Biogeosciences, 14, 1–20, 2017
https://doi.org/10.5194/bg-14-1-2017
© Author(s) 2017. This work is distributed under
the Creative Commons Attribution 3.0 License.

Biogeosciences 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]Computational Sciences and Engineering 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

Correspondence to: Dan Lu (lud1@ornl.gov)

Received: 10 February 2017 – Discussion started: 22 February 2017
Revised: 30 June 2017 – Accepted: 30 August 2017 – Published:

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 work, a differential evolution adaptive Metropolis (DREAM) algorithm is 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 calibration of DREAM results in a better model fit and predictive performance compared to the popular adaptive Metropolis (AM) scheme. Moreover, DREAM indicates that two parameters controlling autumn phenology have multiple modes in their posterior distributions while AM only identifies one mode. The application suggests that DREAM is very suitable to calibrate complex terrestrial ecosystem models, where the uncertain parameter size is usually large and existence of local optima is always a concern. In addition, this effort justifies the assumptions of the error model used in Bayesian calibration according to the residual analysis. The result indicates that a heteroscedastic, correlated, Gaussian error model is appropriate for the problem, and the consequent constructed likelihood function can alleviate the underestimation of parameter uncertainty that is usually caused by using uncorrelated error models.

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.

Two types of Bayesian methods are widely used in parameter estimation of TEMs, variational data assimilation (VAR) methods (Talagrand and Courtier, 1987) and Markov chain Monte Carlo (MCMC) sampling. VAR methods are computationally efficient; however, they assume that the prior parameter values and the observations follow a Gaussian distribution, and they require the model to be differentiable with respect to all parameters for optimization. In addition, VAR methods can only identify a local optimum and approximate the PPDF by a Gaussian function (Rayner et al., 2005; Ziehn et al., 2012). In contrast, MCMC sampling makes no assumptions about the structure of the prior and posterior distributions of model parameters or observation uncertainties. Moreover, the MCMC methods, in principle, can converge to the true PPDF with an identification of all possible optima. Although it is more computationally intensive than VAR approaches, MCMC sampling is being increasingly applied in the land surface modeling community (Dowd, 2007; Zobitz et al., 2011).

One widely used MCMC algorithm is adaptive Metropolis (AM) (Haario et al., 2001). For example, Fox et al. (2009) applied AM in their comparison of different algorithms for the inversion of a terrestrial ecosystem model; Järvinen et al. (2010) utilized AM for estimation of ECHAM5 climate model closure parameters; Hararuk et al. (2014) employed AM for improvement of a global land model against soil carbon data; and Safta et al. (2015) used AM to estimate parameters in the data assimilation linked ecosystem carbon model. The AM algorithm uses a single Markov chain that continuously adapts the covariance matrix of a Gaussian proposal distribution using the information of all previous samples collected in the chain so far (Haario et al., 1999). As a single-chain method, AM has difficulty in traversing multidimensional parameter space efficiently when there are numerous significant local optima, and AM can be inefficient for estimating the PPDFs that exhibit strong correlations, as correlated dimensions are better to be updated together (Vrugt, 2016). In addition, the AM algorithm uses a multivariate Gaussian distribution as the proposal to generate candidate samples and evolve the chain. AM, therefore, is particularly suitable for Gaussian-shaped PPDFs, but it may not converge properly to the distributions with multiple modes. Moreover, AM suffers from uncertainty about how to initialize the covariance of the Gaussian proposal. Poor initialization of the proposal covariance matrix results in slow adaptation and inefficient convergence.

The Gaussian proposal is also widely used in non-AM MCMC studies that involve TEMs. For example, Ziehn et al. (2012) used the Gaussian proposal for the MCMC simulation of the BETHY model (Knorr and Heimann, 2011) and Ricciuto et al. (2008, 2011) utilized the Gaussian proposal in their MCMC schemes to estimate parameters in a terrestrial carbon cycle model. The single-chain and Gaussian-proposal MCMC approaches have limitations in sufficiently exploring the full parameter space and share slow convergence in sampling the non-Gaussian-shaped PPDFs and thus may end up with a local optimum with inaccurate uncertainty representation of the parameters. Therefore, this poses a question on whether the AM and the widely used MCMC algorithms with Gaussian proposal generate a representing sample of the posterior distribution of the underlying model parameters. While we expect that computationally expensive sampling methods for parameter estimation yield a global optimum with an accurate probabilistic description, in reality we may in many cases obtain a local optimum with an inaccurate PPDF due to the limitations of these algorithms.

In this study, we employ the differential evolution adaptive Metropolis (DREAM) algorithm (Vrugt et al., 2008, 2009a; Lu et al., 2014) for an accurate Bayesian calibration of an ecosystem carbon model. The DREAM scheme runs multiple interacting chains simultaneously to explore the entire parameter space globally. During the search, DREAM does 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 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. Recently, Post et al. (2017) reported a successful application of DREAM in estimation of the complex Community Land Model (CLM) using 1-year records of NEE observations. They found that the posterior parameter estimates were superior to their default values in the ability to track and explain the measured NEE data.

While multimodality is a potential feature of parameters in complex models (Kinlan and Gaines, 2003; Stead et al., 2005; 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. In addition, while the importance of likelihood function choice on Bayesian calibration has been well realized (Trudinger et al., 2007), the reasonable usage of an appropriate likelihood function has been barely explored in land surface modeling. Here we apply DREAM and AM to a TEM to estimate the parameter distributions based on a set of synthetic data and real measurement data. In both cases, 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; (3) investigate the effects of model calibration methods on parameter estimation and model performance; and (4) justify the usage of the likelihood function and explore the influence of the likelihood function on the model calibration results. This work should provide ecological practitioners with valuable information on model calibration and understanding of the TEMs.

In the following Sect. 2, we first briefly summarize the general idea of Bayesian calibration and describe the AM and DREAM algorithms. Then in Sect. 3, we apply both algorithms to the DALEC model in a synthetic and a real-data study. Next in Sect. 4, we discuss the influence of the likelihood function on parameter estimation and model performance. Finally in Sect. 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 $p(\boldsymbol{x}|\boldsymbol{D})$ of the model parameters $\boldsymbol{x}$, given observation data $\boldsymbol{D}$, can be obtained from the prior distribution $p(\boldsymbol{x})$ of x and the likelihood function $L(\boldsymbol{x}|\boldsymbol{D})$ using Bayes' theorem (Box and Tiao, 1992) via

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

where c is a distribution represents the prior knowledge about the parameters. It is usually inferred from information of previous studies at similar sites or from expert judgment. In the absence 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 heteroscedastic and uncorrelated variances that are evaluated from the provided daily observation uncertainties. The assumptions of normality and independence are investigated by the residual analysis. In addition, we explore the influence of different choices of the likelihood function on the parameter estimation and model performance. The effect of data correlations on the inferred parameters was also assessed in our previous study (Safta et al., 2015).

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 well-constructed MCMC schemes have been theoretically proven to converge to the appropriate target distribution $p(\boldsymbol{x}|\boldsymbol{D})$ under certain regularity conditions (Robert and Casella, 2004, p. 270). However, in practice the convergence rate is often impractically slow, which suggests that within a limited finite number of iterations, some inefficient schemes may result in an unrealistic distribution. The inefficiency is typically resulted from an inappropriate choice of the proposal distribution used to generate the candidates. Either wide or narrow proposal distribution can cause inefficient chain mixing and slow chain convergence (Geyer, 1992; Tierney, 1994). Hence, the definition of the proposal distribution is crucial and determines the efficiency and the practical applicability of the MCMC simulation.

2.3 AM algorithm

The adaptive Metropolis (AM) algorithm is a modification to the standard Metropolis sampler (Haario et al., 2001). The key feature of the AM algorithm is that it uses a single Markov chain that continuously adapts to the target distribution via its calculation of the proposal covariance using all previous samples in the chain. The proposal distribution employed in the AM algorithm is a multivariate Gaussian distribution with means at the current iteration x_t and having a covariance matrix $\mathbf{C}_t$ that is updated along the chain evolution. To start the chain, AM first selects an arbitrary, strictly positive definite initial covariance $\mathbf{C}_0$ according to the best prior knowledge that may be very poor. Then after a certain number of iterations T, the covariance is updated based on the samples gained so far.

To apply the AM algorithm, an initial covariance $\mathbf{C}_0$ must be defined. The choice of $\mathbf{C}_0$ critically determines the success of the algorithm. For example, in an extreme case, the variance of $\mathbf{C}_0$ is so large that no proposals are accepted within an iteration, and that the chain remains at the initial state without any movement. This situation continues as the chain evolves, and the use of updated $\mathbf{C}_t$ makes no difference because the variances of $\mathbf{C}_t$ are essentially zero since all the previous samples have the same values. Finally, the AM sampler would get stuck in its initial state without exploring the parameter space. To alleviate this problem and start AM fairly efficiently, we can define $\mathbf{C}_0$ 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 $\mathbf{C}_0$ from a test run of 50 000 simulations of a matrix approximation of the community land model in estimating the PPDFs of soil carbon related parameters.

The construction of $\mathbf{C}_t$ is another critical influence on the AM performance. In practice, some adjustments on $\mathbf{C}_t$ are necessary to improve the AM efficiency. For example, when the chain does not have enough movement after a large number of iterations, we can shrink $\mathbf{C}_t$ by some constant to increase acceptance of new samples, and vice versa. The techniques used in the formulation of $\mathbf{C}_0$ and $\mathbf{C}_t$ improve the AM efficiency in some degree for some problems. But, the computational cost spent on applying these techniques is not negligible (such as the test runs used for determining the $\mathbf{C}_0$) and some strategies require some artificial controls (such as manual adjustment of the scaling factor of $\mathbf{C}_t$). Moreover, determining a reasonable $\mathbf{C}_0$ and $\mathbf{C}_t$ becomes difficult for high-dimensional problems.

To improve efficiency in high-dimensional case, Haario et al. (2005) extended the standard AM method to componentwise adaptation. This strategy applies AM on each parameter separately. The proposal distribution of each component is a 1-D normal distribution, which is adapted in a similar manner as in the standard AM algorithm, but the componentwise adaptation does not work very 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 may suffer from some difficulties in judging the convergence. Sometime 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 univariate $\hat{R}$ statistic (Gelman and Rubin, 1992) and the multivariate $\hat{R}$ statistic (Brooks and Gelman, 1998), to check convergence. 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 MCMC algorithm is unable to identify all the modes.

