Biogeosciences Discuss.,
https://doi.org/10.5194/bg-2018-96-AC2, 2018

Biogeosciences
Discussions

Interactive comment on "Linking big models to big data: efficient ecosystem model calibration through Bayesian model emulation" *by* Istem Fer et al.

Istem Fer et al.

fer.istem@gmail.com

Received and published: 17 May 2018

We thank both referees for their comments and suggestions. We would also like thank the associate editor for diligently inviting reviewers who are most well suited. We find the comments very helpful. In the light of these comments, we realized that we were missing an opportunity to improve our workflow, we now highlight this change and report our results below.

Also it is clear from both referee comments that we should enhance our literature citation, and revise text accordingly. Below please find our responses and suggested changes in the manuscript.

Referee's Comment (RC) 1 - Contrary to statements made in the paper, the techniques used by the authors are for the most part not novel. There is in fact a substantial literature on replacing the likelihood function with more efficient calculation methods, and I shall give pointers to the literature below. Overall it seems that the literature is very poorly referenced in this paper. However, in the field of ecosystem modelling, several techniques described by the authors have been used hardly at all, so the paper can be valuable in introducing the ideas to a new audience.

RC 2 - The methods are not novel, but application of the method in the field of biogeosciences is in its infancy and the example experiment provided here may be useful in designing further approaches.

Authors' Comment (AC) - We thank the reviewers for this remark. In terms of our novelty statements, we wanted to explain that this paper is the culmination of work that has started approximately 12 years ago (please see AGU talk abstract by Dietze et al., 2009) and it was rather novel even across disciplines back then. We acknowledge the fact that this is not the case anymore and offer our apologies for missing key papers.

However, as both reviewers highlighted, a decade after they were first introduced, the techniques described in this paper have been used hardly at all in the field of ecosystem modeling. This is not surprising given that applications of these techniques require a non-trivial amount of computational and statistical expertise, not to mention a steep debugging curve of both models and algorithms. In this paper, we report the integration of a standardized ecological application of these methods in an open-source ecological informatics toolbox for the general use of the ecosystem modeling community. It is exactly our hope that the experiments and the implementations provided here may foster more use and development of novel types of model emulators.

We have revised the novelty statements and provided citations from the literature as needed. We appreciate the constructive comments of the reviewers, which have im-

proved both our workflow and manuscript.

RC 1 - To introduce new methodological ideas to people, the language should be clear and consistent, and that is not the case here. There is a worrying lack of understanding of the difference between the concepts of 'error' and 'uncertainty'. The first refers to deviation from truth, the second to incomplete knowledge, but in this paper the terms are occasionally treated as synonyms, which makes the Introduction highly unclear. Proper terminology for these concepts and others can, for example, be found in the review of Bayesian methods by Van Oijen (2017), where also additional references on MCMC, emulation and hierarchical modelling in ecosystem modelling can be found.

The Introduction mentions that "Parameter error refers to the uncertainty about the true values of the model parameters", which is quite wrong. Parameter error means assigning a value to a parameter which differs from reality, e.g. stating that the light- use efficiency is 1 g MJ-1 when in reality it is 2 g MJ-1. Not knowing whether it is 1 or 2 or anything else is uncertainty. It is therefore also incorrect to state, as the authors do, that "parameter error asymptotically goes to zero with enough data". It is the conditional uncertainty that goes to zero, not the error. Every experimentalist knows that having any number of biased measurements makes no parameter converge to its correct value - and all measurements have their hidden or unhidden biases. There is no safe way to "estimate observation error from data".

