Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017


Biogeosciences Discussions Open Access EGU

# Bayesian calibration of terrestrial ecosystem models: A study of advanced Markov chain Monte Carlo methods

**Dan Lu[1,*], Daniel Ricciuto[2], Anthony Walker[2], Cosmin Safta[3], and William Munger[4]**

[1]Computer Science and Mathematics Division, Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN, USA;

[2]Environmental Sciences Division, Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, TN, USA;

[3]Sandia National Laboratories, Livermore, CA, USA;

[4]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA;

*Corresponding author: Dan Lu (**lud1@ornl.gov**)

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017


**Abstract**

Calibration of terrestrial ecosystem models is important but challenging. Bayesian inference implemented by Markov chain Monte Carlo (MCMC) sampling provides a comprehensive framework to estimate model parameters and associated uncertainties using their posterior distributions. The effectiveness and efficiency of the method strongly depend on the MCMC algorithm used. In this study, a Differential Evolution Adaptive Metropolis (DREAM) algorithm was used to estimate posterior distributions of 21 parameters for the data assimilation linked ecosystem carbon (DALEC) model using 14 years of daily net ecosystem exchange data collected at the Harvard Forest Environmental Measurement Site eddy-flux tower. The DREAM is a multi-chain method and uses differential evolution technique for chain movement, allowing it to be efficiently applied to high-dimensional problems, and can reliably estimate heavy-tailed and multimodal distributions that are difficult for single-chain schemes using a Gaussian proposal distribution. The results were evaluated against the popular Adaptive Metropolis (AM) scheme. DREAM indicated that two parameters controlling autumn phenology have multiple modes in their posterior distributions while AM only identified one mode. The calibration of DREAM resulted in a better model fit and predictive performance compared to the AM. DREAM provides means for a good exploration of the posterior distributions of model parameters. It reduces the risk of false convergence to a local optimum and potentially improves the predictive performance of the calibrated model.

**Keywords:** Bayesian calibration, MCMC sampling, AM algorithm, DREAM algorithm, DALEC model, multimodality, terrestrial ecosystem models.

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




## 1 Introduction

Prediction of future climate heavily depends on accurate predictions of the concentration of carbon dioxide ($CO_2$) in the atmosphere. Predictions of atmospheric $CO_2$ concentrations rely on terrestrial ecosystem models (TEMs) to simulate the $CO_2$ exchange between the land surface and the atmosphere. TEMs typically involve a large number of biogeophysical and biogeochemical processes, the representation of which requires knowledge of many process parameters. Some parameters can be determined directly from experimental and measurement data, but many are also estimated through model calibration. Estimating these parameters indirectly from measurements (such as the net ecosystem exchange (NEE) data) is a challenging inverse problem.

Various parameter estimation methods have been applied to TEMs. For an overview, one can refer to the OptIC (Optimization InterComparison) project (Trudinger et al., 2007) and the REFLEX (REgional FLux Estimation eXperiment) project (Fox et al., 2009). In classical optimization based approaches, inverse problems with a large number of parameters can often be ill-posed in that the solution may not be unique or even may not exist (O'Sullivan, 1986). As an alternative approach, the Bayesian framework provides a comprehensive solution to this problem. In Bayesian methods, the model parameters are treated as random variables and their posterior probability density functions (PPDFs) represent the estimation results. The PPDF incorporates prior knowledge of the parameters, mismatch between model and observations, and observation uncertainty (Lu et al., 2012). Thus, compared to other approaches in inverse problems, Bayesian inference not only estimates model parameters, but also quantifies associated uncertainty using a full probabilistic description.

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




64 Two types of Bayesian methods are widely used in parameter estimation of TEMs,
65 variational data assimilation (VAR) methods (Talagrand and Courtier, 1987) and Markov chain
66 Monte Carlo (MCMC) sampling. VAR methods are computationally efficient, however, they
67 assume that the prior parameter values and the observations follow a Gaussian distribution, and
68 they require the model to be differentiable with respect to all parameters for optimization. In
69 addition, VAR methods can only identify a local optimum and approximate the PPDF by a
70 Gaussian function (Rayner et al, 2005; Ziehn et al., 2012). In contrast, MCMC sampling makes
71 no assumptions about the structure of the prior and posterior distributions of model parameters or
72 observation uncertainties. Moreover, the MCMC methods, in principle, can converge to the true
73 PPDF with an identification of all possible optima. Although it is more computationally intensive
74 than VAR approaches, MCMC sampling is being increasingly applied in the land surface
75 modeling community (Dowd, 2007; Zobitz et al, 2011).
76 One widely used MCMC algorithm is adaptive Metropolis (AM) (Haario et al. 2001). For
77 example, Fox et al. (2009) applied the AM in their comparison of different algorithms for the
78 inversion of a terrestrial ecosystem model; Järvinen et al. (2010) utilized the AM for estimation
79 of ECHAM5 climate model closure parameters; Hararuk et al. (2014) employed the AM for
80 improvement of a global land model against soil carbon data; and Safta et al. (2015) used the
81 AM to estimate parameters in the data assimilation linked ecosystem carbon model. The AM
82 algorithm uses a single Markov chain that continuously adapts the covariance matrix of a
83 Gaussian proposal distribution using the information of all previous samples collected in the
84 chain so far (Haario et al., 1999). As a single-chain method, AM has difficulty in traversing
85 multi-dimensional parameter space efficiently when there are numerous significant local optima;
86 and AM can be unreliable for estimating the PPDFs of the parameters that exhibit strong

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




87 correlation. In addition, the AM algorithm uses a multivariate Gaussian distribution as the
88 proposal to generate candidate samples and evolve the chain. AM, therefore, particularly suitable
89 for Gaussian shaped PPDFs, but it may not converge properly to the distributions with multiple
90 modes and those with heavy tails like the Cauchy distributions. Moreover, AM suffers from
91 uncertainty about how to initialize the covariance of the Gaussian proposal. Poor initialization of
92 the proposal covariance matrix results in slow adaptation and inefficient convergence.
93 The Gaussian proposal is also widely used in non-AM MCMC studies that involve TEMs.
94 For example, Ziehn et al. (2012) used the Gaussian proposal for the MCMC simulation of the
95 BETHY model (Knorr and Heimann, 2011) and Ricciuto et al. (2008, 2011) utilized the
96 Gaussian proposal in their MCMC schemes to estimate parameters in a terrestrial carbon cycle
97 model. The single-chain and Gaussian-proposal MCMC approaches have limitations in
98 sufficiently exploring the full parameter space and share low convergence in sampling the non-
99 Gaussian shaped PPDFs and thus may end up with a local optimum with inaccurate uncertain
100 representation of the parameters. Therefore, this poses a question on whether the AM and the
101 widely used MCMC algorithms with Gaussian proposal generate a representing sample of the
102 posterior distributions of the underlying model parameters. While we expect that
103 computationally expensive sampling methods for parameter estimation yield a global optimum
104 with an accurate probabilistic description, in reality, we may in many cases obtain a local
105 optimum with an inaccurate PPDF due to the limitations of these algorithms.
106 In this study, we employ the differential evolution adaptive Metropolis (DREAM)
107 algorithm (Vrugt et al., 2008, 2009a; Zhang et al., 2013; Lu et al., 2014) for an accurate Bayesian
108 calibration of an ecosystem carbon model. The DREAM scheme runs multiple interacting chains
109 simultaneously to explore the entire parameter space globally. During the search, DREAM does

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




not rely on a specific distribution, like the Gaussian distribution used in most MCMC schemes, to move the chains. Instead, it uses the differential evolution optimization method to generate the candidate samples from the collection of chains (Price et al., 2005). This feature of DREAM eliminates the problem of initializing the proposal covariance matrix and enables efficient handling of complex distributions with heavy tails and strong correlations. In addition, as a multi-chain method, DREAM can efficiently sample multimodal posterior distributions with numerous local optima. Thus, the DREAM scheme is particularly applicable to complex and multimodal optimization problems.

While multimodality is a potential feature of parameters in complex models (Thibault et al, 2011; Zhang et al., 2013), its existence has not been well documented in terrestrial ecosystem modeling due to the limitations of methods that have been applied in most previous studies. Here we apply both the DREAM and the AM methods to three benchmark functions and a TEM to estimate the parameter distributions. In the latter case, we estimate the PPDFs of 21 process parameters in the data assimilation linked ecosystem carbon (DALEC) model. The objectives of this study are to (1) present a statistically sound methodology to solve the parameter estimation problems in complex TEMs and to improve the model simulation; (2) characterize parameter uncertainty in detail using accurately sampled posterior distributions; and (3) investigate the effects of model calibration methods on parameter estimation and model performance. This work should provide ecological practitioners with valuable information on model calibration and understanding of the TEMs.

In the following Section 2, we first briefly summarize the general idea of Bayesian calibration, and describe the AM and DREAM algorithms. Then in Section 3, we compare the performance of the two algorithms in sampling three known target distributions. Next in Section

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




4, the application of the DREAM algorithm to the DALEC model is presented and discussed; in comparison, we also discussed the results from the AM sampler. Finally in Section 5, we close this paper with our main conclusions.

## 2 Bayesian calibration and MCMC simulation

### 2.1 Bayesian calibration

Bayesian calibration of a model states that the posterior distribution $p(\mathbf{x}|\mathbf{D})$ of the model parameters $\mathbf{x}$, given observation data $\mathbf{D}$, can be obtained from the prior distribution $p(\mathbf{x})$ of $\mathbf{x}$ and the likelihood function $L(\mathbf{x}|\mathbf{D})$ using Bayes' theorem (Box and Tiao, 1992) via,

$$p(\mathbf{x}\,|\,\mathbf{D}) = cL(\mathbf{x}\,|\,\mathbf{D})\,p(\mathbf{x}) \quad (1)$$

where $c$ is a normalization constant. The prior distribution represents the prior knowledge about the parameters. It is usually inferred from information of previous studies in similar sites or from expert judgment. In the lack of prior information, a common practice is to use uninformative priors within relatively wide parameter ranges such that the prior distribution has little influence on the estimation of the posterior distribution.

The likelihood function measures the model fits to the observations. Selecting a likelihood function suitable to a specific problem is still under study (Vrugt et al., 2009b). A commonly used likelihood function is based on the assumption that the differences between the model simulations and observations are multivariate normally distributed, leading to a Gaussian likelihood such as the work of Fox et al. (2009), Hararuk et al. (2014), and Ricciuto et al. (2008, 2011). In this work, we also use the Gaussian likelihood, with uncorrelated variances that are evaluated from the provided daily observation uncertainties. The effect of data correlations on the inferred parameters was assessed in our previous study (Safta et al., 2015).

