterior distribution. Because the joint pdf of the K chains factorizes to $\pi(\boldsymbol{\theta}_1) \times \ldots \times \pi(\boldsymbol{\theta}_K)$, the states of the individual chains are independent at any iteration after DREAM$_{(ABC)}$ has become independent of its initial value. After this burn-in period, the convergence of DREAM$_{(ABC)}$ can thus be monitored with the $\hat{R}$-statistic of *Gelman and Rubin* [1992].

The jump distribution in Equation (14) of DREAM$_{(ABC)}$ is easily implemented in ABC-PMC to help generate trial moves. This could further improve the scale and orientation of the proposals, but comes at an increased computational cost. The conditional probability to move from $\boldsymbol{\theta}^u_{j-1}$ to $\boldsymbol{\theta}^i_j$ or $q_d(\boldsymbol{\theta}^u_{j-1} \rightarrow \boldsymbol{\theta}^i_j)$ in Equation (4) of ABC-PMC is easy to calculate for a (multi)normal proposal distribution, but requires significantly more CPU-resources if the jumping kernel of DREAM$_{(ABC)}$ is used. Let's assume, for instance, that $\tau = 1$ and $\boldsymbol{\lambda} = \mathbf{0}_\delta$ in Equation (14). The probability to transition in ABC-PMC from the current state, $\boldsymbol{\theta}^u_{j-1}$, to the proposal, $\boldsymbol{\theta}^i_j$, is then equivalent to

$$q(\boldsymbol{\theta}^i_j|\boldsymbol{\theta}^u_{j-1}) = \sum_{m=1}^{N}\sum_{o=1}^{N} \psi(\boldsymbol{\theta}^u_{j-1} + \gamma(\boldsymbol{\theta}^m_{j-1} - \boldsymbol{\theta}^o_{j-1})|\boldsymbol{\Sigma}_d);\ i \neq o \neq m \tag{16}$$

where ψ denotes the normal density with covariance matrix $\boldsymbol{\Sigma}_d = (c^*)^2 I_d$. This Equation is of computational complexity $\mathcal{O}(N^2)$ and becomes particularly CPU-intensive for large N and/or if more than one pair of chains ($\tau > 1$) is used to create proposals. More fundamentally, the lack of subspace sampling (see next section) in ABC-PMC deteriorates search efficiency in high-dimensional spaces (shown later). We therefore do not consider this alternative jumping distribution in ABC-PMC.

3.3. Randomized subspace sampling

Subspace sampling is implemented in DREAM$_{(ABC)}$ by only updating randomly selected dimensions of $\boldsymbol{\theta}_k^{t-1}$ each time a proposal is generated. Following the default of the DREAM suite of algorithms [*Vrugt et al.*, 2008, 2009, 2011; *Laloy and Vrugt*, 2012] we use a geometric series of $n_{\rm CR}$ different crossover values and store this in a vector, CR $= \{\frac{1}{n_{\rm CR}}, \frac{2}{n_{\rm CR}}, \ldots, 1\}$. The prior probability of each crossover value is assumed equal and defines a vector $\mathbf{p}$ with $n_{\rm CR}$ copies of $\frac{1}{n_{\rm CR}}$. We create the set A with selected dimensions to be updated as follows

Algorithm Subspace sampling

```
for k = 1, ..., K do
  Define A to be an empty set, A = ∅
  Sample P from the discrete multinomial distribution, P ~ 𝔉(CR|p)
  Draw d labels from a multivariate uniform distribution, Z ~ 𝒰_d[0, 1]
  for j = 1, ..., d do
    if Z_j > (1 − P) then
      Add dimension j to A
    end if
  end for
  if A = ∅ then
    Choose one index from {1,...,d} and add to A
  end if
end for
```

The number of dimensions stored in A ranges between 1 and d and depends on the sampled value of the crossover. This relatively simple randomized selection strategy enables single-site Metropolis sampling (one dimension at a time), Metropolis-within-Gibbs (one or a group of dimensions) and regular Metropolis sampling (all dimensions). To enhance search efficiency, the probability of each $n_{\rm CR}$ crossover values is tuned adaptively during burn-in by maximizing the normalized Euclidean distance between successive states of the

K chains [*Vrugt et al.*, 2009]. The only algorithmic parameter that needs to be defined by the user is n_{CR}, the number of crossover values used. We use the standard settings of DREAM and use $n_{\mathrm{CR}} = 3$ in all the calculations reported herein. This concludes the algorithmic description of DREAM$_{(\mathrm{ABC})}$.

4. Numerical experiments

The next section compares the efficiency of ABC-REJ, ABC-PMC and DREAM$_{(\mathrm{ABC})}$ for two synthetic and one real-world experiment. These case studies cover a diverse set of problem features, including high-dimensionality, nonlinearity, non-convexity, and numerous local optima. Perhaps not surprisingly, our trials with ABC-REJ show that rejection sampling is highly inefficient when confronted with multidimensional parameter spaces, unless the prior sampling distribution closely mimics the target distribution of interest. In practice, this is an unreasonable expectation and we therefore discard the ABC-REJ algorithm after the first case study and focus our attention on the results of ABC-PMC and DREAM$_{(\mathrm{ABC})}$.

In all our calculations with ABC-PMC we create $N = 1,000$ samples at each iteration, $j = \{1, \ldots, J\}$ using values of $\boldsymbol{\epsilon}$ that are listed in each case study and have been determined through trial-and-error (e.g. *Sadegh and Vrugt* [2013]). In DREAM$_{(\mathrm{ABC})}$ we need to define the number of chains, K and the total number of function evaluations, $M = K \cdot T$. Their values are listed in each individual case study. For all the other algorithmic variables in DREAM$_{(\mathrm{ABC})}$ we use standard settings recommended in *Vrugt et al.* [2009].

4.1. Synthetic benchmark experiments: Gaussian mixture model

We now benchmark the performance of DREAM$_{(ABC)}$ by application to the Gaussian mixture model in Equation (5). We set $\epsilon = 0.025$ and run $K = 10$ different chains using a total of $M = 50,000$ function evaluations. Fig. 2c displays the marginal distribution of the posterior samples using a burn-in of 50%. It is evident that the adaptive capabilities of DREAM$_{(ABC)}$ enables it to track the target distribution. The density of samples in the tails has somewhat improved considerably compared to ABC-PMC. The burn-in required for multi-chain methods such as DREAM$_{(ABC)}$ is relatively costly for this one-dimensional problem, and hence it is not surprising that rejection sampling provides a nicer sample of the mixture distribution. The DREAM$_{(ABC)}$ approximation of the target is easily enhanced by creating more samples.

Table 1 lists the acceptance rate (AR, %), number of function evaluations, and ϵ value used with ABC-REJ, ABC-PMC and DREAM$_{(ABC)}$. It is evident that DREAM$_{(ABC)}$ is most efficient in sampling the target distribution. The acceptance rate of DREAM$_{(ABC)}$ is between 3 to 6 times higher than that ABC-PMC and ABC-REJ, respectively.

4.2. Synthetic benchmark experiments: 20-dimensional bivariate distribution

A more demanding test of the ABC-PMC and DREAM$_{(ABC)}$ algorithms can be devised by using a multi-dimensional target distribution. We consider a set of ten bivariate normal distributions

$$h_i(\mu_{i,1}, \mu_{i,2}) \sim \mathcal{N}_2 \left(\begin{bmatrix} \mu_{i,1} \\ \mu_{i,2} \end{bmatrix}, \begin{bmatrix} 0.01^2 & 0 \\ 0 & 0.01^2 \end{bmatrix} \right), \tag{17}$$

with unknown mean of the ith component, $\boldsymbol{\mu}_i = \{\mu_{i,1}, \mu_{i,2}\}$ and fixed covariance matrix. We now generate $n = 20$ observations by sampling the mean of each bivariate distribution from $\mathcal{U}_2[0, 10]$. The "observed" data are plotted in Figure 3 using the red cross symbols.