AC - As mentioned in our previous short comment, we share the reviewer's concerns about the consistency of concepts. It is important for us that discussions of these concepts, and methods for their analyses become more common practice in ecosystem modeling studies. We completely agree with reviewer's definitions of error and uncertainty, and revised these sections in the introduction as follows:

Authors' changes in the manuscript L64-82 :

> The Bayesian approach also distinguishes between parametric, model structural and data uncertainties, which is critical for ecological forecasting. Parameter uncertainty refers to the uncertainty about the true values of the model parameters due to data deficiency and model simplification (McMahon et al., 2009; Van Oijen, 2017). As models are simplified representations of reality, it is often not possible to measure the true value of an ecosystem model parameter precisely in the field, regardless of the measurement errors (Van Oijen, 2017). However, measurements can still provide estimates for parameter values that makes the model represent the reality better (Van Oijen, 2017). Hence, it is possible to reduce parameter uncertainty with more measurements, conditioned upon the model structure and the measurement error (Van Oijen, 2017; Dietze, 2017a). Therefore, the parameter uncertainty should be reflected by probability distributions and propagated into model predictions. By reducing parameter uncertainties, PDA helps us identify where we need further data collection and improved model representations.
>
> By contrast, process or model structural uncertainty refers to the uncertainty about how to represent ecological processes in models. As every model is a simplification of reality, there will always be underrepresented processes or insufficiently modeled interactions in ecological models (Van Oijen, 2017; McMahon et al., 2009; Clark, 2005). With more observations, we can advance our theoretical understanding and better characterize ecological processes, but process uncertainty does not necessarily decrease with more data, the way parameter uncertainty does (Dietze, 2017a; Gupta et al., 2012; Clark, 2005). As process uncertainty is part of our imperfect models, it is part of the uncertainty associated with the model predictions.
>
> Unlike process and parameter uncertainties, data (observation) uncertainty does not need to be propagated into model predictions. Observation error

is a result of the limited precision and accuracy of the measurement instruments, hence, the uncertainty about it is not part of the process that we are trying to model (Van Oijen, 2017; McMahon et al., 2009). In Bayesian PDA, observation uncertainty should be treated independent of the deviations of model predictions from data as part of the likelihood for observations to inform model predictions without biases (Dietze, 2017a). For a more in depth terminology for these concepts in the context of process-based models and Bayesian methods, see review by Van Oijen (2017).

RC 1 - The treatment of the subject matter in the Introduction is further hampered by poor terminology regarding parameters. Terms like "parameter", "parameter vector", "parameter set[s]" are used arbitrarily and inconsistently. [As an exercise for the reader: show that lines 98 and 147 cannot both apply.] Note that a set is unordered and a vector is ordered, so a point in parameter space can not be a "parameter set". And "covariances among parameters" are not real quantities but statistical quantities that capture part of our uncertainty and that change when more data come in. Therefore the covariances are in no way "accounted for". Please note that your subject matter of Bayesian calibration using MCMC is unfamiliar to many readers, so getting an idea of what is going on requires using precise language. Apologies for these pedantic remarks, but in my experience people stumble over the smallest inconsistency when learning Bayesian methods.

AC - We are grateful for such remarks, and have revised the text accordingly.

L98-99: In the emulator approach, we first propose a set of parameter underlinevectors (knots) according to a statistical design (Fig. 1). Then, we run the full model with this set of underlinedesign points in parameter space, and compare the model outputs with data.

L113-115: Instead of constructing an emulator for the raw model output, we adopt the approach of constructing an emulator of the likelihood – the statistical assessment of the probability of the data given a vector of model parameters which forms the basis for both frequentist and Bayesian inference.

L147-148: (1) Propose initial N_{KNOTS} design points in the parameter space

(2) Run full model with each parameter vector (parallelizable over N_{KNOTS})

L155: (5a) Propose a new vector of process-model parameter values (each parameter vector defines a point in multivariate parameter space)

L177: The second step (2) is to evaluate the full model using the proposed design points in parameter space (knots), and it is the only step where we run the full model.

L189: This allows us to not only accept/reject a proposed parameter vector (5e) but also sample the τ conditional on that parameter vector (step 5f).

L214: In the MCMC, we use the GP to estimate T for both the current and proposed parameter vector (5b).

L216: To propagate this interpolation uncertainty, it is important to draw the T stochastically from the GP, and draw new values for both the current and proposed parameter vector at each iteration.

L221: This is in contrast with traditional optimization and MCMC algorithms that only leverage the current vector of parameter values when proposing new parameters.