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




## 2.2 MCMC sampling

In most environmental problems, the posterior distribution cannot be obtained with an analytical solution and is typically approximated by sampling methods such as MCMC. The MCMC method approximates the posterior distribution by constructing a Markov chain whose stationary distribution is the target distribution of interest. As the chain evolves and approaches the stationary, all the samples after chain convergence are used for posterior distribution approximation, and the samples before convergence, which are affected by the starting states of the chain, are discarded. The earliest MCMC approach is the well-known random walk Metropolis sampler (Metropolis et al., 1953). Assume at iteration $t$ we have obtained the samples $\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_t\}$, where $\mathbf{x}_0$ is the initial sample from a certain distribution (e.g., the parameter prior distribution). Then the Markov chain evolves in the following way. First, a candidate point $\mathbf{z}$ is sampled from a symmetric proposal distribution $q$, which has the property $q(\mathbf{z}\,|\,\mathbf{x}_t) = q(\mathbf{x}_t\,|\,\mathbf{z})$. Next, the candidate point is either accepted or rejected according to a Metropolis ratio α calculated as

$$\alpha = \begin{cases} min\left[\dfrac{p(\mathbf{z}\,|\,\mathbf{D})}{p(\mathbf{x}_t\,|\,\mathbf{D})},\ 1\right] & \text{if } p(\mathbf{x}_t\,|\,\mathbf{D}) > 0 \\ 1 & \text{if } p(\mathbf{x}_t\,|\,\mathbf{D}) = 0 \end{cases} \tag{2}$$

where $p(\mathbf{z}|\mathbf{D})$ and $p(\mathbf{x}_t|\mathbf{D})$ denote the density of the posterior distribution evaluated at $\mathbf{z}$ and $\mathbf{x}_t$, respectively. Lastly, if the candidate is accepted, the chain uses the sample $\mathbf{x}_{t+1} = \mathbf{z}$ at iteration $t$+1; otherwise, it keeps the current sample $\mathbf{x}_{t+1} = \mathbf{x}_t$.

The Metropolis sampler is the basis of many existing MCMC sampling schemes. The well-constructed MCMC schemes have been theoretically proven to converge to the appropriate target distribution $p(\mathbf{x}|\mathbf{D})$ under certain regularity conditions (Robert and Casella, 2004, p.270). However, in practice the convergence rate is often impractically slow, which suggests that within

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




177 a limited finite number of iterations, some inefficient schemes may result in an unrealistic
178 distribution. The inefficiency is typically resulted from an inappropriate choice of the proposal
179 distribution used to generate the candidates. Either wide or narrow proposal distribution can
180 cause inefficient chain mixing and slow chain convergence (Lu et al., 2014). Hence, the
181 definition of the proposal distribution is crucial and determines the efficiency and the practical
182 applicability of the MCMC simulation.

183 **2.3 AM algorithm**

184 The adaptive Metropolis (AM) algorithm is a modification to the standard Metropolis
185 sampler (Haario et al., 2001). The key feature of the AM algorithm is that it uses a single
186 Markov chain that continuously adapts to the target distribution via its calculation of the proposal
187 covariance using all previous samples in the chain. The proposal distribution employed in the
188 AM algorithm is a multivariate Gaussian distribution with means at the current iteration $\mathbf{x}_t$ and a
189 covariance matrix $C_t$ that is updated along the chain evolution. To start the chain, the AM first
190 selects an arbitrary, strictly positive definite initial covariance $C_0$ according to the best prior
191 knowledge that may be very poor. Then after a certain number of iterations $T$, the covariance is
192 updated based on the samples gained so far,

193 $$C_t = \begin{cases} C_0, & t \le T \\ s_d Cov(\mathbf{x}_0, \cdots, \mathbf{x}_t) + s_d \varepsilon I_d, & t > T \end{cases} \quad (3)$$

194 where $\varepsilon$ is a small value chosen to ensure $C_t$ does not become singular, $\boldsymbol{I}_d$ denotes the $d$-
195 dimensional identity matrix and $s_d$ is a scaling parameter depending on the dimensionality of the
196 parameter $\mathbf{x}$ to ensure reasonable acceptance rates. As a basic guideline, Haario et al. (2001)
197 suggested choosing the value $s_d = 2.4^2/d$, which is shown to be optimal for Gaussian targets and
198 Gaussian proposal distributions (Gelman et al., 1995).

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




To apply the AM algorithm, an initial covariance $C_0$ must be defined. The choice of $C_0$ critically determines the success of the algorithm. For example, when the variance of $C_0$ is too large, it is possible that no proposals are accepted within an iteration, and that the chain will remain at the initial state without any movement. This situation continues as the chain evolves, and the use of updated $C_t$ would not make any difference because the variances of $C_t$ are essentially zero since all the previous samples have the same values. In this case, the AM sampler would get stuck in the initial state without exploring the parameter space. To alleviate this problem and start the AM fairly efficiently, we can define $C_0$ based on some prior knowledge about the target distribution. When such information is not available, which is usually the case for complex models, some test simulations are needed. For example, Hararuk et al. (2014) inferred $C_0$ from a test run of 50,000 community land model simulations in estimating the PPDFs of soil carbon related parameters.

The construction of $C_t$ is another critical influence on the AM performance. In practice, some adjustments on $C_t$ are necessary to improve the AM efficiency. For example, on the basis of Eq. (3), the $C_t$ can be shrunk or amplified by some constant according to the chain evolution. When the chain does not have enough movement after a large number of iterations, we shrink $C_t$ a little bit to increase acceptance of new samples, and vice versa. The techniques used in the formulation of $C_0$ and $C_t$ may improve the AM efficiency in some degree for some problems. However, the computational cost spent on applying these techniques is not negligible (such as the test runs used for determining the $C_0$) and some strategies even require artificial control (such as manual adjustment of the scaling factor of $C_t$). Moreover, determining a reasonable $C_0$ and $C_t$ become very difficult for high-dimensional problems.

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




To improve efficiency in high-dimensional case, Haario et al. (2006) extended the standard AM method to componentwise adaptation. This strategy applies the AM on each parameter separately. The proposal distribution of each component is a 1D normal distribution, which is adapted in a similar manner as in the standard AM algorithm. However, the componentwise adaptation does not work well for distributions with a strong correlation. Safta et al. (2015) applied an iterative algorithm to break the original high-dimensional problem into a sequence of steps of increasing dimensionality, with each intermediate step starting with an appropriate proposal covariance based on a test run. This technique provided a rather reasonable proposal distribution, but the computational cost used to define the proposal was rather high.

AM is a single-chain method. As a single chain, it is particularly difficult to judge the convergence. Even the most powerful diagnostics cannot guarantee that the chain has converged to the target distribution (Gelman and Shirley, 2011). One solution to alleviate the problem is running multiple independent chains with widely dispersive starting points and then using the diagnostics for multi-chain schemes, such as the $\hat{R}$ statistic of Gelman and Rubin (1992), to check convergence. The $\hat{R}$ statistic calculates the ratio of between-chain variance to with-in chain variance. When the chain has a good mixing and all the chains converge to the same PPDF, the $\hat{R}$ value is close to one, and in practice the threshold of 1.2 is usually used for convergence diagnosis. On the other hand, when the chain does not mix well and different chains converge to the different portion of the target distribution, it is unlikely that the $\hat{R}$ will reach the value of 1.2 required to declare convergence. Generally, this situation suggests that multiple modes exist in the target PPDF and the AM sampler is unable to identify all the modes using a single chain.

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




### 2.4 DREAM algorithm

The DREAM algorithm is a multi-chain method (Vrugt, 2016). Multi-chain approaches use multiple chains running in parallel for global exploration of the posterior distribution, so they have several desirable advantages over the single-chain methods, particularly when addressing complex problems involving multimodality and having a large number of parameters with strong correlations. In addition, the application of multiple chains allows utilizing a large variety of statistical measures to diagnose the convergence including the $\hat{R}$ statistic mentioned above.

DREAM uses the Differential Evolution Markov Chain (DE-MC) algorithm (ter Braak, 2006) as its main building block. The key feature of the DE-MC scheme is that it does not specify a particular distribution as the proposal, but proposes the candidate point using the differential evolution method based on current samples collected in the multiple chains. Thus, DE-MC can apply to a wide range of problems whose distribution shapes are not necessarily similar to the proposal distribution, and it also removes the requirement of initializing the covariance matrix as in AM. As previously, we denote the sample at iteration $t$ of a single chain by a $d$-dimensional vector $\mathbf{x}_t$, then the samples from $N$ chains at the iteration $t$ construct a $N \times d$ matrix saved in $\mathbf{X}_t$; usually $N$=2$d$. In DE-MC, the candidate point $\mathbf{z}^i$ in each chain $i$={1, 2, …, $N$} are generated by taking a fixed multiple of the difference of two randomly chosen chains of $\mathbf{X}_t$ with indexes $r_1$ and $r_2$, i.e.,

$$\mathbf{z}^i = \mathbf{x}_t^i + \boldsymbol{\gamma}(\mathbf{X}_t^{r_1} - \mathbf{X}_t^{r_2}) + \boldsymbol{\varepsilon}, \quad r_1 \neq r_2 \neq i \tag{4}$$

where the multiplier γ is suggested as $2.4/\sqrt{2d}$ for its optimal performance, and for every $10^{\text{th}}$ iteration γ=1.0 to facilitate jumping between different modes; ε is drawn from a symmetric $d$-dimensional distribution with a small variance compared to the width of the target distribution. By accepting the candidate point with the Metropolis ratio defined in Eq. (2), a Markov chain is

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




obtained with its limiting distribution of the target PPDF, as proved in ter Braak (2006) and Vrugt et al. (2008, 2009a). The DE-MC algorithm can successfully simulate the multimodal distributions, because it directly uses the current location of the multiple chains stored in **X**, instead of Cov(**X**) from a single chain, to generate candidate points, allowing the possibility of direct jumps between different modes. In addition, it has no difficulty in simulating the distributions with heavy tails, because it does not use the Gaussian proposal whose function gets small much faster than the heavy-tailed functions in the tails and thus causes sampling problems as reported in Robert and Casella (2004) and demonstrated in Section 3 of this work.

The DREAM algorithm maintains the nice features of the DE-MC, but greatly accelerates the chain convergence. First, DREAM generates the candidate sample based on the difference of more than one pair of chains, which brings more information about the target distribution and thus accelerates the convergence. Secondly, DREAM uses a subspace sampling strategy that selectively updates only some parameters when generating a candidate sample. This strategy improves efficiency in high-dimensional samplings as the performance of optimization deteriorates exponentially with increasing dimensions. In addition, the subspace sampling strategy also enables the number of chains less than the number of parameters, which reduces the waste of computing resources, as each individual chain requires a certain portion of samples to be discarded before converging to the target distribution. Thirdly, DREAM explicitly handles the unproductive chains stuck in regions of the parameter space that are not contributing to the target PPDF. This strategy of removing outlier chains is very important for multi-chain methods, as the samples from these outlier chains will not only deteriorate the generation of the candidate points, which thus slow down the movement of other robust chains, but even worse they may prevent the convergence to the target distribution. For example, if one chain keeps sampling the area

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




289 isolated from the target distribution that other chains are sampling, it will make the $\hat{R}$ statistic
290 almost impossible to reach the threshold of 1.2 required to declare convergence. Therefore,
291 DREAM detects the outlier chains and forces their current states to the positions of other well-
292 performed chains, which greatly speeds up the convergence. More information about the
293 DREAM algorithm was presented in Vrugt et al. (2008, 2009a, 2016) and Lu et al. (2014).