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 $\hat{R}$ statistics 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 points 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. In addition, the DE-MC can successfully simulate the multimodal distributions, because it directly uses the current location of the multiple chains to generate candidate points, allowing the possibility of direct jumps between different modes.

The DREAM algorithm maintains the nice features of the DE-MC but greatly accelerates the chain convergence. More information about the DREAM algorithm was presented in Vrugt et al. (2008, 2009a), Laloy and Vrugt (2012), Lu et al. (2014), and Vrugt (2016).

2.5 Strategies and capabilities of AM and DREAM in sampling complex problems

Since multimodality is a potential feature of complex problems including terrestrial ecosystem models (Stead et al., 2005; Thibault et al., 2011), it is important to understand the strategies of AM and DREAM and to investigate their capabilities in sampling the multimodal distributions.

The AM sampler is typically tuned for distributions with a single mode. For distributions with closely connected modes, AM can work well with suitable initial values. On the other hand, for distributions consisting of disconnected modes with between regions of low probability, even with a reasonably wide covariance matrix, AM could have a slow convergence and end up with only one mode (e.g., Fig. 5 in Vrugt, 2016). To remedy this problem, AM needs an overly dispersed Gaussian proposal with large initial variances to allow it to transit between the different modes. But this may result in a very low acceptance rate as many of the jumps will fall outside the target distribution with nearly zero densities. To alleviate this problem, Haario et al. (2006) proposed the DRAM algorithm, which combines the delayed rejection (DR) with AM. The DR algorithm allows for a very expansive search at

the beginning by using a large covariance matrix of the proposal, and then the proposal covariance is reduced by a freely chosen scale factor if the parameters do not have significant movement. By creating multiple proposal stages, DRAM enables an extensive search and meanwhile alleviates the overshooting problem and improves the acceptance rate. However, as dimensionality increases, the multimodality becomes more difficult for the algorithms using the Gaussian proposal because it is highly likely different dimensions have different variances and a constant scaling factor can only shrink the covariance simultaneously.

In contrast, DREAM is designed for sampling high-dimensional and multimodal problems by running multiple different chains simultaneously for global exploration. It automatically tunes the scale and orientation of the proposal in randomized subspaces during the search (Vrugt et al., 2009a). As DREAM directly uses the current location of the multiple chains, instead of the covariance of the Gaussian proposal, to generate candidate points, it enables direct jumps between different modes (including the relatively far disconnected modes) as long as the initial samples of the chains are widely distributed over the parameter space. Laloy and Vrugt (2012) demonstrated that DREAM can successfully sample a 25-dimensional trimodal distribution with equal separation of 10 units between modes. However, for the same problem with the same number of function evaluations, AM and DRAM converged to only one mode. Note that to sample a distribution with many modes, one needs to have some prior information about their rough locations; otherwise no methods can guarantee finding all the modes, especially when the distance between the modes is very large and not a constant.

3 Application to a terrestrial ecosystem model

In this section, we applied 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 was 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 used 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.

3.1 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 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 parameters, leaf C : N ratio (which is fixed at a constant of 25 in the simulation) and photosynthetic nitrogen use efficiency (*nue*); and 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

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

	ParName	Nom. Val.	Range	MAP estimates AM LL = −6662.6	MAP estimates 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}	7.39×10^{-3}	6.35×10^{-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	$1/(50\times365)$	$1/(250\times365)$–$1/(10\times365)$	1.98×10^{-5}	1.63×10^{-5}
	troot	$1/(5\times365)$	$1/(25\times365)$–$1/365$	8.55×10^{-4}	7.88×10^{-4}
Decomp.	q10_hr	2	1–4	2.98	2.68
	br_lit	$1/(2\times365)$	$1/(5\times365)$–$10/(5\times365)$	4.97×10^{-3}	5.36×10^{-3}
	br_som	$1/(30\times365)$	$1/(100\times365)$–$1/(10\times365)$	2.79×10^{-5}	2.88×10^{-5}
	dr	10^{-3}	10^{-4}–10^{-2}	2.46×10^{-3}	3.39×10^{-3}
Init. C.	stemc_init	5000	1000–15 000	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	1000–25 000	2029.1	4708.2

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

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*, which 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 years between 1992 and 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. To avoid the influence of prior distributions on the investigation of the posteriors estimated by AM and DREAM, uniform priors were used for all parameters with the ranges specified in Table 1.

3.2 Calibration data

The calibration data consist of the Harvard Forest daily net ecosystem exchange (NEE) values, which were processed for

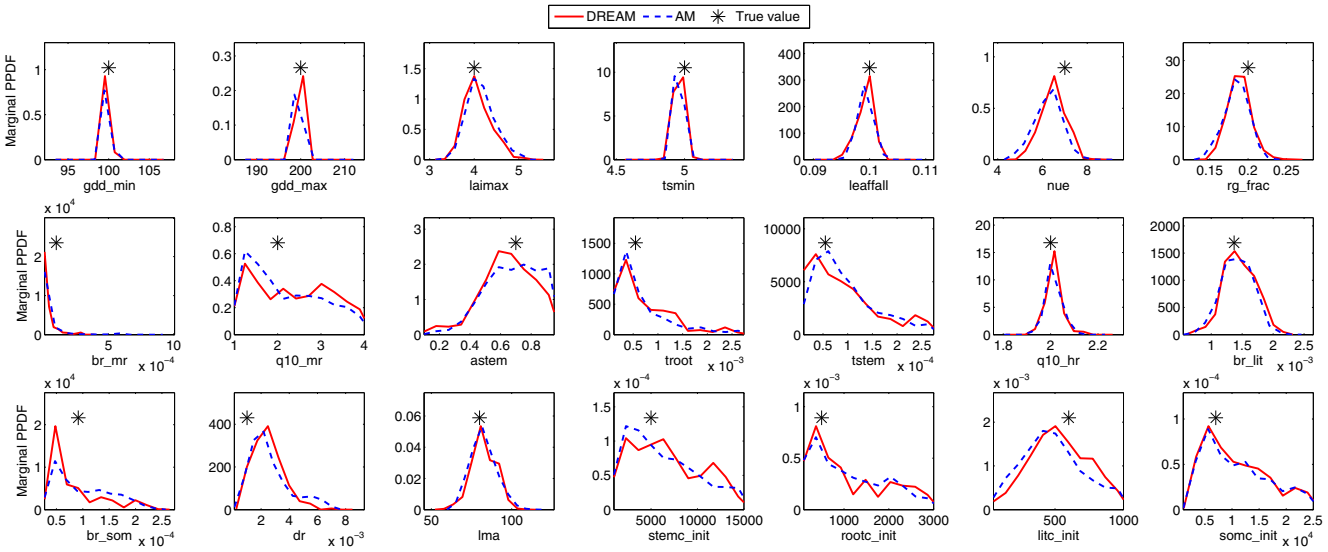

Figure 1. Estimated marginal posterior probability density functions (PPDFs) of the 21 parameters using the AM and DREAM algorithms, along with the true parameter values to generate the pseudo-data in the synthetic case.

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 5114 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 heteroscedastic, diagonal covariance matrix of the Gaussian likelihood function by assuming the data were uncorrelated. In Sect. 4, we examine the independent, Gaussian error assumption using residual analysis and investigate the influence of error models on parameter estimation and model performance.

3.3 Synthetic study with pseudo-data

We first applied AM and DREAM to a synthetic case to evaluate their capability in parameter estimation. The same periods of daily NEE data were generated with the nominal parameter values in Table 1. These synthetic data for calibration were then corrupted with Gaussian errors having means at zero and the same standard deviations with the observed NEEs.

DREAM launched 10 parallel chains starting at values randomly drawn from the parameter prior distributions. AM used one chain and the chain has the same initialization with DREAM. In addition, AM also requires the initialization of the covariance matrix of its Gaussian proposal. We first drew some samples from the parameter space and computed the initial covariance. However, this initialization caused a slow convergence of AM with an extremely small acceptance rate (about 0.01 % after 1×10^5 iterations). The reason could be that for this rather high-dimensional problem with very diverse parameter ranges, the candidate samples are easily outside the target distribution when they are drawn from the Gaussian proposal. To facilitate the AM convergence, we started the chain from the true parameter values and constructed the initial covariance from samples around the true values. This setup can only be done in a synthetic case with information of true parameters available; practically it needs some test runs to get information about the underlying distributions. In addition, this initialization of AM makes an unfair comparison with DREAM that launched chains blindly, but on the other hand, it suggests DREAM's ease of use and setup, its robustness and efficiency.

Chain convergence was assessed via the Gelman–Rubin $\hat{R}$ statistics. Figure 1 presents the estimated marginal PPDFs of the 21 parameters from both AM and DREAM samples after convergence along with their true values. The two algorithms produce very similar distributions that both enclose the true values very well. All the parameters show one mode in their PPDFs and the true values are located or close to the modes. The results indicate that for this unimodal problem both algorithms can successfully infer the underlying parameter distributions, although AM needs a proper initialization for its convergence. To further evaluate the calibration accuracy, we investigate the sum of squared weighted residuals (SSWR) for the optimal parameters. If the parameter optimization is reasonable, the calculated SSWR should follow a chi-squared distribution with its mean equal to the k degrees of freedom, i.e., the number of calibration data

minus the number of calibrated parameters, in this study $k = 5114 - 21 = 5093$. The resulted SSWR is 5044 close to the mean value 5093 of the chi-squared distribution. This once again suggests the accuracy and reasonability of our parameter estimation.

In addition, Fig. 1 indicates that about half of the parameters are well constrained, when we define a well-constrained parameter as its posterior distribution occupying at most half the range of the prior distribution (Keenan et al., 2013). This result is consistent with some of previous studies on DALEC calibration using NEE data alone. For example, in the synthetic study of Fox et al. (2009), their MCMC simulation (M1) showed that 16 of 17 parameters were well constrained. Similarly, the synthetic study in Hill et al. (2012) indicated that 20 of 23 parameters had their 90 % confidence intervals occupy less than half of the prior range.

