Replies to Referee #1 Comments on Manuscript bg-2017-41

We would like to express our sincere gratitude to Prof. Jasper Vrugt for his insightful and constructive comments and suggestions. All comments have been addressed below and are going to be considered when we revise the manuscript.

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 the relatively high-dimensional problem (21 parameters), the $\mathbf{C}_0$ can easily cause the proposed samples of AM outside the parameter space resulting in a low acceptance rate. We tried several independent AM runs, and they all ended up with a single location in the parameter space with little movement.

Safta et al. (2015) applied AM to the similar 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 suggested 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.

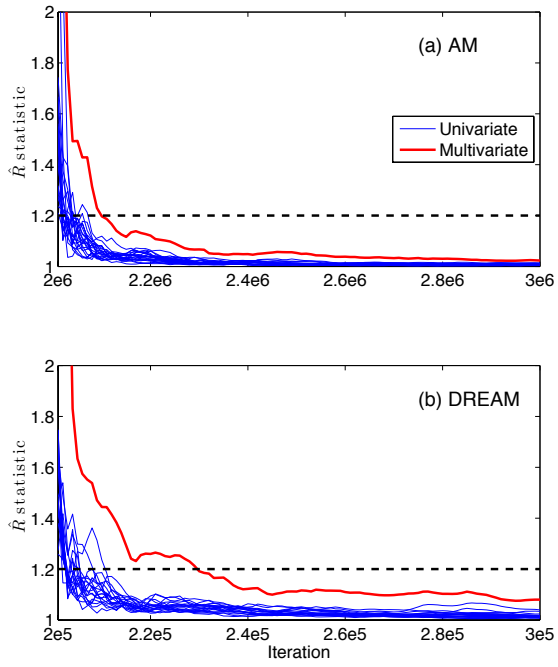

Figure 1. Univariate and multivariate Gelman Rubin $\hat{R}$ statistics (a) for the last 1,000,000 iterations from ten independent AM runs and (b) for the last 100,000 iterations from the DREAM simulation using ten chains. The values less than the threshold of 1.2 suggest chain convergence.

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 suggestions. 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 summarized previous work of AM and DREAM performance. For example, we discussed 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 suggestions. 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 suggestions 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 to justify 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 and found that the residuals of the 5114 NEE data followed a Gaussian distribution but were correlated. As shown in Figure 2 below, the plot of residuals versus simulated NEE in Figure 2(a) justified the assumption of heteroscedastic variances; the density plot of residuals in Figure 2(b) justified the assumption of normality; but the autocorrelation plot of residuals in Figure 2(c) indicated that the residuals were correlated that violated the independence assumption.

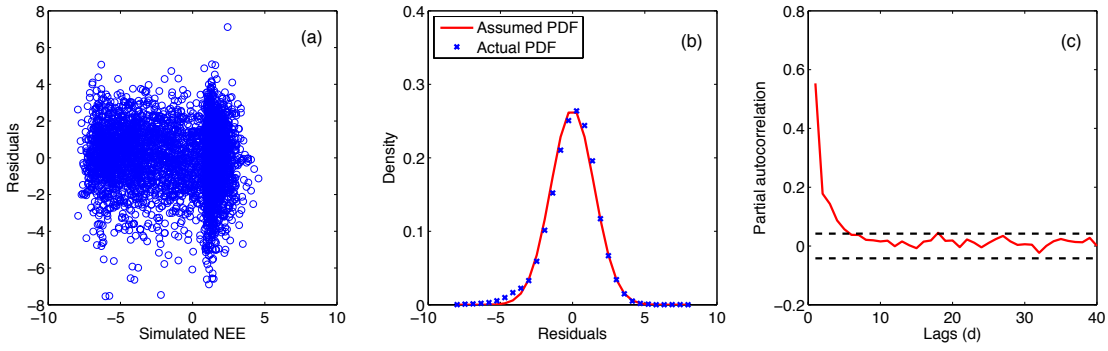

Figure 2. Residual analysis based on the heteroscedastic, uncorrelated, Gaussian error model.

c) The residual analysis suggested 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 can be up to 4. This new error model added 6 extra parameters besides the original 21 TEM parameters, where σ_0 and σ_1 were related to the heteroscedastic error model and ϕ_1, ϕ_2, ϕ_3, and ϕ_4 were 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 indicates that the six error model parameters were 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 indicates that the new heteroscedastic, correlated, Gaussian error model is reasonable where all the a priori assumptions were 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 presents PPDFs of the 21 TEM parameters using the uncorrelated and correlated Gaussian likelihoods. We found that the two error model assumptions produced different PPDFs for most parameters. The most remarkable difference is that the bimodality of parameters *tsmin* and *leaffall* disappeared when using the correlated error assumption. As discussed in 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 accounting for autocorrelation in the errors considers model error probabilistic structures (Lu et al., 2013), resulting in a 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 where we took the simulations in October of 1995 as an example. Although the overall root mean squared errors were similar, but 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.