294 **3 Comparison between AM and DREAM algorithms**

295 It is useful to know the estimated PPDFs from a sampling algorithm are the desired target
296 posterior distributions. So, in this section we evaluate the performance of DREAM in sampling
297 three known target distributions and compare results with the widely used AM algorithm. The
298 three distributions are high-dimensional, heavy-tailed, and multimodal distributions that are
299 notoriously difficult to approximate with MCMC sampling and are typical problems in terrestrial
300 ecosystem modeling (Kinlan and Gaines, 2003; Stead et al., 2005; Thibault et al., 2011). In all
301 studies, the computational efficiency of the algorithms is evaluated by the accuracy of the
302 approximate posterior distribution for the same number of total function evaluations, although
303 DREAM uses multiple parallel chains and AM employs only one chain.

304 **3.1 Case study I: a 50-dimensional Gaussian distribution function**

305 The first case study is a high-dimensional convex problem. The test function is a 50-
306 dimensional (50 parameters) multivariate Gaussian distribution with the mean at the zeros. The
307 covariance matrix was constructed such that the variance of the *i*th dimension is equal to $0.1\times i\times i$
308 and the covariance of *i*th and *j*th variables is calculated as $0.05\times i \times j$. Both the AM and DREAM
309 located the initial states of the chains from a uniform distribution $\mathbf{x}_0 \in U[-50,\ 50]^{50}$. Besides,
310 AM used an identity matrix as its initial covariance $C_0$. DREAM used ten parallel chains and
311 AM run a single chain.

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




312 The simulation results of 100,000 function evaluations for both AM and DREAM are
313 summarized in Figure 1, which depicts the evolution of the sample mean of parameter $x_1$,
314 standard deviations of parameters $x_{10}$ and $x_{50}$, and the covariance between parameters $x_5$ and $x_{30}$.
315 The true values of these statistics were also shown in the figure with different symbols. Figure 1
316 indicates that DREAM can smoothly converge to the true values within the given function
317 evaluations, while AM exhibits difficulty in the approximation of the correct values, although it
318 can eventually converge to the true values with the iterations doubled (results not shown here).
319 This case study suggests that DREAM is more efficient than AM for high-dimensional problems.

320 **3.2 Case study II: a 10-dimensional Cauchy distribution function**

321 The second case study considers a 10-dimensional Cauchy distribution centered at zeros
322 and having identity matrix as its scale matrix. The marginal density in one dimension is shown in
323 Figure 2 (b) and (d) as the red curve. This test function represents the type of problems with
324 heavy tails. It is well known that MCMC schemes with Gaussian-shaped proposals have
325 difficulty in sampling the heavy-tailed distributions because the Gaussian probability density
326 function decays much faster than the heavy-tailed functions. This can cause the AM method to
327 fail to converge due to inefficient sampling. As shown in Figure 2 where we took the parameter
328 samples of one dimension as an example, Figure 2 (a) depicts the trace plot of $x_1$ from the AM
329 chain for the second half of total one million iterations. The figure indicates that in a long period
330 of iterations the chain of AM has no updates and is stuck in certain areas of the parameter space
331 that have negligible probabilities in the target distribution, resulting in a histogram dramatically
332 different from the true distribution in Figure 2 (b). In contrast, the DREAM sampler launched ten
333 chains and each chain evolved 100,000 iterations, so that the total function evaluations were the
334 same as the AM. The trace plot of the ten chains in Figure 2 (c) demonstrates that DREAM

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




exhibits a very good mixing of the individual paths, therefore its sampled histogram displays a near-perfect match with the true distribution in Figure 2 (d). This case study suggests that DREAM is able to sample these heavy-tailed distributions more efficiently than MCMC samplers with a Gaussian proposal such as the AM.

### 3.3 Case study III: a 4-dimensional trimodal distribution function

The third case study is a 4-dimensional trimodal distribution with three well-separated modes at -8, 0, and 8, respectively. The modes have increasing weights in density; from left to right the weights are 0.1, 0.3, and 0.6, respectively. The marginal probability density in one dimension of this distribution is shown in Figure 3 (a)-(d) as the red curve. This type of distribution is known for its difficulty to be approximated by the MCMC methods, because the three modes are so far isolated that the chains need to jump between modes and it is common for chains to get stuck in one mode with a few or even no visits to other modes for many iterations, which greatly increases the number of simulations needed for convergence.

We implemented both the AM and DREAM algorithms using a million function evaluations. The AM used one chain with one million iterations long; we run the AM ten times and at each time the chain initialized with diverse starting points. The DREAM used four parallel chains with each chain evolving 250,000 iterations. Figure 3 (a)-(c) present the histograms of the sampled $x_1$ from three independent runs of AM. As shown, AM exhibits a rather poor performance as each different trial converged to a different mode and none of the ten runs can capture all the three modes, resulting in an unreliable approximation of the target distribution. A single chain is typically incapable of dealing with this multimodal parameter space and providing an accurate characterization within a limited number of iterations. With multiple interacting chains running simultaneously, the simulation results can be significantly improved. As

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




demonstrated in Figure 3 (d), the histogram of the marginal distribution from DREAM agrees with the true mixture distribution very well. The density of the samples in each mode is consistent with the weight of each peak in the distribution. Attributed to the differential evolution technique defined in Eq. (4), DREAM allows for jumps between the different disconnected modes. As illustrated in Figure 3 (e), each different chain jumps back and forth between the three isolated modes with more visits in the mode having higher weight, resulting in a good mix of the individual paths and consequently a relatively fast convergence to the true distribution.

## 4 Application to a terrestrial ecosystem model

In this section, we apply the DREAM algorithm to the data assimilation linked ecosystem carbon (DALEC) model to estimate the posterior distributions of its parameters. In comparison, the AM algorithm is also applied. DALEC is a relatively simple carbon pool and flux model designed specifically to enable parameter estimation in terrestrial ecosystems. We used DALEC to evaluate the performance of AM and DREAM in model calibration; we compared their accurate simulations of the parameter PPDFs, model's goodness-of-fit, and predictive performance of the calibrated models. Previous studies based on MCMC methods that use Gaussian proposals have not reported multimodality in the marginal PPDFs of the model parameters, so it is important to know whether the parameters have multimodality; if the multimodality exists, we assess whether or not DREAM can identify the multiple modes and improve the calibration results and thus the predictive performance.

### 4.1 Description of the model and parameters for optimization

The DALEC v1 model is used here (Williams et al., 2005; Fox et al., 2009) with some structural modifications (Safta et al., 2015). DALEC consists of six process-based submodels that simulate carbon fluxes between five major carbon pools: three vegetation carbon pools for

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




leaf, stem, and root; and two soil carbon pools for soil organic matter and litter. The fluxes calculated on any given day impact carbon pools and processes in subsequent days.

The six submodels in DALEC are photosynthesis, phenology, autotrophic respiration, allocation, litterfall and decomposition. Photosynthesis is driven by the aggregate canopy model (ACM) (Williams et al., 2005), which itself is calibrated against the soil-plant-atmosphere model (Williams et al., 1996). DALEC v1 was modified to incorporate the phenology submodel used in Ricciuto et al. (2011), driven by six parameters. This phenology submodel controls the current leaf area index (LAI) proportion of the seasonal maximum LAI (*laimax*). Spring LAI growth is driven by a linear relationship to growing degree days (gdd), while senescence and LAI loss are driven by mean air temperature. To simplify our model structure, senescence and LAI loss are considered to occur simultaneously. In reality, leaves may still be present on the trees but photosynthetically inactive due to the loss of chlorophyll. Here, this inactive LAI is considered to have fallen and is added to the litter pool. To further reduce model complexity, the plant labile pool in DALEC v1 was removed and a small portion of stem carbon is instead removed to support springtime leaf growth each year. The six phenology parameters are a threshold for leaf out (*gdd_min*), a threshold for maximum leaf area index (*gdd_max*), the temperature for leaf fall (*tsmin*), seasonal maximum leaf area index (*laimax*), the rate of leaf fall (*leaffall*), and leaf mass per unit area (*lma*), respectively. Given the importance of maintenance respiration in other sensitivity analyses (Sargsyan et al., 2014), we expanded the autotrophic respiration submodel to explicitly represent growth respiration (as a fraction of carbon allocated to growth) and maintenance respiration with the base rate and temperature sensitivity parameters.

So for the first three plant submodels, deciduous phenology has six parameters; ACM shares one parameter, *lma*, with the deciduous phenology and employs two additional

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




parameters, leaf C:N ratio (which is fixed at a constant of 25 in the simulation) and photosynthetic nitrogen use efficiency (*nue*); the autotrophic respiration model computes the growth and maintenance respiration components and is controlled by three parameters, the growth respiration fraction (*rg_frac*), the base rate at 25°C (*br_mr*), and temperature sensitivity for maintenance respiration (*q10_mr*).

The allocation model partitions carbon to several vegetation carbon pools. Leaf allocation is first determined by the phenology model, and the remaining available carbon is allocated to the root and stem pools depending on the fractional stem allocation parameter (*astem*). The litter fall model redistributes the carbon content from vegetation pools to litter pools and is based on the turnover times for stem (*tstem*) and root (*troot*). The last submodel is a decomposition model that simulates heterotrophic respiration and the decomposition of litter into soil organic matter (SOM). This model is driven by the temperature sensitivity of heterotrophic respiration (*q10_hr*), the base turnover times for litter (*br_lit*) and SOM (*br_som*) at 25°C, and by the decomposition rate (*dr*) from litter to SOM.

Model parameters are summarized in Table 1. These parameters were grouped according to the six submodels that employ them, except for *lma* that impacts both the deciduous leaf phenology and ACM. The nominal values and numerical ranges for these parameters were designed to reflect average values and broad uncertainties associated with the temperate deciduous forest plant functional type that includes Harvard Forest (Fox et al., 2009; White et al., 2000; Ricciuto et al., 2011). Observed air temperature, solar radiation, vapor pressure deficit, and $CO_2$ concentration were used as boundary conditions for the model.

In order to reduce computational time, we employed transient assumptions for running DALEC. That is, for any given set of parameter values, DALEC was run one cycle only for 15

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




years between 1992-2006 where observation data are available. Under this assumption, four additional parameters were used to describe the initial states of two vegetation carbon pools (*stemc_init* and *rootc_init*) and the two soil carbon pools (*litc_init* and *somc_init*), as also summarized in Table 1. Thus, a total of 21 parameters were considered and estimated in this study.

### 4.2 Calibration data

The calibration data consist of the Harvard Forest daily net ecosystem exchange (NEE) values, which were processed for the NACP site synthesis study (Barr et al., 2013) based on flux data measured at the site (Urbanski et al., 2007). The daily observations cover a period of 15 years starting with the year 1992 and part of the data in the year 2005 is missing. Hill et al. (2012) estimated that daily NEE values followed a normal distribution, with standard deviations estimated by bootstrapping half-hourly NEE data (Papale et al., 2006; Barr et al., 2009). These standard deviations have values between 0.2 and 2.5, with the mean value about 0.7. Total 14 years NEE data (years from 1992 to 2004 and year 2006) were considered here for model calibration and their corresponding standard deviations were used to construct the diagonal covariance matrix of the Gaussian likelihood function by assuming the data are uncorrelated.