Whether a parameter is identifiable depends on the model, model parameters, and the calibration data. When the parameter related processes are necessary to simulate the model outputs whose corresponding observation data are sensitive to the parameters, the parameters can usually be identified and sometimes well constrained. For example, Keenan et al. (2013) showed that in their FöBAAR model with 40 parameters, many parameters could not be constrained even with the consideration of several data streams together. They found that these unidentifiable parameters might be redundant in the model structure representation. Roughly speaking, for a simple model with a few number of parameters, the parameters can be more identifiable than the complex models with a large parameter size (Richardson et al., 2010; Weng and Luo, 2011). On the other hand, if the calibration data are sensitive to the parameters, even a complex model can sometimes be well constrained by using a single type of observations. For example, Post et al. (2017) estimated eight CLM parameters using 1-year records of half-hourly NEE observations at four sites, and found that for most sites the CLM parameters can be well constrained with their 95 % confidence intervals close to the maximum a posteriori estimates. For the only site where the parameter uncertainties were relatively large, they concluded that the simulated NEE was less sensitive to these parameters. In our and those synthetic studies of Fox et al. (2009) and Hill et al. (2012), all the parameter related processes are necessary for DALEC simulation and most parameters were shown to be sensitive to the observation data (Safta et al., 2015), this explains to some extent that many DALEC parameters can be well constrained in these synthetic studies.

3.4 Real-data study

In the real-data study, the measured NEE data with given standard deviations were used for DALEC calibration. Both AM and DREAM algorithms were applied to infer the unknown parameters. Different from the synthetic case, the real-data study involves model structural errors besides the measurement errors. We again use the heteroscedastic, uncorrelated, Gaussian likelihood function for calibration and examine these error assumptions in Sect. 4 through residual analysis.

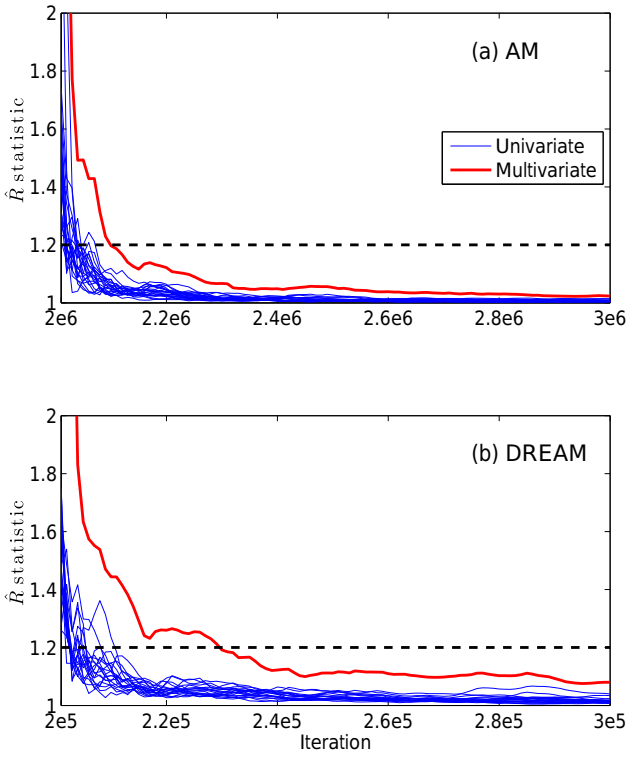

Figure 2. Univariate and multivariate Gelman–Rubin $\hat{R}$ statistics **(a)** for the last 1 000 000 iterations from 10 independent AM runs and **(b)** for the last 100 000 iterations from the DREAM simulation using 10 interacting chains. The values less than the threshold of 1.2 suggest chain convergence.

DREAM launched 10 parallel chains starting at values randomly drawn from the parameter prior distributions, and each chain evolved 300 000 iterations. Chain convergence was assessed via both the univariate and multivariate Gelman–Rubin $\hat{R}$ statistics. Figure 2b plots the $\hat{R}$ values of the 21 parameters for the last 100 000 iterations. The figure suggests that the last 50 000 samples of each chain (i.e., total 500 000 samples from 10 chains) can be used for the PPDF approximation as the $\hat{R}$ has values below the threshold of 1.2.

AM used one chain and the chain has the same initialization of the first sample with DREAM. For the initialization of the Gaussian covariance in the AM proposal, we first drew some samples from the parameter space and constructed the covariance. However, this initialization caused a high rejection rate and ended up with essentially a single parameter state after hundreds of thousands of iterations. To facilitate the convergence of AM, we constructed the initial covariance based on the first 200 000 samples from the DREAM simulation. We conducted 10 independent AM runs, so the same $\hat{R}$ statistics can be used for convergence diagnosis. Each AM

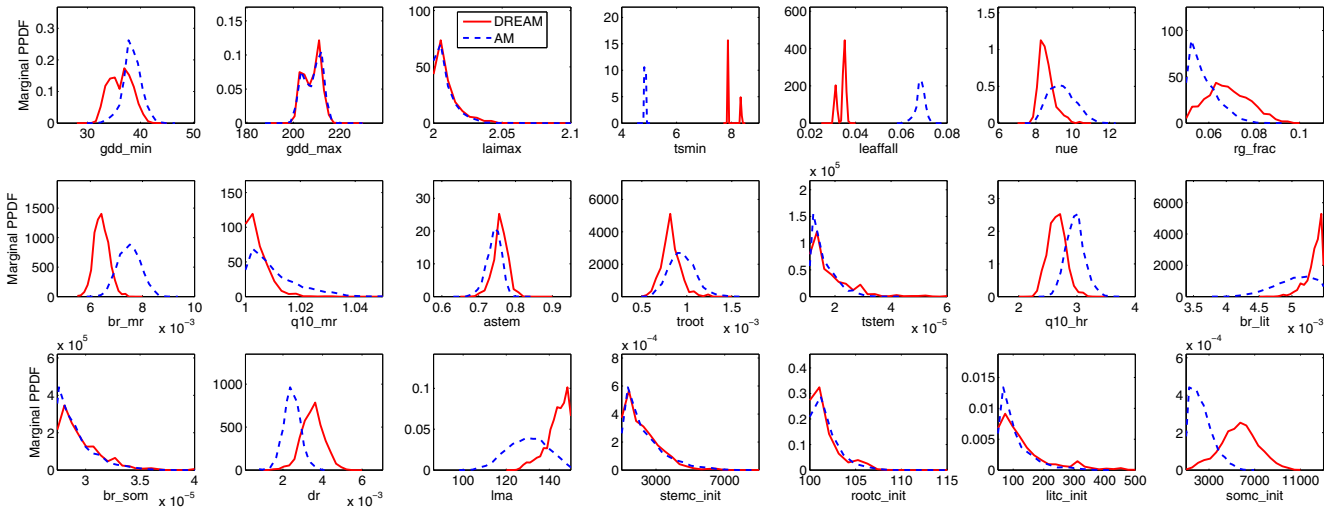

Figure 3. Estimated marginal posterior probability density functions (PPDFs) of the 21 parameters using the AM and DREAM algorithms in the real-data study.

chain simulated 3 000 000 samples, so that the number of function evaluations in one AM chain is the same as that of DREAM using 10 chains. The $\hat{R}$ values of all parameters based on the 10 AM runs for the last 1 000 000 iterations are shown in Fig. 2a. The figure indicates that AM has converged and the last 500 000 samples from one chain were used for the PPDF approximation.

The estimated PPDFs from AM and DREAM are presented in Fig. 3, and the optimal parameter estimates, as represented by the maximum a posteriori (MAP), are summarized in Table 1. Figure 2 shows that more than half of the parameters are constrained and some well-constrained parameters are edge hitting, where the mode of these parameters occur near one of the edges of their allowable ranges and most of the parameter values are clustered near the edge such as *stemc_init*, *rootc_init*, and *litc_init*. As we can see in the synthetic case, these edge-hitting parameters (e.g., *tstem*, *stemc_init*, *rootc_init*, and *litc_init*) have wide confidence intervals that almost occupy the entire allowable ranges, indicating that the NEE data should provide little information about these parameters. This edge-hitting behavior may be caused by a compensation for model structural errors and data biases (Braswell et al., 2005), and we do not consider these edge-hitting parameters to be well constrained despite small posterior uncertainties. The tight uncertainty bounds on these parameters are likely unrealistic and could contribute to overconfidence in model predictions. However, quantifying model structural error is an on-going research topic and no formal results have been published to our knowledge. We will investigate the influence of model structural errors on parameter estimation in future studies.

In comparison of the results between AM and DREAM, Fig. 3 indicates that they produce very similar PPDFs for many parameters, such as *gdd_max*, *laimax*, *br_som*, *stemc_init*, and *rootc_init*; however, for parameters *tsmin* and *leaffall*, their estimated PPDFs are substantially different. This also can be seen in Table 1 where the differences of MAP values for most parameters are relatively small between the two algorithms, the relative difference for *tsmin* and *leaffall* is 38 and 94 %, respectively. The parameter *tsmin* represents the temperature triggering leaf fall and the *leaffall* represents the rate of leaf fall on days when the temperature is below *tsmin*. We further analyze the simulations of these two parameters from AM and DREAM in Fig. 4. Figure 4a and b illustrate two separated modes in the estimated marginal PPDFs of *tsmin* and *leaffall* obtained from DREAM, while AM only identifies one mode for both parameters and they dramatically differ from any modes simulated by DREAM. For example, the single mode of *tsmin* identified by AM gives a lower temperature threshold (meaning a later initiation of senescence) that is compensated for by a higher estimate of *leaffall* rate compared to DREAM. As shown in the trace plots of Fig. 4c and d, all 10 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 10 parallel chains of DREAM, as exhibited in Fig. 4e and f, jumps back and forth between two modes. And the two parameters compensate for 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 Fig. 5, 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