Each of the bivariate means is now subject to inference with ABC, which results in a $d = 20$ dimensional parameter estimation problem. The simulated data, $\mathbf{Y}$ are created by evaluating Equation (17) fifty different times for each proposal, $\boldsymbol{\theta}^* = [\boldsymbol{\mu}_1^*, \ldots, \boldsymbol{\mu}_{10}^*]$. The mean of the fifty samples from each bivariate distribution is stored in $\mathbf{Y}$ (10 by 2 matrix) and compared to the observed data using the following distance function [*Turner and Sederberg*, 2012]

$$\rho\big(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta}^*)\big) = \sqrt{\frac{1}{20}\sum_{i=1}^{10}\sum_{j=1}^{2}\big(\tilde{\mathbf{Y}}_{(i,j)} - \mathbf{Y}_{(i,j)}(\boldsymbol{\theta}^*)\big)^2} \tag{18}$$

We assume a noninformative (uniform) prior, $\boldsymbol{\theta} \sim \mathcal{U}_{20}[0, 10]$ and set $\boldsymbol{\epsilon} = \{3, 2.5, 2.1, 1.8, 1.6, 1.3, 1.1, 0.9, 0.8, 0.7, 0.6\}$ (ABC-PMC) and $\epsilon = 0.025$, $K = 15$, and $M = 200,000$ (DREAM$_{(\mathrm{ABC})}$). The results of the analysis are presented in Table 2 and Figure 3.

Figure 3 plots the posterior samples (plus symbol) derived from (a) ABC-PMC and (b) DREAM$_{(\mathrm{ABC})}$. The sampled solutions cluster around the observed means (red cross) of the bivariate normal distributions. The size of the posterior uncertainty differ markedly between both algorithms. The posterior samples of DREAM$_{(\mathrm{ABC})}$ are in excellent agreement with the bivariate target distributions. The samples group tightly around their observed counterparts, and their structure is in excellent agreement with the covariance matrix of the target distribution. The ABC-PMC samples, on the contrary, exhibit too much scatter. This finding is not surprising. The ABC-PMC sampler terminated its search prematurely with values of $\rho(\cdot)$ between 0.5 and 0.6. This threshold is much larger than the value of $\epsilon = 0.025$ required to converge to the target distribution. Subspace sampling

would significantly enhance the results of ABC-PMC (not shown), but such modifications could affect the theoretical convergence properties.

The DREAM$_{(\mathrm{ABC})}$ algorithm not only better recovers the actual target distribution, but its sampling efficiency is also superior. To illustrate this in more detail, consider Table 2 that lists the acceptance rate (AR ,%), number of function evaluations, and ϵ value of ABC-PMC and DREAM$_{(\mathrm{ABC})}$. The acceptance rate of DREAM$_{(\mathrm{ABC})}$ of about 19.72% is more than $1,000$ times higher than that of ABC-PMC (0.019%). This marks a three-order of magnitude improvement in search efficiency, which in large part is due to the ability of DREAM$_{(\mathrm{ABC})}$ to sample one or groups of variables in turn. This conditional sampling is necessary to traverse multi-dimensional parameter spaces in pursuit of the posterior distribution.

The present case study clearly illustrates the advantages of DREAM$_{(\mathrm{ABC})}$ when confronted with multi-dimensional parameter spaces. The algorithm requires about $M = 200,000$ function evaluations to successfully recover the 20-d target distribution. To illustrate this in more detail, consider Figure 4 which displays trace plots of the $\hat{R}$-statistic of *Gelman and Rubin* [1992] using the last 50% of the samples stored in each of the K chains. This convergence diagnostic compares for each parameter the between- and within-variance of the chains. Because of asymptotic independence, the between-member variance and $\hat{R}$-diagnostic can be estimated consistently from a single DREAM$_{(\mathrm{ABC})}$ trial. Values of $\hat{R}$ smaller than 1.2 indicate convergence to a limiting distribution. The DREAM$_{(\mathrm{ABC})}$ algorithm needs about $40,000$ function evaluations to officially reach convergence and generate a sufficient sample of the posterior distribution. This is much less than the $M = 200,000$ function evaluations used in this study. Obviously, one should be careful

to judge convergence of the sampled Markov chains based on a single diagnostic, yet visual inspection of the sampled trajectories confirms an adequate mixing of the different chains and converge of DREAM$_{(ABC)}$ after about $15,000$ function evaluations. This is substantially lower than the approximately $40,000$ function evaluations estimated with the $\hat{R}$-statistic, simply because the second half of the chain is used for monitoring convergence.

4.3. Hydrologic modeling: Sacramento Soil Moisture Accounting Model

A more realistic case study is now devised and used to illustrate the advantages DREAM$_{(ABC)}$ can offer in real-world modeling problems. We consider simulation of the rainfall-runoff transformation using the SAC-SMA conceptual hydrologic model. This model has been developed by *Burnash et al.* [1973] and is used extensively by the National Weather Service for flood forecasting throughout the United States. The model has been described in detail in many previous publications, and we therefore summarize in Table 3 the fourteen parameters that require calibration and their prior uncertainty ranges.

Daily data of mean areal precipitation, $\tilde{\mathbf{P}} = \{\tilde{p}_1, \ldots, \tilde{p}_n\}$, mean areal potential evaporation and streamflow, $\tilde{\mathbf{Y}} = \{\tilde{y}_1, \ldots, \tilde{y}_n\}$, from the French Broad River basin at Asheville, North Carolina are used in the present study. In keeping with *Vrugt and Sadegh* [2013] we use the annual runoff coefficient, $S_1(\tilde{\mathbf{Y}})$ (-), the annual baseflow index, $S_2(\tilde{\mathbf{Y}})$ (-), and the flow duration curve, $S_3(\tilde{\mathbf{Y}})$ (day/mm) and $S_4(\tilde{\mathbf{Y}})$ (-), as summary metrics of the discharge data. A detailed description of each summary statistic is given by *Vrugt and Sadegh* [2013] and interested readers are referred to this publication for further details. This leaves us with $L = 4$ four summary statistics for three different hydrologic signatures.

We use the following composite distance function to quantify the distance between the observed and simulated summary statistics

$$\rho\big(S(\tilde{\mathbf{Y}}), S(\mathbf{Y}(\boldsymbol{\theta}^*))\big) = \max\big(|S_i(\tilde{\mathbf{Y}}) - S_i(\mathbf{Y}(\boldsymbol{\theta}^*))|\big) \quad i = \{1, \ldots, L\}, \tag{19}$$

and help determine whether to accept $\boldsymbol{\theta}^*$ or not. Model simulations that simultaneously satisfy each of the four summary metrics within their tolerance thresholds are considered behavioral, and hence constitute samples from the posterior distribution. This composite formulation differs fundamentally from multi-criteria model calibration approaches in which the summary statistics (objective functions) are assumed non-commensurate and therefore treated independently from one another. This latter approach gives rise to a Pareto solution set (rather than posterior distribution) and quantifies the trade-offs in the fitting of the different metrics.

