Dear Associate Editor Trevor Keenan,

We are submitting the revised manuscript (bg-2017-41) with the title "Bayesian calibration of terrestrial ecosystem models: A study of advanced Markov chain Monte Carlo methods". We appreciate the constructive comments and suggestions from you and the reviewers.

The manuscript has been substantially revised. The major revisions are summarized below:

(1) The abstract and conclusion have been substantially rewritten to emphasize the significance and contribution of our work, and make the subject of our paper more clearly and interesting to the Biogeosciences audience.
(2) The Methodology Section 2 has been substantially reduced by providing related references.
(3) The original Section 3 with three functional test examples has been entirely removed; instead, we wrote a new Section 2.5 to summarize and discuss previous work of AM and DREAM performance.
(4) A synthetic study of DALEC model has been added in Section 3.3.
(5) A new Section 4 has been added to discuss the influence of likelihood function on parameter estimation and model performance.
(6) More figures related to the new sections have been added.
(7) More references have been included to represent a thorough literature review.

We believe that this work will provide ecological practitioners with valuable information on model calibration and understanding of the terrestrial ecosystem models. We hope that the revised manuscript has met the requirements for publication in Biogeosciences. Should you have any questions regarding the manuscript, please contact me at lud1@ornl.gov.

Yours,

Dan Lu

Replies to Comments on Manuscript bg-2017-41

We would like to express our sincere gratitude to Dr. Trevor Keenan, Prof. Jasper Vrugt, and the other anonymous reviewer for their insightful and constructive comments and suggestions. All comments have been addressed below and considered in the revised manuscript as highlighted in red.

Associate Editor's Evaluation

General Comments:

Comment 1:

Thank you for your thoughtful response to the referees' comments. Their suggestions will no doubt greatly improve your manuscript, potentially making it suitable for publication in Biogeosciences. It is not clear to me whether your response to the comments of Marko Laine will satisfy their concerns, so I will be asking the reviewers to have a look at your revised manuscript once submitted.

Response:

We agree with the associate editor that reviewers' suggestions have greatly improved our manuscript, and we believe that our manuscript is more suitable for the Biogeosciences audience now.

We submitted our responses to the comments of Marko Laine and addressed their comments item by item in detail. Moreover, the reviewer #1, Jasper Vrugt, also had a response to Marko Laine. The reviewers can look up the whole comment-response exchange online if they desire. On the other hand, if there are additional concerns from the reviewers we will be happy to address these. The major concern from Marko Laine is that we did not code AM algorithm right and may get wrong results. To address his concern, we used the AM code provided by Marko Laine and attached the algorithm setup in the response file. By using his code and the described algorithm setup, we (and the reviewers, if they desire) can get the similar results with those presented in the manuscript. This indicates that our application of the AM algorithm is appropriate and correct, and the AM results are trustable.

The comments and suggestions from Marko Laine are also considered in the revised manuscript. We are happy to address any further comments on this work.

Comment 2:

In addition, I am not convinced by your response to comment 4 from referee 2, regarding your claim that DREAM and AM can constrain the majority of parameters in DALEC. I agree with the reviewer that "simply saying the model is simple enough/doesn't have many parameters is insufficient.", yet in your response you argue that "The DALEC model used in this study is a simple model with considering only six processes and five carbon pools, and all the 21 parameters were shown to be

sensitive to the NEE data, despite that some are more sensitive than others (Safta et al, 2015)." But the DALEC model has been shown in multiple studies to be poorly constrained by NEE data alone. See for example the Fox et al. REFLEX paper cited in your manuscript, which also employs the AM approach that you use. And practically all other DALEC papers, of which there are many. The only study to my knowledge that found good constraints was Bloom and Williams, 2015 Biogeosciences, 12, 1299–1315, but they needed to impose additional 'common sense' constraints to do so. You could do a better job of citing the relevant literature regarding DALEC applications, and putting your results in the context of previous studies. You conclude that "Therefore, it is not surprising that both AM and DREAM algorithms can constrain the parameters pretty well." but given that all other studies that use DALEC and EC data alone show poor parameter constraint I find it very surprising indeed. The argument that it is due to low observational uncertainty does not make sense, as previous studies should have used the same NEE uncertainty. Unless of course you have mischaracterized the uncertainty? I really feel you need to get to the bottom of this, as it is a striking difference from most/all previous studies and the implications of such good constraint are large.

Response:

We appreciate the Associate Editor for his insightful comments and suggestions. However, we respectfully disagree with the statement that all other studies using DALEC and NEE data alone show poor parameter constraint. We find studies that use similar methods and constraining data generally yield similar results in terms of parameter constraints. If we define a well-constrained parameter as the posterior distribution occupying at most half the range of the prior distribution (Keenan et al, 2013), a number of previous studies (e.g., Fox et al., 2009 and Hill et al., 2012) showed that the majority of parameters in DALEC can be constrained by NEE alone. Fox et al. (2009) showed in their Fig. 2 (focusing in particular on participant M1, who defined a similar discrepancy between model and data to construct the likelihood and used a similar MCMC method as our study) that only one of 17 parameters (parameter F_{ll}) was poorly constrained; the other 16 parameters were well constrained according to the above definition. Similarly, Hill et al. (2012) showed in their Fig. 3 that only 3 of the 23 calibrated parameters were poorly constrained; the other 20 parameters have their 90% confidence intervals occupy much less than half of the prior ranges when the DALEC was calibrated against 10 years daily NEE.

Methodological differences may account for differences among studies in how well parameters are constrained with NEE data. Although Richardson et al. (2010) showed in their Fig. 5 that almost all the plotted 16 parameters were poorly constrained when the DALEC was calibrated against daytime NEE only, the comparison between their results and ours is not a direct one. Richardson et al. (2010) defined a different likelihood in the MCMC simulation, used a different acceptance rule in the Metropolis algorithm, and estimated the parameter posterior distribution in a different way. From both Fox et al. (2009) and our study, we can see that different algorithms have a substantial influence on the parameter estimation results.

To further justify our parameter optimization, as suggested by referee #2 we implemented a synthetic study where the 14 years daily NEE data were generated based on the given nominal parameter values and then perturbed with normally distributed random noise with mean at zeros and the standard deviation consistent with the measured NEE. In this synthetic case, if the parameter optimization is reasonable, the estimated parameter posterior distributions should enclose the known parameter values within their 90% confidence intervals; and the sum of squared weighted residuals (SSWR) for the optimal parameter estimate 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). We implemented both the AM and DREAM algorithms in the synthetic case in the same way as in the real-data study. The results are shown in the following Figure 1. The figure indicates that all the known parameter values can be enclosed in the estimated posterior distributions very well and are located in the high-probability regions. Moreover, the SSWR value corresponding to the optimal parameter estimate is about 5044, which is close to the mean value 5093 of the chi-squared distribution. This suggests that our parameter estimation is accurate and the results are reasonable.

In addition, Figure 1 shows that about half of the parameters are well constrained according to the same definition as above, and the other half of the parameters have relatively large uncertainty bounds. This synthetic case gives different results from the real-data calibration shown in the following Figure 2. The observation is consistent with the study of Fox et al. (2009) where the measured NEE is an apparently better constraint on the parameters than the synthetic data. Moreover, the SSWR value corresponding to the optimal parameter estimate in the real-data study is about 13157, much larger than the value of 5044 in the synthetic case. One contribution to this difference in results between real and synthetic studies could be model structural error. The real data inevitably include contributions from processes that are either missing or poorly represented in DALEC. A close inspection of Figure 2 shows that 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 carbon pool initial condition parameters, *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 summary, other studies using DALEC and NEE data alone often do show well-constrained parameters (e.g., Fox et al., 2009 and Hill et al., 2012), especially those that use similar methods as our study. There are also some other studies (e.g., Richardson et al., 2010), while using DALEC and similar NEE data, implementing different calibration methods (e.g., likelihood constructions and MCMC algorithms) and a different choice of which parameters to calibrate from our work. Such studies have shown that NEE alone is a relatively poor constraint compared to combined data sources in the context of multi-objective calibration. We do not disagree with that general conclusion, but we note the impact of methodology may be quite large and this has not been extensively studied for multi-objective cases. Our primary purpose is to demonstrate the utility of the DREAM algorithm and show how it could improve upon AM methods, the general results of which we believe to be consistent with past MCMC studies that used only NEE as a constraint. Realizing these differences is important to make a fair comparison, as we can see in our synthetic and real-data calibration cases and from the different results in Fox et al. (2009) when multiple methods were compared.

The above points have been included in the revised manuscript. Besides, we explored possible reasons causing the well-constrained parameters, as highlighted in Sections 3.3 and 3.4 in the revised manuscript.

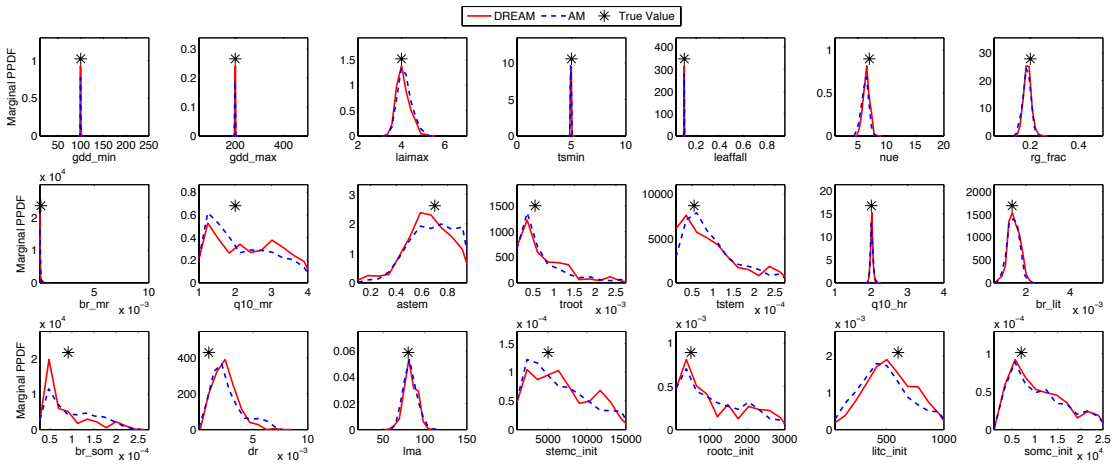


Figure 1. The posterior distributions of 21 DALEC parameters calibrated against the synthetic data generated based on the known parameter values. The limits of the x-axis show the uniform prior distribution ranges.

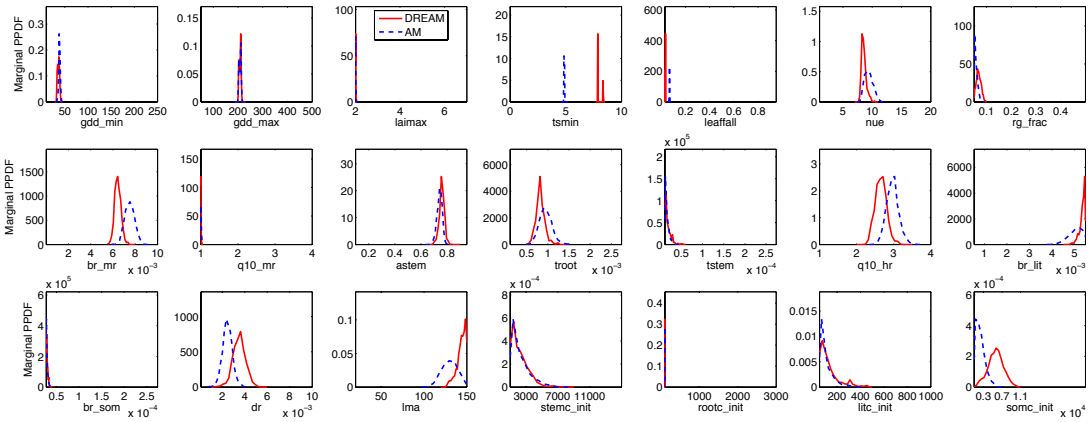


Figure 2. The posterior distributions of 21 DALEC parameters calibrated against the real measurement data, where the limits of the x-axis show the uniform prior distribution ranges.

Comment 3:

I think you might have the wrong reference for Keenan et al (2015). Should be Keenan et al. (2014)? Keenan, T. F., J. Gray, M. A. Friedl, M. Toomey, G. Bohrer, D. Y. Hollinger, J. W. Munger, J. O'Keefe, H. P. Schmid, I. S. Wing, B. Yang, and A. D. Richardson. 2014. Net carbon uptake has increased through warming-induced changes in temperate forest phenology. Nature Climate Change 4: 598–604.