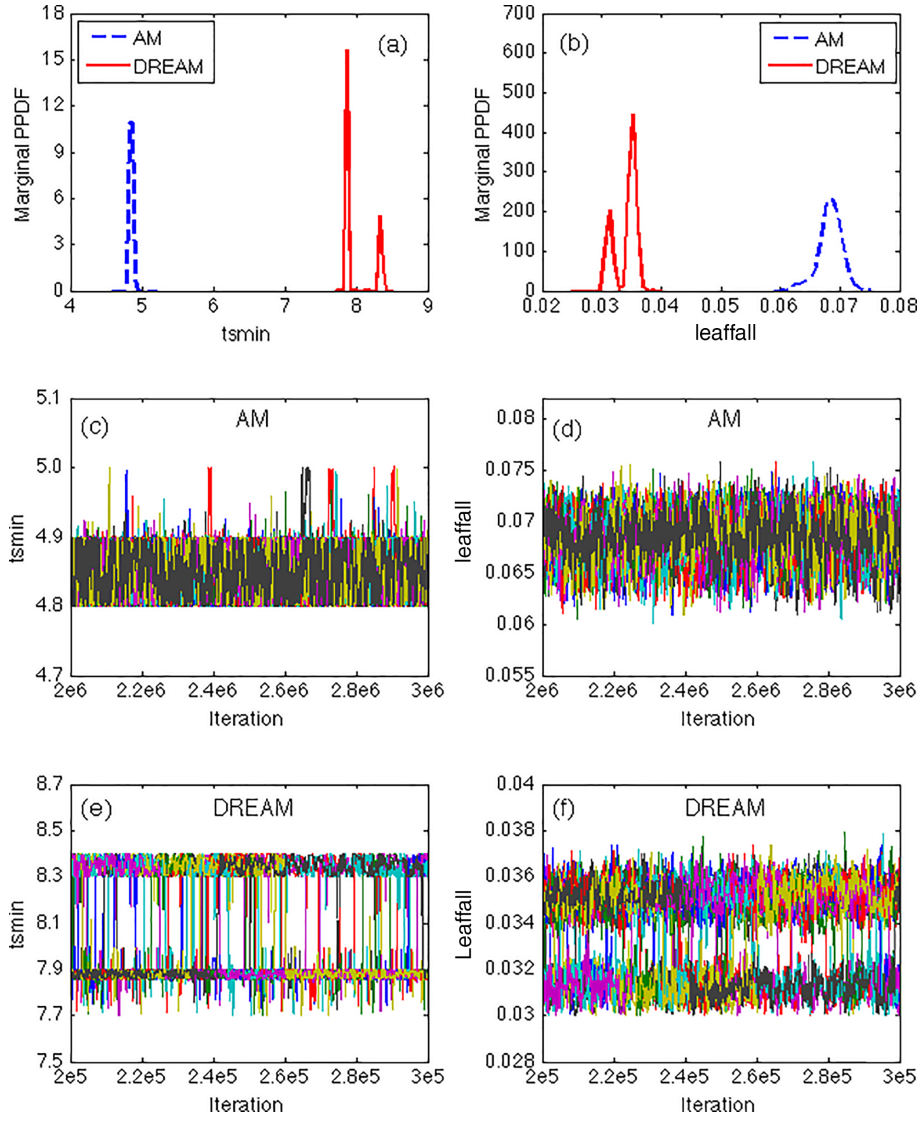

Figure 4. AM and DREAM results for parameters *tsmin* and *leaffall* in the DALEC model. The estimated marginal posterior distributions of **(a)** *tsmin* and **(b)** *leaffall*; Trace plots of **(c)** sampled *tsmin* and **(d)** sampled *leaffall* with AM using 10 independent chains; and trace plots of **(e)** sampled *tsmin* and **(f)** sampled *leaffall* with DREAM using 10 interacting chains. The evolution of each chain is coded with a different color.

with the correlation coefficient of −0.95. As demonstrated in Fig. 5b, 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 is triggered later (i.e., a smaller *tsmin*) but proceeds at a faster rate (i.e., a larger *leaffall*), while in some other years the senescence is triggered earlier (i.e., a larger *tsmin*) but proceeds 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).

Figure 6a highlights the years in red where the model based on the right mode of *tsmin* and the left mode of senescence rate (*leaffall*) has a better fit to the observed NEE, i.e., years 1994, 1995, 1998, 1999, and 2006. The remaining years are highlighted in blue where the left mode of *tsmin* and the right mode of *leaffall* result in a better model fit.

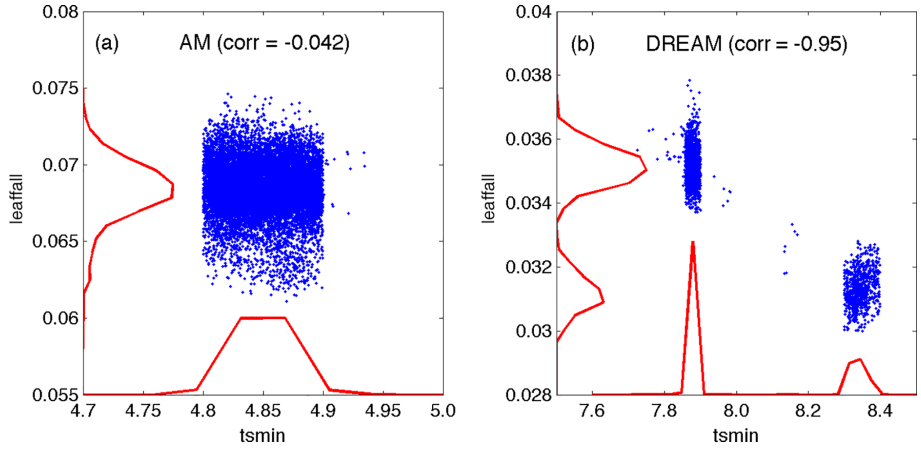

Figure 5. Posterior distributions of parameters *tsmin* and *leaffall* simulated by **(a)** AM and **(b)** DREAM. AM simulation results exhibit a negligible correlation coefficient (corr) between the two parameters with a value of −0.042, while DREAM results show that the two parameters are strongly correlated with the corr value of −0.95.

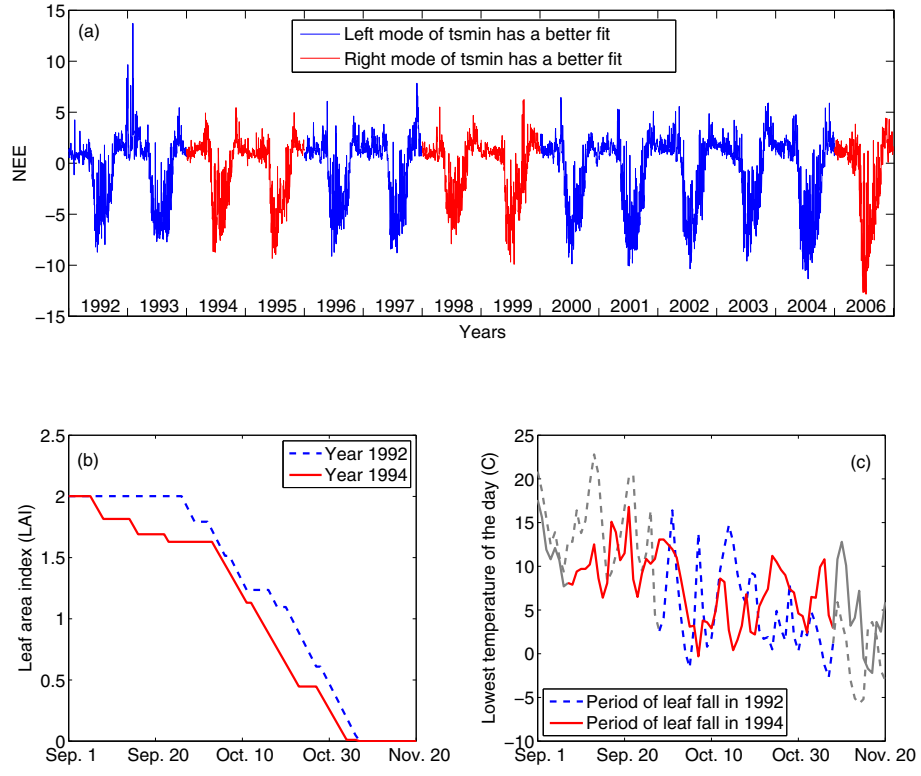

Figure 6. **(a)** Observed NEE with years highlighted in red where the left mode of *tsmin* has a better model fit and years highlighted in blue where the right mode of *tsmin* has a better model fit; **(b)** the simulated leaf area index (LAI) of years 1992 and 1994; and **(c)** the recorded lowest 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 chlorophyll loss rather than leaf drop.

Taking years 1992 and 1994 as an example, we examined the leaf area index (LAI) in the period of senescence. Figure 6b shows that at the first few days of September in both years, the values of LAI were the same around 2.0; after that the timing of senescence during the two years differs dramatically. In year 1994, the value of LAI started decreasing on 7 September, and then decreased slowly over several distinct cool periods during the rest of September and early October until it hit zero in 7 November; the process took about 61 days. In contrast, in year 1992, the value of

LAI remained near the maximum value during all of September, then dropped rapidly in October and hit zero also on 7 November; this process took about 40 days. The changes in the LAI between the two years reflect the variability in the time of year when the leaves start to drop and the rate of leaf drop. Although the leaf fall in 1992 was triggered later than in 1994, the leaves in 1992 dropped at a faster rate, resulting in LAI approaching zero at the same time of the year.

Figure 6c depicts the recorded lowest temperature of the days between 1 September and 20 November for years 1992 and 1994, where the red line highlights the period between the first leaf and the last leaf drops in 1994. The blue line highlights the corresponding period of leaf fall in 1992. Since the senescence was triggered in the early September of 1994, the temperature of triggering leaf fall was relatively high, about 8.1 °C (associated with the higher mode of *tsmin*) as shown in Fig. 6c. In the rest days of September in 1994 following the senescence trigger, temperatures remained warm. The slower leaf fall rate associated with periodic warm conditions (temperatures above *tsmin*) and the lower mode of *leaffall* caused a slow leaf fall in September of 1994 as shown in Fig. 6b. In comparison, in 1992, senescence was triggered at the end of September with a low temperature of 2.6 °C. Then in October with colder temperatures, the leaves drop at a rapid rate associated with the consistent cold temperatures and higher mode of *leaffall*. Especially in late October, the temperatures are consistently below *tsmin*, causing a fast rate of leaf fall, as shown in Fig. 6b, where the decreasing rate of the LAI in the late October of 1992 is very large. This indicates that a higher temperature trigger is usually associated with a lower leaf fall rate and vice versa.

