https://doi.org/10.5194/bg-2020-23
Preprint. Discussion started: 19 February 2020

Biogeosciences Discussions

A Bayesian Approach to Evaluation of Soil Biogeochemical Models

Hua W. Xie[1], Adriana L. Romero-Olivares[2], Michele Guindani[3], and Steven D. Allison[4]

[1]Center for Complex Biological Systems, University of California, Irvine, 2620 Biological Sciences III Irvine, California 92697, United States of America

[2]Department of Natural Resources & the Environment, University of New Hampshire, 114 James Hall, Durham, New Hampshire 03824, United States of America

[3]Department of Statistics, University of California, Irvine, 2241 Donald Bren Hall, Irvine, California 92697, United States of America

[4]Department of Ecology and Evolutionary Biology, Department of Earth System Science, 321 Steinhaus Hall, University of California, Irvine, California 92697, United States of America

Correspondence to: Hua W. Xie (xiehw@uci.edu)

Abstract. To make predictions about the effect of rising global surface temperatures, we rely on mathematical soil biogeochemical models (SBMs). However, it is not clear which models have better predictive accuracy, and a rigorous quantitative approach for comparing and validating the predictions has yet to be established. In this study, we present a Bayesian approach to SBM comparison that can be incorporated into a statistical model selection framework.

We compared the fits of a linear and non-linear SBM to soil respiration CO_2 flux data compiled in a recent meta-analysis of soil warming field experiments. Fit quality was quantified using two Bayesian goodness-of-fit metrics, the Widely Applicable information criterion (WAIC) and Leave-one-out cross-validation (LOO). We found that the linear model generally out-performed the non-linear model at fitting the meta-analysis data set. Both WAIC and LOO computed a higher overfitting penalty for the non-linear model than the linear model, conditional on the data set. Fits for both models generally improved when they were initialized with lower and more realistic steady state soil organic carbon densities.

Testing whether linear models offer definitively superior predictive performance over non-linear models on a global scale will require comparisons with additional site-specific data sets of suitable size and dimensionality. Such comparisons can build upon the approach defined in this study to make more rigorous statistical determinations about model accuracy while leveraging emerging data sets, such as those from long-term ecological research experiments.

1 Introduction

Coupled Earth system models (ESMs) and constituent soil biogeochemical models (SBMs) are used to simulate global soil organic carbon (SOC) dynamics and storage. As global climate changes, some ESM and SBM simulations suggest that substantial SOC losses could occur, resulting in greater soil CO_2 emissions (Crowther et al., 2016). However, there is vast divergence between model predictions. For instance, one ESM predicts a global SOC loss of 72 Pg C over the 21st century, while another predicts a gain of 253 Pg C (Todd-Brown et al., 2014).

Soil biogeochemical models vary in structure (Manzoni and Porporato, 2009), but can be broadly partitioned into two categories: those that implicitly represent soil C dynamics as first-order linear decay processes and those that explicitly represent microbial control over C dynamics with non-linear Michaelis-Menten functions (Wieder et al., 2015). Explicit models typically include more parameters than linear models because multiple microbial parameters are needed for each decay process as opposed to a single rate parameter. The additional parameters allow explicit models to represent microbial mechanisms, but at the expense of greater model complexity.

Rigorous statistical approaches should be applied to investigate how explicit representation of microbial processes affects predictive model performance. ESM and SBM comparisons involving empirical soil C data assimilations have been conducted previously (Allison et al., 2010; Li et al., 2014) but few standardized statistical methods for ESM and SBM benchmarking and comparison have been developed that would allow for rigorous model selection. Prior model comparisons have involved graphical qualitative comparisons or use of basic fit

https://doi.org/10.5194/bg-2020-23
Preprint. Discussion started: 19 February 2020

metrics such as the coefficient of determination, R_2, to judge fit quality. However, these simple approaches are insufficient for comparing an increasing number of complex models (Jiang et al., 2015; Luo et al., 2016; Wieder et al., 2015).

Encouragingly, a rich toolset for quantitative model evaluation and comparison can be drawn from Bayesian statistics. These tools include information criteria and cross-validation, goodness-of-fit metrics designed for the simultaneous comparison of multiple structurally diverse models. Like R_2, information criteria and cross-validation are quantitative measures that estimate the fit quality of a model to a given data set. Differing from R_2, information criteria and cross-validation are relative rather than absolute measures. These metrics evaluate the extent to which the data set supports particular distributions of parameter values and in turn, the uncertainty of parameter estimates. Consequently, if the distribution of Model A outcomes aligns more closely to the data set than the distribution of Model B outcomes, we regard Model A as being more likely to explain the data compared to Model B. Information criteria and cross-validation metrics also typically include terms penalizing for model complexity and overfitting as part of their computation (Gelman et al., 2014). Hence, information criteria and cross-validation are useful tools for model evaluation because they present a comprehensive summary of model fit to data.

In contrast, R_2 provides less information about goodness-of-fit. It quantifies the extent to which the variation of just one model outcome, perhaps the mean outcome for a range of parameter values, corresponds to the variation in the data set. R_2 does not capture model complexity, overfitting, or parameter uncertainty, which is a reason why R_2 by itself is not sufficient for model evaluation. Without accounting for model complexity and parameter count, focusing on optimizing fit by R_2 values alone can easily lead to overfitting.

Well-known examples of information criteria include the Akaike information criterion (AIC) and Deviance information criterion (DIC) (Gelman et al., 2014). However, these two metrics have some limitations. Neither AIC nor DIC use full sampled posterior distributions in their computations. Additionally, the original formulations of AIC and DIC are more limited and less stable in their ability to account for overfitting and parameter count (Gelman et al., 2014).

Two more recently developed metrics, the Widely Applicable information criterion (WAIC) and Leave-one-out cross-validation (LOO), address the stability and parameter count issues and improve upon AIC and DIC by using the full posterior distribution (Gelman et al., 2014; Vehtari et al., 2017). WAIC and LOO also estimate the relative potentials of models for fitting measurements not included within the existing observed data set. Thus, WAIC and LOO can be used as barometers for model predictive accuracy.

The overarching goal of this study was to develop a statistically rigorous and mathematically consistent data assimilation framework for SBM comparison that uses predictive Bayesian goodness-of-fit metrics. We pursued three specific objectives as part of that goal. First, we compared the behaviors of two different models, one linear and one non-linear, following data assimilation with soil respiration data. Second, we characterized the parameter spaces of these models using prior probability distributions of parameter values informed by previous studies and expert judgment. Third, we compared specific Bayesian predictive information criteria, including WAIC and LOO, to the coefficient of determination, R_2, for quantifying goodness-of-fit to data.

2 Methods

2.1 Model Structures

We analyzed the fit of two SBMs, the CON (conventional) and AWB (Allison-Wallenstein-Bradford) models (Allison et al., 2010). CON is a linear ordinary differential equation system, while AWB is a non-linear system (Supplemental Appendix 1). The models were chosen for this study due to their mathematical simplicity and limited data input requirements. Additionally, they were chosen because they are C-only models without nitrogen (N) pools. The increased complexity of N-accounting SBMs will require future studies with coupled N data sets (Manzoni and Porporato, 2009).

