We are grateful to the reviewer for the time and effort he, she, or they has spent on their feedback and suggestions. The comments are insightful and helpful. We address specific reviewer comments below.

> L 86-91: lacks model descriptions; and nitrogen-related increases in complexity has not been addressed in the entire manuscript. I suggest you may discuss in the discussion. Also, figure 1 has not been mentioned in the manuscript

We will describe the model structures in greater detail in the methods section of the paper and add a reference to Figure 1. Models representing N-cycling in addition to C-cycling will be addressed in a revision of the discussion section. We will also consider future Bayesian model comparisons involving N-cycling and other more complex soil biogeochemical models in a new discussion paragraph. Potential challenges arising in future model comparisons of more complex models, such as the difficulty of solving for steady states in some non-linear systems and increased HMC simulation time, will be detailed.

> L 166: Log Pseudomarginal Likelihood (LPML) has popped up without prior introduction

Log Pseudomarginal Likelihood (LPML) is a cross-validation metric that is calculated similarly to Leave-one-out (LOO). As described in the manuscript, LPML does not calculate effective parameter count, which was introduced with LOO. Hence, it was included in the study to serve as a comparison to LOO and underscore the goodness-of-fit calculation without penalizing for overfitting. To make the introduction of LPML less jarring, we will mention it in the paragraph located in the introduction section that provides background on LOO and WAIC.

> L 185: The difference in curve shape (Fig. 3a, b)

We will fix this figure annotation and related ones referring to more than one figure pane.

> L 189: Is it different between 95% confidence interval and 95% model response ratio credible interval?

“Credible interval” should actually be changed to say “posterior predictive interval.” That will be fixed in the revision. The 95% confidence interval and 95% posterior predictive interval are indeed different. In this case, the confidence interval is a frequentist measure corresponding to estimates of the true value of the flux data measurement, while the Bayesian posterior predictive interval indicates the range of model predictions for mean flux response ratio.

> L 196: a bit confused as well as missing figure annotation. It would be better to choose clear points to address why CON and AWB are showing differences

We will revise this paragraph to include the additional appropriate figure citations. The paragraph will be re-structured and re-worded to compare model outputs and data at specific time points.

> L 198: rewording to emphasize how the steady state pool size ratio has been changed-based on increasing MIC; the unit should be mg C g-1(uppercase); Please check other lines as well

We will re-word more simply to emphasize that the pre-warming SOC-to-MIC ratio was changed by increasing MIC while holding SOC steady. g^{-1} and other inverse units show up appropriately as exponents in the Microsoft Word .doc file, but regrettably do not appear as exponents following the *Biogeosciences Discuss* PDF conversion. We will try uploading the file with some other steps to try to fix the exponent issue with our subsequent revision.

> L 199: need to clarify. By the way, what is the function of the trend lines? Have you tried polynomial function? It seems similar patterns between them.

The intent of the posterior predictive mean fit trend lines in Supp. Fig. 1 was to show the moderate, but consistent effect of increasing pre-warming MIC on the AWB model output slopes. In our revision, we can emphasize that the influence of pre-warming MIC on slope magnitude is limited to the AWB model. My writing was not clear for this paragraph. We can also edit the paragraph to remove reporting of the influence of MIC on AWB slope and streamline this section, as this result is not central to the manuscript.

By polynomial function, I assume you mean the polynomial analytic function

$$f(x) = \sum_{n=0}^{\infty} c_n (x)^n = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + ...$$

We could exactly fit the data using a polynomial analytic function with sufficiently many $c_n x^n$ terms, but the goal of the manuscript was not to find the best-fitting arbitrary model. Instead, we sought to demonstrate the feasibility of rigorous Bayesian model comparisons for more complex dynamical ODE systems with the hope that further conclusions regarding model mechanisms and structure could be made following future model comparison results. Fitting analytic polynomial model would not be able to contribute feedback towards the refinement or rejection of elements of dynamical model structure or formulation. Additionally, the polynomial model would be prone to overfitting and effective parameter count penalties from the LOO and WAIC computation.

> L 203-206: Is it possible to replace the supplemental figure 3 to represent SOC loss rather than SOC fraction remaining? It is difficult to interpret.