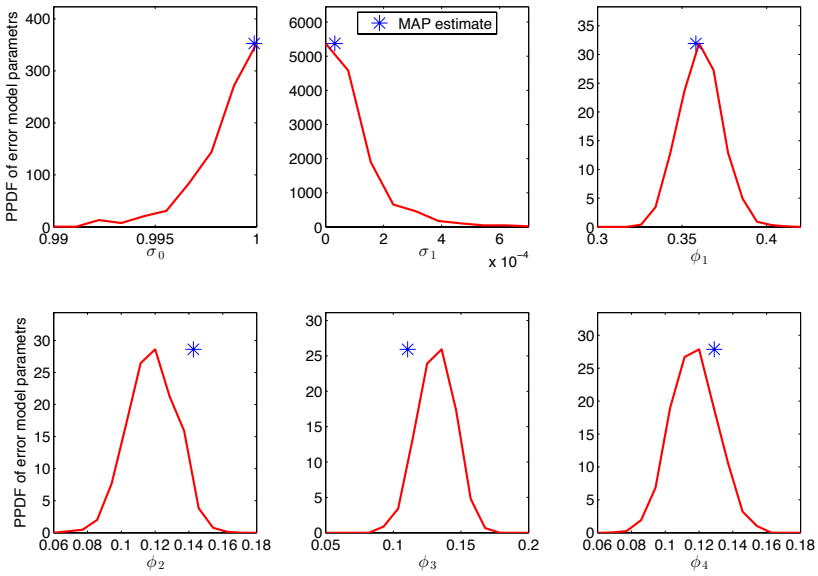

Figure 3. Estimated posterior probability density functions (PPDFs) of the six error model parameters.

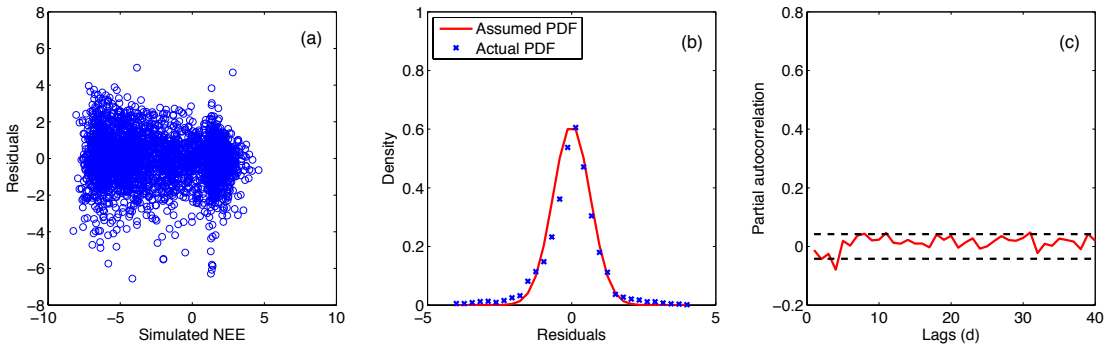

Figure 4. Residual analysis based on the heteroscedastic, correlated, Gaussian error model.

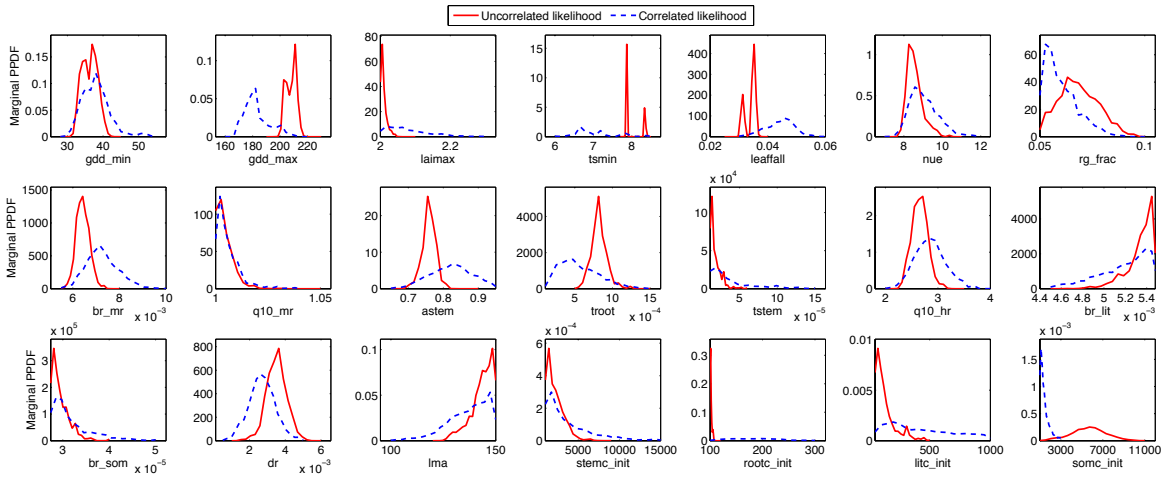

Figure 5. Estimated marginal posterior probability density functions (PPDFs) of the 21 TEM parameters using the uncorrelated and correlated Gaussian likelihoods.

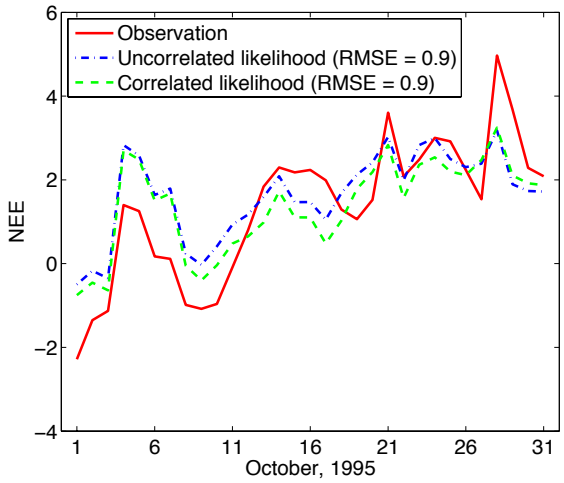

Figure 6. Simulated NEE values based on the MAP estimates from the uncorrelated and correlated Gaussian likelihoods in October 1995. The Root Mean Square Error (RMSE) indicates that the two likelihood functions produced similar results.

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.