2.2 Meta-analysis Data

The data set was based on 27 soil warming studies that measured CO_2 fluxes and were compiled in a recent soil warming meta-analysis (Romero-Olivares et al., 2017). The experiments reported between 1 and 13 years of CO_2 flux measurements following warming perturbation. Models were fit to response ratios calculated by dividing CO_2 fluxes measured in the warming treatments by paired CO_2 fluxes measured in the control treatments. We calculated an annual mean response ratio for each experiment and each year available after treatment began. Using

https://doi.org/10.5194/bg-2020-23
Preprint. Discussion started: 19 February 2020

these annual means, we calculated one overall mean response ratio for each year along with pooled variances and standard deviations. Pooled data points were assumed to be "collected" at the halfway point of each year.

Because the experiments had variable lengths, the sample size for the pooled annual mean declines with increasing time since warming perturbation. The warming perturbation was 3°C on average across all the studies, and this average was used as the magnitude of warming in the model simulations. Model output response ratios were calculated by dividing simulated CO_2 flux following warming perturbation by the CO_2 flux at steady state.

We chose to fit the response ratios rather than raw flux measurements for several reasons. First, there is no need to convert flux measurements from different experiments into a common unit. Second, response ratios represent a standardized metric for warming response across disparate ecosystem types with varying climate, soil, and vegetation properties. Finally, fitting a mean response ratio overcomes data gaps present in individual experiments.

2.3 Markov Chain Monte Carlo Fitting

We performed model fitting using a Markov chain Monte Carlo (MCMC) algorithm called the Hamiltonian Monte Carlo (HMC), using version 2.17 of the RStan interface to the Stan statistical software (Carpenter et al., 2017; Guo et al., 2019) to collect posterior distributions and posterior predictive distributions. Posterior distributions are the distributions of more likely model parameter values conditional on the data. Posterior predictive distributions are the distributions of more likely values for unobserved data points from the data-generating process conditional on the observations. In the case of this study, the experiments constituting the meta-analysis would be the data-generating process.

Differential equation models contain parameters that affect state variables, and model-fitting through MCMC involves iterating through parameter space one set of parameters at a time. HMC is not a random walk algorithm and uses Hamiltonian mechanics to determine exploration steps in parameter space. HMC has been theorized to offer more efficient exploration of high-dimensional parameter space than traditional Random-Walk Metropolis algorithms (Beskos et al., 2013).

In the process of fitting and exploring parameter space with MCMCs, we obtained samples from the posterior distributions of parameter values. Bayesian inference is highly reliant on these distributions, as they provide information about probability densities for parameter values for a given data set. For each HMC run, we ran four chains for 45,000 iterations each, with the first 20,000 iterations being discarded as burn-in in each chain. Hence, our posterior distributions consisted of 100,000 posterior samples per HMC run. To minimize the presence of divergent energy transitions, which indicate issues with exploring the geometry of the parameter space specified by the prior distributions, we set the adaptation and step size HMC parameters respectively to 0.9995 and 0.001. These parameters control how the HMC algorithm proposes new sets of parameters at each step.

We further constrained our HMC runs to characterize parameter regimes corresponding to higher biological realism. Normal informative priors were used to initiate the runs, and the prior distribution parameters were chosen based on expert opinion and previous empirical observations (Allison et al., 2010; Li et al., 2014). Prior distributions had non-infinite supports; supports were truncated to prevent the HMC from exploring parameter space that was unrealistic (Supplemental Table 2).

2.4 Model Steady State Initialization

Because we were mainly interested in testing model predictions of soil warming response, the models were initiated at steady state prior to the introduction of warming perturbation to isolate model warming responses from steady state attraction. We fixed pre-perturbation steady state soil C densities to prevent HMC runs from exploring parameter regimes corresponding to biologically unrealistic C pool densities and mass ratios.

To set pre-warming steady state soil C densities, we first analytically derived steady state solutions of the ordinary differential equations of the models. Then, with the assistance of Mathematica version 12, we re-arranged the equations by moving the steady state pool sizes to the left-hand side (Supplemental Appendix 2), such that we could determine the value of parameters dependent on pool sizes while allowing the rest of the parameters to vary for the HMC. Consequently, we could constrain the pre-warming pool sizes from reaching unrealistic values in the simulations.

2.5 Sensitivity Analysis of C Pool Ratios

https://doi.org/10.5194/bg-2020-23
Preprint. Discussion started: 19 February 2020

Sensitivity analyses examine how the distributions of model input values influence the distributions of model outputs. In our study, we considered pre-warming C-pool densities as a model input. We performed a sensitivity analysis to observe how the choice of pre-warming C pool densities and C-pool ratios would affect the model fits and posterior predictive distribution of C pool ratios.

We compared the model outputs and post-warming response behavior of AWB and CON at equivalent C pool densities and ratios. The fraction of soil microbe biomass C (MIC) density to SOC density has been observed to vary approximately between 0.01 – 0.04 (Anderson and Domsch, 1989; Sparling, 1992), so we used those numbers as guidelines for establishing the ranges of the C pool densities and density ratios explored in our simulations. One portion of the analysis involved running HMC simulations in which we set the pre-warming MIC density at 2 mg C g^{-1} soil and then varied the SOC density from 50 to 200 mg C g^{-1} soil in increments of 25, stepping from 0.04 to 0.01 in terms of MIC-to-SOC fraction. A second portion of the analysis involved observing the effect of varying pre-warming MIC from 1 to 8 mg C g^{-1} soil while holding pre-warming SOC to 100 mg C g^{-1} soil.

For some combinations of the prior distributions and pre-warming steady state C pool densities (Supplemental Table 2), AWB HMC runs wandered into unstable parameter regimes that would prevent the algorithm from reliably running to completion. Consequently, we do not compare simulation results for AWB and CON with pre-warming SOC densities below 50 mg C g^{-1} soil. Other combinations of prior distribution and pre-warming C pool density choices that were not necessarily biologically realistic allowed stable AWB runs with lower pre-warming SOC densities.

2.6 Information Criteria and Cross-validation

In addition to R^2, we used the WAIC, LOO, and Log Pseudomarginal Likelihood (LPML) Bayesian predictive goodness-of-fit metrics to evaluate models with the meta-analysis warming response data. LPML is also an example of cross validation and is calculated similarly to LOO. However, LPML does not account for over-fitting or penalize for parameter count (Christensen et al., 2011). We used the 'loo' package available for R to calculate our WAIC and LOO values (Vehtari et al., 2017). A lower WAIC and LOO and a higher LPML indicate a more likely model for a given data set.