Response:

The citation of Keenan et al (2015) in the manuscript is from reference: 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 Ecology and Management, 352, 9-20, 2015.

In the revised manuscript, the provided reference of Keenan et al. (2014) has also been cited.

Referee #1's Evaluation

General Comment:

This paper advocates the use of Bayesian inference to estimate the parameters of the data assimilation linked ecosystem carbon (DALEC) model. The proposed approach builds on the DREAM algorithm and uses a 14-year data record of daily net ecosystem exchange observations collected at the Harvard Forest Environmental Measurement Site. The DREAM parameter distributions are compared against those obtained using another MCMC method, namely the Adaptive Metropolis (AM) sampler. Results demonstrate a superior performance of DREAM with DALEC parameter estimates that outperform their AM derived counterparts during an independent evaluation period.

The paper is generally well-written and discusses an important topic in ecosystem modeling.

Response:

We appreciate Dr. Vrugt for the concise and nice summary and positive assessment of the manuscript.

Specific Comments:

Comment 1:

Can you not estimate C_0 from the prior parameter ranges? Just create some samples in this space, as DREAM does, and then compute C_0 = cov of these samples? Not to say that C_0 is correctly scaled this way. But it makes comparison with DREAM more fair. If you use prior information to construct C_0 then you should also use this for DREAM to sample the initial states of the chains.

Response:

We thank Dr. Vrugt for the constructive suggestions. We randomly drew 100,000 samples from the parameter space based on which the initial covariance matrix $\mathbf{C}_0$ was computed. According to the parameter ranges listed in Table 1 of the manuscript, the $\mathbf{C}_0$ has relatively large and heteroscedastic variances. For this high-dimensional problem with 21 parameters, the calculated $\mathbf{C}_0$ can easily cause the proposed samples of AM outside the parameter space, resulting in a low acceptance rate. We tried several independent AM runs with the calculated $\mathbf{C}_0$ and they all ended up with a single location in the parameter space with little movement.

Safta et al. (2015) applied AM to the DALEC model and got similar problems. Their strategy was applying the AM gradually, starting with a group of parameters (say 7 parameters) and setting the rest of parameters as constants at their nominal values. The initial proposal covariance for the small group of parameters had variances about 1/16 of the variances of the corresponding normal priors. After several MCMC iterations, the samples were used to compute the $\mathbf{C}_0$. Next, they moved to the second stage to additionally consider another group of parameters. Their iterative process

broke the original high-dimensional problem into a sequence of steps of increasing dimensionality, with each intermediate step starting with a good proposal covariance, finally got AM converged in 6,000,000 MCMC iterations.

We agree with Dr. Vrugt that choosing $\mathbf{C}_0$ specially does not make a fair comparison with DREAM. This has been pointed out in the revised manuscript as "*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.*"

Comment 2:

No surprise that single-site Metropolis does not work well in case of correlated parameters - as correlated dimensions have to be updated together. These arguments have been made in previous DREAM related papers.

Response:

We appreciate and respect the significant efforts that Dr. Vrugt contributes to the development of DREAM algorithms. The related references have been added.

Comment 3:

Page 11: The authors refer to the univariate R_statistic to monitor convergence of the sampled chains. Indeed, this approach is often used in multi-chain methods such as DREAM. Nevertheless, I recommend the authors to look into the multivariate R_stat of Brooks and Gelman. This statistic does not compare parameters one at a time (their between and within-chain variance) but rather assesses the entire posterior distribution. This multivariate R_stat is a single convergence diagnostic and will suggest convergence of the sampled chains at a later time than the univariate R_stat of the parameters. The latest DREAM toolbox in MATLAB returns the multivariate R_stat.

Response:

We appreciate Dr. Vrugt for the suggestions and references. The multivariate R_statistic has been used for convergence diagnostic in the revised manuscript and copied below as Figure 1. As Dr. Vrugt pointed out and we observed in Figure 1, the multivariate R_statistic indeed suggests convergence of the chains at a later time than the univariate R_statistic. Both R statistics indicate that it is reasonable to use the last 50,000 samples of DREAM chains for estimating the posterior distribution.

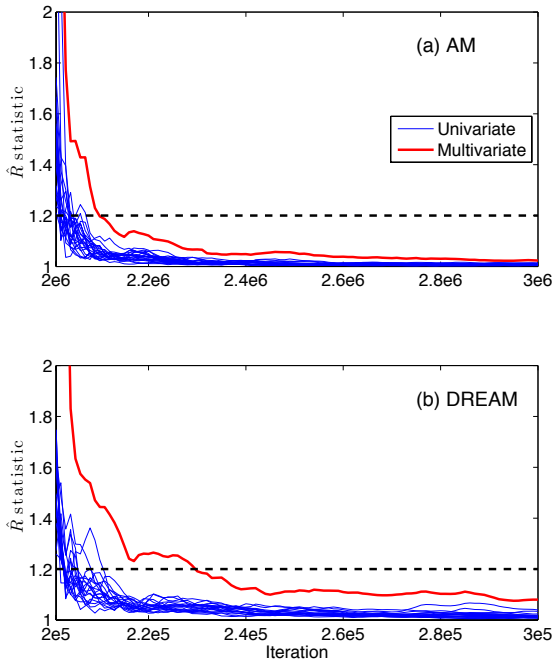


Figure 1. Univariate and multivariate Gelman Rubin $\hat{R}$ statistics (a) for the last 1,000,000 iterations from ten independent AM runs and (b) for the last 100,000 iterations from the DREAM simulation using ten chains. The values less than the threshold of 1.2 suggest chain convergence.

Comment 4:

Section 2.4: This section on DE-MC/DREAM has many similarities with published work; for instance, Vrugt (2016). Similar argumentation. I am not sure whether the authors should repeat all this or that a citation to this DREAM manual paper suffices at some places.

Response:

We thank Dr. Vrugt for the suggestion. Section 2.4 has been substantially reduced in the revised manuscript by providing related references.

Comment 5:

Case study 1: This study is a standard study that has been used in the DREAM literature. I think the authors should reflect this in their writing. They made some adaptations (50d, variance/covariance matrix of target), nevertheless, this type of study has been published before to illustrate DREAM and AM performance. I think the authors should properly discuss related examples in previous publications. As the authors seem to be very familiar with the DREAM body of work I do not think it is necessary that I provide references here. For example, Laloy and Vrugt (2012) do what the authors present in Figure 3 but then in substantially higher dimensions.

Response:

We thank Dr. Vrugt for the suggestions and references. In the revised manuscript, the section presenting the three functional tests has been entirely deleted; instead, we wrote a new section called strategies and capabilities of AM and DREAM in sampling complex problems. In this new section, we summarize previous work of AM and DREAM performance. For example, we discuss the comparison results between AM and DREAM in Figure 5 of Vrugt (2016) and the capability of DREAM in sampling 25D trimodal distribution in Figure 3 of Laloy and Vrugt (2012).

Comment 6:

Overall the benchmark case studies illustrate performance of DREAM but similar studies have appeared in many other papers - not sure if they are needed in this work. Reference to those previous studies might suffice. This includes work in different fields, including the present field of application: biogeosciences.

Response:

We thank Dr. Vrugt for the suggestion. In the revised manuscript, the section presenting the three functional tests has been entirely deleted; instead, we wrote a new section to summarize previous work of AM and DREAM performance. Please see our response to above Comment 5.

Comment 7:

Make sure that the math notation in your figures (and labels) matches exactly symbols used in text. This is not the case presently, for instance, Figure 2, x_1 –> x should be italic. Fig. 4: R_statistic –> \hat{R} as in text, etc.

Response:

We thank Dr. Vrugt for the suggestion of rigorous notations. Now the math notations are consistent in the figures and texts. One example is the y-axis label of Figure 1 shown above.

Comment 8:

The paper is technical - the main theme of this paper is a comparison of two different MCMC methods. This comparison is clear and results are fine. Yet, personally I would appreciate a little bit more focus on what we actually learn from using methods such as DREAM. For example a) the authors assume a Gaussian likelihood. We know that such likelihood function is often too simplistic, that is, the assumptions of normality, independence, and constant variance of the residuals can often not be justified. Indeed, a reader might wonder what the impact of these assumptions is on the final parameters and model behavior (behavior during evaluation period) b) The authors do not investigate the residual properties. Do they satisfy the residual

assumptions made? For instance, a plot of residuals versus NEE (constant variance justified?), histogram of residuals (Gaussian?) and autocorrelation plot of residuals (no serial correlation?). c) Without an adequate check of the residuals we cannot conclude whether the parameters of DALEC are "correctly" estimated. Maybe a Gaussian likelihood is appropriate for the model and data at hand. I would suspect that a more flexible likelihood function, with nuisance variables, would be more appropriate. This would allow a better representation of the residual properties (tails, skew, nonnormality, heteroscedastic variance, etc.). d) With the use of a more complex likelihood function the bimodality of DALEC parameter tsmin might disappear. This is interesting by itself. I do suspect though that the performance of the AM algorithm will further deteriorate (in comparison to DREAM) if a likelihood function is used with nuisance variables; for example the generalized likelihood function of Schoups and Vrugt. This is part of the DREAM toolbox (MATLAB) and DREAM Suite (Windows).

Response:

These are excellent comments. We appreciate Dr. Vrugt for the suggestions.

a) In the revised manuscript, we added a new Section 4 to investigate the assumptions of the heteroscedastic, uncorrelated, Gaussian error model. In addition, we considered a more general likelihood function and discussed the influence of error model assumptions on parameter estimation and model performance.

b) In the original manuscript, we assumed a heteroscedastic, uncorrelated, Gaussian error model. After model calibration, we analyzed the residuals in the revised manuscript and found that the residuals of the 5114 NEE data follow a Gaussian distribution but are correlated. As shown in Figure 2 below, the plot of residuals versus simulated NEE in Figure 2(a) justifies the assumption of heteroscedastic variances; the density plot of residuals in Figure 2(b) justifies the assumption of normality; but the autocorrelation plot of residuals in Figure 2(c) indicates that the residuals are correlated that violates the independence assumption.

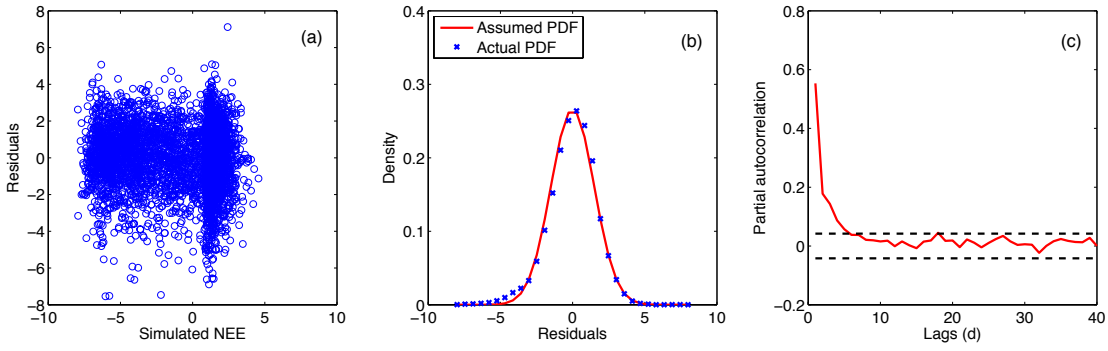


Figure 2. Residual analysis based on the heteroscedastic, uncorrelated, Gaussian error model.

c) The residual analysis suggests that a heteroscedastic, correlated, Gaussian error model would be appropriate. In the revised manuscript, we considered this error model to construct the likelihood functions. Based on 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), and as suggested by Figure 2(c) the p value can be up to 4. This new error model added six extra parameters besides the original 21 TEM parameters, where 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 run to estimate the PPDFs of the 27 parameters and compared the results with those presented in the original manuscript using the uncorrelated error assumption.

d) Figure 3 below indicates that the six error model parameters are well identified. The heteroscedastic parameters σ_0 and σ_1 approach 1 and 0, respectively, which suggests that a constant variance may be reasonable. The nonzero ϕ_1, ϕ_2, ϕ_3, and ϕ_4 values indicate that a AR(4) correlation model is necessary. Figure 4 below indicates that the new heteroscedastic, correlated, Gaussian error model is reasonable where all the a priori assumptions are justified. As it can be seen, the resulted residuals are randomly distributed around the zero line (Figure 4(a)), normally distributed as assumed (Figure 4(b)), and no longer correlated after considering the AR(4) model.