The bimodality identified in the DREAM simulation and examined in the scenarios above reflects the inability of the model structure to predict the observations consistently with a single set of parameters. This bimodality examined in DREAM may be caused in part by an incomplete representation of the senescence process. Using a temperature threshold (parameter *tsmin*) and a constant rate of leaf fall (parameter *leaffall*) to predict senescence is almost certainly an oversimplification. In reality, the process of senescence is also affected by day length. Longer days and warmer temperatures cause a relatively slow rate of leaf fall, whereas shorter days and cooler temperatures accelerate the rate that the leaves fall (Leigh et al., 2002; Saxena, 2010). The higher mode of *tsmin* means that senescence is initiated earlier, when day lengths are still relatively long. This may partially explain why this mode is associated with a lower mode of the *leaffall* parameter. Other factors not represented in DALEC are also likely to play a role such as soil moisture, or a more complex relationship with spring phenology (Keenan et al., 2014, 2015).

The difference in estimated parameters between AM and DREAM causes different simulations of NEE, especially during the autumn. As an example, Fig. 7 illustrates the comparison of the simulated NEE to observations for a month in autumn of the year 1995 based on MAP estimates obtained

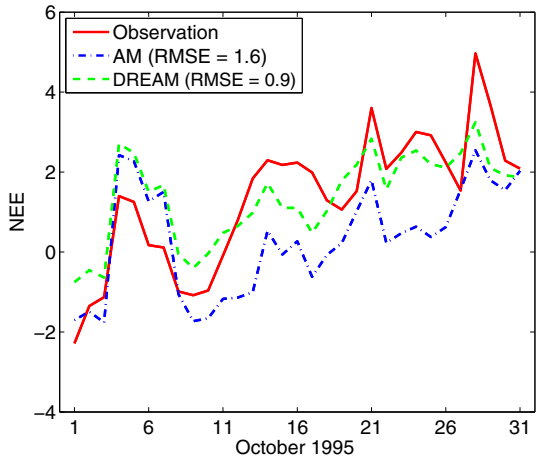

Figure 7. Simulated NEE values based on the optimal parameters (i.e., the MAP values listed in Table 1) estimated by the AM and DREAM algorithms in October 1995. The root mean square error (RMSE) indicates that DREAM produces a better model fit than AM.

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

3.5 Assessment of predictive performance

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

$$p(\boldsymbol{y}|\boldsymbol{D}) = \int p(\boldsymbol{y}|\boldsymbol{x})\, p(\boldsymbol{x}|\boldsymbol{D}) dx. \quad (2)$$

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

From the estimated $p(\boldsymbol{y}|\boldsymbol{D})$, we extracted the 95 % confidence intervals for daily NEE values in the year 1995 and presented the results in Fig. 8. The top panel corresponds to

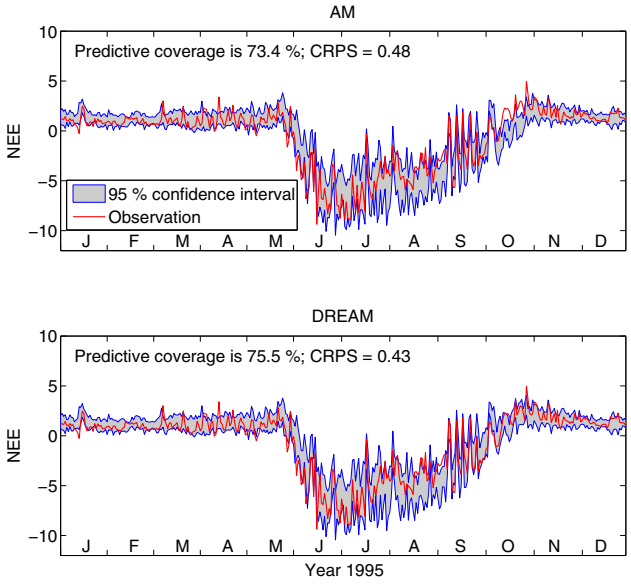

Figure 8. 95 % confidence intervals of the simulated NEE values in year 1995 based on the parameter samples from AM and DREAM. Two measures of predictive performance, CRPS statistic and predictive coverage, indicate that DREAM outperforms AM in prediction.

the results of AM and the bottom panel to DREAM. Overall, the predictive intervals from both algorithms cover well the observed NEE for the entire time range with occasional spikes outside the intervals. Closer visual inspection indicates that DREAM produces better predictive performance than AM. As seen during the period in October, the predictive interval of DREAM can enclose most of the observed NEE while AM actually has under-prediction, causing the observations outside the intervals.

In order to quantitatively compare the predictive performance of the calibrated models based on AM and DREAM, we defined two metrics, a probabilistic score called CRPS and predictive coverage. The CRPS (Gneiting and Raftery, 2007) measures the difference between the cumulative distribution function (CDF) of the observed data and that of the predicted data. The lower the value of the CRPS is, the better the predictive performance. The predictive coverage measures the percent of observations that fall within a given predictive interval. A larger value of the predictive coverage suggests better predictive performance. Figure 8 shows that AM gives a CRPS value of 0.48, while the value of DREAM is 0.43. The lower value of DREAM indicates that, on average, DREAM produces tighter marginal predictive CDF that are better centered around the NEE data, suggesting its superior predictive performance to AM in terms of both accuracy and precision. In addition, the predictive coverage of DREAM is larger than that of AM, attesting once again to its superior performance in prediction.

3.6 Investigation of reliability of the algorithms

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

In this section, we investigate the influence of the proposal initialization on the computational efficiency and reliability of AM. In above analysis, the initial covariance matrix of AM was constructed based on DREAM samples *before* convergence. This setting facilitated the convergence of AM but resulted in AM false convergence to inaccurate PPDFs, leading to a relatively poor calibration and predictive performance. We implemented another AM simulation here for further examination. In this new simulation, we constructed two independent AM chains; both chains initialized $\mathbf{C}_0$ using the DREAM samples *after* convergence, but one chain only used *tsmin* samples around its left mode and *leaffall* samples around its right mode, and the other chain used *tsmin* samples around its right mode and *leaffall* samples around its left mode. Each chain evolved 3 000 000 iterations, and for the last 1 000 000 iterations the convergence diagnostic $\hat{R}$ values were calculated and shown in Fig. 9a. The figure indicates that most parameters have $\hat{R}$ less than the threshold of 1.2 except parameters *tsmin* and *leaffall*, whose values are far above 1.2 and no signs show that they are going significantly smaller in the following 1 000 000 iterations. This suggests that the two chains converged to different optima for these two parameters. We then estimated PPDFs using the last 500 000 samples from each chain respectively. The results for *tsmin* and *leaffall* are shown in Fig. 9b–e. The figures illustrate that the samples from one AM chain can only identify one mode, and this mode is consistent with the samples used to construct the initial covariance matrix $\mathbf{C}_0$.

As a single-chain sampler, it is conceptually possible for AM to become trapped in a single mode (Jeremiah et al., 2011). Consider a distribution with two far-separated modes and assume that the chain is initialized near one of the two modes (both samples initialization and proposal covariance initialization). At the beginning of the sampling, AM will explore the area around the mode where it is initialized and start identifying the first mode. Since the candidate samples generated by the Gaussian proposal have higher Metropolis ratios (Eq. 2) in the nearby area than in the far-away regions of the identified mode, the chain is hardly to move to the other mode. When the Gaussian proposal covariance matrix $\mathbf{C}_t$ begins to update, the chance of the chain jumping to the other mode depends on the relative scale of the proposal covariance and the distance between the two modes. When the modes' separation exceeds the range of the proposal, AM is less likely to escape the identified local mode.

Although the two AM chains can only simulate one of the two modes for *tsmin* and *leaffall*, the estimated PPDFs for

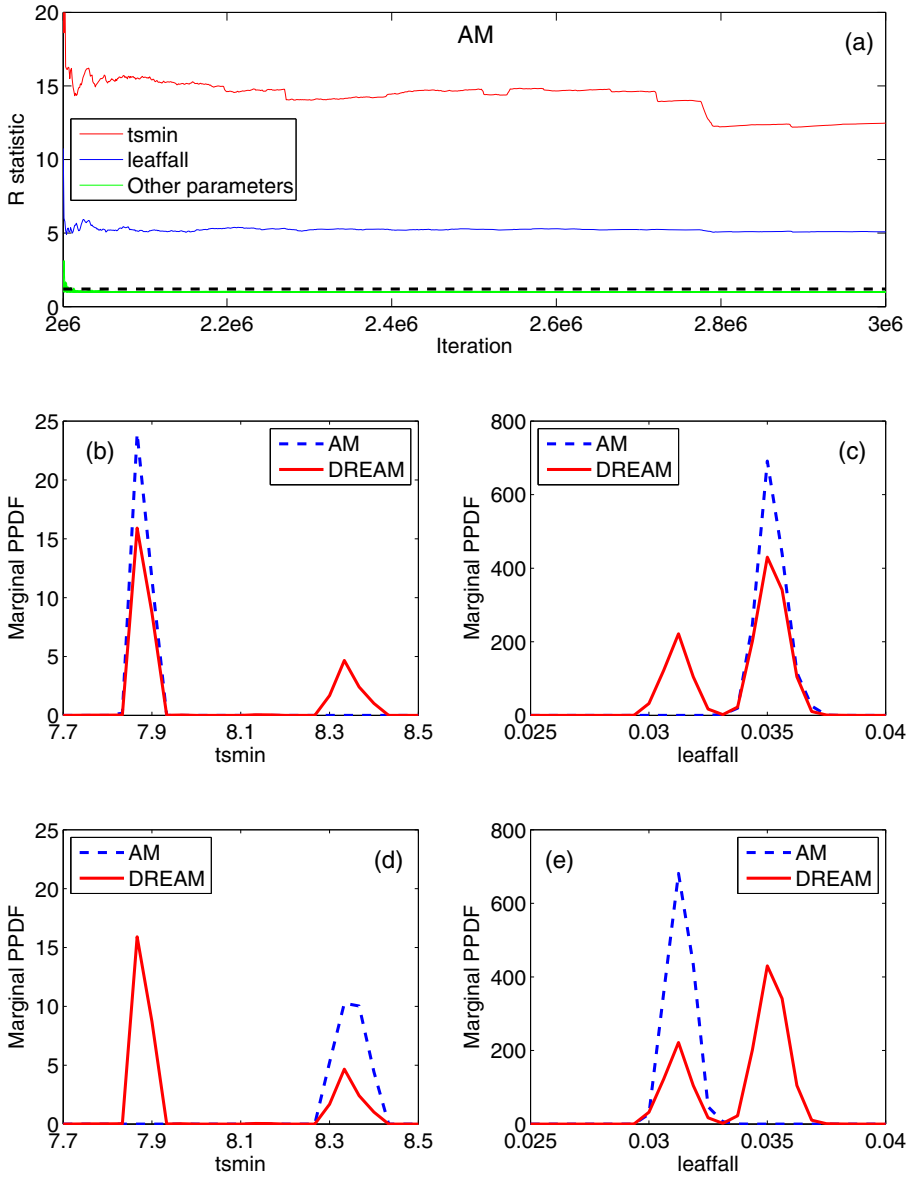

Figure 9. Results of two independent chains of AM with the initial covariance matrix constructed using the converged DREAM samples. The $\hat{R}$ statistic in **(a)** suggests that different AM chains converged to different *tsmin* and *leaffall* values. One chain captures **(b)** the left mode of *tsmin* 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.