3 Results

3.1 Parameter Posterior Distributions

We obtained posterior parameter distributions and fits to the univariate response ratio data for both AWB and CON across different pre-warming MIC-to-SOC ratios. Sampler diagnostics for the HMC runs generally indicated convergence for the Markov chains and usable posteriors (Supplemental Fig 5 – 7). We also tracked divergent transitions that indicate the presence of regions of parameter space that are too geometrically confined and difficult to explore by the HMC. Divergent transitions occurred in the AWB HMC runs (Supplemental Fig 9), though the ratios of divergent transitions to sampled iterations was relatively low for all runs, with none exceeding 0.025. There were no divergent transitions in the CON runs. Effective sample proportion for parameters was generally satisfactory and greater than 0.3 for parameters across various MIC-to-SOC ratios, with total posterior sample sizes of 75,000 to 100,000 iterations (Supplemental Table 4).

3.2 Model Behaviors

The CON curve monotonically decreases in response ratio over time, whereas the AWB curve displays changes in slope sign (Fig 2). The difference in curve shape is in line with CON's linear system and AWB's non-linear formulation with more parameters (Allison et al., 2010). By 50 years after warming, mean fit curves for AWB and CON return to 1.0 after their initial increase (Fig 3c-d), consistent with prior observations and expectations at steady state (van Gestel et al., 2018; Romero-Olivares et al., 2017).

The 95% confidence interval of first the data point mean does not include the AWB mean, which could negatively impact AWB's quantitative goodness-of-fit and information criteria metrics. However, the 95% model response ratio credible interval suggests that AWB is able to replicate the trend of response ratio increase 1-3 years following warming perturbation, which CON does not. The mean AWB fit also matches the data points after eight years more closely than CON. Visually, though, it is not clear which model provides the better fit.

3.3 Sensitivity Analysis of Parameter Distributions to Pre-warming C Pool Densities and Density Ratios

https://doi.org/10.5194/bg-2020-23
Preprint. Discussion started: 19 February 2020

For both AWB and CON, higher pre-warming SOC corresponds to lower initial response ratio (Fig 3a-b). For CON, higher initial SOC reduces the magnitude of the mean fit slope and slows the return of the response curve to 1.0. For AWB, more time is needed to reach the peak response ratio and return to pre-warming response ratios. Changing the pre-warming MIC-to-SOC steady state pool size ratio by increasing MIC has a subtle effect on the fit curve; the magnitude and severity of slope changes decreases from MIC = 1 to MIC = 8 mg C g^{-1} soil (Supplemental Fig 1). Increasing MIC did not have an appreciable qualitative effect on CON fit.

In addition to response ratio fit, we observed the influence of pre-warming MIC-to-SOC ratios on fractional SOC loss for AWB and CON following warming. The fractional SOC loss at 12.5 years for CON and AWB decreased as pre-warming SOC was increased (and hence, MIC-to-SOC ratio decreased). For CON, SOC loss ranged from 27.1% at SOC = 50 to 9.2% at SOC = 200 (Supplemental Fig 3). For AWB, it ranged from 17.2% at SOC = 50 to 8.1% at SOC = 200. Similarly, AWB SOC loss decreased from 16.3% to 11.3% as MIC was reduced from 8 to 1. In contrast, the CON SOC loss increased from 17.4% to 18.8% when MIC was reduced from 8 to 1.

Truncation of prior supports, or distribution domains, generally did not prevent posterior densities from retaining normal distribution shapes. Deformation away from Gaussian shapes was observed at SOC = 50 mg C g^{-1} soil and SOC = 75 mg C g^{-1} soil for the densities of E_{aS} for CON and E_{aV}, E_{aK}, and E_{Cref} for AWB. All CON and AWB parameter posterior densities were otherwise observed to be Gaussian from SOC = 100 mg C g^{-1} soil to SOC = 200 mg C g^{-1} soil. Example posterior densities and means for select model parameters at pre-warming SOC = 100 mg C g^{-1} are plotted in Fig 4. Parameter posterior means corresponding to other pre-warming C pool densities and ratios are presented in Supplemental Table 3.

3.4 Sensitivity Analysis of Quantitative Fit Metrics to Pre-warming C Pool Densities and Density Ratios

Fit metrics generally worsened as pre-warming steady state SOC increased for both CON and AWB (Fig 5). However, LOO, WAIC, and R^2 agree that fit quantitatively improved from SOC = 50 to SOC = 75, with LOO and WAIC suggesting a more pronounced improvement in fit than R^2 due to overfitting penalties (Supplemental Fig 8). From SOC = 50 to 75, LOO improved from -5.04 to -6.23, and WAIC improved from -5.73 to -9.85. LOO, WAIC, LPML, and R^2 unanimously agree on trends of worsening fit quality from SOC = 125 to SOC = 200.

Varying pre-warming steady state MIC appeared to slightly reduce fit quality across the various metrics as MIC ranged from 1 to 8 mg C g^{-1} soil (Supplemental Fig 4), though the trend was not consistent in LOO and WAIC. Since increasing MIC has the opposite effect on MIC-to-SOC ratio compared to increasing SOC, these results indicate no consistent effect for absolute changes to MIC-to-SOC ratio.

4 Discussion

Our study develops a quantitative, data-driven framework for model comparison that could be applied across different research questions, ecosystems, and scales. We demonstrated the novel deployment of WAIC and LOO, two more recently developed Bayesian goodness-of-fit metrics that estimate model predictive accuracy, to evaluate SBMs using data from longitudinal soil warming experiments. WAIC and LOO improve upon older and more frequently used metrics, such as AIC and DIC, by accounting for model complexity and overfitting of data in a more comprehensive, stable, and accurate fashion.

We constrained the fitting of AWB and CON to biologically reasonable parameter space by fixing pre-warming steady state C pool densities and establishing prior distributions informed by expert judgment (Supplemental Table 2). We observed that CON and AWB can both explain the soil response to warming in the meta-analysis data set (Fig 2) and that certain pre-warming soil C densities and density ratios for SOC and MIC correspond to better warming response fits (Fig 5).

4.1 Model Responses to Warming over Time

CON and AWB both displayed similar general trends in the progression of their response ratio curves following soil warming (Fig 2). The return of the curves to their pre-warming steady states aligns with previous literature which demonstrates that the magnitude of CO_2 flux falls following a post-warming peak (Crowther et al., 2016; Romero-Olivares et al., 2017).

AWB, unlike CON, displays oscillations in its response ratios following warming due to its non-linear dynamics. However, it is unclear whether oscillations quantitatively aid AWB with its fit to our response ratio data

https://doi.org/10.5194/bg-2020-23
Preprint. Discussion started: 19 February 2020

Biogeosciences Discussions
Open Access

set. The presence of respiration oscillations has been observed in long-term warming experiments, such as the one taking place at Harvard Forest (Melillo et al., 2017). It is possible AWB would be quantitatively rewarded in goodness-of-fit metrics over CON for its ability to replicate oscillations in site-specific data sets such as those from Harvard Forest.