Figure 5 below presents PPDFs of the 21 TEM parameters using the uncorrelated and correlated Gaussian likelihoods. We found that the two error model assumptions produce 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 the original manuscript, 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 *tsmi*n and uni-modal PPDF of *leaffall*. In addition, Figure 5 indicates that parameter uncertainty is larger in the correlated likelihood than the uncorrelated likelihood for most parameters. The reason can be that accounting for error correlation reduces the data information for calibrating parameters. 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).

The difference in the parameter PPDFs from the two likelihood functions results in different model performance as shown in Figure 6 below where we took the simulations in October of 1995 as an example. Although the overall root mean squared errors are 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.

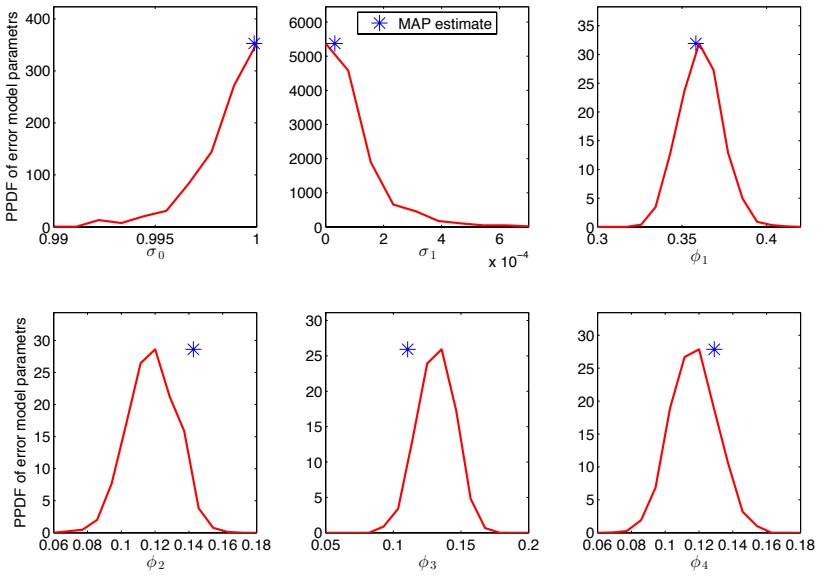


Figure 3. Estimated posterior probability density functions (PPDFs) of the six error model parameters.

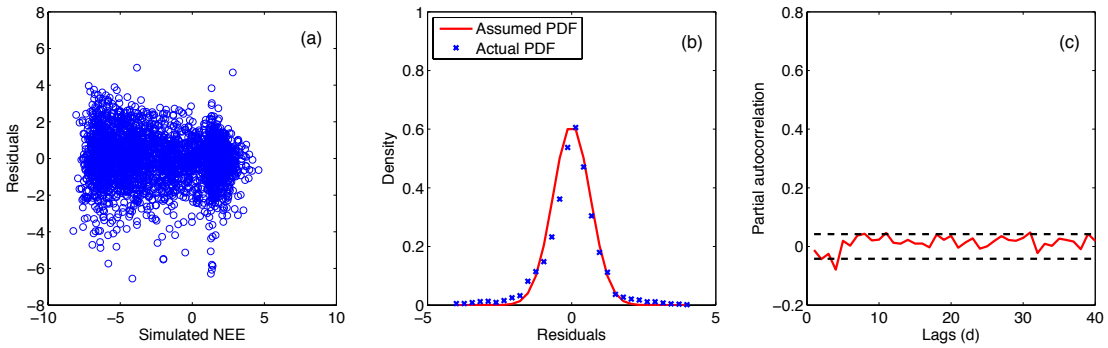


Figure 4. Residual analysis based on the heteroscedastic, correlated, Gaussian error model.

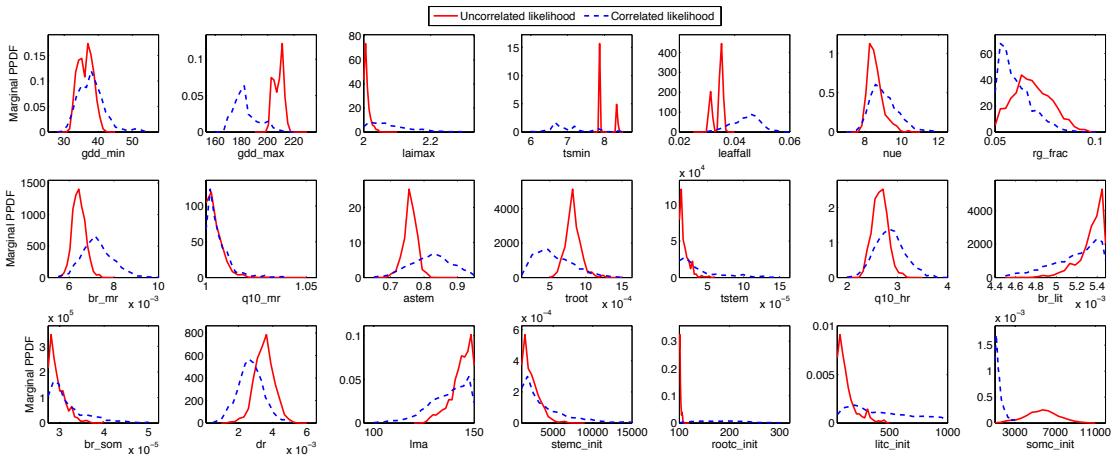


Figure 5. Estimated marginal posterior probability density functions (PPDFs) of the 21 TEM parameters using the uncorrelated and correlated Gaussian likelihoods.

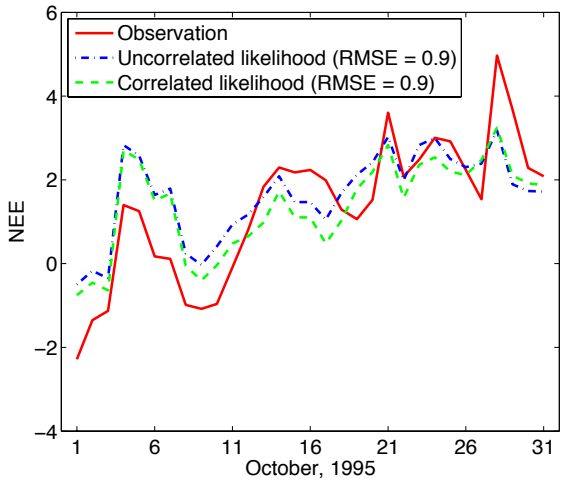


Figure 6. Simulated NEE values based on the MAP estimates from the uncorrelated and correlated Gaussian likelihoods in October 1995.

Comment 9:

Indeed, I think some focus on the choice of likelihood function, and the properties of the residuals would significantly enhance this paper without too much additional work. Otherwise, the paper is merely an important demonstration for the need of robust MCMC methods in ecosystem modeling; simpler methods might get stuck in local minima. This is an important message for the ecosystem modeling community, yet similar studies/messages have appeared elsewhere, in other journals using different Earth system models.

Response:

We appreciate Dr. Vrugt for the suggestions. The residual analysis has been added in the revised manuscript. In addition, we considered a more general likelihood function and discussed its influence on parameter estimation and model performance. Please see our responses to the above Comment 8 for details.

Comment 10:

Note: Figure 8 is very nice. An excellent demonstration of the effect of inadequate inference of AM and consequence of bimodality.

Response:

We thank Dr. Vrugt for the positive assessment.

Comment 11:

A few editorial suggestions Line 143: ...at similar sites...? Line 144: In the absence of prior information, ... Equation (2) –> min should not be italicized. Line 180 –> many studies have demonstrated this - way before Lu et al. (2014). In fact, this is

justification why better MCMC methods have been developed in past two decades. Line 189 –> covariance matrix, C_t, should be bold. It is a matrix of size d x d, where d is number of elements of x, the parameters to be estimated Equation (3) –> C should be bold, and function Cov as well. Also no need to place s_d in front of e*I_d, as last term is just for small perturbation to avoid singularity of C_t Line 328: x_1 –> x should be italic. Please carefully check your math notation. scalars italic, vectors lower case bold, matrices, upper case bold.

Response:

We thank Dr. Vrugt for the suggestions. All the editorial typos have been corrected.

Comment 12:

Altogether, I would recommend a major revision. Comments should be relatively easy to address - but will require more work (investigate residual assumptions) and DALEC simulations (to test another likelihood function).

Response:

We appreciate Dr. Vrugt for the excellent suggestions and comments. The comments have been addressed item by item above in this response and all the suggestions have been considered in the revised manuscript.

Referee #2's Evaluation

General Comments:

Comment 1:

This paper describes the testing of the performance of the Differential Evolution Adaptive Metropolis (DREAM) MCMC algorithm versus the Adaptive Metropolis (AM) algorithm in two benchmarking exercises and with the Data Assimilation Linked Ecosystem Carbon (DALEC) model using Harvard Forest flux tower data.

The manuscript is clear and well written, and highlighting the good performance of the DREAM algorithm is of interest to others addressing the issues associated with parameterizing ecosystem models.

Response:

We appreciate the reviewer for the concise and nice summary and positive assessment of the manuscript.

Comment 2:

This is a highly technical manuscript, detailing the implementation of two algorithms, and I note the interactive comments of Vrugt and Laine, both far more qualified than me to assess the technical aspects of this study. Therefore, I'll concentrate on my concern that this manuscript is too technical, or at least focused in the wrong area, for the scope of Biogeosciences.

For publication here, I would suggest some major revisions are required, shifting the focus of the manuscript to make it more relevant to this audience. This would involve: (i) relying more on referencing previous work when discussing the technicalities of the algorithms and their implementation and testing; (ii) bringing in an observing system simulation experiment (OSSE) approach; and (iii) concentrating more model and ecological insights these implementation of DREAM/AM and DALEC provide.

Response:

We thank the reviewer for the great suggestions. The manuscript has been substantially revised to be interested to the general audience of Biogeosciences. In summary:

(i) We have substantially reduced the technical Section 2 and completely removed the functional comparison study in Section 3 by providing related references and discussing previous work.

(ii) The OSSE study with known parameters and pseudo data has been added in Section 3.3 of the revised manuscript. Results indicate that (1) our parameter estimation is accurate and reasonable as the approximated parameters posterior distributions enclose the known values very well, and (2) the problem can be constrained by the NEE data along as about half parameters are well identified. For more information about this study, please see our responses to the specific comments

below and the discussion in Section 3.3 of the revised manuscript.

(iii) More discussion about the model and ecological insights has been added in the revised manuscript. For example, we added the OSSE study and did not observe the bimodality of the posterior distributions, which provides a good evidence to support the statement that the bimodality is partially caused by model structural uncertainty (i.e., incomplete representation of the senescence process). In addition, we added the residual analysis and investigated the impact of observation error assumptions on the parameter estimation and model performance. More details please see our responses to the specific comments below and the corresponding discussion in the revised manuscript.

Specific Comments:

Comment 1:

Given the large literature and other information there is already available describing DREAM, and the DE-MC Section 2.4 is overly long, and repetitive of much existing work.

Response:

We thank the reviewer for the suggestion. The Section 2.4 has been substantially reduced in the revised manuscript by providing related references and discussing previous work.

Comment 2:

For the benchmarking exercises described in Sections 3, similar tests have been carried out in the extensive existing literature on both DREAM and AM, and it doesn't seem that further benchmarking like this is relevant to the Biogeosciences audience.

Response:

We thank the reviewer for the suggestion. The original Section 3 has been entirely removed in the revised manuscript; instead, we wrote a new section called strategies and capabilities of AM and DREAM in sampling complex problems. In this new section, we briefly summarized previous work of AM and DREAM performance.

Comment 3:

Section 4, the application of the MCMC algorithms to an ecosystem model seems to be more pertinent. Given the nature of the comparison between algorithms, I would perhaps prefer to see an OSSE-type experiment using the model with known parameters to generate pseudo-observations with realistic uncertainties that are then used to try estimate the (known) values, rather than the more standard benchmarks described in Section 3.