To maximize the search efficiency of ABC-PMC we use $\boldsymbol{\epsilon} = \{1, 0.3, 0.15, 0.1, 0.07, 0.06, 0.04, 0.025\}$. This sequence is determined from a preliminary run of ABC-PMC with adaptive selection of ϵ (see Appendix B of *Sadegh and Vrugt* [2013]). The DREAM$_{(\mathrm{ABC})}$ sampler is executed using default values of the algorithmic parameters and $\epsilon = 0.025$, $K = 15$, and $M = 200,000$. Tables 3 and 4 and Figures 5, 6, and 7 summarize our main findings.

Table 4 compares the computational efficiency of ABC-PMC and DREAM$_{(\mathrm{ABC})}$. For completeness, we also list the results of DREAM using a residual-based Gaussian likelihood function, hereafter referred to as DREAM$_{(\mathrm{RBGL})}$. The DREAM$_{(\mathrm{ABC})}$ algorithm has an acceptance rate (AR, %) of about 3.14% and requires $200,000$ SAC-SMA model evaluations to generate $40,000$ posterior samples. The ABC-PMC sampler, on the contrary, is far less efficient (AR = 0.046%) and needs about 2.2 million function evaluations

to produce 1,000 posterior samples. This constitutes a more than ten times difference in sampling efficiency, and favors the use of DREAM$_{(ABC)}$ for diagnostic inference of complex and CPU-intensive models.

The acceptance rate of DREAM$_{(RBGL)}$ of 4.54% is much lower than the theoretical (optimal) value of about 23.4% for the considered dimensionality of the target distribution. This finding is not surprising and can be explained by the non-ideal properties of the SAC-SMA response surface [*Duan et al.*, 1992], which, to a large extent, are inflicted by poor numerics [*Clark and Kavetski*, 2010; *Kavetski and Clark*, 2010; *Schoups and Vrugt*, 2010]. The use of an explicit, Euler-based, integration method introduces pits and local optima (amongst others) on the response surface, and their presence deteriorates the search efficiency of MCMC methods. An implicit, time-variable, integration method would give a smoother response surface but at the expense of an increase in CPU-time. This increase in computational cost, will however, be balanced by a decrease in the number of model evaluations needed for a MCMC algorithm to converge to a limiting distribution.

Figure 5 presents histograms of the marginal posterior distributions derived with ABC-PMC (top panel), DREAM$_{(ABC)}$ (middle panel) and DREAM$_{(RBGL)}$ (lower panel). We display the results of a representative set of six SAC-SMA parameters and plot, from left to right across each panel, the posterior distributions of PCTIM, ADIMP, LZFSM, LZFPM, and LZPK. The x-axis matches exactly the ranges of each parameter used in the (uniform) prior distribution.

The marginal distributions derived from both sampling methods are in good agreement, and exhibit similar functional shapes. This inspires confidence in the ability of DREAM$_{(ABC)}$ to correctly sample the target distribution. Most histograms extent a large

part of the prior distribution, which suggests that the parameters are poorly defined by calibration against the four different summary statistics. This finding is perhaps not surprising. The four metrics used in this study are not sufficient, and extract only a portion of the information available in the discharge calibration data set. We will revisit this issue in the final paragraph of this section. Information theory can help to determine an approximate set of sufficient statistics, but this is beyond the scope of the present paper.

We can further constrain the behavioral (posterior) parameter space by adding other signatures of catchment behavior to the current set of summary metrics. But, it is not particulary clear whether this would actually support the purpose of diagnostic model evaluation in which the (our) goal is not to just find the best possible fit of some model to some data set, but rather to detect and pinpoint (epistemic) errors arising from inadequate or incomplete process representation. The chosen metrics appear relatively insensitive to rainfall data errors (not shown herein) and therefore exhibit useful diagnostic power. The bottom panel illuminates what happens to the SAC-SMA parameters if a least-squares type likelihood function is used for posterior inference. The parameters appear to be much better resolved by calibration against the observed discharge data but the remaining error residuals violate assumptions of homoscedasticity, normality and independence (not shown in detail). In part, this is due to a lack of treatment of rainfall data errors, whose probabilistic properties are difficult to accurately represent in a likelihood function. The generalized likelihood function of *Schoups and Vrugt* [2010] provides ways to handle nontraditional residual distributions, nevertheless, this approach does not separate the contribution of individual error sources, and is therefore unable to provide insights into model malfunctioning. Note that the histograms of PCTIM, ADIMP, LZFSM, LZFPM,

and LZPK are relatively tight and well described by a normal distribution, except for PCTIM which is hitting its lower bound.

To illustrate how the SAC-SMA posterior parameter uncertainty translates into modeled discharge uncertainty, please consider Figure 6 that presents time series plots of the 95% streamflow simulation uncertainty ranges (gray region) for a selected 500-day portion of the 3-year evaluation period derived from the posterior samples of ABC-PMC (top panel) and DREAM$_{(ABC)}$ (middle panel). The observed discharge data are indicated with solid circles. Both time-series plots are in excellent agreement with simulated discharge dynamics that appear visually very similar and uncertainty ranges that envelop a large majority of the streamflow observations. Epistemic errors are not readily visible, yet this requires much further analysis possibly with the use of additional summary metrics. Previous results for this data set presented in *Vrugt and Sadegh* [2013], demonstrated an inability of the 7-parameter hmodel [*Schoups and Vrugt*, 2010] to simulate accurately the immediate response of the watershed to rainfall. This structural error can be resolved by model correction, a topic that will be studied in future publications. Note that the posterior mean RMSE derived with DREAM$_{(ABC)}$ (0.72 mm/day) is somewhat lower than its counterpart from ABC-PMC (0.81 mm/day). This inconsistency conveys a difference in sampling density and posterior approximation.

For completeness, the bottom panel of Figure 6 plots the 95% streamflow simulation uncertainty ranges derived from formal Bayes using a least-squares likelihood function. To enable a direct comparison with the results for diagnostic inference in the top two panels, we only consider the effect of parameter uncertainty on simulated discharge dynamics. The coverage has decreased substantially to about 13%, which is hardly surprising given

the relatively small width of the marginal distributions shown in Fig. 5. The RMSE of the posterior mean SAC-SMA simulation (0.55 mm/day) is considerably lower than its counterparts derived from ABC-PMC and DREAM$_{(ABC)}$. This finding is not alarming but warrants some discussion. The four summary metrics used for diagnostic inference only extract partial information from the available discharge observations. This insufficiency makes it difficult to find a posterior model that "best" fits, in least-squares sense, the streamflow data, which is expected from a Gaussian likelihood function with homoscedastic measurement error. Also, the main purpose of diagnostic model evaluation with ABC is not that of model calibration, but rather to provide insights into model malfunctioning. Residual-based model calibration approaches provide little guidance on this issue which limits our ability to learn from the calibration data.

Table 5 presents summary variables (coverage, width, root mean square error, bias, and correlation coefficient) of the performance of the posterior mean SAC-SMA discharge simulation derived from the samples of ABC-PMC, DREAM$_{(ABC)}$ and DREAM$_{(RBGL)}$. We list results for the 5-year calibration and 3-year evaluation period. These statistics confirm our previous findings. The two different ABC sampling methods provide very similar results, and exhibit a better coverage of the discharge observations, larger width of the 95% simulation uncertainty ranges, and higher posterior mean RMSE than least-squares inference. The performance of the SAC-SMA model does not deteriorate during the evaluation period. In fact, the RMSE of the ABC derived posterior mean simulation substantially improves during the evaluation period, whereas this is not the case with least-squares fitting. This is a heartening prospect, and suggests (among others) that the

chosen summary metrics at least partially represent the underlying signatures of watershed behavior.