For an additional check on model realism, we tallied SOC loss percentages from pre-warming SOC stocks after 12.5 years for AWB and CON. SOC losses ranged from 8.14% to 27.1% across both models (Supplemental Fig 3). These results aligned with a recent comprehensive meta-analysis of 143 soil warming studies (Supplemental Fig 10). The largest loss of 27.1%, occurring in CON at SOC = 50, is sizable, but the van Gestel et al. meta-analysis included 7 studies measuring losses greater than 20%, with the maximum loss observed at 54.4% (van Gestel et al., 2018).

For both AWB and CON, increasing pre-warming SOC reduced C loss fraction following the perturbation. Varying pre-warming MIC more prominently affected the fraction of SOC lost from AWB compared to CON, with soil C loss increasing as MIC increased. In CON's case, there was a minimal decline in SOC loss as MIC was increased. The larger effect of increasing MIC on the fraction of SOC lost in AWB is likely due to MIC influence on SOC-to-DOC turnover, which is not a feedback included in the CON model.

4.2 Sensitivity Analysis of C Pool Densities and Density Ratios

We performed a sensitivity analysis to check whether the response ratio trends stayed consistent, biologically realistic, and interpretable across a range of pre-warming, steady state soil C densities and pool-to-pool density ratios. For instance, we imposed constraints to reflect that MIC-to-SOC density ratios range between 0.01 and 0.04 across various soil types (Anderson and Domsch, 1989; Sparling, 1992). CON and AWB response ratio curves exhibited realistic values and qualitatively consistent shapes across all pre-warming SOC and MIC steady state densities, even at less realistic SOC densities above 100 mg C g^{-1} soil (Fig 3). There was enough uncertainty in the data that the 95% posterior predictive intervals for the model output always overlapped with the 95% confidence intervals of each fitted data point (Fig 2). In most cases, the posterior mean response ratio curve also fell within the 95% data confidence interval.

We were unable to initiate our pre-warming SOC steady state below 50 mg SOC g^{-1} soil with the priors and MIC-to-SOC ratios used for AWB. Under 50 mg SOC g^{-1} soil, AWB HMC runs would not reliably run to conclusion and would terminate due to ODE instabilities. Even at 50 mg SOC g^{-1} soil, we saw a reduction in independent and effective samples for certain parameters, namely E_{aV} and E_{aK} (Supplementary Table 13). We did not drop under 50 mg SOC g^{-1} soil for CON, as we sought to compare AWB and CON at similar MIC-to-SOC ranges. Similarly, we were unable to drop our pre-warming MIC steady state below 1 mg SOC g^{-1} soil. Our experience underscores the challenge of choosing realistic steady state soil C densities, density ratios, and prior distributions to obtain valid model comparisons limited to biologically realistic regimes.

The information criteria and cross-validation fit metrics generally indicated higher relative probability and predictive performance for the data at lower pre-warming SOC values for AWB and CON (Fig 5). The fit results suggest that SOC density of the soil at the sites included in the meta-analysis was likely closer to the lower end of the SOC density ranges examined in our sensitivity analysis. A less pronounced trend toward better fits was observed as pre-warming MIC density was decreased while pre-warming SOC density was held constant (Supplemental Fig 4). No clear relationship was observed between MIC-to-SOC ratio and goodness-of-fit in the AWB and CON models.

The worsening IC and CV results at higher SOC densities support the notion that pre-warming steady state soil C densities should not be initialized over 100 mg C g^{-1} soil in AWB and CON when fitting to this meta-analysis data set. The majority of the CO_2 respired by soil microbes is sourced from surface soil (Fang and Moncrieff, 2005), and it is well-documented that the highest SOC densities are in the top 20 centimeters of soil (Jobbágy and Jackson, 2000). Pre-warming SOC density was not observed to exceed 50 mg SOC g^{-1} soil at sites included in the meta-analysis, reaching a maximum of 45 mg SOC g^{-1} soil for the top 20 cm in one study with alpine wetland soil (Zhang et al., 2014). ^{14}C measurements of CO_2 fluxes suggest that SOC densities representing the source of most heterotrophic respiration in topsoil range between 40 to 80 mg SOC g^{-1} soil (Trumbore, 2000).

4.3 Parameter Space Exploration

Truncating prior and posterior parameter distributions proved useful for establishing biological constraints and modestly deformed posterior densities for AWB and CON. From pre-warming SOC = 100 to SOC = 200, CON and AWB posterior densities showed little or no deformation from typical normal distribution shapes. Moderate posterior density deformation was observed for some parameters in both models at SOC = 50 and 75 (E_{Cref} for AWB and Ea_S for CON). Even so, most of the other parameter posterior densities still remained undeformed at those SOC values. Thus, prior truncation generally did not prevent posterior means from falling within biologically realistic intervals, suggesting that priors were appropriately informed and chosen.

A small frequency of divergent transitions was detected for the AWB HMC runs. A more thorough description of the theory, computation, and implications of divergent transitions can be found in literature focusing on the Hamiltonian Monte Carlo algorithm (Betancourt, 2016, 2017). The number of divergent transitions generally increased as the pre-warming MIC-to-SOC steady state ratio was reduced (Supplemental Fig 9). Prior truncation and the fixing of select parameters to constrain the pre-warming steady state mass values for biological realism could have played a combined role in generating the Markov chain divergences by hindering the smooth exploration of parameter space. We were unable to eliminate divergent transitions by adjusting HMC parameter proposal step size, suggesting that other methods, such as modification of the HMC algorithm itself or introduction of auxiliary parameters to AWB that reduce correlation between existing model parameters may be more applicable in reducing divergent transitions in our case (Betancourt and Girolami, 2015). Additionally, the interaction between the ranges of values used for the prior distributions and the limited number of observations in the data set could have contributed to the shaping of geometric inefficiencies (Betancourt, 2017).

4.4 Applying and Interpreting Bayesian Predictive Fit Metrics

With respect to the IC and CV metrics, in both Fig 5 and Supplementary Fig 5, there is disagreement between LOO and WAIC versus LPML. LPML displays more consistent trends for CON and AWB across the range of pre-warming SOC values with a unidirectional change in slope. LPML is calculated similarly to LOO but does not account for overfitting and parameter count (Gelfand and Dey, 1994; Gelman et al., 2014). The computational difference accounts for the divergence between the results of LPML and those of LOO and WAIC. The effective parameter count and penalty for overfitting in both the WAIC and LOO calculations generally increases as pre-warming SOC is reduced (Supplemental Fig 8a and 8b). Thus, while the LPML results appear clearer, we do not recommend use of LPML by itself to quantitatively compare model fits because it does not fully account for the impacts of differing model structure, parameterization, and parameter count on overfitting for a data set.

General agreement between WAIC, LOO, and LPML reinforces the usage of IC and CV metrics alongside usage of R_2. R_2 is not suitable as sole quantitative metric for model evaluation and selection. The traditional unadjusted R_2 calculation does not have a cost function for parameter counts. R_2 estimates the strength of the relationship between a linear model and a dependent variable and is calculated from the variance in data and residuals separating model outputs from observations. The metric cannot be applied to nonlinear models. Model selection involves a relative comparison of models, but the value of R_2 can result in misleading conclusions regarding absolute goodness of fit of a model to data. For instance, a model appropriate for a data set can correspond to a low R_2 calculation, while a flawed model can correspond to a high R_2 (Spiess and Neumeyer, 2010). Adjusted R_2 accounts for model parameter count, but still shares other pitfalls with non-adjusted R_2.