Response:

We appreciate the reviewer for the constructive suggestion. A synthetic study with known parameters and pseudo-observations has been added in Section 3.3 of the revised manuscript. The estimation results based on both AM and DREAM are presented in following Figure 1. As we can see, (1) the estimated parameter PPDFs can enclose the known (true) values very well; (2) the bimodality identified in the real-data study has disappeared in this synthetic case, which once again suggests that the bimodality may be caused by the model structural uncertainty; (3) for the single-modal problems, with proper initialization AM can produce the similar results as DREAM, and (4) about half parameters are well constrained when we define the constraint as the posterior distribution occupies at most half the range of the prior distribution according to Keenan et al. (2013). We believe the added synthetic study will be interesting to the Biogeosciences audience.

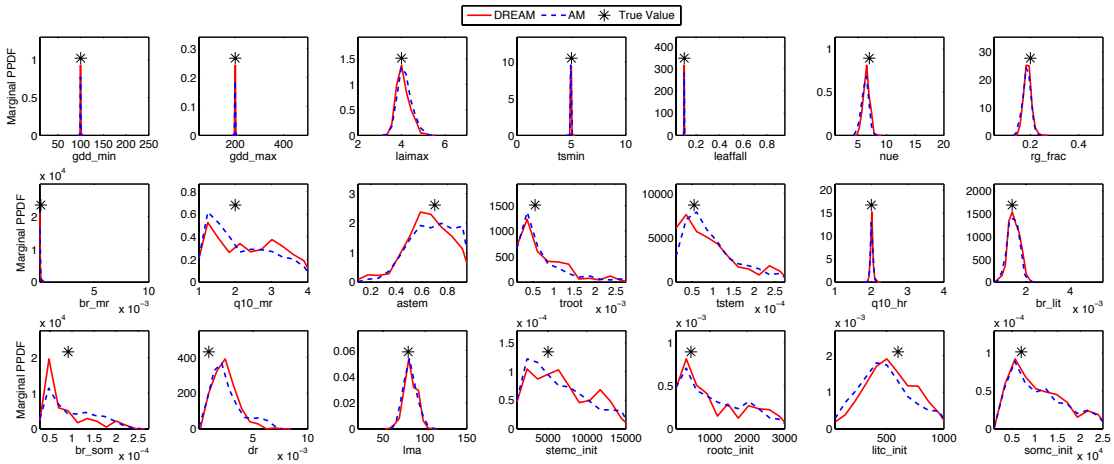


Figure 1. The posterior distributions of 21 DALEC parameters calibrated against the synthetic data generated based on the known parameter values. The limits of the x-axis show the range of uniform prior distributions.

Comment 4:

This is in part motivated by being a little surprised NEE alone has allowed all the parameters to be "successfully" determined when using flux tower data from Harvard Forest. This seems to run counter to many (most?) studies that suggest constraining slow turnover rates and a large pool size from NEE data alone is problematic. With such an experiment you might hope to both demonstrate that this result is feasible (in the absence of model structural and initial condition error) and provide a tool to enable a more detailed analysis of why this seems to be case – simply saying the model is simple enough/doesn't have many parameters is insufficient. For example – how important are the data themselves to this conclusion? Is the length of the record and quality of the observations important?

Response:

We appreciate the reviewer for the excellent suggestion. We find studies that use similar methods and constraining data generally yield similar results in terms of parameter constraints. If we define a well-constrained parameter as the posterior distribution occupying at most half the range of the prior distribution (Keenan et al., 2013), a number of previous studies (e.g., Fox et al., 2009 and Hill et al., 2012) showed that the majority of parameters in DALEC can be constrained by NEE alone. Fox et al. (2009) showed in their Fig. 2 (focusing in particular on participant M1, who defined a similar discrepancy between model and data to construct the likelihood and used a similar MCMC method as our study) that only one of 17 parameters (parameter F_{ll}) was poorly constrained; the other 16 parameters were well constrained according to the above definition. Similarly, Hill et al. (2012) showed in their Fig. 3 that only 3 of the 23 calibrated parameters were poorly constrained; the other 20 parameters have their 90% confidence intervals occupy much less than half of the prior ranges when the DALEC was calibrated against 10 years daily NEE.

Methodological differences may account for differences among studies in how well parameters are constrained with NEE data. Although Richardson et al. (2010) showed in their Fig. 5 that almost all the plotted 16 parameters were poorly constrained when the DALEC was calibrated against daytime NEE only, the comparison between their results and ours is not a direct one. Richardson et al. (2010) defined a different likelihood construction in the MCMC simulation, used a different acceptance rule in the Metropolis algorithm, and estimated the parameter posterior distribution in a different way. From both Fox et al. (2009) and our study, we can see that different algorithms have a substantial influence on the parameter estimation results.

As shown in above Figure 1, the synthetic study indicates that about half of the parameters are well constrained by using the NEE alone, and the other half of the parameters have relatively large uncertainty bounds. This synthetic case gives different results from the real-data calibration. The observation is consistent with the study of Fox et al. (2009) where the measured NEE is an apparently better constraint on the parameters than the synthetic data. In the real-data calibration results (Figure 3 of the manuscript), 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 carbon pool initial condition parameters, *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 summary, other studies using DALEC and NEE data alone often do show well-constrained parameters (e.g., Fox et al., 2009 and Hill et al., 2012), especially those that use similar methods as our study. There are still some other studies (e.g., Richardson et al., 2010), while using DALEC and similar NEE data, implementing different calibration methods (e.g., likelihood constructions and MCMC algorithms) and a different choice of which parameters to calibrate from our work. Such studies have shown that NEE alone is a relatively poor constraint compared to combined data sources in the context of multi-objective calibration. We do not disagree with that general conclusion, but we note the impact of methodology may be quite large and this has not been extensively studied for multi-objective cases. Realizing these differences is important to make a fair comparison, as we can see in our synthetic and real-data calibration cases and from the different results in Fox et al. (2009) when multiple methods were compared. In addition, we agree with the reviewer that the length and quality of the observations, the quality of the models, and the sensitivity of the model parameters to the observations are all important factors on parameter estimation. According to Hill et al. (2012) and Ricciuto et al. (2011), the parameter estimation results are sensitive to observation data length and increasing observation length results in tight parameter uncertainty bounds.

In Sections 3.3 and 3.4 of the revised manuscript, we discussed in detail the parameter estimation results of the synthetic and real-data studies, and explored possible reasons causing the well-constrained parameters.

Comment 5:

Post et al, 2017 JGR-Biogeosciences used DREAM to optimize a set of parameters in the Community Land Model, an ecosystem model massively more complex than DALEC, using flux tower data. Given the similarities, you should draw analogies and make comparisons as appropriate.

Response:

We thank the reviewer for the insightful suggestion and reference. The discussion about the Post et al. (2017) has been added in several places in the revised manuscript. For example, in the Introduction, we added the following sentences: "*Recently, Post et al. (2017) reported a successful application of DREAM in estimation of the complex Community Land Model (CLM) using one-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.*" In addition, we discussed Post et al. (2017)'s work when we analyzed the synthetic study results in the new Section 3.3. Please see the revised manuscript for detail.

Comment 6:

Parameter estimation using MCMC techniques remains very challenging for complex ecosystems models such as CLM for many practical reasons, including computational costs. Again, focusing on the readership of Biogeosciences, it would be useful to provide a comparison of the algorithms not just in terms of intrinsic performance given unlimited resource, but also most importantly their efficiency and also their ease of use and set up.

Response:

We thank the reviewer for the insightful suggestion. In several places of the revised manuscript, we compared the two algorithms' efficiency and ease of use. For example, when we described the implementation of the synthetic study, we added the following sentences: "*To facilitate the convergence of AM, we started the chain from the true parameter values and constructed the initial covariance from samples around the true parameter 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 of 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.*"

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;

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

For publication in Biogeosciences

June 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 work, 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 calibration of DREAM resulted in a better model fit and predictive performance compared to the popular Adaptive Metropolis (AM) scheme. Moreover, DREAM indicated that two parameters controlling autumn phenology have multiple modes in their posterior distributions while AM only identified 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 justified the assumptions of the error model used in Bayesian calibration and investigated their influence on parameter estimation and model performance.

Keywords: Bayesian calibration, MCMC sampling, AM algorithm, DREAM algorithm, DALEC model, multimodality, terrestrial ecosystem 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 the AM in their comparison of different algorithms for the inversion of a terrestrial ecosystem model; Järvinen et al. (2010) utilized the AM for estimation of ECHAM5 climate model closure parameters; Hararuk et al. (2014) employed the AM for improvement of a global land model against soil carbon data; and Safta et al. (2015) used the 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 multi-dimensional 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 one-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. 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) 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 Section 2, we first briefly summarize the general idea of Bayesian calibration and describe the AM and DREAM algorithms. Then in Section 3, we apply both algorithms to the DALEC model in a synthetic and a real-data study. Next in Section 4, we discuss the influence of the likelihood function on parameter estimation and model performance. 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,

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

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 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 $p(\mathbf{x}|\mathbf{D})$ under certain regularity conditions (Robert and Casella, 2004, p.270). However, in practice the convergence rate is often impractically slow, which suggests that within 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.

176 **2.3 AM algorithm**

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

186 To apply the AM algorithm, an initial covariance $\mathbf{C}_0$ must be defined. The choice of $\mathbf{C}_0$
187 critically determines the success of the algorithm. For example, in an extreme case where the
188 variance of $\mathbf{C}_0$ is so large that no proposals are accepted within an iteration, and that the chain
189 remains at the initial state without any movement. This situation continues as the chain evolves,
190 and the use of updated $\mathbf{C}_t$ makes no difference because the variances of $\mathbf{C}_t$ are essentially zero
191 since all the previous samples have the same values. Finally, the AM sampler would get stuck in
192 its initial state without exploring the parameter space. To alleviate this problem and start the AM
193 fairly efficiently, we can define $\mathbf{C}_0$ based on some prior knowledge about the target distribution.
194 When such information is not available, which is usually the case for complex models, some test
195 simulations are needed. For example, Hararuk et al. (2014) inferred $\mathbf{C}_0$ from a test run of 50,000
196 community land model simulations in estimating the PPDFs of soil carbon related parameters.

197 The construction of $\mathbf{C}_t$ is another critical influence on the AM performance. In practice,
198 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 $\mathbf{C}_t$ become 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 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, 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 the AM could have a slow convergence and end up with only one mode (e.g., Figure 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 that combines the delayed rejection (DR) with the 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, the 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

312 out (*gdd_min*), a threshold for maximum leaf area index (*gdd_max*), the temperature for leaf fall
313 (*tsmin*), seasonal maximum leaf area index (*laimax*), the rate of leaf fall (*leaffall*), and leaf mass
314 per unit area (*lma*), respectively. Given the importance of maintenance respiration in other
315 sensitivity analyses (Sargsyan et al., 2014), we expanded the autotrophic respiration submodel to
316 explicitly represent growth respiration (as a fraction of carbon allocated to growth) and
317 maintenance respiration with the base rate and temperature sensitivity parameters.
318 So for the first three plant submodels, deciduous phenology has six parameters; ACM
319 shares one parameter, *lma*, with the deciduous phenology and employs two additional
320 parameters, leaf C:N ratio (which is fixed at a constant of 25 in the simulation) and
321 photosynthetic nitrogen use efficiency (*nue*); the autotrophic respiration model computes the
322 growth and maintenance respiration components and is controlled by three parameters, the
323 growth respiration fraction (*rg_frac*), the base rate at 25°C (*br_mr*), and temperature sensitivity
324 for maintenance respiration (*q10_mr*).
325 The allocation model partitions carbon to several vegetation carbon pools. Leaf allocation
326 is first determined by the phenology model, and the remaining available carbon is allocated to
327 the root and stem pools depending on the fractional stem allocation parameter (*astem*). The litter
328 fall model redistributes the carbon content from vegetation pools to litter pools and is based on
329 the turnover times for stem (*tstem*) and root (*troot*). The last submodel is a decomposition model
330 that simulates heterotrophic respiration and the decomposition of litter into soil organic matter
331 (SOM). This model is driven by the temperature sensitivity of heterotrophic respiration (*q10_hr*),
332 the base turnover times for litter (*br_lit*) and SOM (*br_som*) at 25°C, and by the decomposition
333 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 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. 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 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 Section 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. This synthetic data for calibration was then corrupted with Gaussian errors having means at zero and the same standard deviations with the observed NEEs.