the other 19 parameters from the two chains are close to each other and both similar to the DREAM results. This finding once again shows the reasonable existence of the two separated modes and their equivalent importance. With an improved initialization of $\mathbf{C}_0$ in the new simulation, the performance of AM also improved as it can accurately simulate unimodal PPDFs and capture one mode for the multimodal PPDFs. This investigation suggests that for AM an appropriate initialization of its Gaussian proposal has a significant impact on its performance. We made several test runs of AM, and only when we initialized $\mathbf{C}_0$ using the complete set of converged DREAM samples was AM able to produce PPDFs similar to the ones resulted from DREAM with identifying all the possible optima. However, the information of a reasonable $\mathbf{C}_0$ in practice is either unavailable or very computationally expensive to obtain.

4 Discussion

The choice of likelihood function plays an important role in the Bayesian parameter estimation, and the likelihood construction depends on the error model assumption. In this

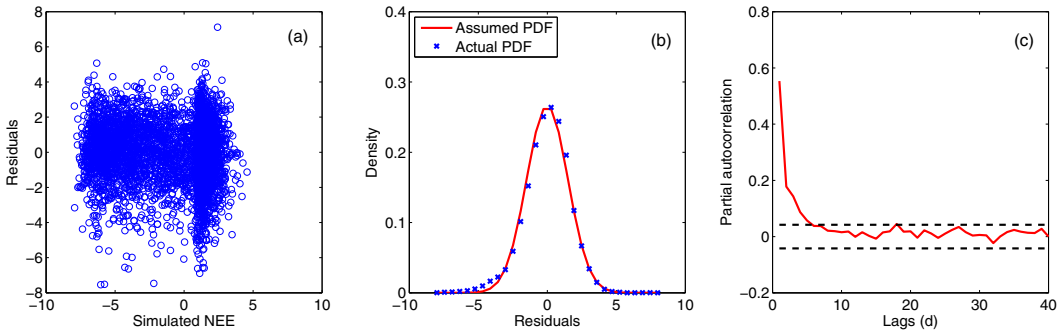

Figure 10. Residual analysis of the calibration using Gaussian likelihood with heteroscedastic and *uncorrelated* errors: **(a)** residuals vs. simulated NEE; **(b)** assumed and actual probability density functions of residuals; and **(c)** partial autocorrelation coefficients of residuals with 95 % significance levels (black dashed lines).

study, we assumed a heteroscedastic, uncorrelated, Gaussian error model. However, this simplistic assumption may not be realistic for complex TEMs. In this section, we examine whether the assumed error model provides an accurate representation of residuals between the simulated and observed NEEs. If the assumptions are not satisfied, we consider a more flexible error model and investigate the influence of the corresponding likelihood function on parameter estimation and model performance.

Figure 10 presents results of residual analysis based on the heteroscedastic, uncorrelated, Gaussian assumption. The plot of residuals versus simulated NEE in Fig. 10a justifies the assumption of heteroscedastic variances; the density plot of residuals in Fig. 10b justifies the assumption of normality, but the autocorrelation plot of residuals in Fig. 10c indicates that the errors are significantly correlated at a lag of 4, which violates the independence assumption. This violation has been reported in several time-series data models, such as the TEM in Ricciuto et al. (2008), the rainfall–runoff model in Feyen et al. (2007), and the groundwater reactive transport model in Lu et al. (2013). The correlated errors are likely to be observed in models where systematic model errors exist like the DALEC model in this study.

According to the residual analysis, we consider a heteroscedastic, *correlated*, Gaussian error model and construct the likelihood function correspondingly. Similar to Schoups and Vrugt (2010), the heteroscedasticity was explicitly accounted for using a linear model $\sigma_t = \sigma_0 + \sigma_1 E_t$, where σ_t represents the error standard deviation, σ_0 and σ_1 are parameters to be inferred from the data and E_t is the mean value of NEE. The correlation was simulated by the pth-order autoregressive model AR(p). This new error model adds six extra parameters besides the original 21 TEM parameters, where parameters σ_0 and σ_1 are related to the heteroscedastic error model and ϕ_1, ϕ_2, ϕ_3, and ϕ_4 are from the AR(4) correlation model. We set up a DREAM simulation to estimate the PPDFs of the 27 parameters and compared the results with those using the uncorrelated error assumption.

Figure 11 indicates that the six error model parameters are well identified in current parameter ranges. The heteroscedastic parameters σ_0 and σ_1 approach 1 and 0, respectively, which suggests that a constant variance may be reasonable. This finding contradicts what we usually assumed – that the data errors are heteroscedastic. The reason for this could be the epistemic error or forcing data errors. Alternatively, an extended prior distribution of σ_0 and σ_1 may give different results. More work is needed to find out the underlying reasons. The nonzero ϕ_1, ϕ_2, ϕ_3, and ϕ_4 values indicate that a AR(4) correlation model is necessary. This new heteroscedastic, correlated, Gaussian error model is appropriate as the resulted residuals demonstrate consistent features with the a priori assumptions. As it is shown in Fig. 12, the residuals are randomly distributed around the zero line (Fig. 12a), normally distributed as assumed (Fig. 12b), and no longer correlated after considering the AR(4) model (Fig. 12c).

The PPDFs of the 21 TEM parameters using the correlated Gaussian likelihood are presented in Fig. 13, associated with the results from the uncorrelated Gaussian likelihood. In comparison, we found that the two error model assumptions produced different PPDFs for most parameters. The most remarkable difference is that the bimodality of parameters *tsmin* and *leaffall* disappeared when using the correlated error assumption. As discussed in Sect. 3.4, the identified bimodality from the uncorrelated likelihood may be caused in part by the model structural error with an incomplete representation of the senescence process. The new likelihood function considers model error probabilistic structures (Lu et al., 2013) and somehow alleviates the effect of model errors on the parameter estimation, resulting in a relatively flat PPDF of *tsmin* and unimodal PPDF of *leaffall*. In addition, Fig. 13 indicates that parameter uncertainty is larger in the correlated likelihood than the uncorrelated one for most parameters, and fewer parameters are well constrained in the correlated likelihood than the uncorrelated case. For example, *rootc_init* and *litc_init* have much wider uncertainty bounds in the correlated likelihood. The synthetic

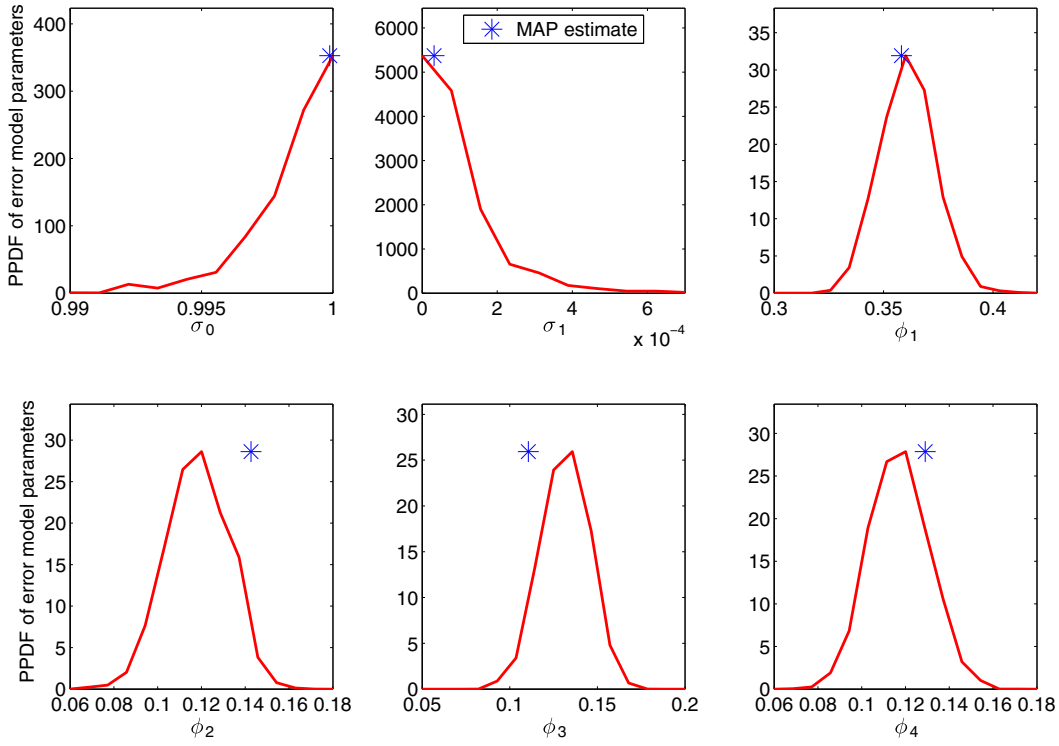

Figure 11. Estimated posterior probability density functions (PPDFs) of the six error model parameters.