4.5 Conclusion and Future Directions

Recent SBM comparisons have been unable to demonstrate the superiority of one model over another because the uncertainty boundaries of the data were not sufficient for distinguishing model outcomes (Sulman et al., 2018; Wieder et al., 2018). Similar to Sulman et al., our results indicate that more data is needed to constrain model outputs and to verify the strengths and limitations of linear versus non-linear SBMs in Earth system modeling.

Consequently, future SBM comparisons would benefit from additional data collection efforts sourced from long-term ecological research experiments. The limited number of longitudinal soil warming studies presents a challenge for facilitating site-specific model comparisons. We addressed this issue by using meta-analysis data to aggregate warming responses across sites, but this approach does not provide site-specific parameters. Additional data from ongoing and future field warming studies in the vein of the Harvard Forest and Tropical Responses to Altered Climate experiments will be of critical importance for model testing (Melillo et al., 2017; Wood et al., 2019). Model parameters could also be better constrained through the use of multivariate data sets, for example microbial biomass dynamics in addition to soil respiration.

Our approach can also be used to compare the predictive accuracy of linear models that only implicitly represent microbial activity to that of more complex non-linear SBMs that explicitly represent the Michaelis-Menten dynamics of soil microbial processes, such as CORPSE (Sulman et al., 2014) and MIMICS (Wieder et al., 2015).

Such comparisons will help determine if inclusion of more detailed microbial dynamics in models offers predictive advantages that can overcome the overfitting burdens associated with an increase in parameter count.

Despite limited data availability, the development of our formalized, statistically rigorous approach for model comparison and evaluation is a critical step toward the goal of improving the forecasting of global SOC levels and soil emissions through the rest of the 21st century. Our initial results indicate promise in continued development of our approach to better evaluate a range of models that vary widely in structure and parameter count.

Code and Data Availability

The R scripts, Stan code, and respiration data set used for HMC model fitting along with the original soil respiration meta-analysis data set (Romero-Olivares et al., 2017) are available from the directory located at https://osf.io/7mey8/?view_only=af1d54f858c34c41ab4854551d015896 (Xie et al., 2019).

Author contribution

SDA and HWX designed the study with assistance from MG. HWX and ALR performed the data cleaning and analysis. HWX wrote the necessary code for the study with assistance from SDA. SDA and HWX prepared the paper with suggestions from MG.

Competing interests

The authors declare they have no conflict of interest.

Acknowledgments

We would like to thank Stan development team members Bob Carpenter (Columbia University) and Michael Betancourt for their patient help with Stan code implementation and troubleshooting.

Financial support

This research was supported by funds from The U.S. Department of Energy Office of Science BER-TES program under grant DESC0014374 and the National Institutes of Health T32 Training Program under grant EB009418.

References

Allison, S. D., Wallenstein, M. D. and Bradford, M. A.: Soil-carbon response to warming dependent on microbial physiology, Nat. Geosci., 3(5), 336–340, doi:10.1038/ngeo846, 2010.

Anderson, T.-H. and Domsch, K. H.: Ratios of microbial biomass carbon to total organic carbon in arable soils, Soil Biol. Biochem., 21(4), 471–479 [online] Available from: papers2://publication/uuid/90F0F273-FB35-4093-B8BB-67E2429E94BF, 1989.

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J. M. and Stuart, A.: Optimal tuning of the hybrid Monte Carlo algorithm, Bernoulli, 19(5 A), 1501–1534, doi:10.3150/12-BEJ414, 2013.

Betancourt, M.: Diagnosing Suboptimal Cotangent Disintegrations in Hamiltonian Monte Carlo, [online] Available from: http://arxiv.org/abs/1604.00695, 2016.

Betancourt, M.: A Conceptual Introduction to Hamiltonian Monte Carlo, [online] Available from: http://arxiv.org/abs/1701.02434, 2017.

Betancourt, M. and Girolami, M.: Hamiltonian Monte Carlo for Hierarchical Models, Curr. Trends Bayesian

https://doi.org/10.5194/bg-2020-23
Preprint. Discussion started: 19 February 2020

Methodol. with Appl., 79–101, doi:10.1201/b18502-5, 2015.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P. and Riddell, A.: Stan: A probabilistic programming language, J. Stat. Softw., 76(1), doi:10.18637/jss.v076.i01, 2017.

Christensen, R., Johnson, W., Branscum, A. and Hanson, T. E.: Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians., 2011.

Crowther, T. W., Todd-Brown, K. E. O., Rowe, C. W., Wieder, W. R., Carey, J. C., MacHmuller, M. B., Snoek, B. L., Fang, S., Zhou, G., Allison, S. D., Blair, J. M., Bridgham, S. D., Burton, A. J., Carrillo, Y., Reich, P. B., Clark, J. S., Classen, A. T., Dijkstra, F. A., Elberling, B., Emmett, B. A., Estiarte, M., Frey, S. D., Guo, J., Harte, J., Jiang, L., Johnson, B. R., Kroël-Dulay, G., Larsen, K. S., Laudon, H., Lavallee, J. M., Luo, Y., Lupascu, M., Ma, L. N., Marhan, S., Michelsen, A., Mohan, J., Niu, S., Pendall, E., Peñuelas, J., Pfeifer-Meister, L., Poll, C., Reinsch, S., Reynolds, L. L., Schmidt, I. K., Sistla, S., Sokol, N. W., Templer, P. H., Treseder, K. K., Welker, J. M. and Bradford, M. A.: Quantifying global soil carbon losses in response to warming, Nature, 540(7631), 104–108, doi:10.1038/nature20150, 2016.

Fang, C. and Moncrieff, J. B.: The variation of soil microbial respiration with depth in relation to soil carbon composition, Plant Soil, 268(1), 243–253, doi:10.1007/s11104-004-0278-4, 2005.

Gelfand, A. E. and Dey, D. K.: Bayesian Model Choice : Asymptotics and Exact Calculations Author (s): A . E . Gelfand and D . K . Dey Published by : Blackwell Publishing for the Royal Statistical Society Stable URL : http://www.jstor.org/stable/2346123, J. R. Stat. Soc. Ser. B, 56(3), 501–514, 1994.

Gelman, A.: Conservative prior distributions for variance parameters in hierarchical models, Bayesian Anal., 1(3), 515–533, doi:10.1002/cjs.5550340302, 2006.

Gelman, A., Hwang, J. and Vehtari, A.: Understanding predictive information criteria for Bayesian models, Stat. Comput., 24(6), 997–1016, doi:10.1007/s11222-013-9416-2, 2014.

van Gestel, N., Shi, Z., van Groenigen, K. J., Osenberg, C. W., Andresen, L. C., Dukes, J. S., Hovenden, M. J., Luo, Y., Michelsen, A., Pendall, E., Reich, P. B., Schuur, E. A. G. and Hungate, B. A.: Predicting soil carbon loss with warming, Nature, 554(7693), E4–E5, doi:10.1038/nature25745, 2018.