DREAM launched ten 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 uni-modal 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, Figure 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 couldn't 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 one 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

426 besides the measurement errors. We again use the heteroscedastic, uncorrelcted, Gaussian
427 likelihood function for calibration, and examine these error assumptions in Section 4 through
428 residual analysis.
429 DREAM launched ten parallel chains starting at values randomly drawn from the
430 parameter prior distributions, and each chain evolved 300,000 iterations. Chain convergence was
431 assessed via both the univariate and multivariate Gelman Rubin $\hat{R}$ statistics. Figure 2 (b) plots
432 the $\hat{R}$ values of the 21 parameters for the last 100,000 iterations. The figure suggests that the last
433 50,000 samples of each chain (i.e., total 500,000 samples from ten chains) can be used for the
434 PPDF approximation as the $\hat{R}$ has values below the threshold of 1.2.
435 AM used one chain and the chain has the same initialization of the first sample with
436 DREAM. For the initialization of the Gaussian covariance in the AM proposal, we first drew
437 some samples from the parameter space and constructed the covariance. However, this
438 initialization caused a high rejection rate and ended up with essentially a single parameter state
439 after hundred thousands of iterations. To facilitate the convergence of AM, we constructed the
440 initial covariance based on the first 200,000 samples from the DREAM simulation. We
441 conducted ten independent AM runs, so the same $\hat{R}$ statistics can be used for convergence
442 diagnosis. Each AM chain simulated 3,000,000 samples, so that the number of function
443 evaluations in one AM chain is the same with that of DREAM using ten chains. The $\hat{R}$ values of
444 all parameters based on the ten AM runs for the last 1,000,000 iterations are shown in Figure 2
445 (a). The figure indicates that AM has converged and the last 500,000 samples from one chain
446 were used for the PPDF approximation.
447 The estimated PPDFs from AM and DREAM are presented in Figure 3, and the optimal
448 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, Figure 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 Figure 4. Figure 4 (a) 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 by a higher estimate of *leaffall* rate compared to DREAM. As shown in the trace plots of Figure 4 (c) and (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 4 (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 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 with the correlation coefficient of -0.95. As demonstrated in Figure 5 (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 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 6 (a) 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. Taking years 1992 and 1994 as an example, we examined the leaf area index (LAI) in the period of senescence. Figure 6 (b) 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 September 7th, and then decreased slowly over several distinct cool periods during the rest of September and early October until it hit zero in November 7th; 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 November 7th; 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 6 (c) depicts the recorded lowest temperature of the days between September 1st and November 20th 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

518 temperature of triggering leaf fall was relatively high, about 8.1^{o}C (associated with the higher
519 mode of *tsmin*) as shown in Figure 6 (c). In the rest days of September in 1994 following the
520 senescence trigger, temperatures remained warm. The slower leaf fall rate associated with
521 periodic warm conditions (temperatures above *tsmin*) and the lower mode of *leaffall* caused a
522 slow leaf fall in September of 1994 as shown in Figure 6 (b). In comparison, in 1992, senescence
523 was triggered at the end of September with a low temperature of 2.6^{o}C. Then in October with
524 colder temperatures, the leaves drop at a rapid rate associated with the consistent cold
525 temperatures and higher mode of *leaffall*. Especially in late October, the temperatures are
526 consistently below *tsmin*, causing a fast rate of leaf fall, as shown in Figure 6 (b) where the
527 decreasing rate of the LAI in the late October of 1992 is very large. This indicates that a higher
528 temperature trigger is usually associated with a lower leaf fall rate and vice versa.
529 The bimodality identified in the DREAM simulation and examined in the scenarios above
530 reflects the inability of the model structure to predict the observations consistently with a single
531 set of parameters. This bimodality examined in DREAM may be caused in part by an incomplete
532 representation of the senescence process. Using a temperature threshold (parameter *tsmin*) and a
533 constant rate of leaf fall (parameter *leaffall*) to predict senescence is almost certainly an
534 oversimplification. In reality, the process of senescence is also affected by day length. Longer
535 days and warmer temperatures cause a relatively slow rate of leaf fall, whereas shorter days and
536 cooler temperatures accelerate the rate that the leaves fall (Leigh et al, 2002; Saxena, 2010). The
537 higher mode of *tsmin* means that senescence is initiated earlier, when day lengths are still
538 relatively long. This may partially explain why this mode is associated with a lower mode of the
539 *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; Keenan et al., 2015).

The difference in estimated parameters between AM and DREAM causes different simulations of NEE, especially during the Autumn. As an example, Figure 7 illustrates the comparison of the simulated NEE to observations for a month in Autumn of the year 1995 based on MAP estimates obtained 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 (RMSE). 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, $p(\mathbf{y}|\mathbf{D})$, is represented by marginalization of the likelihood over the posterior distribution of model parameters $\mathbf{x}$ as

$$p(\mathbf{y}\,|\,\mathbf{D}) = \int p(\mathbf{y}\,|\,\mathbf{x})p(\mathbf{x}\,|\,\mathbf{D})d\mathbf{x}\,. \qquad (2)$$

In approximation of $p(\mathbf{y}|\mathbf{D})$, we used the converged MCMC samples from $p(\mathbf{x}|\mathbf{D})$. The last 500 samples of each chain (total 500×10=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 $p(\mathbf{y}|\mathbf{D})$.

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

586 existence of local optima. Thus, a robust and efficient MCMC algorithm is desired to give
587 reliable probabilistic descriptions of the TEM parameters.

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

606 As a single-chain sampler, it is conceptually possible for AM to become trapped in a single
607 mode (Jeremiah et al., 2009). Consider a distribution with two far-separated modes and assume
608 that the chain is initialized near one of the two modes (both samples initialization and proposal

609 covariance initialization). At the beginning of the sampling, AM will explore the area around the
610 mode where it is initialized and start identifying the first mode. Since the candidate samples
611 generated by the Gaussian proposal have higher Metropolis ratios (Eq. (2)) in the nearby area
612 than in the far-away regions of the identified mode, the chain is hardly to move to the other
613 mode. When the Gaussian proposal covariance matrix $\mathbf{C}_t$ begins to update, the chance of the
614 chain jumping to the other mode depends on the relative scale of the proposal covariance and the
615 distance between the two modes. When the modes separation exceeds the range of the proposal,
616 AM is less likely to escape the identified local mode.

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

629 ## 4 Discussion

630 The choice of likelihood function plays an important role in the Bayesian parameter
631 estimation, and the likelihood construction depends on the error model assumption. In this 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 Figure 10(a) justifies the assumption of heteroscedastic variances; the density plot of residuals in Figure 10(b) justifies the assumption of normality; but the autocorrelation plot of residuals in Figure 10(c) 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

655 estimate the PPDFs of the 27 parameters and compared the results with those using the
656 uncorrelated error assumption.
657 Figure 11 indicates that the six error model parameters are well identified. The
658 heteroscedastic parameters σ_0 and σ_1 approach 1 and 0, respectively, which suggests that a
659 constant variance may be reasonable. The nonzero ϕ_1, ϕ_2, ϕ_3, and ϕ_4 values indicate that a AR(4)
660 correlation model is necessary. This new heteroscedastic, correlated, Gaussian error model is
661 appropriate as the resulted residuals demonstrate consistent features with the a priori
662 assumptions. As it is shown in Figure 12, the residuals are randomly distributed around the zero
663 line (Figure 12 (a)), normally distributed as assumed (Figure 12 (b)), and no longer correlated
664 after considering the AR(4) model (Figure 12 (c)).
665 The PPDFs of the 21 TEM parameters using the correlated Gaussian likelihood are
666 presented in Figure 13, associated with the results from the uncorrelated Gaussian likelihood. In
667 comparison, we found that the two error model assumptions produced different PPDFs for most
668 parameters. The most remarkable difference is that the bimodality of parameters *tsmin* and
669 *leaffall* disappeared when using the correlated error assumption. As discussed in Section 3.4, the
670 identified bimodality from the uncorrelated likelihood may be caused in part by the model
671 structural error with an incomplete representation of the senescence process. The new likelihood
672 function considers model error probabilistic structures (Lu et al., 2013) and somehow alleviates
673 the effect of model errors on the parameter estimation, resulting in a relatively flat PPDF of
674 *tsmi*n and uni-modal PPDF of *leaffall*. In addition, Figure 13 indicates that parameter uncertainty
675 is larger in the correlated likelihood than the uncorrelated one for most parameters. The reason
676 can be that consideration of the error correlation reduces the data information for calibrating

parameters. 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).

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

699 superior predictive performance, and perhaps identifying structural errors or process differences
700 between the model and ecosystem from which observations were used for calibration.
701 In Bayesian calibration, the choice of likelihood function plays an important role in
702 parameter estimation. In this effort, we justify the assumptions of error model used in
703 constructing the likelihood function and find that a heteroscedastic, correlated, Gaussian error
704 model is reasonable for this problem as supported by the residual analysis.

705 **Acknowledgement**

706 This research was conducted by the Terrestrial Ecosystem Science – Science Focus Area
707 (TES-SFA) project, supported by the Office of Biological and Environmental Research in the
708 DOE Office of Science. The Harvard Forest flux tower is part of the AmeriFlux network
709 supported by Office of Biological and Environmental Research in the DOE Office of Science
710 and is additionally supported by National Science Foundation as part of the Harvard Forest
711 Long-Term Ecological Research site. The NACP site-synthesis activity supported assembling the
712 data set. Oak Ridge National Laboratory is managed by UT-BATTELLE for DOE under contract
713 DE-AC05-00OR22725. Sandia National Laboratories is a multi-program laboratory managed
714 and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin
715 Corporation, for the DOE's National Nuclear Security Administration under contract DE-AC04-
716 94-AL85000.

717 **References**

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

722 Active Archive Center, doi:10.3334/ORNLDAAC/1178, 2013.
723 Box, E.P., and Tiao, G.C.: Bayesian inference in statistical analysis, Wiley, New York, 588pp,
724 1992.
725 Braswell, B. H., William, J. S., Linder, E., and Scheimel, D. S.: Estimating diurnal to annual
726 ecosystem parameters by synthesis of a carbon flux model with eddy covariance net
727 ecosystem exchange observations, Global Change Biology, 11, 335-355, 2005.
728 Brooks, S. P., and Gelman, A.: General methods for monitoring convergence of iterative
729 simulations, J. Comput. Graph. Stat., 7, 434-455, 1998.
730 Dowd, M.: Bayesian statistical data assimilation for ecosystem models using Markov Chain
731 Monte Carlo, Journal of Marine Systems, 68, 439-456, 2007.
732 Feyen, L., Vrugt, J. A., Nuallain, B. O., van der Knijff, and de Roo, A.: Parameter optimization
733 and uncertainty assessment for large-scale stream flow forecasting, J. Hydrol., 332, 276-
734 289, 2007.
735 Fox, A., et al.: The REFLEX project: Comparing different algorithms and implementations for
736 the inversion of a terrestrial ecosystem model against eddy covariance data, Agric. For.
737 Meteorol., 149(10), 1597-1615, 2009.
738 Gelman, A., and Rubin, D.B.: Inference from iterative simulation using multiple sequences,
739 Statistical science 7(4), 457-472 1992.
740 Gelman, A., and Shirley, K.: Inference from simulations and monitoring convergence, Handbook
741 of Markov Chain Monte Carlo, CRC Press; Boca Raton, FL: 2011.
742 Geyer, C. J.: Practical Markov chain Monte Carlo, Statistical Science, 7(4), 473-511, 1992.
743 Gneiting, T., and Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation, J. Am.
744 Statist. Assoc., 102, 359-378, 2007.