L61: As opposed to piecewise evaluation of different parts of the model against different data sets, a Bayesian framework allows the evaluation of the whole model at once against all data sources, reflecting the connections between variables and the covariances among parameters (Dietze, 2017a).

RC 1 - Can you elaborate on the limitations of your approach? What is the maximum number of parameters (p) that can be calibrated in general, and for your two models in particular? You set the number of model-runs at p^3. Does that mean that calibrating 100 parameters is unfeasible because it would require 10^6 model evaluations just to build the emulator?

AC - With the current (p^3) scheme calibrating 100 parameters would be infeasible as it would require 10^6 model evaluations just to build the emulator. With ED2, running the model 10^6 times is not feasible at all, unless iterative emulator rounds are massively parallelized. With SIPNET, the Cholesky decomposition within the GP, rather than the model evaluations, would become limiting for 10^6 design points. In that case, emulators other than GP (e.g. NNGP) could be considered as we discuss in the manuscript.

That said, the (p^3) scheme is just the rule-of-thumb that we employed in these experiments, and not an inherent limit of the emulator approach itself. The calibration of 100 parameters might be possible with a much smaller number of knots ($\ll 10^6$) depending on the model. For example, our scaling experiment (Figure 7b) shows that, in terms of deviance, it was possible to constrain 6 SIPNET parameters to a reasonable extent with 120 knots in total (likewise, 8 and 10 SIPNET parameters with 240 parameters in total). A common recommendation in computer experiments with GP is to use a sample size about 10 times (n = 10d) the input dimension (Loeppky, Sacks Welch, 2009). Others found this is often too small and suggest 20 times (n = 20d) larger sample size (Erickson, Ankenman Sanchez, 2018).

Therefore, calibrating 100 model parameters with 100 x 20 design points could be possible in theory. In practice, we would advocate for performing an uncertainty analysis to reduce the dimensionality of the problem to the subset these 100 parameters that contribute most to model uncertainty. In addition, the data would need to be strong enough to actually constrain 100 parameters. We would be happy to extend the discussion in section 4.6 to explicitly report these numbers (about sample sizes) with references.

RC 1 - How exactly does PEcAn calculate the contributions of different parameters to overall uncertainty, i.e. what was the screening algorithm?

AC - The uncertainty analysis in PEcAn uses a one-at-a-time (OAT) approach. An OAT approach involves multiple model runs while holding all parameters at their median except one each time, and evaluating how it translates to differences in model outputs. The parameters are varied at their parameter data assimilation (PDA) analysis priors' (which could be original priors or, if the parameter was constrained by the meta analysis, they could be meta analysis posteriors in PEcAn) median and at six PDA prior quantiles equivalent to $\pm[1, 2, 3]\sigma$ in the standard normal. Details are given in previous papers as cited (LeBauer et al., 2013; Dietze et al., 2014). Plans are in place to develop a more general multivariate uncertainty analysis in the future once the multivariate version of our trait meta-analysis is in place (Shiklomanov et al in review).

RC 1 - There are linguistic errors (plural subjects with singular verbs, missing definite articles etc.) on lines 54, 55, 92, 93, 100, 183, 201, 248, 294, 306 (twice), 309, 323, 351, 372, 418, 434, 436, 443, 454-455, 482 (twice), 483, 484, 485, 507, 511, 520, 539 (twice), 581.

We thank the reviewer for noticing and noting these errors. We went through the text more carefully and believe we have corrected these errors.

> L54: In Bayesian calibration it is possible to use more than one type of data to simultaneously constrain multiple output variables in a model.
>
> L55: Using multiple data constraints is particularly helpful because model errors can compensate for each other and single variables often do not provide robust constraints.
>
> L92-93: Thus, it is particularly advantageous to consider techniques that are both parallel in nature and which have substantial "memory"
>
> L100: Next, we fit a statistical approximation through the design points

where we evaluated the model.

L183: In this case, T for a Gaussian likelihood would be the sum of squared residuals, $\Sigma(y-\mu)^2$, where y is the observation and μ is the model prediction.