To provide more insights into the convergence behavior of DREAM$_{(ABC)}$, Figure 7 plots the evolution of the $\hat{R}$-statistic of *Gelman and Rubin* [1992]. Each of the SAC-SMA parameters is coded with a different color. About $160,000$ SAC-SMA model evaluations are required to converge to a limiting distribution. This marks a significant improvement in sampling efficiency over the ABC-PMC sampler which requires about 2.2 million function evaluations to create $1,000$ posterior samples.

We now turn our attention to the simulated values of the summary metrics. Figure 8 plots histograms of the posterior summary statistics derived with ABC-PMC (top panel) and DREAM$_{(ABC)}$ (middle panel). The observed values of summary statistics are separately indicated in each plot with a red cross. The marginal distributions of summary metrics generally center around their measured values with the exception of the histogram of S_1 (annual runoff coefficient) that appears heavily skewed to the right. This points to a potential deficiency in the SAC-SMA model structure, yet this requires further analysis. For completeness, the bottom panel plots the posterior summary metric distributions derived from a residual-based likelihood function. This approach provides the closest fit to the observed streamflow data, but at the expense of summary metrics S_3 and S_4 (flow duration curve) that deviate considerably from their observed values. This finding highlights a profound difference in methodology between diagnostic model evaluation with ABC and residual-based inference. Of course, we could have used a different, and perhaps more reasonable, likelihood function for the error residuals [e.g. *Schoups and Vrugt* [2010]].

This would affect some of our findings in Figure 8. Nevertheless, this is outside the scope of the present paper, and we leave such comparison for future work.

Finally, Figure 9 presents bivariate scatter plots of the posterior samples generated with DREAM$_{(ABC)}$. We display the results for a representative set of all SAC-SMA parameter pairs including (A) PCTIM - LZPK, (B) LZSK - LZPK, (C) PCTIM - LZFPM, (D) PCTIM - LZSK, (E) LZPK - PFREE, and (F) ADIMP - LZPK. The axes in each of the six plots are in agreement with those used in the prior distribution. The bivariate posterior samples are confined to a densely-sampled rectangular (or square) space, and occupy a significant portion of the prior distribution. The white area immediately outside of the sampled space is made up of nonbehavioral solutions with fitness values smaller than zero (at least one summary metric is ϵ removed from its measured counterpart). The binary acceptance rule used in DREAM$_{(ABC)}$ introduces a rather sharp demarcation of the behavioral solution space, nevertheless the sampled posterior distribution is in excellent agreement with its counterpart derived from ABC-PMC (see Figure 5). One could argue that that for this type of target distribution uniform random sampling should suffice. However, this method, which is at the heart of ABC-REJ, is highly inefficient in multi-dimensional parameter spaces. Many millions of function evaluations would be needed to provide a sufficient sample of the posterior distribution. This is rather cumbersome, particularly if, as in the present case (not shown), the posterior samples exhibit parameter correlation.

The focus of the present paper has been on improving ABC sampling efficiency to permit diagnostic inference of complex system models involving multi-dimensional parameter and summary metric spaces. Subsequent work can now focus on the intended purpose of

643 diagnostic model evaluation and that is to help detect, diagnose, and resolve model struc-
644 tural deficiencies [*Vrugt and Sadegh*, 2013]. Commonly used model-data fusion approaches
645 provide limited guidance on this important issue, in large part because of their aggregated
646 treatment of input (forcing) data and epistemic errors. The use of summary statistics for
647 statistical inference holds great promise, not only because signatures of system behav-
648 ior are much less sensitive to, for instance, precipitation data errors than residual-based
649 model fitting approaches, but also because the metrics can be devised in such a way that
650 they relate directly to individual process descriptions and thus model components. This
651 has important diagnostic advantages. Failure to fit one or more summary metrics can be
652 directly addressed through correction of the responsible model component(s). This itera-
653 tive process of inference, adjustment and refinement constitutes the basis of the scientific
654 method. This also leaves the possibility to collect additional data to help validate new
655 components of the model.

656 A recurrent issue with the application of diagnostic inference will be sufficiency of the
657 summary metrics. Ideally, the summary metrics contain as much information as the
658 original data itself. Unfortunately, for most systems it will be rather difficult to find
659 a set of sufficient summary statistics, unless each calibration data measurement is used
660 as independent metric (e.g. *Sadegh and Vrugt* [2013]) but this defeats the purpose of
661 diagnostic inference. Actually, it is not particularly clear whether sufficiency of the metrics
662 is required to help detect and resolve epistemic errors. If deemed necessary, then one
663 possible solution is to adapt formal Bayes and to use the summary metrics as an explicit
664 prior. This type of approach has shown to significantly enhance the results of geophysical
665 inversion [*Lochbühler et al.*, 2014].

5. Summary and Conclusions

The paper by *Vrugt and Sadegh* [2013] has introduced approximate Bayesian computation (ABC) as vehicle for diagnostic model evaluation. Successful application of this methodology requires availability of an efficient sampling method that rapidly explores the space of behavioral models. Commonly used rejection sampling approaches adopt a boxcar kernel (0/1) to differentiate between behavioral ("1") and non-behavioral ("0") solutions, and use full-dimensional updating in pursuit of the posterior parameter distribution. This approach might work well for low-dimensional problems (e.g. $d \leq 10$) but is not particularly efficient in high-dimensional parameter spaces which require partial (subspace) sampling to rapidly locate posterior solutions.

In this paper, we have introduced DREAM$_{(ABC)}$ to permit diagnostic inference of complex system models. This approach uses Metropolis-within-Gibbs simulation with DREAM [*Vrugt et al.*, 2008, 2009] to delineate the space of behavioral (posterior) models. Three different case studies involving a simple one-dimensional toy problem, a 20-dimensional mixture of bivariate distributions, and a 14-dimensional hydrologic model calibration problem illustrate that DREAM$_{(ABC)}$ is about 3 - 1,000 times more efficient than commonly used ABC sampling approaches. This gain in sampling efficiency increases with dimensionality of the parameter space.

The source code of DREAM$_{(ABC)}$ is written in MATLAB and available upon request from the second author: `jasper@uci.edu`. This code includes (amongst others) the three different case studies considered herein and implements many different functionalities (postprocessing and visualization tools, convergence and residual diagnostics) to help users analyze their results.

688 **Acknowledgments.** Both authors highly appreciate the support and funding from the
689 UC-Lab Fees Research Program Award 237285. The comments of the three anonymous
690 referees have improved the current version of this manuscript.

References

Barnes, C., S. Filippi, M.P.H. Stumpf, and T. Thorne (2011), Considerate approaches to achieving sufficiency for ABC model selection, `http://arxiv.org/pdf/1106.6281v2.pdf`, 1–21.

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

Beaumont, M.A., J.M. Cornuet, J.M., Marin, and C.P. Robert (2009), Adaptive approximate Bayesian computation, *Biometrika*, *asp052*, 1–8.

Beaumont, M.A. (2010), Approximate Bayesian Computation in Evolution and Ecology, *Annual Review of Ecology, Evolution, and Systematics*, *41*, 379–406.

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

Blum, M.G.B., and O. François (2010), Non-linear regression models for approximate Bayesian computation, *Statistics and Computing*, *20*, 63–73.

Burnash, R.J., R.L. Ferral, R.L., and R.A. McGuire (1973), A generalized streamflow simulation system: Conceptual modeling for digital computers, Joint Federal-State River Forecast Center, Sacramento, CA, USA.

Clark, M.P., and D. Kavetski (2010), Ancient numerical daemons of conceptual hydrological modeling: 1. Fidelity and efficiency of time stepping schemes, *Water Resources Research*, *46*, W10510, doi:10.1029/2009WR008894.