### 4.3 Numerical experiments and calibration results

Both AM and DREAM were implemented to estimate the 21 parameters of the DALEC model using the PPDFs. To avoid the influence of prior distributions on the investigation of the posteriors estimated by the two algorithms, uniform priors were used for all parameters with the ranges specified in Table 1. DREAM launched ten parallel chains starting at values randomly drawn from the parameter prior distributions; and each chain evolved 300,000 iterations. Chain convergence was assessed via the Gelman Rubin $\hat{R}$ statistic. Figure 4 (b) plots the $\hat{R}$ values of

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




450 the 21 parameters for the last 100,000 iterations. The figure suggests that the last 50,000 samples
451 of each chain (i.e., total 500,000 samples from ten chains) can be used for the PPDF
452 approximation as the $\hat{R}$ has values below the threshold of 1.2.
453 AM used one chain and the chain has the same initialization with DREAM. In addition,
454 AM also requires the initialization of the covariance matrix of its Gaussian proposal. To facilitate
455 the convergence of AM, we constructed the initial covariance $C_0$ based on the first 200,000
456 samples from the DREAM simulation. We conducted ten independent AM runs, so the same $\hat{R}$
457 statistic can be used for convergence diagnosis. As AM is a single-chain algorithm, to make a
458 fair comparison with the multi-chain algorithm of DREAM, each AM chain simulated 3,000,000
459 samples, so that the number of function evaluations in one AM chain is the same with that of
460 DREAM using ten chains. The $\hat{R}$ values of all parameters based on the ten AM runs for the last
461 1,000,000 iterations are shown in Figure 4 (a). The figure indicates that AM has converged and
462 the last 500,000 samples from one chain were used for the PPDF approximation.
463 The estimated PPDFs from AM and DREAM are presented in Figure 5, and the optimal
464 parameter estimates, as represented by the maximum a posteriori (MAP), are summarized in
465 Table 1. Both AM and DREAM results show that all the 21 parameters can be well constrained
466 by the calibration data, although some studies reported that eddy-covariance observations along
467 could not identify all the model parameters with their posterior distributions significantly smaller
468 than their priors (Wang et al., 2007, Keenan et al., 2012, 2013). Whether a parameter is
469 identifiable depends on the model, model parameters, and the calibration data. When the
470 parameter related processes are necessary to simulate the model outputs whose corresponding
471 observation data are sensitive to the parameters, the parameters can usually be identified and
472 sometimes well constrained. For example, Keenan et al. (2013) showed that in their FöBAAR



473 model with 40 parameters, many parameters couldn't be constrained even with the consideration
474 of several data streams together. They found that these unidentifiable parameters might be
475 redundant in the model structure representation. Roughly speaking, for a simple model with a
476 few number of parameters, the parameters can be more identifiable than the complex models
477 with a large parameter size (Richardson et al., 2010, Weng and Luo, 2011). The DALAC model
478 used in this study is a simple model with considering only six processes and five carbon pools,
479 and all the 21 parameters were shown to be sensitive to the NEE data, despite that some are more
480 sensitive than others (Safta et al, 2015). So it is not surprising that both AM and DREAM
481 algorithms can constrain the parameters pretty well.

482 In comparison of the results between AM and DREAM, Figure 5 indicates that they
483 produced very similar PPDFs for many parameters, such as *gdd_max*, *astem*, *br_som*, *rootc_init*,
484 and *litc_init*. However, for parameters *tsmin* and *leaffall*, the estimated PPDFs of the two
485 algorithms are substantially different. This also can be seen in Table 1, where the differences of
486 MAP values for most parameters are relatively small between the two algorithms, but for *tsmin*
487 and *leaffall*, the relative difference was 38% and 94%, respectively. The parameter *tsmin*
488 represents the temperature triggering leaf fall and the *leaffall* represents the rate of leaf fall on
489 days when the temperature is below *tsmin*. We further analyzed the simulations of these two
490 parameters from AM and DREAM in Figure 6. Figure 6 (a) and (b) illustrate two separated
491 modes in the estimated marginal PPDFs of *tsmin* and *leaffall* obtained from DREAM, while AM
492 only identifies one mode for both parameters and they dramatically differ from any modes
493 simulated by DREAM. For example, the single mode of *tsmin* identified by AM gives a lower
494 temperature threshold (meaning a later initiation of senescence) that is compensated by a higher
495 estimate of *leaffall* rate compared to DREAM. As shown in the trace plots of Figure 6 (c) and

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




(d), all ten independent runs of AM converged to a single mode, with values of *tsmin* between 4.8 to 5.0 and values of *leaffall* between 0.06 and 0.075. In contrast, each of the ten parallel chains of DREAM, as exhibited in Figure 6 (e) and (f), jumps back and forth between two modes. And the two parameters compensate each other by jumping in opposite directions, where *tsmin* is more likely to be near the mode with a smaller value of 7.9 than that of 8.35 and *leaffall* is more likely to be near the mode of a larger value of 0.035 than that of 0.031.

In addition, the simulated joint PPDFs of the two parameters, *tsmin* and *leaffall*, are different between AM and DREAM. As illustrated in Figure 7, AM results exhibit a negligible correlation between the two parameters with the correlation coefficient of -0.042, while DREAM results show that the two parameters are strongly negatively correlated with the correlation coefficient of -0.95. As demonstrated in Figure 7 (b), the samples of *tsmin* and *leaffall* from DREAM fall almost perfectly on the line with slope of -1, where the mode with smaller *tsmin* values corresponds to the mode of larger *leaffall* and the similar correspondence can be found for the other pair of modes.

The existence of two modes for *tsmin* and *leaffall* and the negative correlation between the two parameters are not unreasonable as we used multiple years of observations for parameter estimation. It is possible that in some years the senescence was triggered later (i.e., a smaller *tsmin*) but proceded at a faster rate (i.e., a larger *leaffall*), while in some other years the senescence was triggered earlier (i.e., a larger *tsmin*) but proceded at a slower rate (i.e., a smaller *leaffall*). Given our model simplification of concurrent senescence and leaf fall and our use of NEE rather than LAI observations as a constraining variable, we note that these optimized parameters are more likely to reflect the process of chlorophyll loss than actual leaf loss. Cool temperatures are a key driver of senescence at this site (Richardson et al., 2006).

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




519 Figure 8 (a) highlights the years in red where the model based on the right mode of *tsmin*
520 and the left mode of senescence rate (*leaffall*) has a better fit to the observed NEE, i.e., years
521 1994, 1995, 1998, 1999, and 2006. The remaining years are highlighted in blue where the left
522 mode of *tsmin* and the right mode of *leaffall* result in a better model fit. Taking years 1992 and
523 1994 as an example, we examined the leaf area index (LAI) in the period of senescence. Figure 8
524 (b) shows that at the first few days of September in both years, the values of LAI were the same
525 around 2.0; after that the timing of senescence during the two years differs dramatically. In year
526 1994, the value of LAI started decreasing on September 7$^{th}$, and then decreased slowly over
527 several distinct cool periods during the rest of September and early October until it hit zero in
528 November 7$^{th}$; the process took about 61 days. In contrast, in year 1992, the value of LAI
529 remained near the maximum value during all of September, then dropped rapidly in October and
530 hit zero also on November 7$^{th}$; this process took about 40 days. The changes in the LAI between
531 the two years reflect the variability in the time of year when the leaves start to drop and the rate
532 of leaf drop. Although the leaf fall in 1992 was triggered later than in 1994, the leaves in 1992
533 dropped at a faster rate, resulting in LAI approaching zero at the same time of the year.
534 Figure 8 (c) depicts the recorded lowest temperature of the days between September 1$^{st}$ and
535 November 20$^{th}$ for years 1992 and 1994, where the red line highlights the period between the
536 first leaf and the last leaf drops in 1994. The blue line highlights the corresponding period of
537 leaf fall in 1992. Since the senescence was triggered in the early September of 1994, the
538 temperature of triggering leaf fall was relatively high, about 8.1$^{o}$C (associated with the higher
539 mode of *tsmin*) as shown in Figure 8 (c). In the rest days of September in 1994 following the
540 senescence trigger, temperatures remained warm. The slower leaf fall rate associated with
541 periodic warm conditions (temperatures above *tsmin*) and the lower mode of *leaffall* caused a

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




542 slow leaf fall in September of 1994 as shown in Figure 8 (b). In comparison, in 1992, senescence
543 was triggered at the end of September with a low temperature of 2.6$^{o}$C. Then in October with
544 colder temperatures, the leaves drop at a rapid rate associated with the consistent cold
545 temperatures and higher mode of *leaffall*. Especially in late October, the temperatures are
546 consistently below *tsmin*, causing a fast rate of leaf fall, as shown in Figure 8 (b) where the
547 decreasing rate of the LAI in the late October of 1992 is very large. This indicates that a higher
548 temperature trigger is usually associated with a lower leaf fall rate and vice versa.
549 The bimodality identified in the DREAM simulation and examined in the scenarios above
550 reflects the inability of the model structure to predict the observations consistently with a single
551 set of parameters. This bimodality examined in DREAM may be caused in part by an incomplete
552 representation of the senescence process. Using a temperature threshold (parameter *tsmin*) and a
553 constant rate of leaf fall (parameter *leaffall*) to predict senescence is almost certainly an
554 oversimplification. In reality, the process of senescence is also affected by day length. Longer
555 days and warmer temperatures cause a relatively slow rate of leaf fall, whereas shorter days and
556 cooler temperatures accelerate the rate that the leaves fall (Leigh et al, 2002; Saxena, 2010). The
557 higher mode of *tsmin* means that senescence is initiated earlier, when day lengths are still
558 relatively long. This may partially explain why this mode is associated with a lower mode of the
559 *leaffall* parameter. Other factors not represented in DALEC are also likely to play a role such as
560 soil moisture, or a more complex relationship with spring phenology (Keenan et al., 2015).
561 The difference in estimated parameters between AM and DREAM causes different
562 simulations of NEE, especially during the autumn. As an example, Figure 9 illustrates the
563 comparison of the simulated NEE to observations for a month in Autumn of the year 1995 based
564 on MAP estimates obtained under AM and DREAM. Visual inspection indicates that the

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




565 simulated NEE from the DREAM-calibrated parameters provides a better fit to the observations,
566 as also indicated by the smaller root mean squared errors (RMSE). In addition, the maximum log
567 likelihoods listed in Table 1 suggest that overall the DREAM-estimated parameters produce a
568 better model fit to the observations, comparing -6578.3 with the smaller AM value of -6662.6.

569 **4.4 Assessment of predictive performance**

570 To further compare the calibration results between AM and DREAM, we explore their
571 predictive skill based on the sampled PPDFs of model parameters. We employed the Bayesian
572 posterior predictive distribution (Lynch and Western, 2004) to assess the adequacy of the
573 calibrated models. Specifically, the posterior distribution for the predicted NEE data, $p(\mathbf{y}|\mathbf{D})$, is
574 represented by marginalization of the likelihood over the posterior distribution of model
575 parameters **x** as

576 $$p(\mathbf{y}\,|\,\mathbf{D}) = \int p(\mathbf{y}\,|\,\mathbf{x})p(\mathbf{x}\,|\,\mathbf{D})d\mathbf{x}\,. \qquad (5)$$

577 In approximation of $p(\mathbf{y}|\mathbf{D})$, we used the converged MCMC samples from $p(\mathbf{x}|\mathbf{D})$. The last 500
578 samples of each chain (total 500×10=5000 samples) were considered; for each parameter sample
579 we drew 20 samples of the 14 years NEE data from their normal distributions, where the mean
580 values are the model simulations. Then the total 100,000 prediction samples were used to
581 approximate the posterior predictive density $p(\mathbf{y}|\mathbf{D})$.