745 Haario, H., Saksman, E., and Tamminen, J.: Adaptive proposal distribution for random walk
746 Metropolis algorithm. Comput. Statist., 14, 375-395, 1999.
747 Haario, H., Saksman, E., and Tamminen, J.: An adaptive Metropolis algorithm, Bernoulli, 7(2),
748 223-242, 2001.
749 Haario, H., Saksman, E., and Tamminen, J.: Componentwise adaptation for high dimensional
750 MCMC. Comput. Stat., 20, 265-274, 2005.
751 Haario, H., Laine, M., Mira, A, and Saksman, E.: DRAM: Efficient adaptive MCMC, Statistics
752 and Computing, 16, 339-354, 2006.
753 Hararuk, O., Xia, J., and Luo, Y.: Evaluation and improvement of a global land model against
754 soil carbon data using a Bayesian Markov chain Monte Carlo method, J. Geophys. Res.
755 Biogeosci., 119, 403-417, 2014.
756 Hill, T. C., Ryan, E., and Williams, M.: The use of CO2 flux time series for parameter and
757 carbon stock estimation in carbon cycle research, Global Change Biol., 18, 179–193,
758 2012.
759 Järvinen, H., Raisanen, P., Laine, M., Tamminen, J., Ilim, A., Oja, E., Solonen, A., and Haario,
760 H.: Estimation of ECHAM5 climate model closure parameters with adaptive MCMC,
761 Atmos. Chem. Phys., 10, 9993-10002, 2010.
762 Jeremiah, E., Sisson, S., Marshall, L., Mehrotra, R., and Sharma, A.: Bayesian calibration and
763 uncertainty analysis of hydrological models: A comparison of adaptive Metropolis and
764 sequential Monte Carlo samplers, Water Resour. Res., 47, W07547,
765 doi:10.1029/2010WR010217, 2011.
766 Keenan R.J., Reams, G.A., Achard, F., de Freitas, J.V., Grainger A., and Lindquist E.: Dynamics
767 of global forest area: Results from the FAO global forest resources assessment 2015,

768 Forest Ecology and Management, 352, 9-20, 2015.
769 Keenan, T. F., Davidson, E., Moffat, A., Munger, W., and Richardson, A.D.: Using model-data
770 fusion to interpret past trends, and quantify uncertainties in future projections, of
771 terrestrial ecosystem carbon cycling, Global Change Biology, 18, 2555-2569, 2012.
772 Keenan, T. F., Davidson, E., A., Munger, W., and Richardson, A.D.: Rate my data: quantifying
773 the value of ecological data for the development of models of the terrestrial carbon
774 cycle, Ecological Applications, 23(1), 273-286, 2013.
775 Keenan, T. F., Gray, J., Friedl, M. A., Toomey, M., Bohrer, G., Hollinger, D. Y., Munger, J. W.,
776 O'Keefe, J., Schmid, H. P., Wing, I. S., Yang, B., and Richardson, A. D.: Net carbon
777 uptake has increased through warming-induced changes in temperate forest phenology.
778 Nature Climate Change 4: 598-604, 2014.
779 Kinlan, B. P, and Gaines, S. D.: Propagule dispersal in marine and terrestrial environments: A
780 community perspective, Ecology, 84(8), 2007-2020, 2003.
781 Knorr, W., and Heimann M.: Uncertainties in global terrestrial biosphere modeling: 1. A
782 comprehensive sensitivity analysis with a new photosynthesis and energy balance
783 scheme, Global Biogeochem. Cycles, 15, 207-225, 2001.
784 Laloy, E., and Vrugt, J. A.: High-dimensional posterior exploration of hydrologic models using
785 multiple-try DREAM(ZS) and high-performance computing, Water Resour. Res., 48,
786 W01526, 2012.
787 Leigh, M., Nihevia, N., Covich, E., and Kehn, D.: How temperature and daylength effect
788 seasonal leaf change in honeysuckle plants, online article,
789 http://jrscience.wcp.muohio.edu/nsfall01/labpacketArticles/Howtemperatureanddayleng
790 t.html, 2002.

791 Lu, D., Ye, M., and Hill, M.C.: Analysis of regression confidence intervals and Bayesian
792 credible intervals for uncertainty quantification, Water Resour. Res. 48, W09521.
793 http://dx.doi.org/10.1029/2011WR011289, 2012.
794 Lu, D., Ye, M., Meyer, P.D., Curtis, G.P., Shi, X., Niu, X.-F., and Yabusaki, S.B.: Effects of
795 error covariance structure on estimation of model averaging weights and predictive
796 performance, Water Resour. Res., 49, doi:10.1002/wrcr.20441, 2013.
797 Lu, D., Ye, M., Hill, M.C., Poeter, E.P., and Curtis, G.P.: A computer program for uncertainty
798 analysis integrating regression and Bayesian methods, Environmental Modelling &
799 Software, 60, 45-56, 2014.
800 Lynch, S.M, and Western, B.: Bayesian posterior predictive checks for complex models, Sociol.
801 Meth. Res., 32, 301-335, 2004.
802 O'Sullivan, F.: A statistical perspective on ill-posed inverse problems, Statistical Science, 1(4),
803 502-518, 1986.
804 Papale, D., Reichstein, M., Aubinet, M., Canfora, E., Bernhofer, C., Kutsch, W., Longdoz, B.,
805 Rambal, S., Valentini, R., Vesala, T., and Yakir, D.: Towards a standardized processing
806 of Net Ecosys- tem Exchange measured with eddy covariance technique: algorithms and
807 uncertainty estimation, Biogeosciences, 3, 571-583, 2006.
808 Post, H., Vrugt, J. A., Fox, A., Vereecken, H., and Hendricks Franssen H.-J.: Estimation of
809 Community Land Model parameters for an improved assessment of net carbon fluxes at
810 European sites, J. Geophys. Res. Biogeosci., 122, 1-29, 2017.
811 Price, K.V., Storn, R.M., and Lampinen, J.A.: Differential Evolution, a practical approach to
812 global optimization, Springer, Berlin, 2005.
813 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. Cycles, 19, GB2026, doi: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. Cycles, 22, GB2030, doi: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, doi: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, Global Change Biology, 12, 1174-1188, doi: 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, Springer. 2rd ed. 645pp, 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, 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, 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, Journal of Animal Ecology, 74, 475–487, 2005.

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.

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

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

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

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

860 Urbanski, S., Barford, C., Wofsy, S. C., Kucharik, C., Pyle, E., Budney, J., McKain, K.,
861 Fitzjarrald, D., Czikowsky, M., and Munger, J. W.: Factors controlling CO2 exchange
862 on timescales from hourly to decadal at Harvard Forest, J. Geophys. Res.
863 Biogeosciences, 112, 1-25, 2007.
864 Wang, Y. P., Baldocchi, D., Leuning, R., Falge, E., and Vesala, T.: Estimating parameters in a
865 land-surface model by applying nonlinear inversion to eddy covariance flux
866 measurements from eight FLUXNET sites, Global Change Biology, 13(3), 652-670,
867 2007.
868 Weng, E., and Luo, Y.: Relative information contributions of model vs. data to short- and long-
869 term forecasts of forest carbon dynamics, Ecological Applications, 21, 1490-1505,
870 2011.
871 White, M. A., Thornton, P. E., Running, S. W., and Nemani, R. R.: Parameterization and
872 sensitivity analysis of the BIOME-BGC terrestrial ecosystem model: net primary
873 production controls, Earth Interactions, 4, 1-85, 2000.
874 Williams, M., et al.: Modelling the soil-plant-atmosphere continuum in a Quercus' Acer stand at
875 Harvard Forest: the regulation of stomatal conductance by light, nitrogen and soil/plant
876 hydraulic properties, Plant, Cell Environ., 19, 911-927, 1996.
877 Williams, M., Schwarz, P. A., Law, B. E., Irvine, J., and Kurpius, M. R.: An improved analysis
878 of forest carbon dynamics using data assimilation, Global Change Biol., 11, 89-105,
879 2005.
880 Vrugt, J.A., ter Braak, C.J.F., Clark, M.P., Hyman, J.M., Robinson, B.A.,: Treatment of input
881 uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain
882 Monte Carlo simulation, Water Resour. Res., 44, W00B09,

883 doi:10.1029/2007WR006720, 2008.
884 Vrugt, J.A., ter Braak, C.J.F., Diks, C.G.H., Robinson, B.A., Hyman, J.M., and Higdon, D.:
885 Accelerating Markov Chain Monte Carlo simulation by differential evolution with self-
886 adaptive randomized subspace sampling, International Journal of Nonlinear Sciences &
887 Numerical Simulation, 10(3), 271-288, 2009a.
888 Vrugt, J.A., ter Braak, C.J.F., Gupta, H.V., Robinson, B.A.: Equifinality of formal (DREAM)
889 and informal (GLUE) Bayesian approaches in hydrologic modeling, Stoch. Environ.
890 Res. Risk Assess., 23 (7), 1011-1026, 2009b.
891 Vrugt, J.A.: Markov chain Monte Carlo simulation using the DREAM software package: Theory,
892 concepts, and MATLAB implementation, Environmental modeling & Software, 75,
893 273-316, 2016.
894 Zhang, G., Lu, D., Ye, M., Gunzburger, M., and Webster, C.: An adaptive sparse-grid high-order
895 stochastic collocation method for Bayesian inference in groundwater reactive transport
896 modeling, Water Resour. Res., 49, http://dx.doi.org/ 10.1002/wrcr.20467, 2013.
897 Ziehn, T., Scholze, M., and Knorr, W.: On the capability of Monte Carlo and adjoint inversion
898 techniques to derive posterior parameter uncertainties in terrestrial ecosystem models,
899 Global Biogeochem. Cycles, 26, GB3025, doi:10.1029/2011GB004185, 2012.
900 Zobitz, J. M., Desai, A. R., Moore, D. J. P., and Chadwick, M. A.: A primer for data assimilation
901 with ecological models using Markov Chain Monte Carlo, Oecologia, 167, 599-611,
902 2011.

903 **List of Tables**

904 Table 1. Nominal values and ranges of the 21 parameters for optimization in the DALEC model,
905 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}	7.39×10^{-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/(50×365)	1/(250×365) – 1/(10×365)	1.98×10^{-5}	1.63×10^{-5}
	troot	1/(5×365)	1/(25×365) – 1/365	8.55×10^{-4}	7.88×10^{-4}
Decomp.	q10_hr	2	1–4	2.98	2.68
	br_lit	1/(2×365)	1/(5×365) – 10/(5×365)	4.97×10^{-3}	5.36×10^{-3}
	br_som	1/(30×365)	1/(100×365) – 1/(10×365)	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 – 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

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

List of Figures

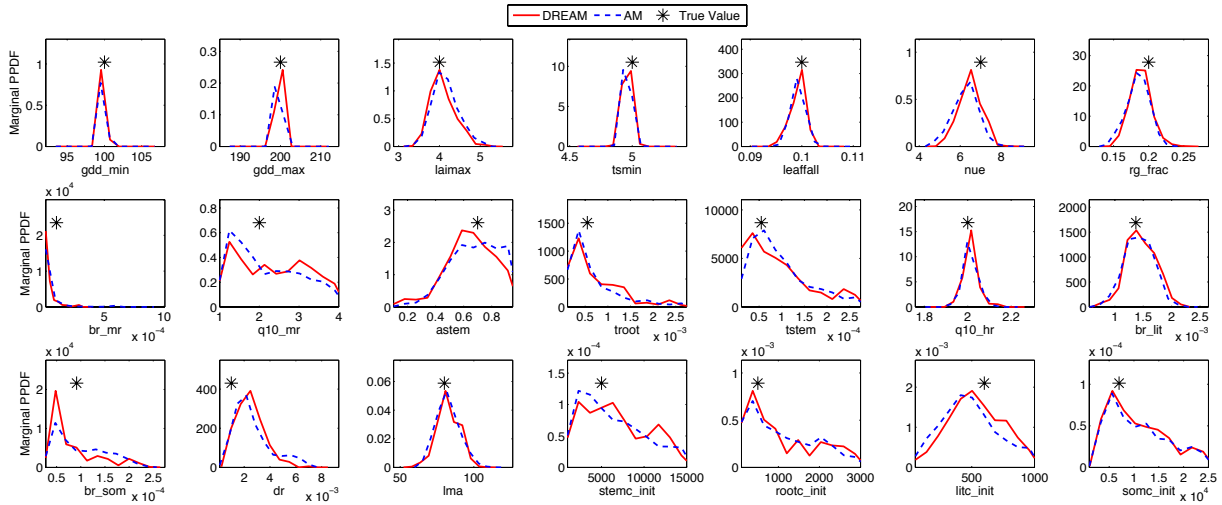
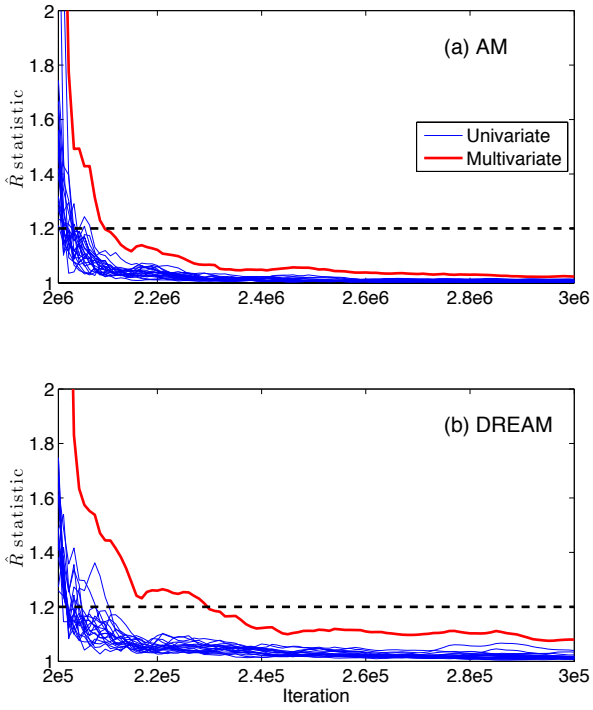


Figure 1. Estimated marginal posterior probability density functions (PPDFs) of the 21 parameters using the AM and DREAM algorithms, along with the true parameter values to generate the pseudo data in the synthetic case.