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

Del Moral, P., A. Doucet, and A. Jasra (2008), An adaptive sequential Monte Carlo method for approximate Bayesian computation, Technical Report, Imperial College London.

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

Duan, Q., S. Sorooshian, and V. Gupta (1992), Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resources Research*, *28*(4), 1015–1031.

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

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

Gelman, A., D.B. Rubin (1992), Inference from iterative simulation using multiple sequences, *Statistical Science*, *7*, 457–472.

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

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

Haario, H., E. Saksman, and J. Tamminen (1999), Adaptive proposal distribution for random walk Metropolis algorithm, *Comp. Stat., 14*(3), 375–395.

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

Hastings, H. (1970), Monte Carlo sampling methods using Markov chains and their applications, *Biometrika, 57*, 97–109.

Higuchi, T. (1997), Monte Carlo filter using the genetic algorithm operators, *Journal of Statistical Computation and Simulation*, *5*9, 1–23.

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

Kavetski, D., and M.P. Clark (2010), Ancient numerical daemons of conceptual hydrological modeling: 2. Impact of time stepping schemes on model analysis and prediction, *Water Resources Research*, *46*, W10511, doi:10.1029/2009WR008896.

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

Lochbühler, T., J.A. Vrugt, M. Sadegh, and N. Linde (2014), Summary Statistics from Training Images as Prior Information in Probabilistic Inversion, *Geophysical Research Letters*, Under Review.

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

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller (1953), Equation of state calculations by fast computing machines, *Journal of Chemical Physics*, *21*, 1087–1092.

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

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

Pritchard, J.K., M.T. Seielstad, A. Perez-Lezaun, and M.T. Feldman (1999), Population Growth of Human Y Chromosomes: A Study of Y Chromosome Microsatellites, *Molecular Biology and Evolution*, *16*(12), 1791–1798.

Ratmann, O., C. Andrieu, C. Wiuf, and S. Richardson (2009), 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*, *106*, 1–6.

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

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

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

Smith, T., A. Sharma, L. Marshall, R Mehrotra, and S. Sisson (2010), Development of a formal likelihood function for improved Bayesian inference of ephemeral catchments, *Water Resources Research*, *46*(12), W12551.

Sunnåker M, A.G. Busetto, E. Numminen J. Corander M. Foll, and C. Dessimoz (2013), Approximate Bayesian Computation, *Plos Computational Biology*, *9*(1), e1002803, 1–10, doi:10.1371/journal.pcbi.1002803.

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

Toni, T., D. Welch, N. Strelkowa, A. Ipsen, and M.P. Stumpf (2009), Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems, *Journal of the Royal Society Interface*, *6*, 187–202.

Turner, B.M, and T. van Zandt (2012), A tutorial on approximate Bayesian computation, *Journal of Mathematical Psychology*, *56*, 69–85.

Turner, B.M, and P.B. Sederberg (2012), Approximate Bayesian computation with differential evolution, *Journal of Mathematical Psychology*, *56*(5), 375–385, doi:10.1016/j.jmp.2012.06.004.

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

Vrugt, J.A., C.J.F. ter Braak, C.G.H. Diks, D. Higdon, B.A. Robinson, and J.M. Hyman (2009), Accelerating Markov chain Monte Carlo simulation by differential evolution

with self-adaptive randomized subspace sampling, *International Journal of Nonlinear Sciences and Numerical Simulation, 10(3)*, 273–290.

Vrugt, J.A., and C.J.F. ter Braak, (2011), 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*, *15*, 3701–3713, doi:10.5194/hess-15-3701-2011.

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

Wegmann, D., C. Leuenberger, and L. Excoffier (2009), Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood, *Genetics*, *182*(4), 1207–1218.

Table 1: Case study I: 1-dimensional toy example. We list the (final) epsilon value, acceptance rate, AR (%) and number of function evaluations needed to sample the target distribution.

	ϵ	AR(%)	Func. Eval.
ABC-REJ	0.025	0.259	386,742
ABC-PMC	0.025	0.585	170,848
DREAM$_{(ABC)}$	0.025	1.708	50,000

Table 2: Case study II: 20-dimensional bivariate Gaussian distribution. We list the final epsilon value, acceptance rate, AR (%) and number of function evaluations needed to sample a stationary distribution.

	ϵ	AR(%)	Func. Eval.
ABC-PMC	0.600	0.019	5,143,989
DREAM$_{(ABC)}$	0.025	19.717	200,000

Table 3: Prior ranges of the SAC-SMA model parameters and their posterior mean values derived for the French Broad river basin data using the ABC-PMC and DREAM$_{(ABC)}$ algorithms. We also summarize the results of a residual-based Gaussian likelihood function, DREAM$_{(RBGL)}$.

Parameter	Range	Posterior Parameter Mean		
		ABC-PMC	DREAM$_{(ABC)}$	DREAM$_{(RBGL)}$
UZTWM	1-150	81.812	82.676	39.725
UZFWM	1-150	73.722	67.945	10.531
UZK	0.1-0.5	0.299	0.301	0.387
PCTIM	0-0.1	0.007	0.010	0.003
ADIMP	0-0.4	0.118	0.121	0.189
ZPERC	1-250	132.350	130.883	109.207
REXP	1-5	2.972	2.847	4.880
LZTWM	1-500	396.792	363.329	456.619
LZFSM	1-1000	282.049	246.326	146.584
LZFPM	1-1000	796.113	728.855	733.093
LZSK	0.01-0.25	0.145	0.142	0.124
LZPK	0.0001-0.025	0.006	0.006	0.009
PFREE	0-0.6	0.292	0.304	0.530
RRC	0-1	0.630	0.581	0.321

Table 4: Case study III: 14-dimensional SAC-SMA model calibration problem. We list the final epsilon value, acceptance rate, AR (%) and number of function evaluations needed to sample the posterior distribution. We also include the results of DREAM using a residual-based Gaussian likelihood function.

	ϵ	AR(%)	Func. Eval.
ABC-PMC	0.025	0.046	2,173,490
DREAM$_{(ABC)}$	0.025	3.135	200,000
DREAM$_{(RGBL)}$	N/A	4.543	200,000

Table 5: Performance of the SAC-SMA model for the calibration and evaluation data period of the French Broad river basin. We summarize the coverage (%) and average width (mm/d) of the 95% simulation intervals (due to parameter uncertainty), and the RMSE (mm/d), bias (%) and correlation coefficient, R (-) of the posterior mean SAC-SMA simulation.