L201: GP assumes that the covariance between any set of points in parameter space is multivariate Gaussian, and that the correlation between points decreases as the distance between them increases

L248: The error distribution of flux data is known to be both heteroskedastic, with variance increasing with the magnitude of the flux, and to have a double exponential distribution rather than a normal (Richardson et al., 2006).

L294: Unlike SIPNET, it is possible to run ED2 simulations with more than one competing PFT.

L306: The use of literature constraints ensures that the posterior parameter estimates fall within underlinea biologically plausible range, and reduces the problem of equifinality, as parameters that are already well constrained cannot vary much, and thus cannot trade-off with poorly constrained parameters.

L309: The scaling factors used for common ED2 PFT parameters all have Beta(1,1) prior distributions.

L323: In the end, 9 and 10 parameters were targeted in SIPNET and ED2, respectively (i.e. in the case of ED2, 9 model parameters are shown in Fig. 2, plus the multiplicative bias parameter)

L351: In our scaling experiment, we evaluate the trade-off between the number of model runs and the approximation error by comparing the 8-parameter SIPNET bruteforce calibration to emulator calibrations with varying numbers of k knots ($k = 120, 240, 480, 960$).

L372: Shaded distributions are the posteriors obtained after each round of emulation.

L418: While root-mean-square-error (RMSE) scores evaluate the deviations of model predictions from data, deviance (-2 x log-likelihood) scores evaluate predictive ability.

L434: However, the time-series plot of LE for SIPNET (Fig. 4, middle panel) shows that SIPNET largely overestimates the winter moisture fluxes whereas ED2 does not (Fig. 6, middle panel).

L436: Both pre- and post-PDA ED2 performance for SoilResp were better than SIPNET (bottom panels).

L443: As expected, the post-PDA ensemble CI approaches the bruteforce post-PDA CI. In other words, the RCI asymptotically converges to zero, while the clock time increases with the number of knots.

L454-455: With a lower number of knots fewer parameters were well-constrained, but with too few parameters we traded-off the ability to get a good fit.

L482-485: First, we ran the full MCMC in between the adaptive sampling steps, and on the final response surface, instead of an optimization search. Hence, we were able to provide full posterior probability distributions of the parameters targeted for calibration, instead of point estimates of optimum values as Li et al. (2018). The ASMO scheme has also been recently updated for distribution estimation using full MCMC runs (ASMO-PODE) and has been tested with Common Land Model (Gong and Duan, 2017).

L507: In addition to just fitting the model, emulators make it practical to implement different hypotheses within a model, re-calibrate the model, and test them against data repeatedly.

L511: For example, it is a known issue that site-level calibrations are not easily transferable to new sites or to larger scales (Post et al., 2017).

L520: A lack of independence in observation errors causes overfitting of the model parameters and underestimates prediction uncertainty.

L539: This experiment showed that the emulator method with SFs could constrain ED2 PFT parameters and improve model predictions.

L581: Future directions may include exploring alternative emulators, such as the Nearest-Neighbor Gaussian Process model (which takes advantage of the fact that nearest neighbors contribute the most information while fitting a GP model), and could help reduce computational costs substantially for bigger datasets and larger numbers of parameters.

RC 1 - The last sentence of the Abstract (l. 34-36) can be deleted without loss of content.

AC - We can delete this sentence.

RC 1 - How is the "Euclidean distance between confidence intervals" determined?

AC - Please also see the previously posted short comment for more details. Realizing that this was not clear in the manuscript, we added the following text:

L354:

To do this, we compared the post-emulator PDA ensemble confidence interval errors relative (RCI) to the post-bruteforce PDA ensemble CI in terms of mean Euclidean distance between their 2.5% - 97.5% CIs. For each experiment with k different knots and variable $(CI_{E,L,k} - CI_{B,L,k})^2$ values were calculated where E stands for emulator, B stands for bruteforce ensemble

and L stands for lower CI limit. The same is calculated for upper CI limit and sum of their mean is used as a score for relative confidence interval (RCI) coverage per variable:

$$RCI_{VAR,k} = mean((CI_{E,L,k} - CI_{B,L,k})^2) + mean(CI_{E,U,k} - CI_{B,U,k})^2)]$$

Next, each RCI vectors ($RCI_{VAR} = RCI_{VAR,960}, RCI_{VAR,480}, RCI_{VAR,240}, RCI_{VAR,120}$) are normalized by dividing by their mean to obtain values independent of the units. Then, the sum over variables (in our case, $RCI_{FINAL} = RCI_{NEE} + RCI_{LE} + RCI_{SoilResp}$) gives us the final RCI score.