582 From the estimated $p(\mathbf{y}|\mathbf{D})$, we extracted the 95% confidence intervals for daily NEE
583 values in the year 1995 and presented the results in Figure 10. The top panel corresponds to the
584 results of AM and the bottom panel to DREAM. Overall, the predictive intervals from both
585 algorithms cover well the observed NEE for the entire time range with occasional spikes outside
586 the intervals. Closer visual inspection indicates that DREAM produces better predictive
587 performance than AM. As seen during the period in October, the predictive interval of DREAM

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




588 can enclose most of the observed NEE while AM actually has under-prediction, causing the
589 observations outside the intervals.

590 In order to quantitatively compare the predictive performance of the calibrated models
591 based on AM and DREAM, we defined two metrics, a probabilistic score called CRPS and
592 predictive coverage. The CRPS (Gneiting and Raftery, 2007) measures the difference between
593 the cumulative distribution function (CDF) of the observed data and that of the predicted data.
594 The lower the value of the CRPS is, the better the predictive performance. The predictive
595 coverage measures the percent of observations that fall within a given predictive interval. A
596 larger value of the predictive coverage suggests better predictive performance. Figure 10 shows
597 that AM gives a CRPS value of 0.48 while the value of DREAM is 0.43. The lower value of
598 DREAM indicates that, on average, DREAM produces tighter marginal predictive CDF that are
599 better centered around the NEE data, suggesting its superior predictive performance to AM in
600 terms of both accuracy and precision. In addition, the predictive coverage of DREAM is larger
601 than that of AM, attesting once again to its superior performance in prediction.

602 **4.5 Investigation of reliability of the algorithms**

603 Bayesian calibration of TEMs is challenging due to high model nonlinearity, high
604 computational cost, a large number of model parameters, large observation uncertainty, and the
605 existence of local optima. So, a robust and efficient MCMC algorithm is desired to give reliable
606 probabilistic descriptions of the TEM parameters.

607 In this section, we investigate the influence of the proposal initialization on the
608 computational efficiency and reliability of AM. In above analysis, the initial covariance matrix
609 $C_0$ of AM was constructed based on DREAM samples *before* convergence. This setting
610 facilitated the convergence of AM but resulted in AM false convergence to inaccurate PPDFs,

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




611 leading to a relatively poor calibration and prediction performance. We implemented another
612 AM simulation here for further examination. In this new simulation, we constructed two
613 independent AM chains; both chains initialized $C_0$ using the DREAM samples *after*
614 convergence, but one chain only used *tsmin* samples around its left mode and *leaffall* samples
615 around its right mode, and the other chain used *tsmin* samples around its right mode and *leaffall*
616 samples around its left mode. Each chain evolved 3,000,000 iterations, and for the last 1,000,000
617 iterations the convergence diagnostic $\hat{R}$ values were calculated and shown in Figure 11 (a). The
618 figure indicates that most parameters have $\hat{R}$ less than the threshold of 1.2 except parameters
619 *tsmin* and *leaffall* whose values are far above 1.2 and no signs show that they are going
620 significantly smaller in the following one million iterations. This suggests that the two chains
621 converged to different optima for these two parameters. We then estimated PPDFs using the last
622 500,000 samples from each chain respectively. The results for *tsmin* and *leaffall* are shown in
623 Figure 11 (b)-(e). The figures illustrate that the samples from one AM chain can only identify
624 one mode, and this mode is consistent with the samples used to construct the initial covariance
625 matrix $C_0$.

626 As a single-chain sampler, it is conceptually possible for AM to become trapped in a single
627 mode (Jeremiah et al., 2009). Consider a distribution with two far-separated modes and assume
628 that the chain is initialized near one of the two modes (both samples initialization and proposal
629 covariance initialization). At the beginning of the sampling, AM will explore the area around the
630 mode where it is initialized and start identifying the first mode. Since the candidate samples
631 generated by the Gaussian proposal have higher Metropolis ratios (Eq. (2)) in the nearby area
632 than in the far-away regions of the identified mode, the chain is hardly to move to the other
633 mode. When the Gaussian proposal covariance matrix $C_t$ begins to update, the chance of the



634 chain jumping to the other mode depends on the relative scale of the proposal covariance and the
635 distance between the two modes. When the modes separation exceeds the range of the proposal,
636 AM is less likely to escape the identified local mode.

637 Although the two AM chains can only simulate one of the two modes for *tsmin* and
638 *leaffall*, the estimated PPDFs for the other 19 parameters from the two chains are close to each
639 other and both similar to the DREAM results. This finding once again shows the reasonable
640 existence of the two separated modes and their equivalent importance. With an improved
641 initialization of $C_0$ in the new simulation, the performance of AM also improved as it can
642 accurately simulate uni-modal PPDFs and capture one mode for the multi-modal PPDFs. This
643 investigation suggests that for AM an appropriate initialization of its Gaussian proposal has a
644 significant impact on its performance. We made several test runs of AM and only when we
645 initialized $C_0$ using the complete set of converged DREAM samples, was the AM able to
646 produce PPDFs similar to the ones resulted from DREAM with identifying all the possible
647 optima. However, the information of a reasonable $C_0$ in practice is either unavailable or very
648 computationally expensive to obtain.

649 **5 Conclusions**

650 In this study we compare the performance of two MCMC algorithms, namely AM and
651 DREAM, for sampling high-dimensional multi-modal posterior distributions for models that
652 exhibit non-linear behavior. We present results obtained in canonical, manufactured
653 configurations, as well as the terrestrial ecosystem model DALEC. The AM algorithm is
654 sensitive to initial conditions when sampling multi-modal posteriors. By comparison, DREAM's
655 performance does not depend on initialization of the algorithm. Thus, DREAM is particularly
656 suitable to calibrate complex terrestrial ecosystem models, where the existence of local optima is



657 always a concern and where often little information about the parameters is known in order to
658 optimize the algorithm implementation. In this study, the application in an ecosystem carbon
659 model indicates that, compared to the AM, the DREAM can accurately simulate the posterior
660 distributions of the model parameters, resulting in a better model fit, superior predictive
661 performance, and perhaps identifying structural errors or process differences between the model
662 and ecosystem from which observations were used for calibration.

663 **Acknowledgement**

664 This research was conducted by the Terrestrial Ecosystem Science – Science Focus Area
665 (TES-SFA) project, supported by the Office of Biological and Environmental Research in the
666 DOE Office of Science. The Harvard Forest flux tower is part of the AmeriFlux network
667 supported by Office of Biological and Environmental Research in the DOE Office of Science
668 and is additionally supported by National Science Foundation as part of the Harvard Forest
669 Long-Term Ecological Research site. The NACP site-synthesis activity supported assembling the
670 data set. Oak Ridge National Laboratory is managed by UT-BATTELLE for DOE under contract
671 DE-AC05-00OR22725. Sandia National Laboratories is a multi-program laboratory managed
672 and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin
673 Corporation, for the DOE's National Nuclear Security Administration under contract DE-AC04-
674 94-AL85000.

675 **References**

676 Barr, A., Hollinger, D., and Richardson, A. D.: CO2 flux measurement uncertainty estimates for
677 NACP, AGU Fall Meeting, December 2009, abstract number B54A-04B, 2009.
678 Barr, A. et al.: NACP site: tower meteorology, flux observations with uncertainty, and ancillary
679 data, available at: http://daac.ornl.gov from Oak Ridge National Laboratory Distributed

680 Active Archive Center, doi:10.3334/ORNLDAAC/1178, 2013.
681 Box, E.P., and Tiao, G.C.: Bayesian inference in statistical analysis, Wiley, New York, 588pp,
682 1992.
683 Dowd, M.: Bayesian statistical data assimilation for ecosystem models using Markov Chain
684 Monte Carlo, Journal of Marine Systems, 68, 439-456, 2007.
685 Fox, A., et al.: The REFLEX project: Comparing different algorithms and implementations for
686 the inversion of a terrestrial ecosystem model against eddy covariance data, Agric. For.
687 Meteorol., 149(10), 1597-1615, 2009.
688 Gelman, A., and Rubin, D.B.: Inference from iterative simulation using multiple sequences,
689 Statistical science 7(4), 457-472 1992.
690 Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B.: Bayesian data analysis, Chapman and
691 Hall, London, 696pp, 1995.
692 Gelman, A., and Shirley, K.: Inference from simulations and monitoring convergence, Handbook
693 of Markov Chain Monte Carlo, CRC Press; Boca Raton, FL: 2011.
694 Gneiting, T., and Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation, J. Am.
695 Statist. Assoc., 102, 359-378, 2007.
696 Haario, H., Saksman, E., and Tamminen, J.: Adaptive proposal distribution for random walk
697 Metropolis algorithm. Comput. Statist., 14, 375-395, 1999.
698 Haario, H., Saksman, E., and Tamminen, J.: An adaptive Metropolis algorithm, Bernoulli, 7(2),
699 223-242, 2001.
700 Haario, H., Laine, M., Mira, A, and Saksman, E.: DRAM: Efficient adaptive MCMC, Stat Comp
701 16, 339-354, 2006.
702 Hararuk, O., Xia, J., and Luo, Y.: Evaluation and improvement of a global land model against

703 soil carbon data using a Bayesian Markov chain Monte Carlo method, J. Geophys. Res.
704 Biogeosci., 119, 403-417, 2014.
705 Hill, T. C., Ryan, E., and Williams, M.: The use of CO2 flux time series for parameter and
706 carbon stock estimation in carbon cycle research, Global Change Biol., 18, 179–193,
707 2012.
708 Järvinen, H., Raisanen, P., Laine, M., Tamminen, J., Ilim, A., Oja, E., Solonen, A., and Haario,
709 H.: Estimation of ECHAM5 climate model closure parameters with adaptive MCMC,
710 Atmos. Chem. Phys., 10, 9993-10002, 2010.
711 Jeremiah, E., Sisson, S., Marshall, L., Mehrotra, R., and Sharma, A.: Bayesian calibration and
712 uncertainty analysis of hydrological models: A comparison of adaptive Metropolis and
713 sequential Monte Carlo samplers, Water Resour. Res., 47, W07547,
714 doi:10.1029/2010WR010217, 2011.
715 Keenan R.J., Reams, G.A., Achard, F., de Freitas, J.V., Grainger A., and Lindquist E.: Dynamics
716 of global forest area: Results from the FAO global forest resources assessment 2015,
717 Forest Ecology and Management, 352, 9-20, 2015.
718 Keenan, T. F., Davidson, E., Moffat, A., Munger, W., and Richardson, A.D.: Using model-data
719 fusion to interpret past trends, and quantify uncertainties in future projections, of
720 terrestrial ecosystem carbon cycling, Global Change Biology, 18, 2555-2569, 2012.
721 Keenan, T. F., Davidson, E., A., Munger, W., and Richardson, A.D.: Rate my data: quantifying
722 the value of ecological data for the development of models of the terrestrial carbon
723 cycle, Ecological Applications, 23(1), 273-286, 2013.
724 Kinlan, B. P, and Gaines, S. D.: Propagule dispersal in marine and terrestrial environments: A
725 community perspective, Ecology, 84(8), 2007-2020, 2003.