	Coverage (%)		Width (mm/d)		RMSE (mm/d)		Bias (%)		R (-)	
	Calib.	Eval.	Calib.	Eval.	Calib.	Eval.	Calib.	Eval.	Calib.	Eval.
ABC-PMC	70.991	65.328	1.301	1.185	0.932	0.812	3.281	-3.721	0.863	0.832
DREAM$_{(ABC)}$	71.100	66.515	1.408	1.272	0.831	0.722	3.731	-3.245	0.893	0.866
DREAM$_{(RBGL)}$	18.829	12.500	0.155	0.141	0.539	0.545	2.767	-0.042	0.956	0.924

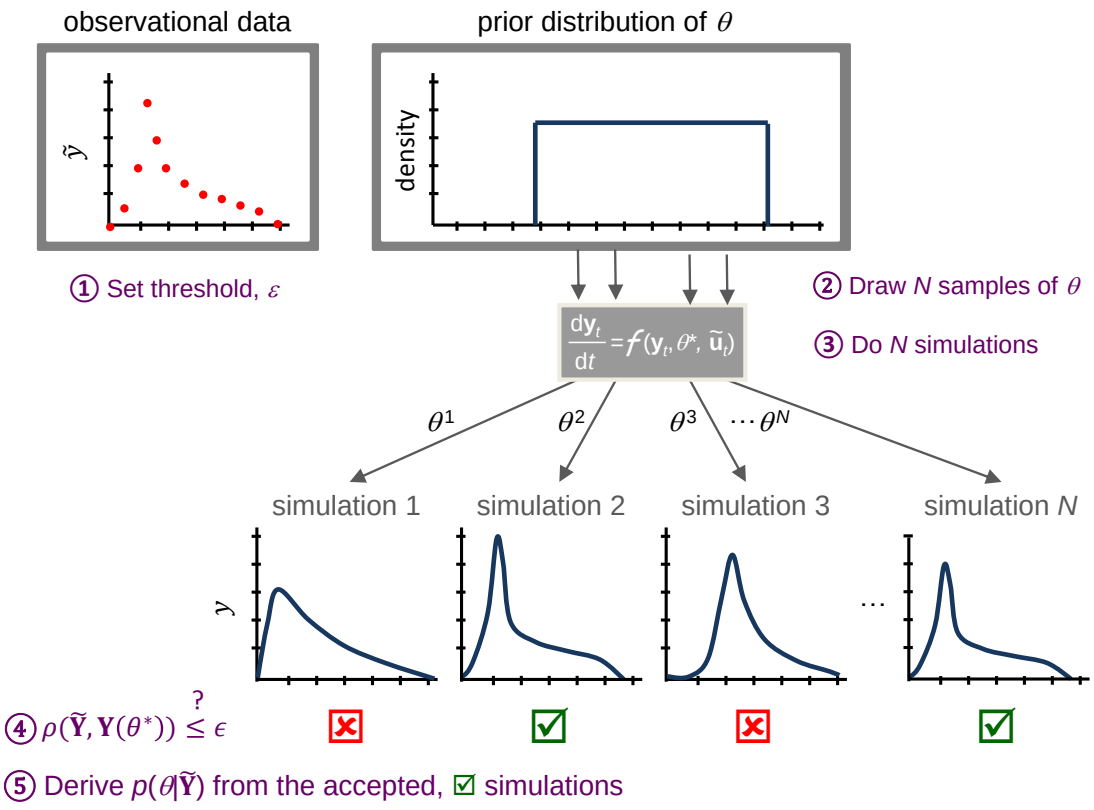

Figure 1: Conceptual overview of approximate Bayesian computation (ABC) for a hypothetical one-dimensional parameter estimation problem. First, N samples are drawn from a user-defined prior distribution, $\theta^* \sim p(\theta)$. Then, this ensemble is evaluated with the model and creates N model simulations. If the distance between the observed and simulated data, $\rho(\tilde{\mathbf{Y}}, \mathbf{Y}(\theta^*))$ is smaller than or equal to some nominal value, ϵ then θ^* is retained, otherwise the simulation is discarded. The accepted samples are then used to approximate the posterior parameter distribution, $p(\theta|\tilde{\mathbf{Y}})$. Note that for sufficiently complex models and large data sets the probability of happening upon a simulation run that yields precisely the same simulated values as the observations will be very small, often unacceptably so. Therefore, $\rho(\tilde{\mathbf{Y}}, \mathbf{Y}(\boldsymbol{\theta}^*))$ is typically defined as a distance between summary statistics of the simulated, $S(\mathbf{Y}(\boldsymbol{\theta}^*))$ and observed, $S(\tilde{\mathbf{Y}})$ data, respectively. Modified after *Sunnåker et al.* [2013].

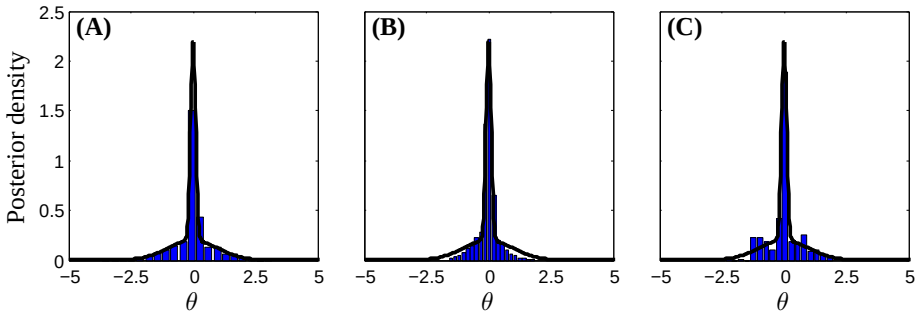

Figure 2: One-dimensional mixture distribution (solid black line) and histogram of the posterior samples derived from (A) Rejection sampling (ABC-REJ) , (B) Population Monte Carlo sampling (ABC-PMC), and (C) MCMC simulation with DREAM$_{(ABC)}$.

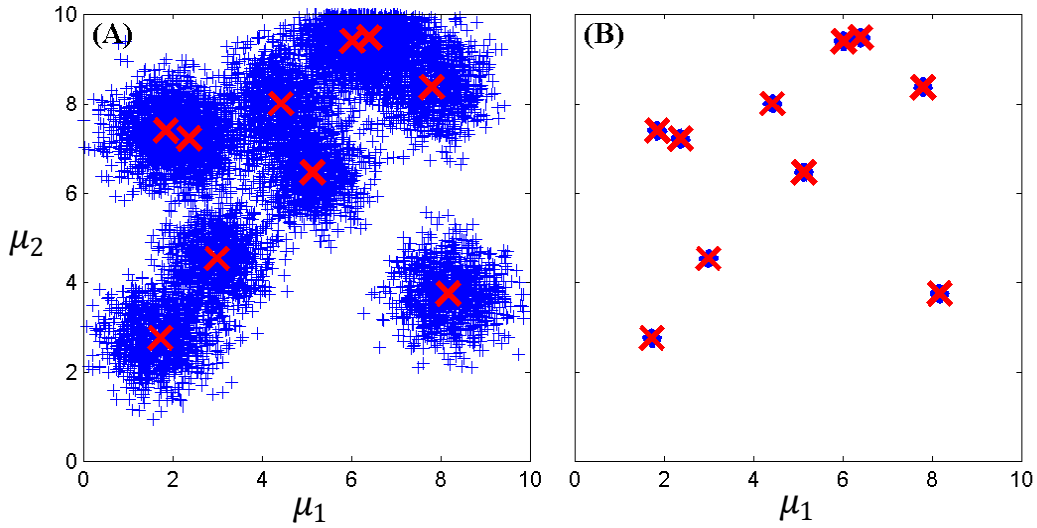

Figure 3: Two-dimensional scatter plots of the posterior samples, "+" generated with (A) Population Monte Carlo sampling, and (B) MCMC simulation with DREAM$_{(ABC)}$. The true values of μ_1 and μ_2 are indicated with a cross, "×". The ABC-PMC samples are very dispersed and significantly over estimate the actual width of the target distribution.