We wanted the y-axis of Sup. Fig. 3 to align with Sup. Fig. 10 to show that the SOC remaining in our models was in the range of change of SOC observed in soil warming experiments across various soil types. This indicates that our models were within the realm of biological realism. However, we can see that the labeling would make Sup. Fig. 3 confusing. We do feel that changing Sup. Fig. 3 to show an absolute level of SOC loss in terms of density would make it less useful for direct comparison to Sup. Fig. 10. The fraction change is what we want to observe; soils vary dramatically in SOC concentration and the fraction of change in SOC provides more information about the biological realism of the model following preturbation from initial conditions than the change in absolute SOC density.

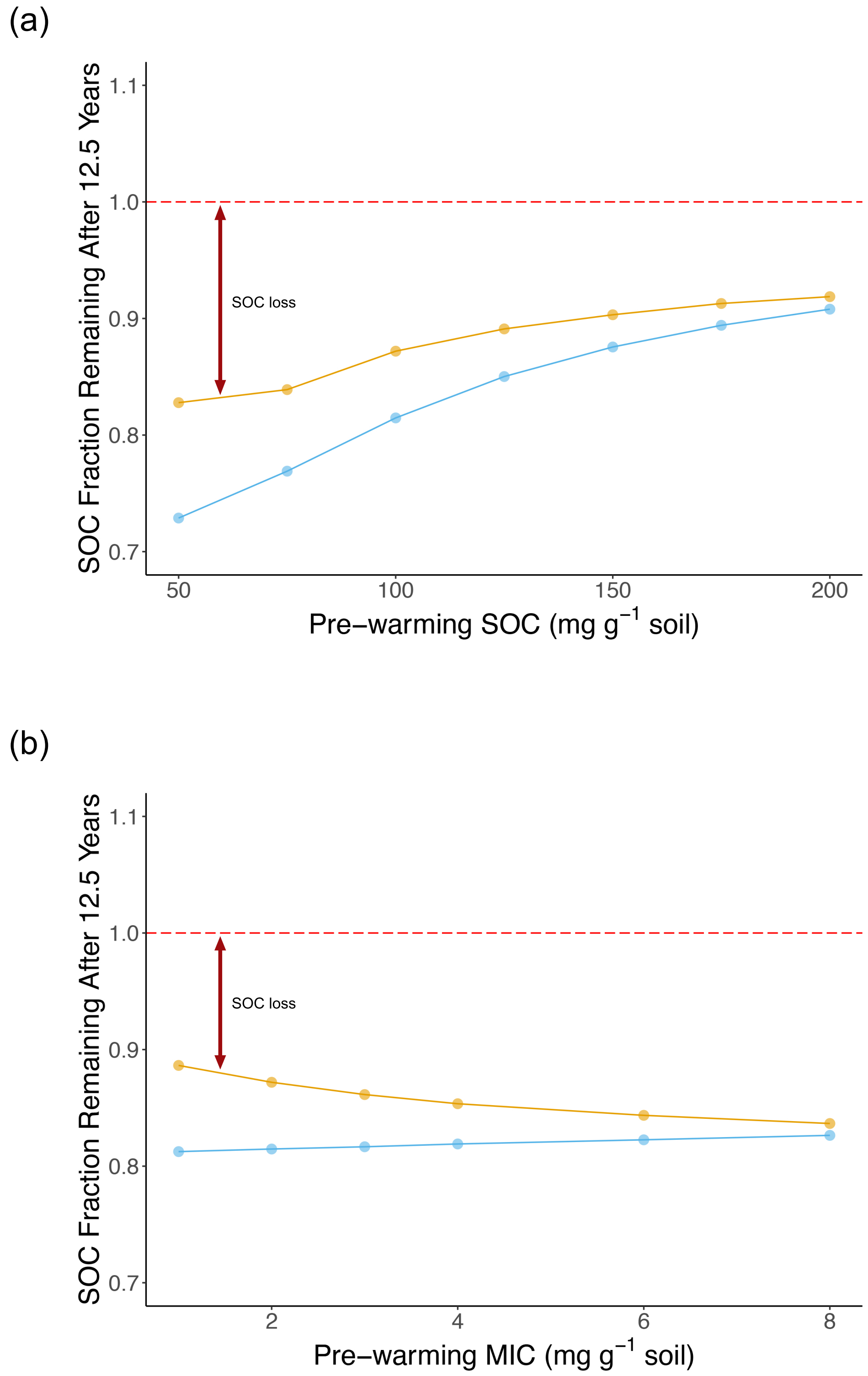
Consequently, we propose the following changes to Sup. Fig. 3 and Sup. Fig. 10:

- We will re-arrange and re-number the figures in the supplement so that Sup. Fig. 3 and Sup. Fig. 10 follow each other to reduce confusion.