L443: As expected, the post-PDA ensemble CI approaches to bruteforce post-PDA CI, in other words the RCI asymptotically converges to zero, while the clock time to increases with the number of knots (Fig. 7a; also see Fig. S6 for time-series plot that shows emulator CI coverage approaching the bruteforce CI coverage with increasing number of knots).

We also include additional supplementary figures showing coverage convergence, similar to the one we presented in the short comment.

RC 1 - Why were 729 knots used for $p = 8$ parameters of SIPNET, given that you state the need for p^3 knots ($729 = 9^3$, not 8^3)?

AC - Because we counted the multiplicative bias parameter in the p. So, 8 SIPNET parameters plus the multiplicative bias parameter, $p = 9$ for SIPNET. 9 ED2 parameters (6 of 9 being scaling factors for common PFT parameters) plus the multiplicative bias parameter, $p = 10$ for ED2. Thank you for pointing this out. We will state this in the text more explicitly and add figures for bias parameter posteriors as well.

L323: In the end, 9 and 10 parameters were targeted in SIPNET and ED2, respectively. To be more specific, the 8 (9) model parameters for SIPNET

(ED2) that are shown in Fig. 2, plus the multiplicative bias parameter were targeted in the PDA, therefore 9^3 (10^3) knots were proposed iteratively with the emulator approach.

RC 1 - Two of the references are not placed in their proper alphabetical position, and the reference to Hartig et al. (2012) is missing.

AC - We corrected these in the revised manuscript.

RC 1 - Can you explain the results shown in Tables A2 and A5? How can posterior distributions for parameters following MCMC neatly fall into parameterised probability distributions (which also are often of different type than their priors)? And what were the posterior covariances?

AC - As explained in the short comment, the results reported in Tables S2 and S5 are fitted parametric distributions to the marginal MCMC samples. We wanted to provide an approximate parametric distribution for the reader for ease of use. Otherwise, all the raw MCMC samples are accessible via PEcAn for more interested readers. We will now extend the explanation before Table S6 to:

Table S6 caption:

Links to the Workflow IDs. The input/output files associated with each workflow can be accessed via the history table at https://pecan2.bu.edu/pecan/history.php. Or each workflow can be accessed directly by replacing the workflowID in the following link:

https://pecan2.bu.edu/pecan/08-finished.php?workflowid=1000008503
(please note that this takes a while to load)

The left frame on the page can be used to navigate through PEcAn settings, input and output files. If you wish to conduct further visualization or analysis on the MCMC samples, you can first select the "mcmc.list.pda***.Rdata" file

(*** being the ensemble IDs given by the workflow) under the "PEcAn Files" dropdown menu on the left frame. By clicking "Show File" button you can download the raw MCMC outputs to your own machines.

If you would like to display posterior density distributions, first select either soil or plant physiology the under the "PFTs/PFT" menu on the left frame. Next, under the "PFTs/Output" dropdown menu select "posteriors.pda.***.pdf" files and click "Show PFT Output". The red line would be the posterior density plot and the black line would be the approximated parametric distributions (such as the ones reported in Table S2 and S5) fitted by PEcAn's approx.posterior function that can be found under pecan/modules/meta.analysis/R/approx.posterior.R

L428: Fitted parametric posterior distributions of ED2 are given in the supplement (Fig. S1, Table S5.) In addition all raw MCMC samples ("mcmc.list.pda***.Rdata") and posterior distribution plots ("posteriors.pda.***.pdf") are available from the respective workflow directories (see Table S6).

L397: The strongest correlations between leaf growth and leaf turnover rate, and soil respiration rate and soil respiration Q10 parameters were also detectable in emulator posteriors (emulator Fig S4, bruteforce Fig S5).