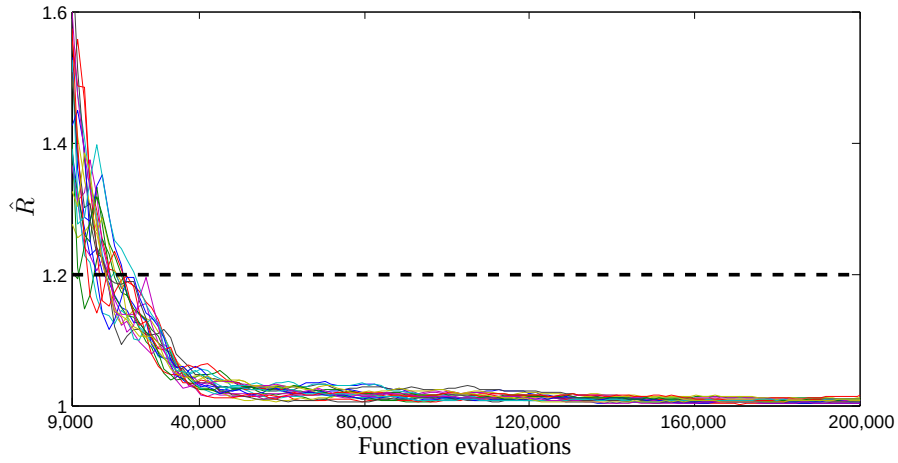

Figure 4: Trace plots of the $\hat{R}$-statistic of the sampled Markov chains with DREAM$_{(ABC)}$ for the 20-dimensional bivariate normal distribution. Each parameter is coded with a different color. The dashed line denotes the default threshold used to diagnose convergence to a limiting distribution.

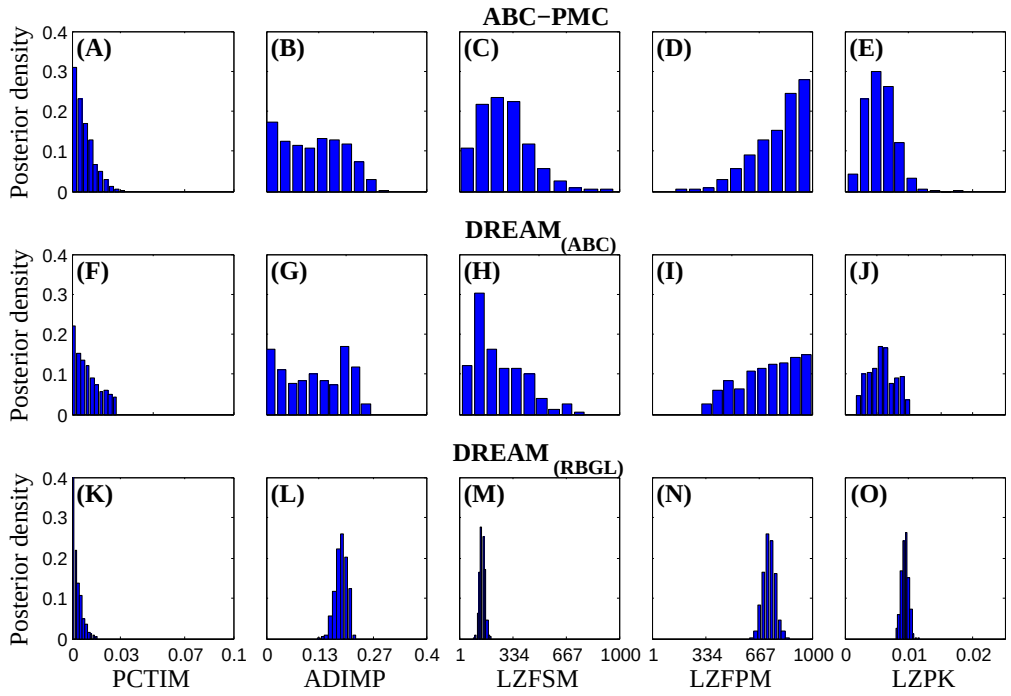

Figure 5: Marginal distributions of PCTIM, ADIMP, LZFSM, LZFPM and LZPK derived from the posterior samples created with ABC-PMC (upper panel) and DREAM$_{(ABC)}$ (middle panel). The histograms of the five SAC-SMA parameters occupy almost their entire prior distribution, which suggests that they are not particularly well identifiable by calibration against the observed annual baseflow index, annual runoff coefficient, and flow duration curve, respectively. The results of both ABC sampling methods are in good (visual) agreement, which inspires confidence in the ability of DREAM$_{(ABC)}$ to correctly sample the underlying target distribution. The bottom panel displays histograms of the SAC-SMA parameters derived using a classical residual-based (Gaussian) likelihood function. The SAC-SMA parameters are much better resolved, but these results cannot be justified given (amongst others) a lack of treatment of rainfall data errors.

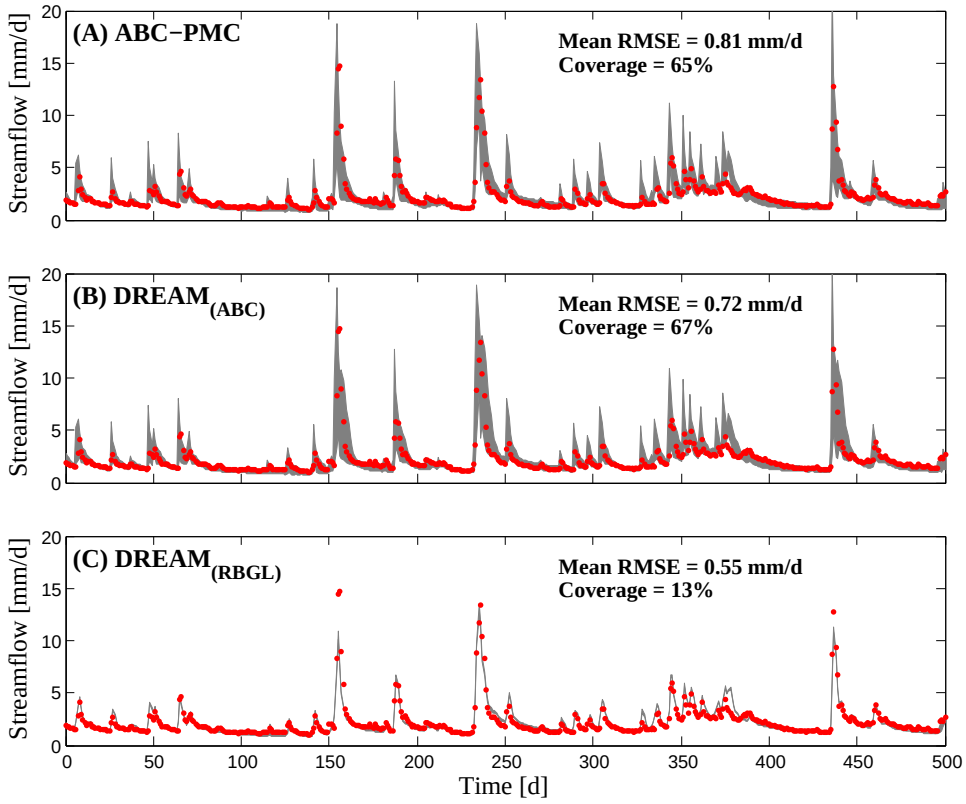

Figure 6: 95% posterior simulation uncertainty ranges (gray region) of the SAC-SMA model for a selected portion of the evaluation data set of the French Broad watershed. The top two panels displays the results of diagnostic model evaluation using (A) ABC-PMC, and (B) DREAM$_{(ABC)}$, whereas the bottom panel depicts the results of DREAM with a classical residual-based Gaussian likelihood function. The observed discharge values are indicated with the red dots. The SAC-SMA simulation intervals derived from ABC-PMC and DREAM$_{(ABC)}$ are very similar and encapsulate a large part of the discharge observations. The DREAM$_{(RBGL)}$ uncertainty ranges, on the other hand, exhibit a much lower coverage, but closer track the observed discharge data.