Guo, J., Gabry, J. and Goodrich, B.: RStan: the R interface to Stan, 2019.

Jiang, L., Yan, Y., Hararuk, O., Mikle, N., Xia, J., Shi, Z., Tjiputra, J., Wu, T. and Luo, Y.: Scale-dependent performance of CMIP5 earth system models in simulating terrestrial vegetation carbon, J. Clim., 28(13), 5217–5232, doi:10.1175/JCLI-D-14-00270.1, 2015.

Jobbágy, E. and Jackson, R. B.: The Vertical Distribution of Soil Organic Carbon and Its Relation to Climate and Vegetation, Ecol. Appl., 10(April), 423–436, doi:Doi 10.2307/2641104, 2000.

Li, J., Wang, G., Allison, S. D., Mayes, M. A. and Luo, Y.: Soil carbon sensitivity to temperature and carbon use

https://doi.org/10.5194/bg-2020-23
Preprint. Discussion started: 19 February 2020

efficiency compared across microbial-ecosystem models of varying complexity, Biogeochemistry, 119(1–3), 67–84, doi:10.1007/s10533-013-9948-8, 2014.

Luo, Y., Ahlström, A., Allison, S. D., Batjes, N. H., Brovkin, V., Carvalhais, N., Chappell, A., Ciais, P., Davidson, E. A., Finzi, A., Georgiou, K., Guenet, B., Hararuk, O., Harden, J. W., He, Y., Hopkins, F., Jiang, L., Koven, C., Jackson, R. B., Jones, C. D., Lara, M. J., Liang, J., McGuire, A. D., Parton, W., Peng, C., Randerson, J. T., Salazar, A., Sierra, C. A., Smith, M. J., Tian, H., Todd-Brown, K. E. O., Torn, M. T., van Groenigen, K. J., Wang, Y. P., West, T. O., Wei, Y. W., Wieder, W. R., Xia, J., Xu, X., Xu, X. and Zhou, T.: Towards more realistic projections of soil carbon dynamics by earth system models, , (February), 40–56, doi:10.1002/2015GB005239.Received, 2016.

Manzoni, S. and Porporato, A.: Soil carbon and nitrogen mineralization: Theory and models across scales, Soil Biol. Biochem., 41(7), 1355–1379, doi:10.1016/j.soilbio.2009.02.031, 2009.

Melillo, J. M., Frey, S. D., DeAngelis, K. M., Werner, W. J., Bernard, M. J., Bowles, F. P., Pold, G., Knorr, M. A. and Grandy, A. S.: Long-term pattern and magnitude of soil carbon feedback to the climate system in a warming world, Science (80-.)., 358(6359), 101–105, doi:10.1126/science.aan2874, 2017.

Romero-Olivares, A. L., Allison, S. D. and Treseder, K. K.: Soil microbes and their response to experimental warming over time: A meta-analysis of field studies, Soil Biol. Biochem., 107, 32–40, doi:10.1016/j.soilbio.2016.12.026, 2017.

Sparling, G. P.: Ratio of microbial biomass carbon to soil organic carbon as a sensitive indicator of changes in soil organic matter, Aust. J. Soil Res., 30(2), 195–207, doi:10.1071/SR9920195, 1992.

Spiess, A. N. and Neumeyer, N.: An evaluation of R2as an inadequate measure for nonlinear models in pharmacological and biochemical research: A Monte Carlo approach, BMC Pharmacol., 10, 1–11, doi:10.1186/1471-2210-10-6, 2010.

Sulman, B. N., Phillips, R. P., Oishi, A. C., Shevliakova, E. and Pacala, S. W.: Microbe-driven turnover offsets mineral-mediated storage of soil carbon under elevated CO 2, Nat. Clim. Chang., 4(12), 1099–1102, doi:10.1038/nclimate2436, 2014.

Sulman, B. N., Moore, J. A. M., Abramoff, R., Averill, C., Kivlin, S., Georgiou, K., Sridhar, B., Hartman, M. D., Wang, G., Wieder, W. R., Bradford, M. A., Luo, Y., Mayes, M. A., Morrison, E., Riley, W. J., Salazar, A., Schimel, J. P., Tang, J. and Classen, A. T.: Multiple models and experiments underscore large uncertainty in soil carbon dynamics, Biogeochemistry, 141(2), 109–123, doi:10.1007/s10533-018-0509-z, 2018.

Todd-Brown, K. E. O., Randerson, J. T., Hopkins, F., Arora, V., Hajima, T., Jones, C., Shevliakova, E., Tjiputra, J., Volodin, E., Wu, T., Zhang, Q. and Allison, S. D.: Changes in soil organic carbon storage predicted by Earth system models during the 21st century, Biogeosciences, 11(8), 2341–2356, doi:10.5194/bg-11-2341-2014, 2014.

Trumbore, S.: Age of soil organic matter and soil respiration: Radiocarbon constraints on belowground C dynamics, Ecol. Appl., 10(2), 399–411, doi:10.1890/1051-0761(2000)010[0399:AOSOMA]2.0.CO;2, 2000.

Vehtari, A., Gelman, A. and Gabry, J.: Practical Bayesian model evaluation using leave-one-out cross-validation and

WAIC, Stat. Comput., 27(5), 1413–1432, doi:10.1007/s11222-016-9696-4, 2017.

Wieder, W. R., Allison, S. D., Davidson, E. A., Georgiou, K., Hararuk, O., He, Y., Hopkins, F., Luo, Y., Smith, M. J., Sulman, B., Todd-Brown, K., Wang, Y. P., Xia, J. and Xu, X.: Explicitly representing soil microbial processes in Earth system models, Global Biogeochem. Cycles, 29(10), 1782–1800, doi:10.1002/2015GB005188, 2015.

Wieder, W. R., Hartman, M. D., Sulman, B. N., Wang, Y. P., Koven, C. D. and Bonan, G. B.: Carbon cycle confidence and uncertainty: Exploring variation among soil biogeochemical models, Glob. Chang. Biol., 24(4), 1563–1579, doi:10.1111/gcb.13979, 2018.

Wood, T. E., González, G., Silver, W. L., Reed, S. C. and Cavaleri, M. A.: On the shoulders of giants: Continuing the legacy of large-scale ecosystem manipulation experiments in Puerto Rico, Forests, 10(3), 1–18, doi:10.3390/f10030210, 2019.

Xie, H. W., Romero-Olivares, A. L., Treseder, K. K. and Allison, S. D.: A Bayesian Approach to Evaluation of Soil Biogeochemical Models R and Stan Code, doi:10.17605/OSF.IO/7MEY8, 2019.