913

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

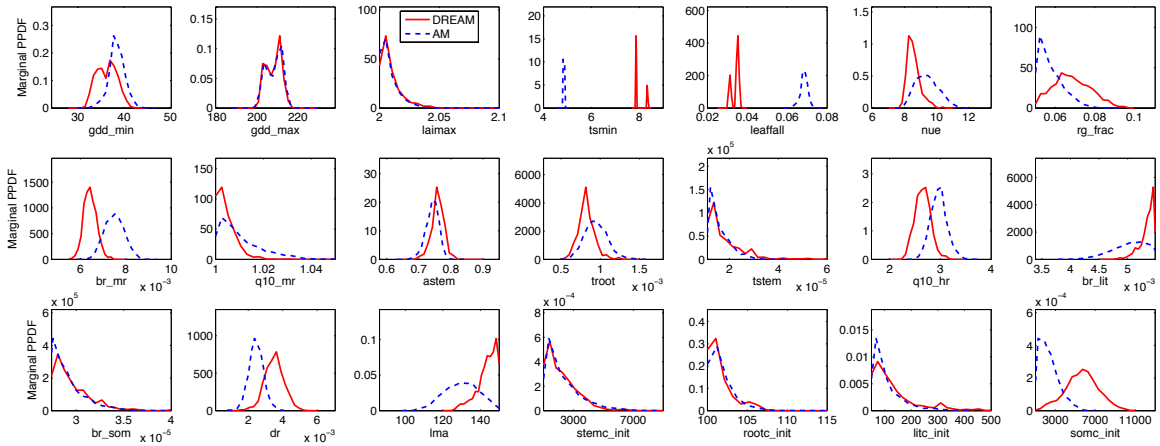
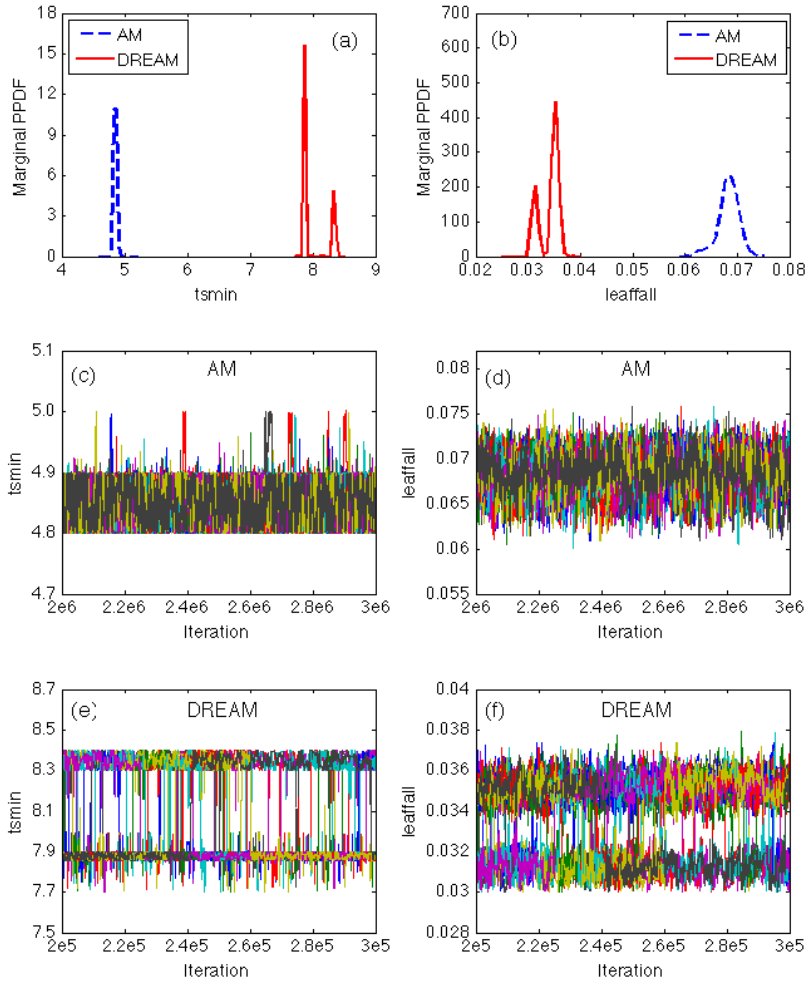
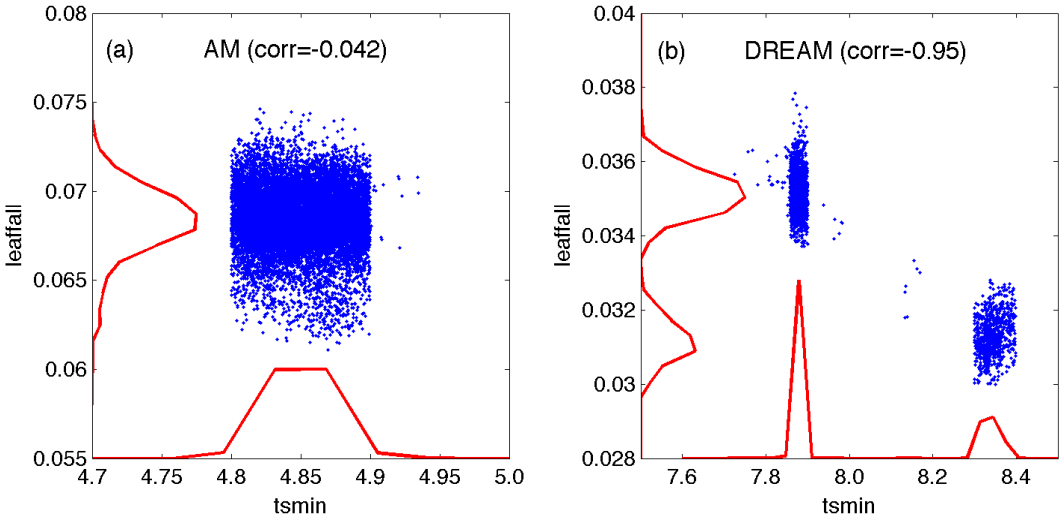


Figure 3. Estimated marginal posterior probability density functions (PPDFs) of the 21 parameters using the AM and DREAM algorithms in the real data study.



922

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



928

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

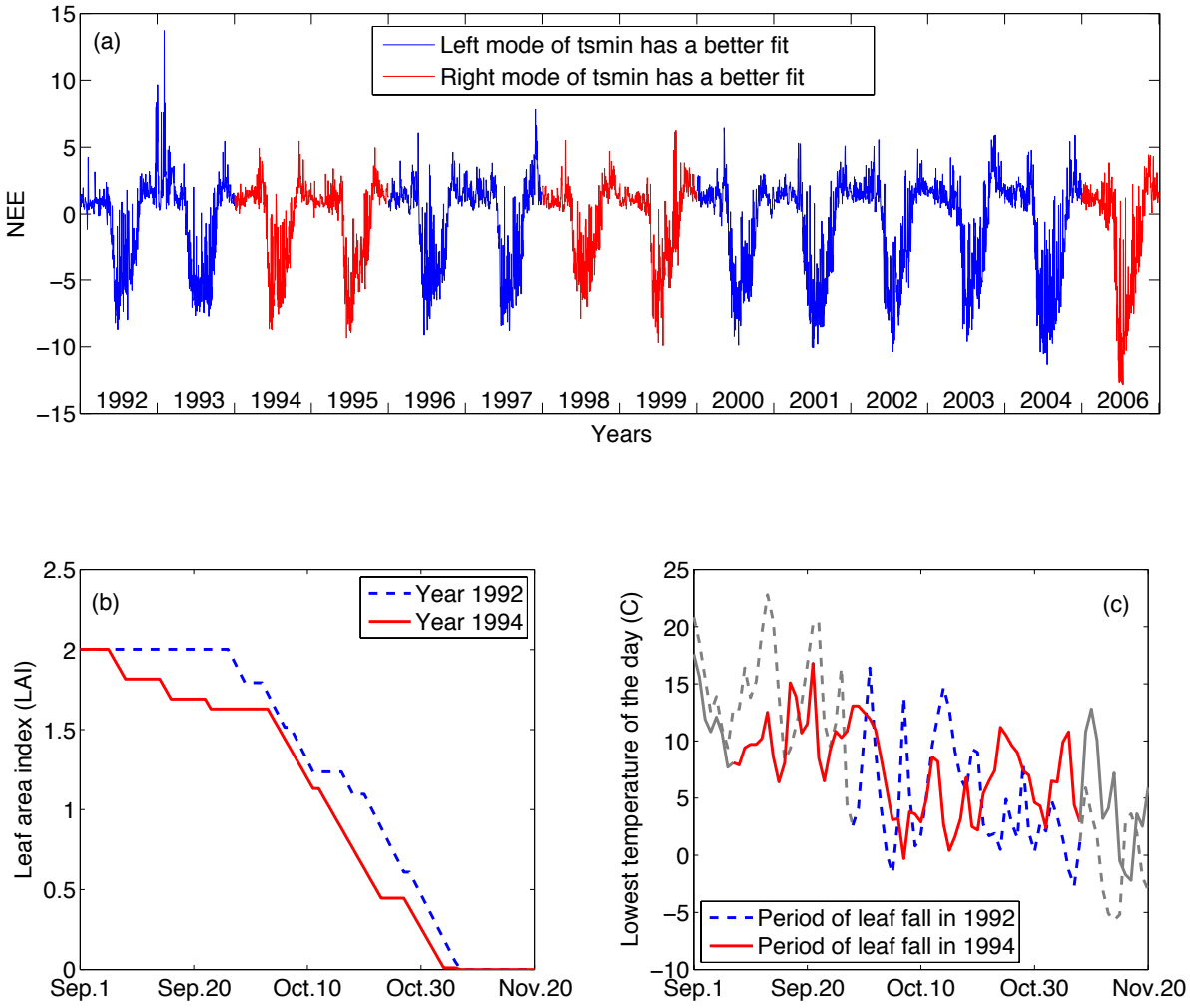
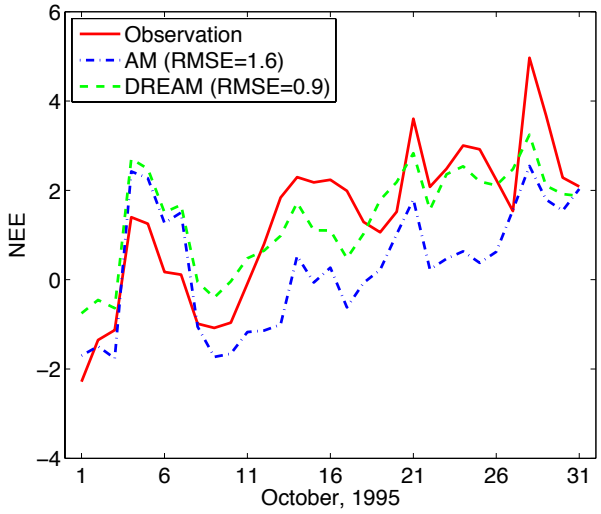
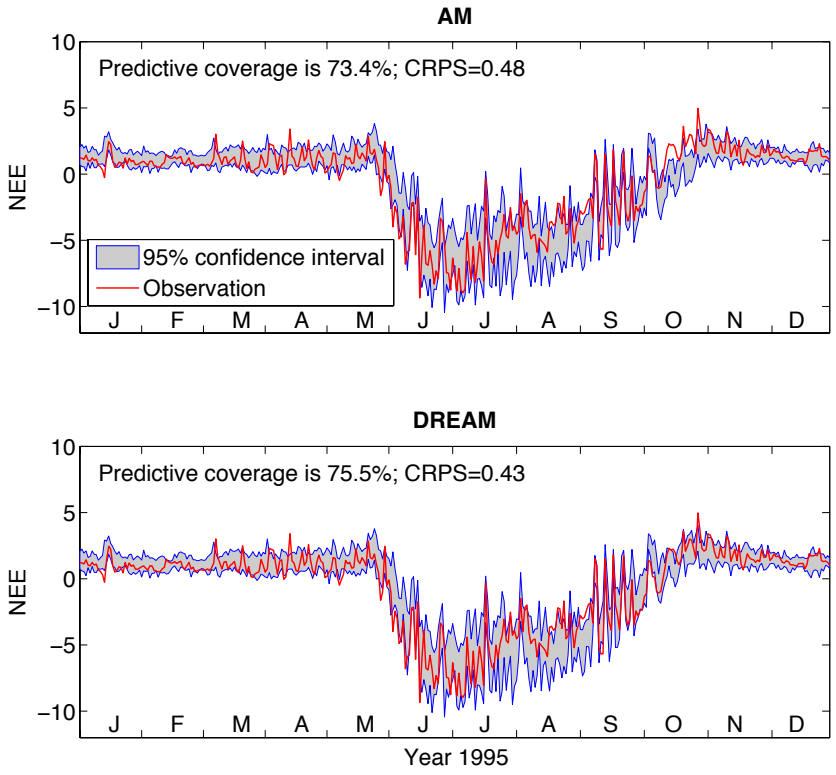