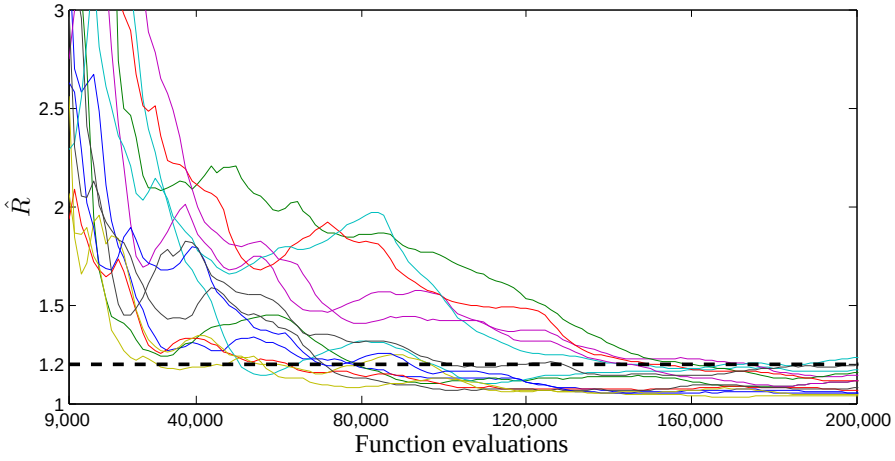

Figure 7: Evolution of the $\hat{R}$-statistic for the parameters of the SAC-SMA model using DREAM$_{(ABC)}$ and discharge data from the French Broad watershed. Each of the parameters is coded with a different color. The dashed line denotes the default threshold used to diagnose convergence to a limiting distribution.

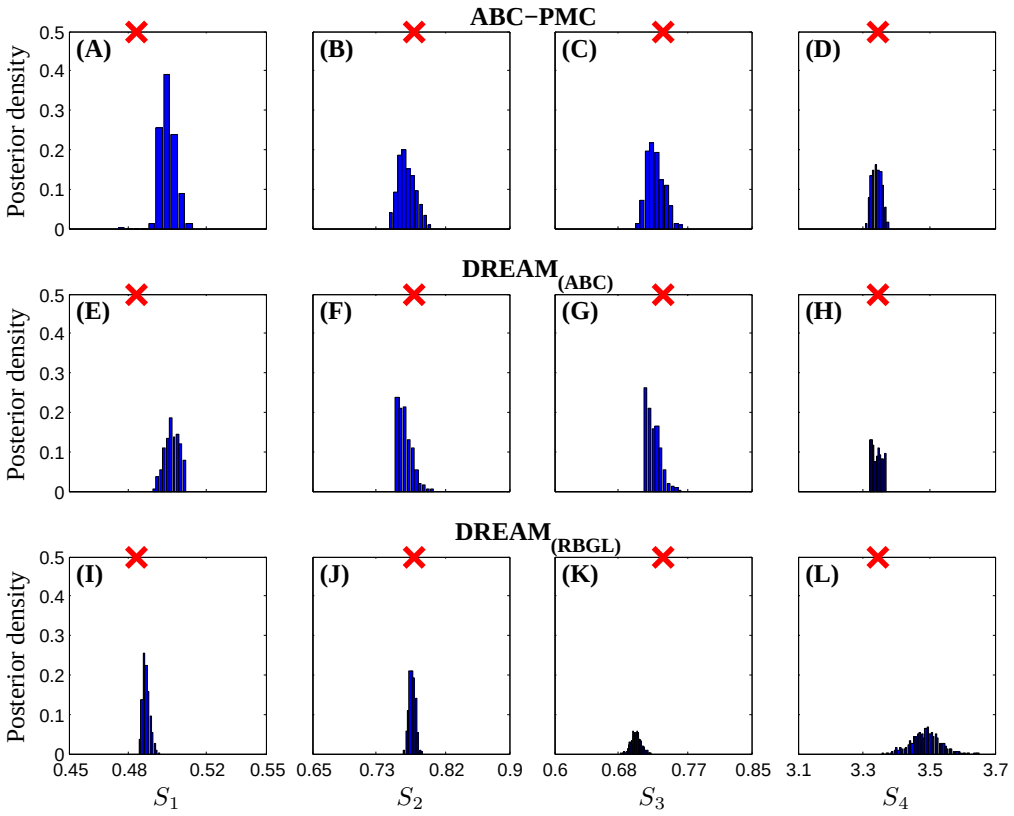

Figure 8: Histograms of the SAC-SMA derived summary statistics, S_1 (runoff), S_2 (baseflow), S_3 and S_4 (flow duration curve) of the posterior samples from ABC-PMC (top panel) and DREAM$_{(ABC)}$ (middle panel). The bottom panel displays the results of DREAM with a residual-based Gaussian likelihood function. The observed values of the summary metrics are separately indicated in each plot using the "×" symbol. While $S_2 \rightarrow S_4$ center around their observed value (×) for the ABC analysis, the marginal posterior distribution of S_1 is skewed to the right and does not encapsulate its measured value. This demonstrates that model is unable to simultaneously satisfy all the four different summary metrics used herein. This points to a structural deficiency in the SAC-SMA model structure, which will be investigated in more detail in subsequent papers.

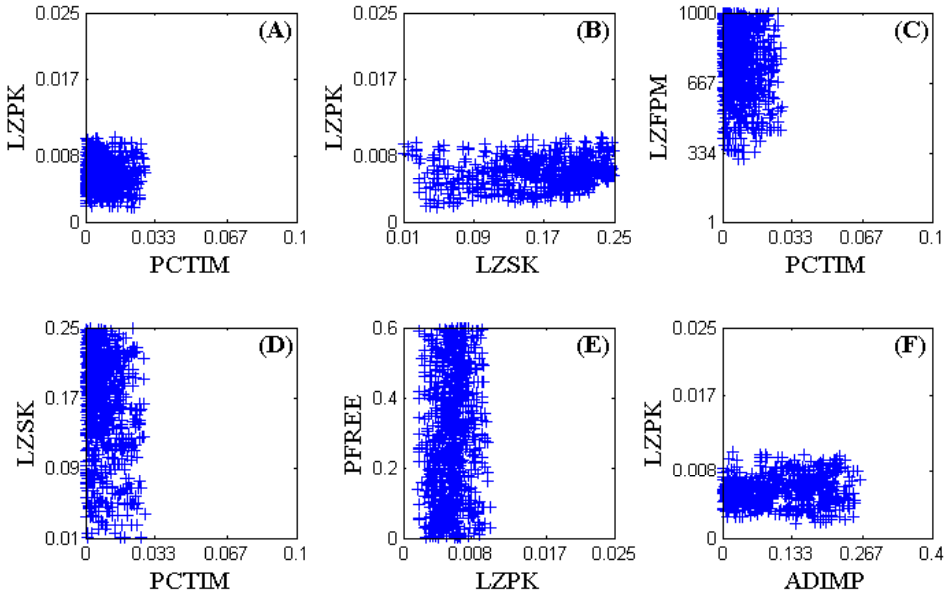

Figure 9: Two-dimensional scatter plots of the posterior samples, "+" generated with DREAM$_{(ABC)}$ for six different pairs of parameters of the SAC-SMA model. We restrict our attention to the (A) PCTIM - LZPK, (B) LZSK - LZPK, (C) PCTIM - LZFPM, (D) PCTIM - LZSK, (E) LZPK - PFREE, and (F) ADIMP - LZPK space, respectively. The bivariate posterior samples occupy a well-defined hypercube interior to the prior distribution.