Zhang, B., Chen, S., He, X., Liu, W., Zhao, Q., Zhao, L. and Tian, C.: Responses of soil microbial communities to experimental warming in alpine grasslands on the Qinghai-Tibet Plateau, PLoS One, 9(8), doi:10.1371/journal.pone.0103859, 2014.

https://doi.org/10.5194/bg-2020-23
Preprint. Discussion started: 19 February 2020

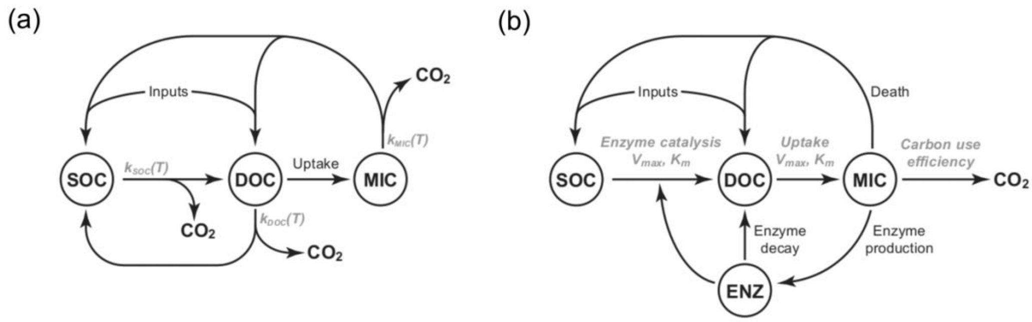


Figure 1: Diagrams of the pool structures of the **(a)** CON model; and **(b)** AWB model. Pools are shown within circles including soil organic carbon (SOC), dissolved organic carbon (DOC), and microbial (MIC) pools. AWB has SOC, DOC, and MIC pools as in CON, but also an extra enzymatic (ENZ) pool. AWB additionally differs from CON in its non-linear feedbacks and assumption that MIC can influence SOC-to-DOC turnover through the ENZ pool.

https://doi.org/10.5194/bg-2020-23
Preprint. Discussion started: 19 February 2020

Biogeosciences
Discussions
Open Access

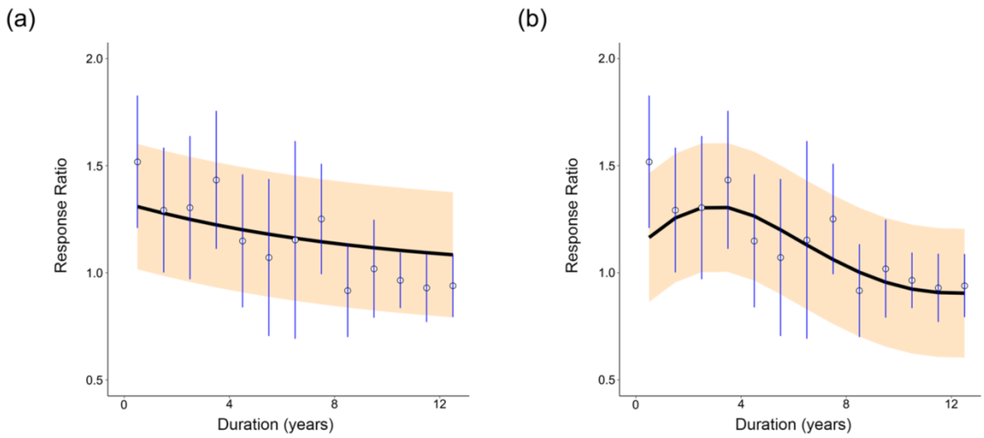


Figure 2: Distribution of fits of **(a)** CON; and **(b)** AWB to the meta-analysis data from Romero-Olivares et al., 2017. Open circles show the meta-analysis data points. Blue vertical lines mark the 95% confidence interval for each data point calculated from the pooled standard deviation. The black line indicates the mean (and median) model response ratio fit. The orange shading marks the 95% posterior predictive interval for the fit. For **(a)**, pre-warming steady state soil C densities were set at SOC = 100 mg C g$_{-1}$ soil, MIC = 2 mg C g$_{-1}$ soil, DOC = 0.2 mg C g$_{-1}$ soil. For **(b)**, pre-warming steady state soil C densities were set at SOC = 100 mg C g$_{-1}$ soil, MIC = 2 mg C g$_{-1}$ soil, DOC = 0.2 mg C g$_{-1}$ soil, and ENZ = 0.1 mg C g$_{-1}$ soil.

https://doi.org/10.5194/bg-2020-23
Preprint. Discussion started: 19 February 2020

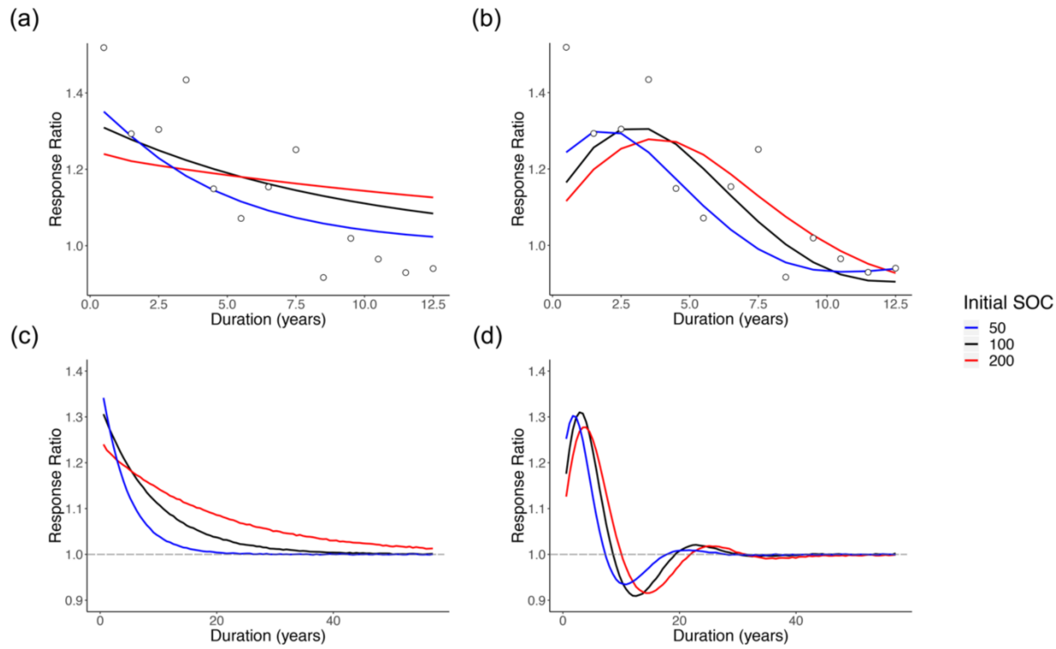


Figure 3: Intra-model comparisons of mean posterior predictive response ratio fits for AWB and CON across different MIC-to-SOC ratios. Open circles show the meta-analysis data points for reference. The blue, black, and red lines indicate model mean fits corresponding to different pre-warming-perturbation steady state SOC values of 50 mg C g^{-1} soil, 100 mg C g^{-1} soil, and 200 mg C g^{-1} soil. The dashed gray line indicates the steady state expectation at the response ratio of 1.0. Mean fits are plotted in order of **(a)** CON; and **(b)** AWB over the time span of the data and **(c)** CON; and **(d)** AWB over 57 years.