Important note on an improvement/fix from the authors: Before, we were using these fitted parametric distributions to i) propose new knots in an iterative round, ii) produce post-PDA ensembles. In other words, we were sampling from the marginal distributions, and missing further constraint from covariances. We are now sampling the joint posterior distributions for both proposing new knots and generating post-PDA ensembles. At the end of responses, some of our figures are redrawn with new results (please note that other figures will also be redrawn in the revised manuscript). We also include the following explanation at the end of section 2.1 Emulator-based calibration:

L226: In this study, new points were added by proposing 20% of the new knots from the original prior distribution and 80% from the joint posterior of the previous emulator round (via re-sampling the MCMC samples in between rounds).

Enhancing Literature references:

RC 1 - Missing references to the literature include the following.

AC - We thank the reviewer for going the extra mile and briefly summarizing relevant aspects in all these references. We will include most of them in the manuscript in regarding places.

RC 1 - Further, as perhaps an unmentioned predecessor of calibrating data-scaling parameters, see the ecosystem model Bayesian calibration approach of Van Oijen et al. (2011), where every separate data stream came with its own bias parameter.

L276-278: The bias term is included to account for the scaling from the discrete soil collars to the stand as a whole (Van Oijen et al., 2011). This term was also introduced because observed soil chamber fluxes were typically over twice the ecosystem respiration estimated from the eddy-covariance tower (Phillips et al., 2017). As in previous studies, this parameter is also estimated in the calibration (Van Oijen et al., 2011).

RC 1 - Jandarov et al. (2014) used the same refinement employed in the present paper, of emulating sufficient statistics instead of the overall likelihood directly.

AC - If we understood their study correctly, Jandarov et al.'s (2014) approach is related but different than ours. As the spatiotemporal data they were dealing with was high dimensional, likelihood-based inference for their model was becoming intractable. Their

approach consists of obtaining summary statistics from forward simulator runs, and emulating the Euclidean distances between the summary statistics of their simulated data and the summary statistics of the real data. In other words, they compared model and data on a more aggregated level in their calibration. They chose these key summary statistics by expert opinion to capture important characteristics of their modeled process (disease dynamics). The emulated Euclidean distances is then treated as the likelihood function in their study. Whereas in this study, we compared model and data directly, and emulated sufficient statistics of the likelihood. Here, sufficient statistics has a formal mathematical definition (Fisher, 1992; Mikusheva, 2011).

L179: Next (step 3), a sufficient statistic (T) is calculated by comparing each model output to each data set (Fig. 1). Statistic T is sufficient for the job of estimating the unknown parameters "when no other statistic calculated from the same sample provides any additional information" (Fisher, 1992).

RC 1- Oakley Youngman (2017) showed many of the same methods as the present authors do.

L113-115: Instead of constructing an emulator for the raw model output, we adopt the approach of constructing an emulator of the likelihood – the statistical assessment of the probability of the data given a set of model parameters which forms the basis for both frequentist and Bayesian inference. Emulating the likelihood has the advantage that likelihood surfaces are generally smooth and univariate (Oakley and Youngman, 2017).

L471-473: The efficiency of this workflow could potentially be increased further by other adaptive sampling designs, and this remains an important area for further research. For example, Oakley and Youngman (2017) used an initial set of simulator runs to screen-out low likelihood regions to reduce the parameter space before the calibration. For a review of adaptive

sampling methods, and emulator design methodologies in general, see Forrester and Keane (2009).

RC 1 - For many examples of likelihood-emulation using Gaussian processes etc. in cosmology, see Aslanyan et al. (2015) and references 7-24 therein (which also tend to focus on how much computations are made faster by likelihood-emulation).