726 Knorr, W., and Heimann M.: Uncertainties in global terrestrial biosphere modeling: 1. A
727 comprehensive sensitivity analysis with a new photosynthesis and energy balance
728 scheme, Global Biogeochem. Cycles, 15, 207-225, 2001.
729 Leigh, M., Nihevia, N., Covich, E., and Kehn, D.: How temperature and daylength effect
730 seasonal leaf change in honeysuckle plants, online article,
731 http://jrscience.wcp.muohio.edu/nsfall01/labpacketArticles/Howtemperatureanddayleng
732 t.html, 2002.
733 Lu, D., Ye, M., and Hill, M.C.: Analysis of regression confidence intervals and Bayesian
734 credible intervals for uncertainty quantification, Water Resour. Res. 48, W09521.
735 http://dx.doi.org/10.1029/2011WR011289, 2012.
736 Lu, D., Ye, M., Hill, M.C., Poeter, E.P., and Curtis, G.P.: A computer program for uncertainty
737 analysis integrating regression and Bayesian methods, Environmental Modelling &
738 Software, 60, 45-56, 2014.
739 Lynch, S.M, and Western, B.: Bayesian posterior predictive checks for complex models, Sociol.
740 Meth. Res., 32, 301-335, 2004.
741 Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E.: Equations of
742 state calculations by fast computing machines, J. Chem. Phys. 21, 1087-1091, 1953.
743 O'Sullivan, F.: A statistical perspective on ill-posed inverse problems, Statistical Science, 1(4),
744 502-518, 1986.
745 Papale, D., Reichstein, M., Aubinet, M., Canfora, E., Bernhofer, C., Kutsch, W., Longdoz, B.,
746 Rambal, S., Valentini, R., Vesala, T., and Yakir, D.: Towards a standardized processing
747 of Net Ecosys- tem Exchange measured with eddy covariance technique: algorithms and
748 uncertainty estimation, Biogeosciences, 3, 571-583, 2006.

749 Price, K.V., Storn, R.M., and Lampinen, J.A.: Differential Evolution, a practical approach to
750 global optimization, Springer, Berlin, 2005.
751 Rayner, P. J., Scholze, M., Knorr, W., Kaminski, T., Giering, R. and Widmann, H.: Two decades
752 of terrestrial carbon fluxes from a carbon cycle data assimilation system (CCDAS),
753 Global Biogeochem. Cycles, 19, GB2026, doi:10.1029/2004GB002254, 2005.
754 Ricciuto, D. M., Davis, K. J., and Keller, K.: A Bayesian calibration of a simple carbon cycle
755 model: The role of observations in estimating and reducing uncertainty, Global
756 Biogeochem. Cycles, 22, GB2030, doi:10.1029/2006GB002908, 2008.
757 Ricciuto, D. M., King, A. W., Dragoni, D., and Post, W. M.: Parameter and prediction
758 uncertainty in an optimized terrestrial carbon cycle model: Effects of constraining
759 variables and data record length, J. Geophys. Res., 116, G01033,
760 doi:10.1029/2010JG001400, 2011.
761 Richardson, A., Bailey, A.S., Denny, E.G., Martin, C.W., and O'Keefe, J.: Phenology of a
762 northern hardwood forest canopy, Global Change Biology, 12, 1174-1188, doi:
763 10.1111/j.1365-2486.2006.01164.x, 2006.
764 Richardson, A.D., Williams, M., Hollinger, D.Y., Moore, D.J.P., Dail, D.B., Davidson, E.A.,
765 Scott, N.A., Evans, R.S., Hughes, H., Lee, J.T., Rodrigues, C., and Savage, K.:
766 Estimating parameters of a forest ecosystem C model with measurements of stocks and
767 fluxes as joint constraints, Oecologia, 164, 25-40, 2010.
768 Robert, C., and Casella, G.: Monte Carlo statistical method, Springer. 2rd ed. 645pp, 2004.
769 Safta, C., Ricciuto, D. M., Sargsyan, K., Debusschere, B., Najm, H. N., Williams, M., and
770 Thornton, P. E.: Global sensitivity analysis, probabilistic calibration, and predictive
771 assessment for the data assimilation linked ecosystem carbon model, Geosci. Model Dev., 8,

772 1899-1918, 2015.

773 Sargsyan, K., Safta, C., Najm, H. N., Debusschere, B., Ricciuto, D. M., and Thornton, P. E.:
774 Dimensionality reduction for complex models via Bayesian compressive sensing, Int. J.
775 Uncert. Quant., 4, 63-93, 2014.

776 Saxena, N.P.: Objective Botany for all medical entrance examinations, Krishna Prakashan Media
777 Ltd., 2010.

778 Stead, T. K., Schmid-Araya, J. M., Schmid, P. E., and Hildrew, A.G.: The distribution of body
779 size in a stream community: one system, many patterns, Journal of Animal Ecology, 74,
780 475–487, 2005.

781 ter Braak, C.J.F.: A Markov chain Monte Carlo version of the genetic algorithm differential
782 evolution: easy Bayesian computing for real parameter spaces, Stat. Comput. 16, 239-
783 249, 2006.

784 Talagrand, O., Courtier, P.: Variational assimilation of meteorological observations with the
785 adjoint vorticity equation – Part I. Theory. Q. J. R. Meteorol. Soc. 113, 1311-1328,
786 1987.

787 Thibault K. M., White, E. P., Hurlbert, A. H., and Morgan Ernest S. K.: Multimodality in the
788 individual size distributions of bird communities, Global Ecology and Biogeography,
789 20, 145-153, 2011.

790 Trudinger, C. M., et al.: OptIC project: An intercomparison of optimization techniques for
791 parameter estimation in terrestrial biogeochemical models, J. Geophys. Res., 112,
792 G02027, doi:10.1029/2006JG000367, 2007.

793 Urbanski, S., Barford, C., Wofsy, S. C., Kucharik, C., Pyle, E., Budney, J., McKain, K.,
794 Fitzjarrald, D., Czikowsky, M., and Munger, J. W.: Factors controlling CO2 exchange

795 on timescales from hourly to decadal at Harvard Forest, J. Geophys. Res.
796 Biogeosciences, 112, 1-25, 2007.
797 Wang, Y. P., Baldocchi, D., Leuning, R., Falge, E., and Vesala, T.: Estimating parameters in a
798 land-surface model by applying nonlinear inversion to eddy covariance flux
799 measurements from eight FLUXNET sites, Global Change Biology, 13(3), 652-670,
800 2007.
801 Weng, E., and Luo, Y.: Relative information contributions of model vs. data to short- and long-
802 term forecasts of forest carbon dynamics, Ecological Applications, 21, 1490-1505,
803 2011.
804 White, M. A., Thornton, P. E., Running, S. W., and Nemani, R. R.: Parameterization and
805 sensitivity analysis of the BIOME-BGC terrestrial ecosystem model: net primary
806 production controls, Earth Interactions, 4, 1-85, 2000.
807 Williams, M., et al.: Modelling the soil-plant-atmosphere continuum in a Quercus' Acer stand at
808 Harvard Forest: the regulation of stomatal conductance by light, nitrogen and soil/plant
809 hydraulic properties, Plant, Cell Environ., 19, 911-927, 1996.
810 Williams, M., Schwarz, P. A., Law, B. E., Irvine, J., and Kurpius, M. R.: An improved analysis
811 of forest carbon dynamics using data assimilation, Global Change Biol., 11, 89-105,
812 2005.
813 Vrugt, J.A., ter Braak, C.J.F., Clark, M.P., Hyman, J.M., Robinson, B.A.,: Treatment of input
814 uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain
815 Monte Carlo simulation, Water Resour. Res., 44, W00B09,
816 doi:10.1029/2007WR006720, 2008.
817 Vrugt, J.A., ter Braak, C.J.F., Diks, C.G.H., Robinson, B.A., Hyman, J.M., and Higdon, D.:

Accelerating Markov Chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling, International Journal of Nonlinear Sciences & Numerical Simulation, 10(3), 271-288, 2009a.

Vrugt, J.A., ter Braak, C.J.F., Gupta, H.V., Robinson, B.A.: Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling, Stoch. Environ. Res. Risk Assess., 23 (7), 1011-1026, 2009b.

Vrugt, J.A.: Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation, Environmental modeling & Software, 75, 273-316, 2016.

Zhang, G., Lu, D., Ye, M., Gunzburger, M., and Webster, C.: An adaptive sparse-grid high-order stochastic collocation method for Bayesian inference in groundwater reactive transport modeling, Water Resour. Res., 49, http://dx.doi.org/ 10.1002/wrcr.20467, 2013.

Ziehn, T., Scholze, M., and Knorr, W.: On the capability of Monte Carlo and adjoint inversion techniques to derive posterior parameter uncertainties in terrestrial ecosystem models, Global Biogeochem. Cycles, 26, GB3025, doi:10.1029/2011GB004185, 2012.

Zobitz, J. M., Desai, A. R., Moore, D. J. P., and Chadwick, M. A.: A primer for data assimilation with ecological models using Markov Chain Monte Carlo, Oecologia, 167, 599-611, 2011.

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017




836 **List of Tables**

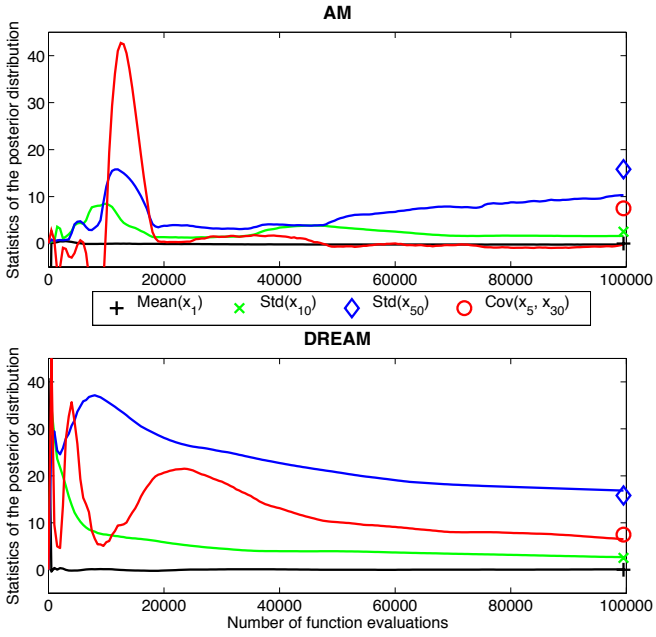
837 Table 1. Nominal values and ranges of the 21 parameters for optimization in the DALEC model,
838 and the maximum a posteriori (MAP) estimates based on the AM and DREAM samplers.