https://doi.org/10.5194/bg-2020-23
Preprint. Discussion started: 19 February 2020

Biogeosciences Discussions
Open Access
EGU

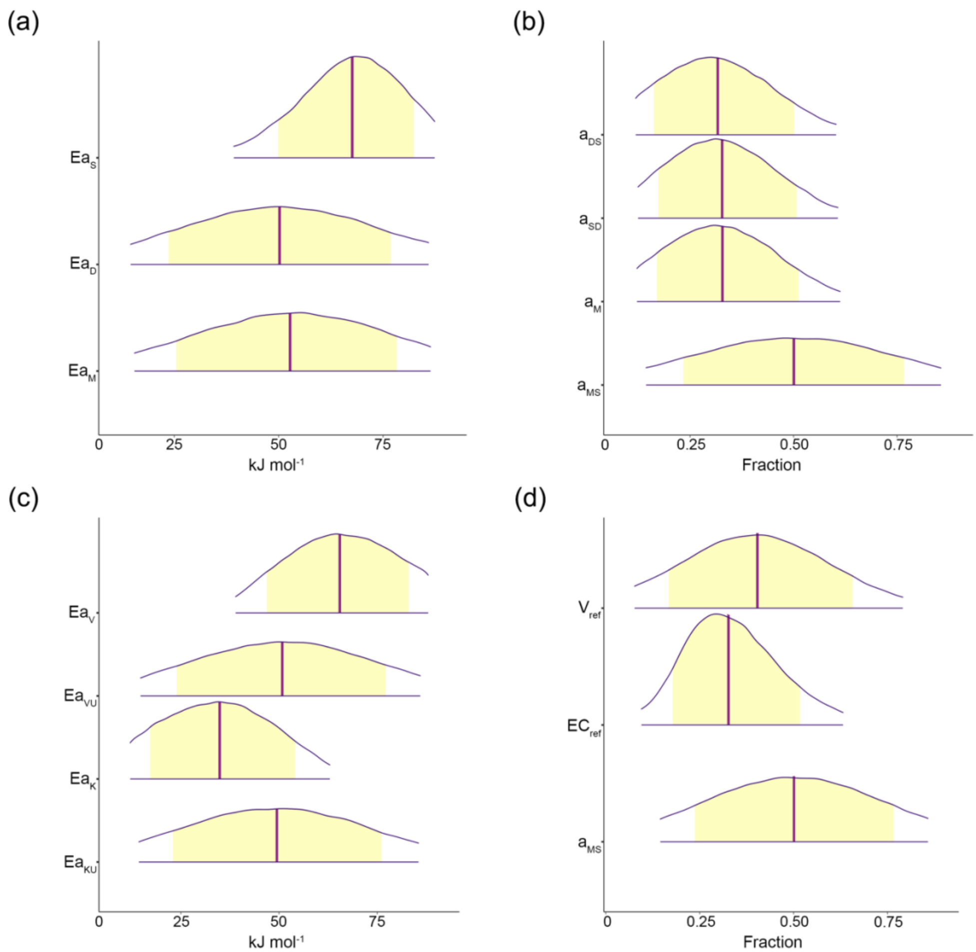
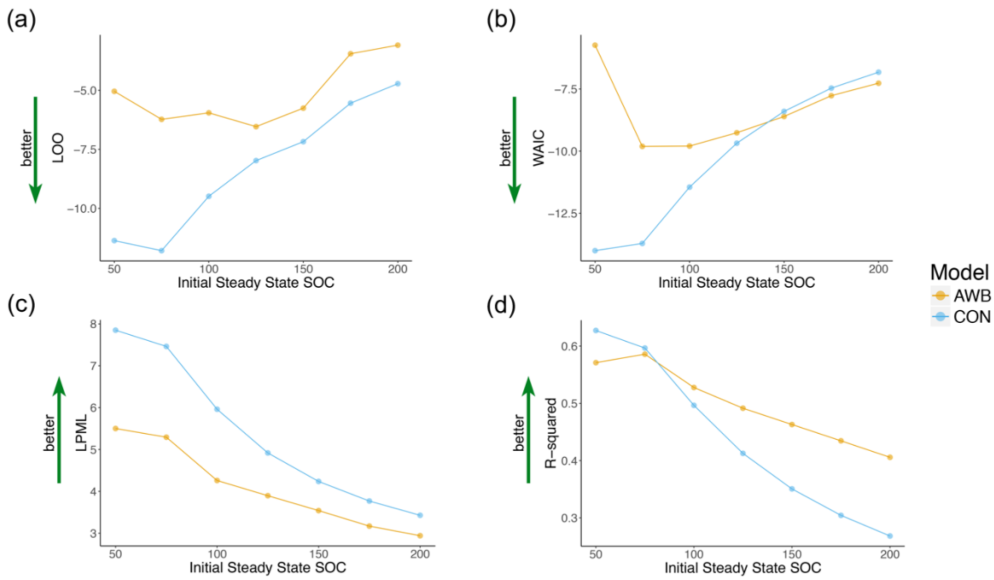


Figure 4: 95% credible areas for model parameters corresponding to pre-warming steady state SOC = 100 mg C g^{-1} soil, DOC = 0.2 mg C g^{-1} soil, MIC = 2 mg C g^{-1} soil, and (for AWB) ENZ = 0.1 mg C g^{-1} soil. Yellow shaded regions represent 80% credible areas and vertical purple lines indicate distribution mean. **(a)** CON activation energy parameters E_{aS}, E_{aD}, E_{aM}; **(b)** CON C pool partition fraction parameters a_{DS}, a_{SD}, a_{M}, and a_{MS}; **(c)** AWB activation energy parameters Ea_{V}, Ea_{VU}, Ea_{K}, Ea_{KU}; **(d)** AWB parameters V_{ref}, E_{Cref}, and a_{MS}. V_{ref} is the SOC V_{max} at the reference temperature 283.15 K, E_{Cref} is the carbon use efficiency fraction at the reference temperature, and a_{MS} is the fraction parameter representing the proportion of dead microbial biomass C transferred to the SOC pool. Credible areas for AWB parameters V_{Uref} and m_{t} are shown in Supplemental Fig 2 because of differing horizontal axes scales.

https://doi.org/10.5194/bg-2020-23
Preprint. Discussion started: 19 February 2020

Biogeosciences
Discussions
Open Access



526

528 **Figure 5:** Fit metric versus initial steady state SOC for AWB and CON models for **(a)** LOO; **(b)** WAIC cross-validation; **(c)** LPML; and **(d),** R_2 values. Pre-perturbation steady state MIC, DOC, and ENZ (for AWB) is held constant as pre-perturbation SOC is varied.