L476-480 (section 4.2): In this study, we focused on calibrating process-based mechanistic simulators (ecosystem models) using computationally cheaper emulators. Variations of emulator approach are many, and can be found in Jandarov et al., (2014); Aslanyan et al. (2015), Huang et al. (2016), Oakley and Youngman (2017) and the references therein. Here we adopted the version which emulates the likelihood surface with a Gaussian process, similar to previous studies including applications with a cosmological likelihood function (Aslanyan et al., 2015), a stochastic natural history model (Oakley and Youngman, 2017), the Hartman function and a hydrologic model (Wang et al., 2014) and two land surface models (Li et al., 2018). Our scheme also resembles the adaptive surrogate modelling-based optimization approach (ASMO; Wang et al., 2014; Li et al., 2018) in terms of both the nature of the problem (calibration of a process-based mechanistic simulator) and the general scheme of the calibration algorithm. However, aside from differences in initial sampling designs and error characterizations in these studies, there are two main differences of our scheme from ASMO.

RC 1 - Kandasamy Schneider show that instead of emulating the likelihood, it is also possible to emulate the product of prior and likelihood (i.e. the posterior up to a constant), an approach not mentioned by the present authors.

AC - We might be looking at the wrong paper (because we found a paper from Kandasamy, Schneider and Poczos by the same name and year, not from Kandasamy Schneider), but this paper also emulates the likelihood surface (they estimate posteriors through emulated likelihoods). However, a paper we are already citing (Gong and Duan, 2017) does emulate posterior surface. Both papers are now cited (see next comment).

(this is a comment by R2, included here as well for completeness) RC 2 - The argumentation behind the sufficient statistics is not sustained by the experiment.

AC - We thank the reviewer for pointing this out. We will now extend the following section in the discussion. Please also see our next response.

L493-496: A second addition to our scheme was that we included a further generalization of emulation of the sufficient statistics (T) surface. T is, by definition, sufficient to estimate the simulator (process model) parameters in the MCMC. Unlike emulating the likelihood (this study, Oakley and Youngman, 2017; Kandasamy, Schneider and Poczos, 2015) or the posteriors (Gong and Duan, 2017), emulating T allows us to estimate parameters that are not part of the process model but are part of the statistical data model (the likelihood) as well. In this study, we tested the sufficient statistics emulation for the SoilResp data and updated Gaussian likelihood precision parameter in the MCMC together with other process model parameters. This residual parameter includes both data error and model structural error, and it is not possible to distinguish one from the other with this approach (Van Oijen, 2017). However, when we apply the same calibration scheme to different process models at the same site, because the observation error in the data are the same, the difference in the posteriors of this residual parameter (Fig. S2)* could give us clues about the model structural errors of models relative to each other, as we demonstrate in this study as a proof-of-

concept. However, in our study, use of multiplicative bias parameter further obscures the difference between observation and model structural error.

*Please note that Fig S2 will be redrawn with the revised workflow and will be mentioned explicitly in the results section.

RC 1 - Published methods for Bayesian calibration increasingly take into account that models are imperfect. There is a discrepancy between model output and reality, even at the best possible setting of model parameter values. This discrepancy is often modeled as a Gaussian Process for which - in the Bayesian calibration - the hyperparameters are estimated together with the regular model parameters. Likelihood-emulation precludes including discrepancy-estimation because model outputs are not calculated during the MCMC. Please add a discussion of this limitation of your approach.

AC - This is an important point. First, it is worth noting that our current scheme does allow the inclusion of this discrepancy in terms of a bias and variance terms that are estimated together with the regular model parameters. Indeed, the ability to fit the variance term at the same time as the parameters is precisely why we switched to emulating summary statistics. That said, it is true that in the current implementation the bias term is assumed to be a fixed constant, not varying dynamically, and the soil respiration variance is assumed to be homoskedastic (though the tower fluxes are not). Second, we would argue that our approach does not preclude a more flexible bias specification. Indeed, while beyond the scope of the current paper, conceptually it should be possible to use a bias-variance decomposition to separate our single emulator of the error surface into two separate emulators for bias and variance terms. Similarly, our approach does not preclude specifying a likelihood with a temporally autocorrelated error (which is functionally equivalent to a GP error model in the time dimension), and augmenting the emulator with the autocorrelation parameter similar to how we augmented the emulator with the bias term. As discussed in the paper we instead chose to approxi-

mate this as an effective sample size correction, both for computational efficiency and because accounting for autocorrelation in an asymmetric heteroskedastic Laplace is more complicated than doing so in a multivariate Normal. Finally, in PEcAn, we are working towards a more general framework for model-data integration that takes into account initial condition / driver / parameter / model structural uncertainty in calibration and prediction. However, this is still work in progress.