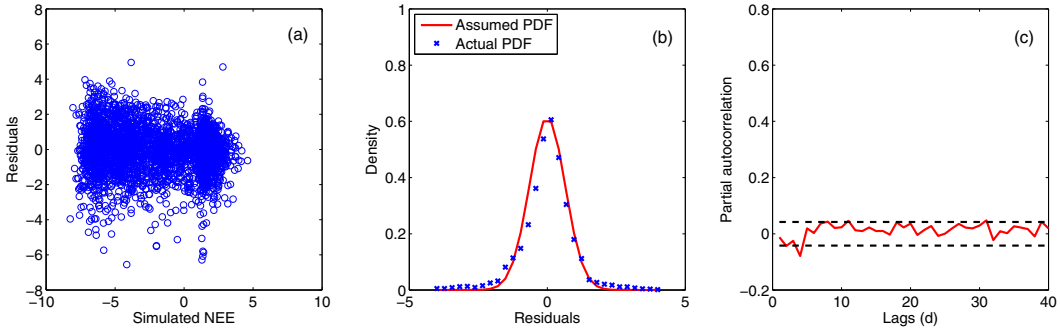

Figure 12. Residual analysis of the calibration using Gaussian likelihood with heteroscedastic and *correlated* errors: **(a)** residuals vs. simulated NEE; **(b)** assumed and actual probability density functions of residuals; and **(c)** partial autocorrelation coefficients of residuals with 95 % significance levels (black dashed lines).

study shows that these two parameters have wide confidence intervals that almost occupy the entire allowable ranges, indicating that the NEE data should provide little information about these parameters. The tight uncertainty bounds resulting from the uncorrelated error assumption are likely unrealistic and could contribute to overconfidence in model predictions. The appropriate correlated error assumption considers the error correlation that reduces the data information for calibrating parameters, thus alleviating the problem of underestimation of parameter uncertainties. The underestimation of parameter uncertainty using uncorrelated error model was also reported in Ricciuto et al. (2008), Schoups and Vrugt (2010), and Lu et al. (2013). Moreover, Fig. 13 indicates that some parameters have similar PPDFs for the two different likelihood choices, such as *gdd_min* and *q10_mr*. Those parameters that are not much affected by the model error assumptions should, in theory, be reasonably well determined in parameter estimation. And according to Safta et al. (2015), these less changed parameters are indeed sensitive parameters.

The difference in the parameter PPDFs from the two likelihood functions results in different model performance as shown in Fig. 14, where we took the simulations in October of 1995 as an example. Although the overall RMSEs are

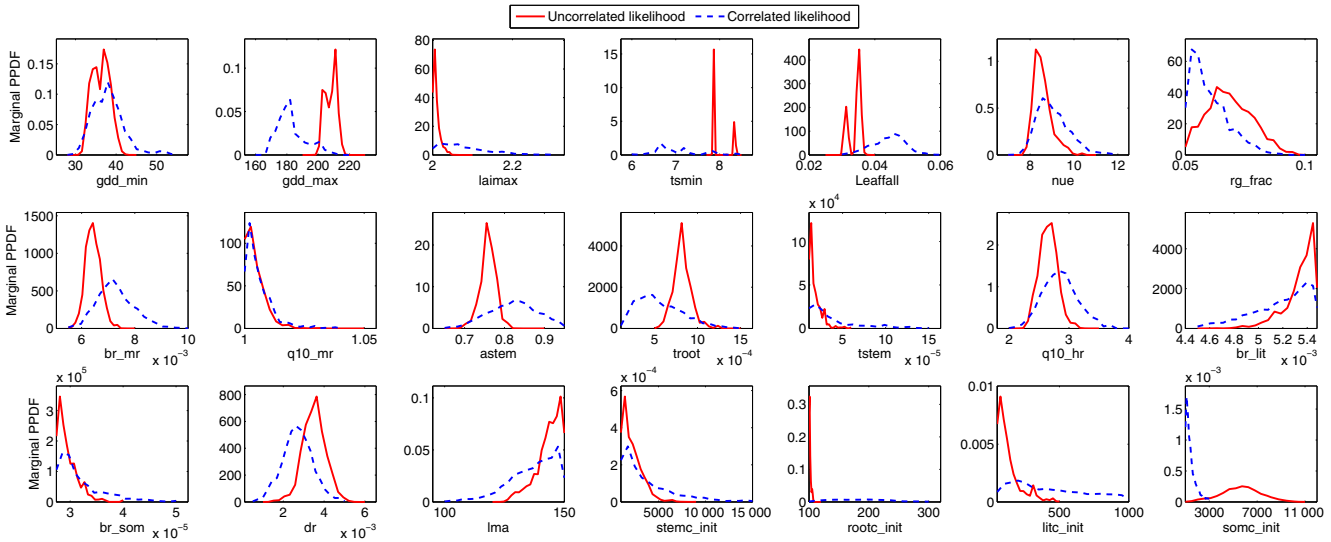

Figure 13. Estimated marginal posterior probability density functions (PPDFs) of the 21 TEM parameters using the uncorrelated and correlated Gaussian likelihoods.

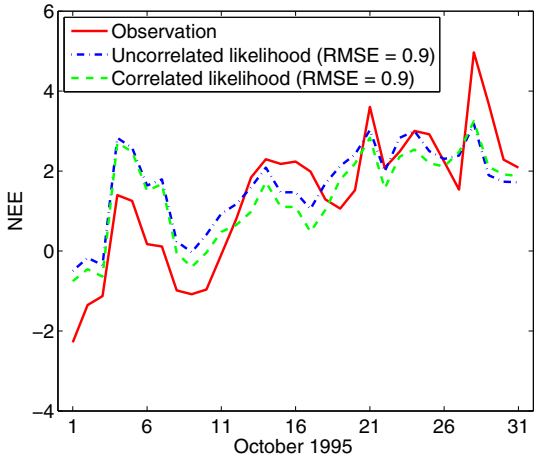

Figure 14. Simulated NEE values based on the MAP estimates from the uncorrelated and correlated Gaussian likelihoods in October 1995.

similar, the simulations on a single day are different. This is not surprising, as MCMC is a Bayesian calibration and the calibration results depend on the choice of the likelihood function, mainly the assumptions of the error model. In this study, the heteroscedastic, correlated, Gaussian error model is more reasonable than the uncorrelated one.

5 Conclusions

In this work, we apply two advanced MCMC algorithms, AM and DREAM, in the Bayesian calibration of the terrestrial ecosystem model DALEC. In both synthetic and real-data studies, we found that AM is sensitive to the algorithm initializations. When it starts with a proper initialization, through prior information or some test runs or even some dimension-reduction strategies, AM can produce reasonable approximation of the parameter posterior distributions. However, AM still shows some difficulties in sampling multimodal distributions with the Gaussian proposal. By comparison, DREAM's performance does not depend on initialization of the algorithm and can fast converge to the high-dimensional and multimodal distributions. Thus, DREAM is particularly suitable to calibrate complex terrestrial ecosystem models, where the uncertain parameter size is usually large and existence of local optima is always a concern. The application indicates that, compared to AM, DREAM can accurately simulate the posterior distributions of the model parameters, resulting in a better model fit, superior predictive performance, and perhaps identifying structural errors or process differences between the model and ecosystem from which observations were used for calibration.

In Bayesian calibration, the choice of likelihood function plays an important role in parameter estimation. In this effort, we justify the assumptions of error model used in constructing the likelihood function and find that a heteroscedastic, correlated, Gaussian error model is reasonable for this problem as supported by the residual analysis.

Data availability. The NEE observation data used in this study are available from Oak Ridge National Laboratory Distributed Active Archive Center (https://doi.org/10.3334/ORNLDAAC/1183).

Competing interests. The authors declare that they have no conflict of interest.

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

Edited by: Trevor Keenan
Reviewed by: Jasper Vrugt and one anonymous referee

References

Barr, A., Hollinger, D., and Richardson, A. D.: CO_2 flux measurement uncertainty estimates for NACP, AGU Fall Meeting, December 2009, abstract number B54A-04B, 2009.

Barr, A. G., Ricciuto, D. M., Schaefer, K., Richardson, A., Agarwal, D., Thornton, P. E., Davis, K. J., Cook, R. B., Hollinger, D. Y., van Ingen, C., Amiro, B., Andrews, A. E., Arain, M. A., Baldocchi, D., Black, T. A., Bolstad, P., Curtis, P., Desai, A., Dragoni, D., Flanagan, L., Gu, L., Katul, G., Law, B. E., Lafleur, P. M., Margolis, H., Matamala, R., Meyers, T., McCaughey, J. H., Monson, R., Munger, J. W., Oechel, W., Oren, R., Roulet, N. T., Torn, M., and Verma, S. B.: NACP Site: Tower Meteorology, Flux Observations with Uncertainty, and Ancillary Data, ORNL DAAC, Oak Ridge, Tennessee, USA, https://doi.org/10.3334/ORNLDAAC/1178, 2013.

Box, E. P. and Tiao, G. C.: Bayesian inference in statistical analysis, Wiley, New York, 588 pp., 1992.

Braswell, B. H., William, J. S., Linder, E., and Scheimel, D. S.: Estimating diurnal to annual ecosystem parameters by synthesis of a carbon flux model with eddy covariance net ecosystem exchange observations, Glob. Change Biol., 11, 335–355, 2005.

Brooks, S. P. and Gelman, A.: General methods for monitoring convergence of iterative simulations, J. Comput. Graph. Stat., 7, 434–455, 1998.

Dowd, M.: Bayesian statistical data assimilation for ecosystem models using Markov Chain Monte Carlo, J. Marine Syst., 68, 439–456, 2007.

Feyen, L., Vrugt, J. A., Nuallain, B. O., van der Knijff, J., and De Roo, A.: Parameter optimization and uncertainty assessment for large-scale stream flow forecasting, J. Hydrol., 332, 276–289, 2007.

Fox, A., Williams, M., Richardson, A. D., Cameron, D., Gove, J. H., Quaife, T., Ricciuto, D., Reichstein, M., Tomelleri, E., Trudinger, C. M., and Van Wijk, M. T.: The REFLEX project: Comparing different algorithms and implementations for the inversion of a terrestrial ecosystem model against eddy covariance data, Agric. For. Meteorol., 149, 1597–1615, 2009.

Gelman, A. and Rubin, D. B.: Inference from iterative simulation using multiple sequences, Stat. Sci. 7, 457–472 1992.

Gelman, A. and Shirley, K.: Inference from simulations and monitoring convergence, Handbook of Markov Chain Monte Carlo, CRC Press, Boca Raton, FL, 2011.

Geyer, C. J.: Practical Markov chain Monte Carlo, Stat. Sci., 7, 473–511, 1992.

Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, J. Am. Statist. Assoc., 102, 359–378, 2007.

Haario, H., Saksman, E., and Tamminen, J.: Adaptive proposal distribution for random walk Metropolis algorithm, Comput. Statist., 14, 375–395, 1999.

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

Haario, H., Saksman, E., and Tamminen, J.: Componentwise adaptation for high dimensional MCMC, Comput. Stat., 20, 265–274, 2005.

Haario, H., Laine, M., Mira, A., and Saksman, E.: DRAM: Efficient adaptive MCMC, Stat. Comput., 16, 339–354, 2006.

Hararuk, O., Xia, J., and Luo, Y.: Evaluation and improvement of a global land model against soil carbon data using a Bayesian Markov chain Monte Carlo method, J. Geophys. Res.-Biogeo., 119, 403–417, 2014.

Hill, T. C., Ryan, E., and Williams, M.: The use of CO_2 flux time series for parameter and carbon stock estimation in carbon cycle research, Glob. Change Biol., 18, 179–193, 2012.

Järvinen, H., Räisänen, P., Laine, M., Tamminen, J., Ilin, A., Oja, E., Solonen, A., and Haario, H.: Estimation of ECHAM5 climate model closure parameters with adaptive MCMC, Atmos. Chem. Phys., 10, 9993–10002, https://doi.org/10.5194/acp-10-9993-2010, 2010.

Jeremiah, E., Sisson, S., Marshall, L., Mehrotra, R., and Sharma, A.: Bayesian calibration and uncertainty analysis of hydrological models: A comparison of adaptive Metropolis and sequential Monte Carlo samplers, Water Resour. Res., 47, W07547, https://doi.org/10.1029/2010WR010217, 2011.

Keenan, T. F., Davidson, E. A., Munger, J. W., and Richardson, A. D.: Rate my data: quantifying the value of ecological data for the development of models of the terrestrial carbon cycle, Ecol. Appl., 23, 273–286, 2013.

Keenan, T. F., Gray, J., Friedl, M. A., Toomey, M., Bohrer, G., Hollinger, D. Y., Munger, J. W., O'Keefe, J., Schmid, H. P., Wing, I. S., Yang, B., and Richardson, A. D.: Net carbon uptake has increased through warming-induced changes in temperate forest phenology, Nat. Clim. Change, 4, 598–604, 2014.

Keenan, R. J., Reams, G. A., Achard, F., de Freitas, J. V., Grainger, A., and Lindquist, E.: Dynamics of global forest area: Results from the FAO global forest resources assessment 2015, Forest Ecol. Manag., 352, 9–20, 2015.

Kinlan, B. P. and Gaines, S. D.: Propagule dispersal in marine and terrestrial environments: A community perspective, Ecology, 84, 2007–2020, 2003.

Knorr, W. and Heimann M.: Uncertainties in global terrestrial biosphere modeling: 1. A comprehensive sensitivity analysis with a new photosynthesis and energy balance scheme, Global Biogeochem. Cy., 15, 207–225, 2001.

Laloy, E. and Vrugt, J. A.: High-dimensional posterior exploration of hydrologic models using multiple-try DREAM(ZS) and

high-performance computing, Water Resour. Res., 48, W01526, https://doi.org/10.1029/2011WR010608, 2012.

Leigh, M., Nihevia, N., Covich, E., and Kehn, D.: How temperature and daylength effect seasonal leaf change in honeysuckle plants, available at: http://jrscience.wcp.muohio.edu/nsfall01/labpacketArticles/Howtemperatureanddaylengt.html, 2002.

Lu, D., Ye, M., and Hill, M. C.: Analysis of regression confidence intervals and Bayesian credible intervals for uncertainty quantification, Water Resour. Res., 48, W09521, https://doi.org/10.1029/2011WR011289, 2012.

Lu, D., Ye, M., Meyer, P. D., Curtis, G. P., Shi, X., Niu, X.-F., and Yabusaki, S. B.: Effects of error covariance structure on estimation of model averaging weights and predictive performance, Water Resour. Res., 49, 6029–6047, https://doi.org/10.1002/wrcr.20441, 2013.

Lu, D., Ye, M., Hill, M. C., Poeter, E. P., and Curtis, G. P.: A computer program for uncertainty analysis integrating regression and Bayesian methods, Environ. Modell. Softw., 60, 45–56, 2014.

Lynch, S. M. and Western, B.: Bayesian posterior predictive checks for complex models, Sociol. Meth. Res., 32, 301–335, 2004.

O'Sullivan, F.: A statistical perspective on ill-posed inverse problems, Stat. Sci., 1, 502–518, 1986.

Papale, D., Reichstein, M., Aubinet, M., Canfora, E., Bernhofer, C., Kutsch, W., Longdoz, B., Rambal, S., Valentini, R., Vesala, T., and Yakir, D.: Towards a standardized processing of Net Ecosystem Exchange measured with eddy covariance technique: algorithms and uncertainty estimation, Biogeosciences, 3, 571–583, https://doi.org/10.5194/bg-3-571-2006, 2006.

Post, H., Vrugt, J. A., Fox, A., Vereecken, H., and Hendricks Franssen, H.-J.: Estimation of Community Land Model parameters for an improved assessment of net carbon fluxes at European sites, J. Geophys. Res.-Biogeo., 122, 1–29, 2017.

Price, K. V., Storn, R. M., and Lampinen, J. A.: Differential Evolution, a practical approach to global optimization, Springer, Berlin, 539 pp., 2005.

Rayner, P. J., Scholze, M., Knorr, W., Kaminski, T., Giering, R., and Widmann, H.: Two decades of terrestrial carbon fluxes from a carbon cycle data assimilation system (CCDAS), Global Biogeochem. Cy., 19, GB2026, https://doi.org/10.1029/2004GB002254, 2005.

Ricciuto, D. M., Davis, K. J., and Keller, K.: A Bayesian calibration of a simple carbon cycle model: The role of observations in estimating and reducing uncertainty, Global Biogeochem. Cy., 22, GB2030, https://doi.org/10.1029/2006GB002908, 2008.

Ricciuto, D. M., King, A. W., Dragoni, D., and Post, W. M.: Parameter and prediction uncertainty in an optimized terrestrial carbon cycle model: Effects of constraining variables and data record length, J. Geophys. Res., 116, G01033, https://doi.org/10.1029/2010JG001400, 2011.

Richardson, A., Bailey, A. S., Denny, E. G., Martin, C. W., and O'Keefe, J.: Phenology of a northern hardwood forest canopy, Glob. Change Biol., 12, 1174–1188, https://doi.org/10.1111/j.1365-2486.2006.01164.x, 2006.

Richardson, A. D., Williams, M., Hollinger, D. Y., Moore, D. J. P., Dail, D. B., Davidson, E. A., Scott, N. A., Evans, R. S., Hughes, H., Lee, J. T., Rodrigues, C., and Savage, K.: Estimating parameters of a forest ecosystem C model with measurements of stocks and fluxes as joint constraints, Oecologia, 164, 25–40, 2010.

Robert, C. and Casella, G.: Monte Carlo statistical method, 2nd Edn., Springer, 645 pp., 2004.

Safta, C., Ricciuto, D. M., Sargsyan, K., Debusschere, B., Najm, H. N., Williams, M., and Thornton, P. E.: Global sensitivity analysis, probabilistic calibration, and predictive assessment for the data assimilation linked ecosystem carbon model, Geosci. Model Dev., 8, 1899–1918, https://doi.org/10.5194/gmd-8-1899-2015, 2015.

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

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

Schoups, G. and Vrugt, J. A.: A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic and non-gaussian errors, Water Resour. Res., 46, W10531, https://doi.org/10.1029/2009WR008933, 2010.

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

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

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

Thibault, K. M., White, E. P., Hurlbert, A. H., and Morgan Ernest S. K.: Multimodality in the individual size distributions of bird communities, Global Ecol. Biogeogr., 20, 145–153, 2011.

Tierney, L.: Markov chains for exploring posterior distributions, Ann. Stat., 22, 1701–1728, 1994.

Trudinger, C. M., Raupach, M. R., Rayner, P. J., Kattge, J., Liu, Q., Pak, B., Reichstein, M., Renzullo, L., Richardson, A. D., Roxburgh, S. H., Styles, J., Wang, Y. P., Briggs, P., Barrett, D., and Nikolova, S.: OptIC project: An intercomparison of optimization techniques for parameter estimation in terrestrial biogeochemical models, J. Geophys. Res., 112, G02027, https://doi.org/10.1029/2006JG000367, 2007.

Urbanski, S., Barford, C., Wofsy, S. C., Kucharik, C., Pyle, E., Budney, J., McKain, K., Fitzjarrald, D., Czikowsky, M., and Munger, J. W.: Factors controlling CO_2 exchange on timescales from hourly to decadal at Harvard Forest, J. Geophys. Res.-Biogeo., 112, 1–25, 2007.

Weng, E. and Luo, Y.: Relative information contributions of model vs. data to short- and long-term forecasts of forest carbon dynamics, Ecol. Appl., 21, 1490–1505, 2011.

White, M. A., Thornton, P. E., Running, S. W., and Nemani, R. R.: Parameterization and sensitivity analysis of the BIOME-BGC terrestrial ecosystem model: net primary production controls, Earth Interact., 4, 1–85, 2000.

Williams, M., Rastetter, E. B., Fernandes, D. N., Goulden, M. L., Wofsy, S. C., Shaver, G. R., Melillo, J. M., Munger, J. W., Fan, S.-M., and Nadelhoffer, K. J.: Modelling the soil-plant-atmosphere continuum in a Quercus' Acer stand at Harvard Forest: the regulation of stomatal conductance by light, nitrogen and soil/plant hydraulic properties, Plant Cell Environ., 19, 911–927, 1996.

Williams, M., Schwarz, P. A., Law, B. E., Irvine, J., and Kurpius, M. R.: An improved analysis of forest carbon dynamics using data assimilation, Glob. Change Biol., 11, 89–105, 2005.

Vrugt, J. A.: Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation, Environ. Model. Softw., 75, 273–316, 2016.

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

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, Int. J. Nonlin. Sci. Num., 10, 271–288, 2009a.

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

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, 6871–6892, https://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. Cy., 26, GB3025, https://doi.org/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.