| | ParName | Nom. Val. | Range | MAP estimates | |
|---|---|---|---|---|---|
| | | | | AM LL=-6662.6 | DREAM LL=-6578.3 |
| Decid. Phen. | gdd_min | 100 | 10–250 | 37.90 | 39.53 |
| | gdd_max | 200 | 50–500 | 203.44 | 201.77 |
| | tsmin | 5 | 0–10 | 4.88 | 7.87 |
| | laimax | 4 | 2–7 | 2.01 | 2.00 |
| | leaffall | 0.1 | 0.03–0.95 | 0.067 | 0.035 |
| | lma | 80 | 20–150 | 136.81 | 147.45 |
| ACM | nue | 7 | 1–20 | 8.90 | 8.21 |
| A. R. | q10_mr | 2 | 1–4 | 1.00 | 1.00 |
| | br_mr | $10^{-4}$ | $10^{-5}$–$10^{-2}$ | $7.39\times10^{-3}$ | $6.35\times10^{-3}$ |
| | rg_frac | 0.2 | 0.05–0.5 | 0.06 | 0.066 |
| A. | astem | 0.7 | 0.1–0.95 | 0.75 | 0.74 |
| Lit. Fal. | tstem | 1/(50×365) | 1/(250×365) – 1/(10×365) | $1.98\times10^{-5}$ | $1.63\times10^{-5}$ |
| | troot | 1/(5×365) | 1/(25×365) – 1/365 | $8.55\times10^{-4}$ | $7.88\times10^{-4}$ |
| Decomp. | q10_hr | 2 | 1–4 | 2.98 | 2.68 |
| | br_lit | 1/(2×365) | 1/(5×365) – 10/(5×365) | $4.97\times10^{-3}$ | $5.36\times10^{-3}$ |
| | br_som | 1/(30×365) | 1/(100×365) – 1/(10×365) | $2.79\times10^{-5}$ | $2.88\times10^{-5}$ |
| | dr | $10^{-3}$ | $10^{-4}$–$10^{-2}$ | $2.46\times10^{-3}$ | $3.39\times10^{-3}$ |
| Init. C. | stemc_init | 5000 | 1000 – 15000 | 1070.9 | 1417.8 |
| | rootc_init | 500 | 100 – 3000 | 100.56 | 100.61 |
| | litc_init | 600 | 50 – 1000 | 60.74 | 66.77 |
| | somc_init | 7000 | 1000 – 25000 | 2029.1 | 4708.2 |

839 Parameter units refer to Table 1 of Safta et al. (2015). The LL represents the log likelihood
840 evaluated at the MAP parameter estimates; the larger the value is, the better the model fit.

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017


841 **List of Figures**




842

843 Figure 1. Simulated traces for the 50-dimensional Gaussian distribution based on the AM and
844 DREAM sampler. The black line denotes the evolution of the mean of sampled $x_1$; the green and
845 blue lines depict the evolution of the sampled standard deviation (Std) of $x_{10}$ and $x_{50}$,
846 respectively; and the red line denotes the evolution of the covariance between $x_5$ and $x_{30}$. The
847 true values of these four statistics are indicated with different symbols at the right hand side in
848 each plot.

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017

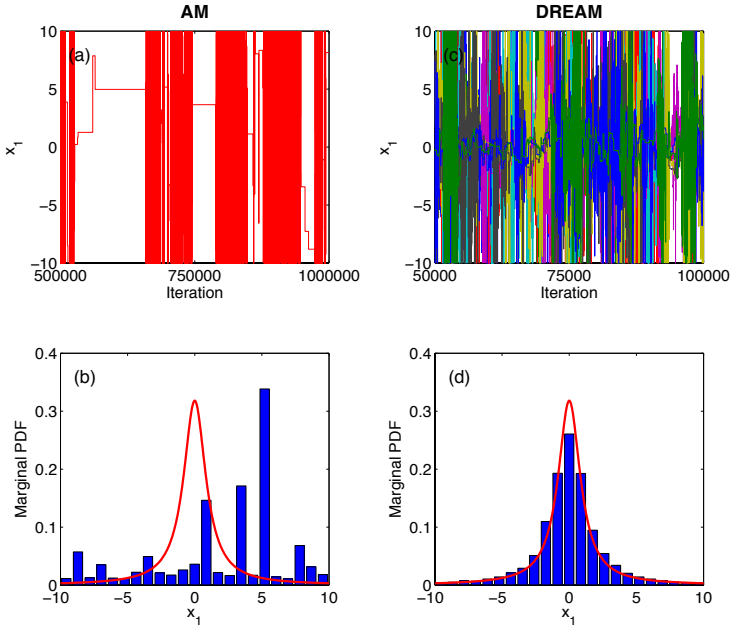




Figure 2. Results for the 10-dimensional Cauchy distribution from AM and DREAM. Evolution of sampled $x_1$ values (a) in one chain of AM and (c) in ten interacting chains of DREAM, where each chain is coded with a different color; and approximated (histogrm) and actual (red curve) marginal posterior distributions of $x_1$ from (b) AM and (d) DREAM.

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017

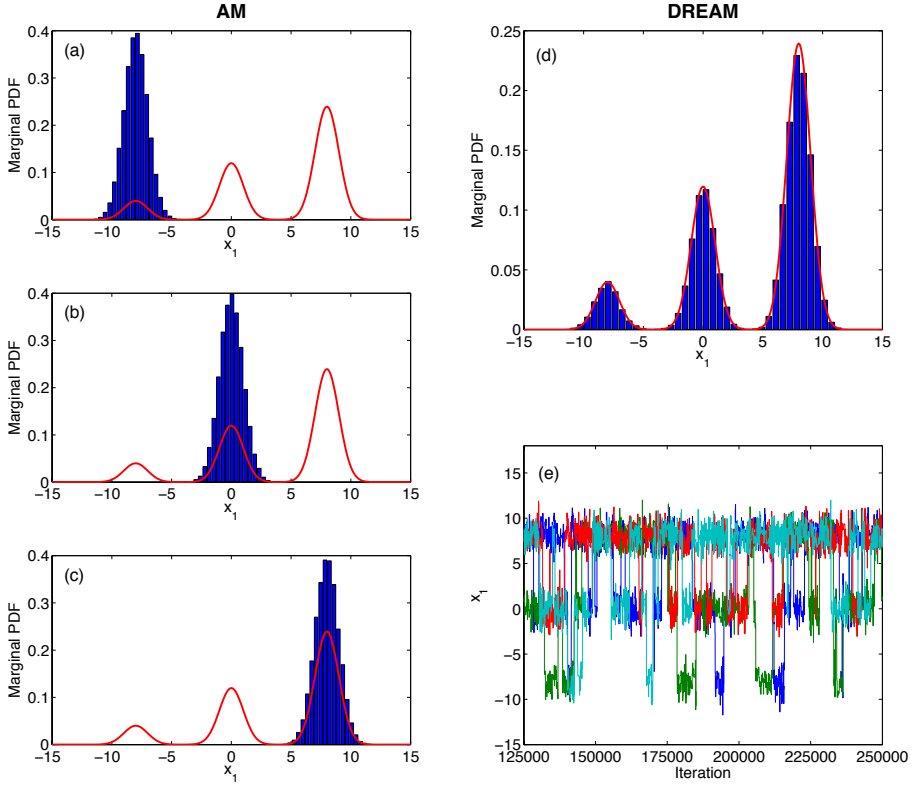




Figure 3. Results for the 4-dimensional trimodal distribution based on AM and DREAM. Approximated (histogrm) and actual (red curve) marginal posterior distributions of $x_1$ from (a-c) three different runs of AM and (d) four interacting chains of DREAM; and (e) evolution of sampled $x_1$ values with DREAM in the four chains, where each chain is coded with a different color.

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017

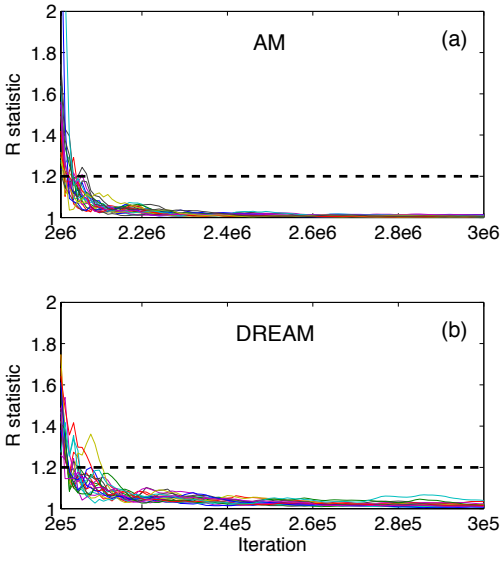




Figure 4. Gelman Rubin $\hat{R}$ statistics of all the 21 parameters in the DALEC model (a) for the last 1,000,000 iterations from ten independent AM runs and (b) for the last 100,000 iterations from the DREAM simulation using ten chains. The values less than the threshold of 1.2 suggest chain convergence.

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017

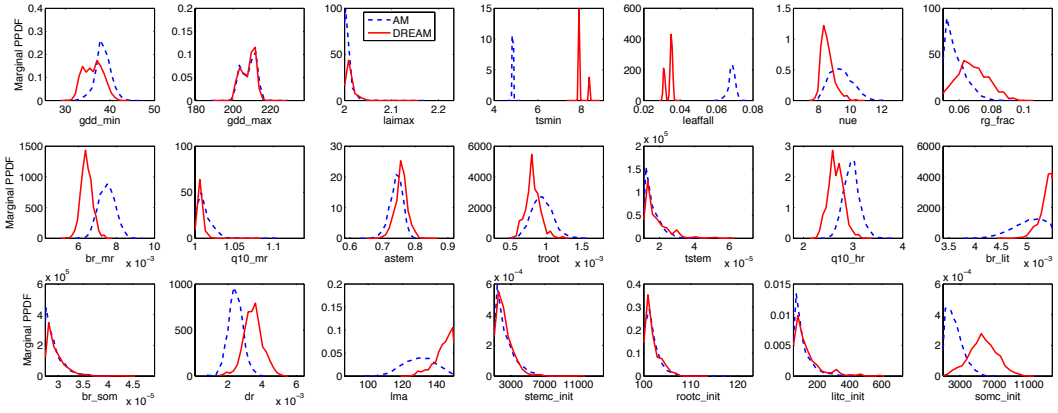




Figure 5. Estimated marginal posterior probability density functions (PPDF) of the 21 parameters in DALEC model based on the AM and DREAM algorithms.

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017

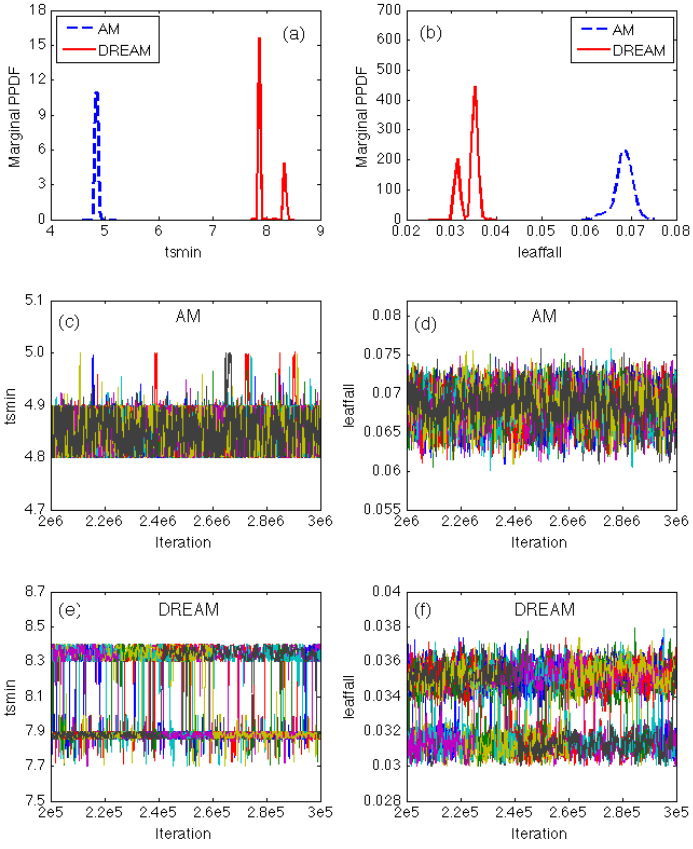




Figure 6. AM and DREAM results for parameters *tsmin* and *leaffall* in the DALEC model. The estimated marginal posterior distributions of (a) *tsmin* and (b) *leaffall*; Trace plots of (c) sampled *tsmin* and (d) sampled *leaffall* with AM using ten independent chains; and trace plots of (e) sampled *tsmin* and (f) sampled *leaffall* with DREAM using ten interacting chains. The evolution of each chain is coded with a different color.