We will include the following in the discussion, at the end of section 4.2 after the paragraph in the response above:

> L497: Indeed, implementation of a more formal way of accounting for model structural error (also called the discrepancy between model output and reality) in our emulator scheme is one of our planned next steps. Explicitly specifying a model discrepancy term and estimating it through MCMC would allow us to account for all sources of model predictive uncertainty (Van Oijen, 2017). However, determining the expected form of discrepancy in order to learn about model parameters realistically could be difficult due to lack of mechanistic knowledge of the underlying processes (Brynjarsdottir and O'Hagan, 2014). In that sense, accounting for discrepancy in model calibration is not an emulator approach specific issue. For a novel approach investigating model structural uncertainty through a modular modeling framework see Walker et al. (2018), which could be useful for modeling prior knowledge about discrepancy in ecosystem models in the future. Because of the unknowns about the discrepancy functions, it is common to use Gaussian processes to model the discrepancy (Kennedy and O'Hagan, 2001). Even then, only with realistic prior constraints about the process, calibrated model predictions will be unbiased (Brynjarsdottir and O'Hagan, 2014). For an example of addressing discrepancy in calibration that combines likelihood-emulation approach with importance sampling, see Oakley and Youngman (2017) where they inflated simulator un-

certainty to account for simulator discrepancy instead of explicitly specifying a prior for it in order to make the likelihood tractable. When likelihood function becomes intractable, techniques using likelihood-free inference could also be a remedy (Gutmann and Corander, 2016).

Cited references (that were not mentioned in previous correspondence or in the manuscript):

Brynjarsdottir and O'Hagan, 2014, Learning about physical parameters: the importance of model discrepancy, 30, IOP, doi:10.1088/0266-5611/30/11/114007

Dietze et al., Beyond MCMC: Data-constraint and error propagation in a dynamic terrestrial biosphere model through Bayesian model emulation, American Geophysical Union, Fall Meeting 2009, http://adsabs.harvard.edu/abs/2009AGUFM.B44A..02D

Erickson, C.B., Ankenman, B.E., Sanchez, S.M., 2018, Comparison of Gaussian process modeling software, European Journal of Operational Research, 266(1), 179-192 https://doi.org/10.1016/j.ejor.2017.10.002

Fisher, R., A., On the mathematical foundations of theoretical statistics, Philosophical Transactions Of The Royal Society A.222:309-368, doi:10.1098/rsta.1922.0009

Gupta, H. V., M. P. Clark, J. A. Vrugt, G. Abramowitz, and M. Ye (2012), Towards a comprehensive assessment of model structural adequacy, Water Resour. Res., 48, W08301, doi:10.1029/2011WR011044.

Loeppky, J. L., Sacks, J., Welch, W. J. (2009). Choosing the sample size of a computer experiment: A practical guide. Technometrics, 51(4), 366–376. doi:10.1198/TECH.2009.08040.

Mikusheva, A., Lecture 4 Sufficient Statistics, course materials for 14.381 Statistical Methods in Economics, Fall 2011. MIT OpenCourseWare (http://ocw.mit.edu), Massachusetts Institute of Technology. Last access on 14/05/2018.

Rasmussen, C. E. (1996). Evaluation of Gaussian Processes and Other Methods for Non-linear Regression. PhD thesis, Dept. of Computer Science, University of Toronto. http://www.kyb.mpg.de/publications/pss/ps2304.ps

Walker, A. P., Ye, M., Lu, D., De Kauwe, M. G., Gu, L., Medlyn, B. E., Rogers, A., and Serbin, S. P.: The Multi-Assumption Architecture and Testbed (MAAT v1.0): Code for ensembles with dynamic model structure including a unified model of leaf-scale C3 photosynthesis, Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-71, in review, 2018.

Interactive comment on Biogeosciences Discuss., https://doi.org/10.5194/bg-2018-96, 2018.