Figure 6. (a) Observed NEE with years highlighted in red where the left mode of *tsmin* has a better model fit and years hightlighted in blue where the right mode of *tsmin* has a better model fit; (b) the simulated leaf area index (LAI) of years 1992 and 1994; and (c) the recorded lowest temperature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

941



942 Figure 7. Simulated NEE values based on the optimal parameters (i.e., the MAP values listed in
943 Tablel 1) estimated by the AM and DREAM algorithms in October 1995. The Root Mean Square
944 Error (RMSE) indicates that DREAM produces a better model fit than AM.



945

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

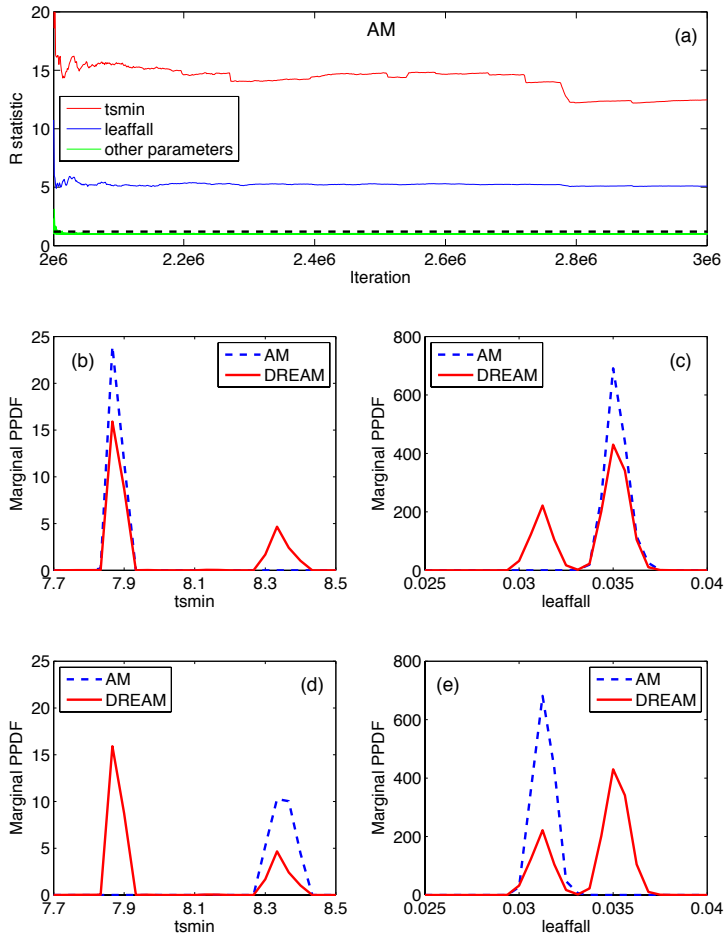
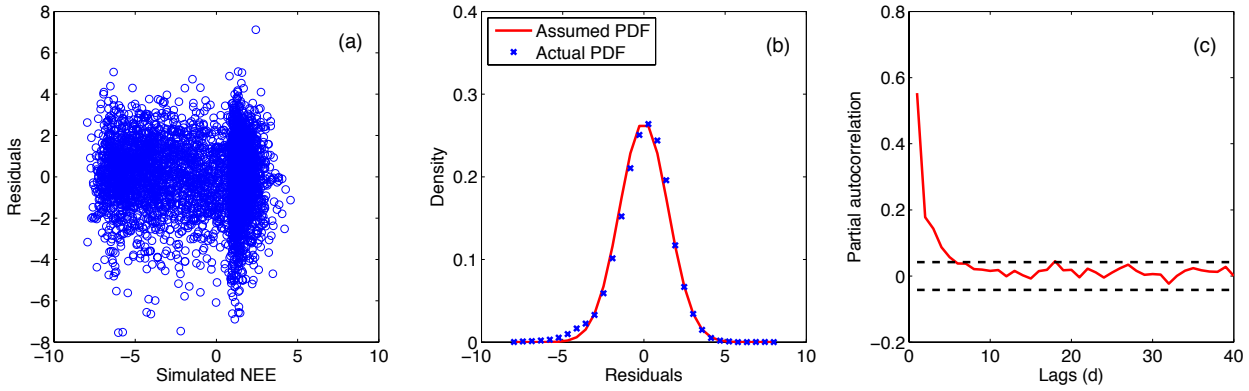


Figure 9. Results of two independent chains of AM with the initial covariance matrix constructed using the converged DREAM samples. The $\hat{R}$ statistic in (a) suggests that different AM chains converged to different *tsmin* and *leaffall* values. One chain captures (b) the left mode of *tsmin* and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

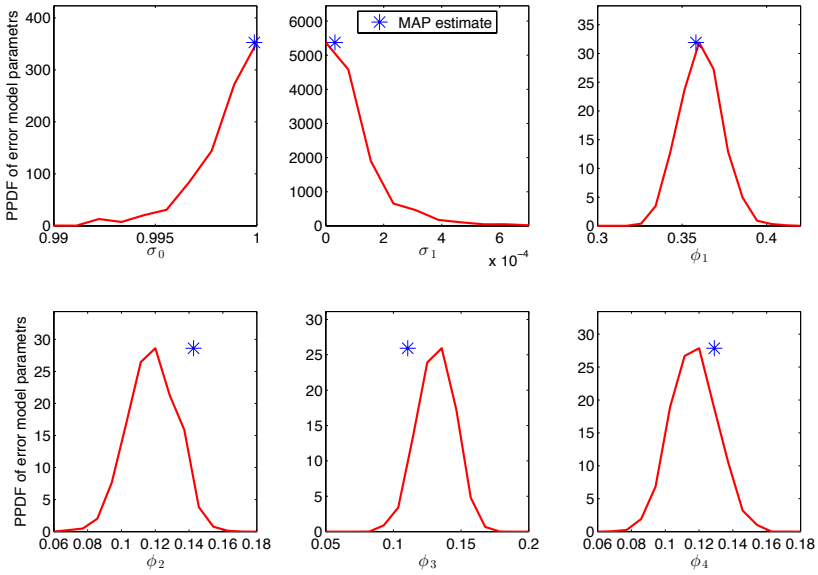
957



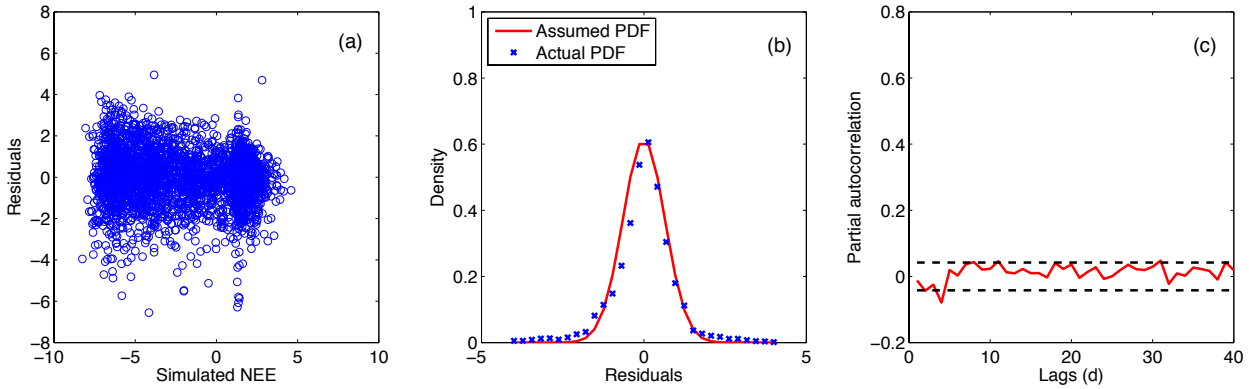
958

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

963

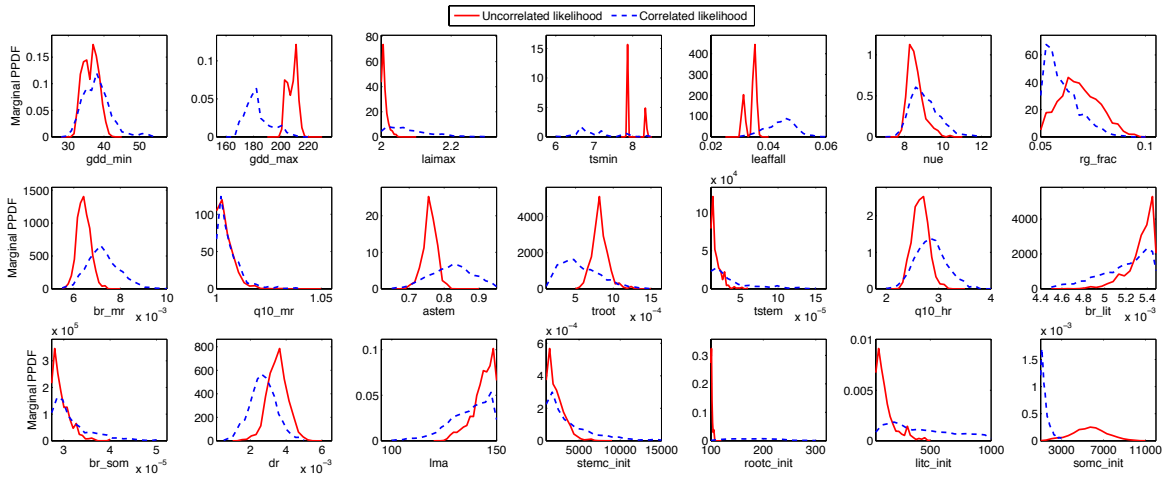


964 Figure 11. Estimated posterior probability density functions (PPDFs) of the six error model
965 parameters.



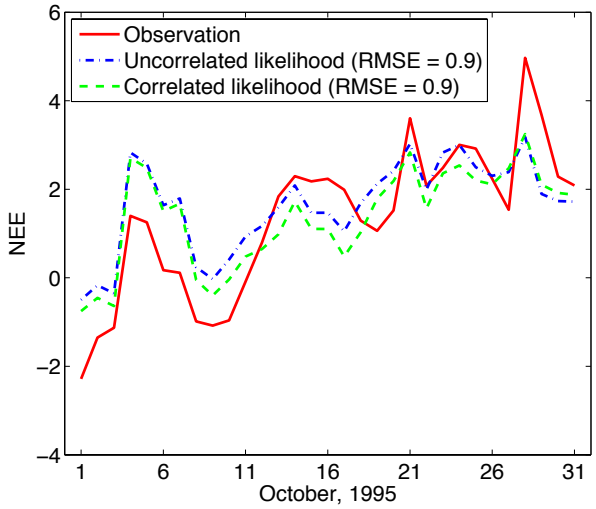
966

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



971

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



974

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