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017

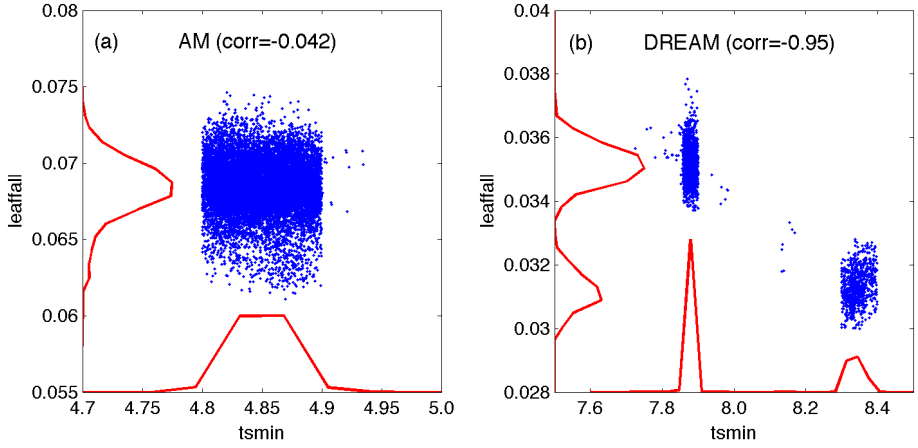




Figure 7. Posterior distributions of parameters *tsmin* and *leaffall* simulated by (a) AM and (b) DREAM. AM simulation results exhibit a negligible correlation coefficient (corr) between the two parameters with a value of -0.042, while DREAM results show that the two parameters are strongly correlated with the corr value of -0.95.

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017

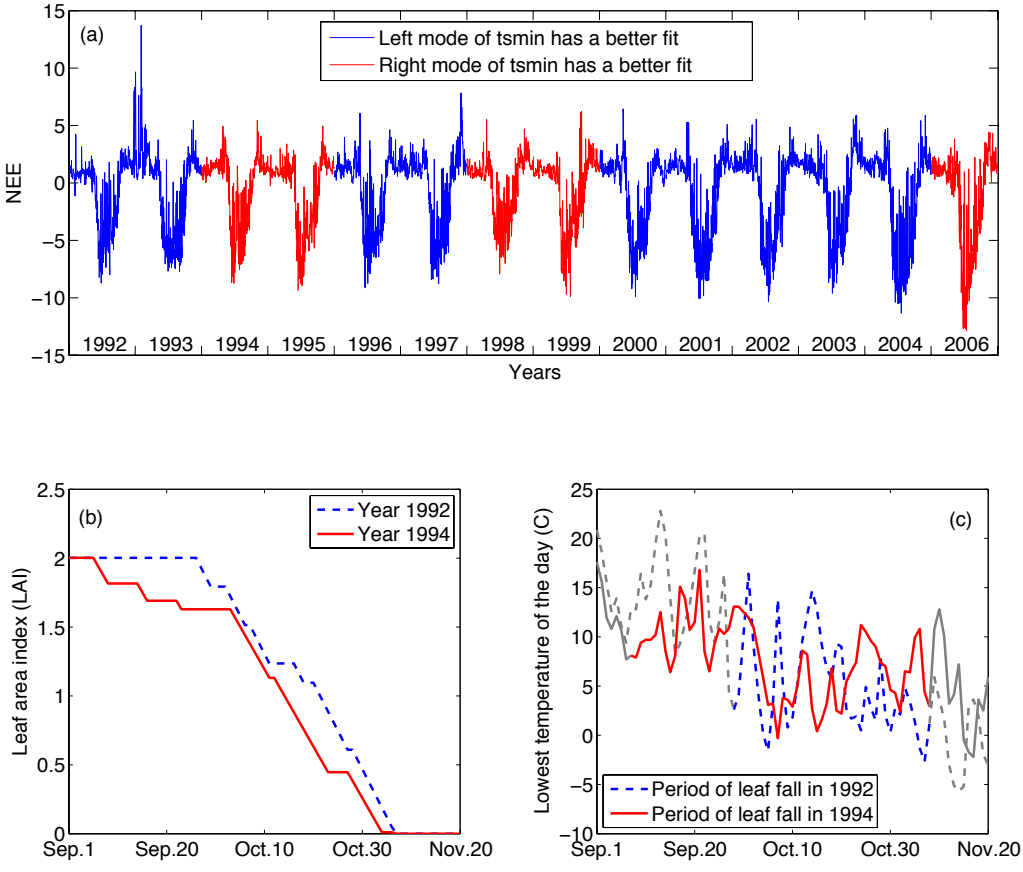
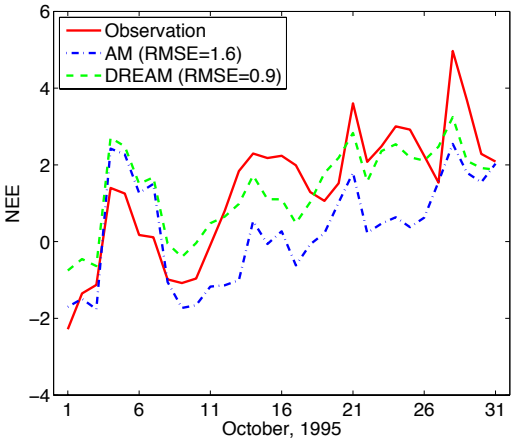




Figure 8. (a) Observed NEE with years highlighted in red where the left mode of *tsmin* has a better model fit and years hightlighted in blue where the right mode of *tsmin* has a better model fit; (b) the simulated leaf area index (LAI) of years 1992 and 1994; and (c) the recorded lowest temperature of years 1992 (blue) and 1994 (red). The blue and red lines in (c) highlight the corresponding periods of leaf fall until LAI becomes zero for 1992 and 1994, respectively. The color scheme is synchronized between (a), (b), and (c) frames. Note that decreases in LAI as predicted by our simplified version of DALEC reflect cholorphyll loss rather than leaf drop.

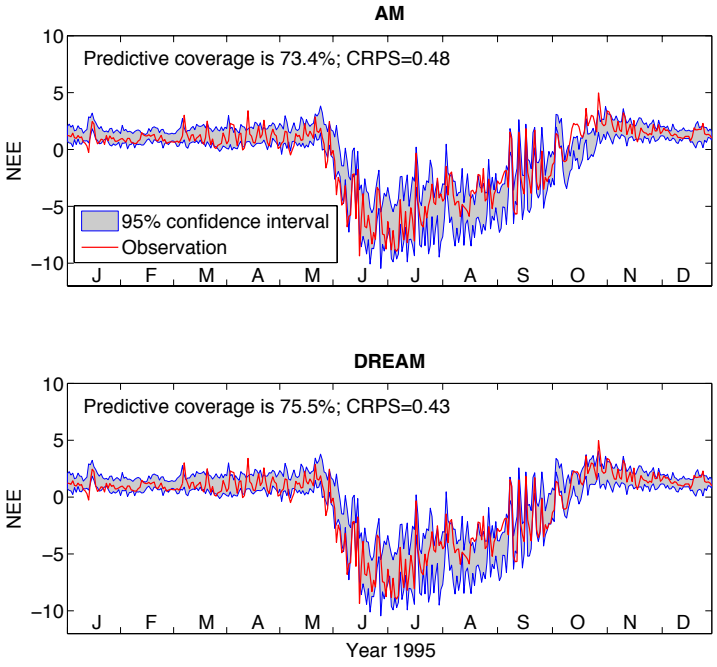
Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017


888

889 Figure 9. Simulated NEE values based on the optimal parameters (i.e., the MAP values listed in
890 Tablel 1) estimated by the AM and DREAM algorithms for October, 1995. The Root Mean
891 Square Error (RMSE), indicates that DREAM produces a better model fit than AM.

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017


892

893 Figure 10. 95% confidence intervals of the simulated NEE values in year 1995 based on the
894 parameter samples from AM and DREAM. Two measures of predictive performance, CRPS
895 statistic and predictive coverage, indicate that DREAM outperforms AM in prediction.

Biogeosciences Discuss., doi:10.5194/bg-2017-41, 2017
Manuscript under review for journal Biogeosciences
Discussion started: 22 February 2017

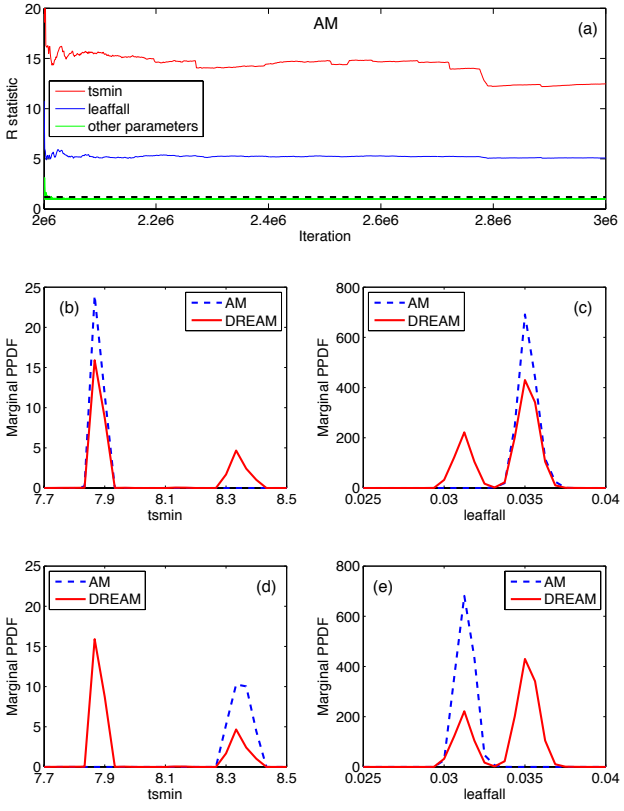




Figure 11. Results of two independent chains of AM with the initial covariance matrix constructed using the converged DREAM samples. The $\hat{R}$ statistic in (a) suggests that different AM chains converged to different *tsmin* and *leaffall* values. One chain captures (b) the left mode of *tsmin* and (c) the corresponding right mode of *leaffall*; and the other chain identifies (d) the right mode of *tsmin* and (e) the corresponding left mode of *leaffall*. No single AM chain can capture all the modes of the two parameters within a reasonable number of MCMC iterations.