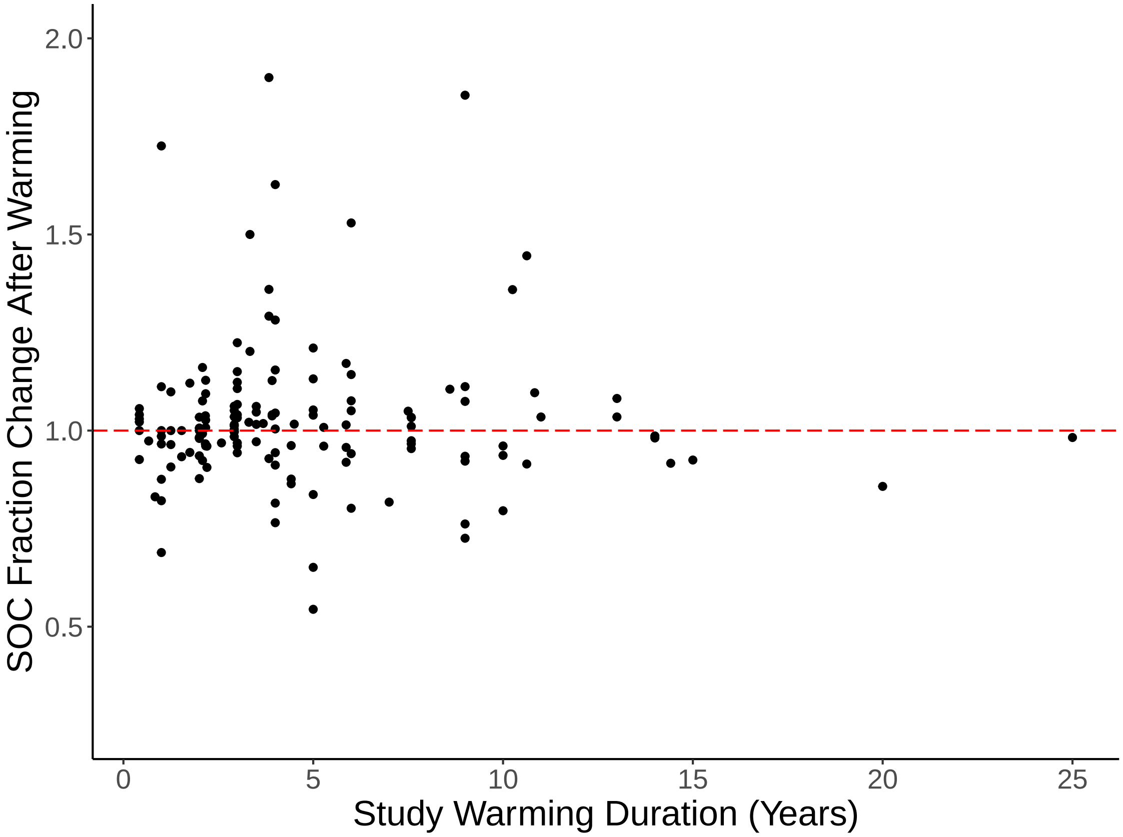
- We will change the vertical limits and add a dashed horizontal line at 1.0 in Sup. Fig. 3 and Sup. Fig. 10 to reflect the divide between SOC gain and loss after warming.

- We will add vertical arrows going down from the horizontal line in Sup. Fig. 3 labeled with “SOC loss”to clarify the direction of greater SOC losses.

Updated Supp. Fig. 3. To be re-numbered.



Updated Supp. Fig. 10. To be re-numbered.



> L 205: decreased from 16.3 to 11.3 %. Please check other lines as well

We will fix occurrences of redundant symbols in the revision.

> L 218: R2; Annotation for varying SOC: “SOC = 50 -> SOC50”

We will look into and fix the aforementioned exponent rendering issues and standardize the notation for pre-warming steady state densities. We commend this effective notation suggestion.

> L 231: not sure this manuscript compared models through AIC and DIC with WAIC and LOO

Yes, AIC and DIC were not computed and compared for reasons of redundancy, stability, and algorithmic limitations. AIC relies on a maximum likelihood estimate which cannot be calculated from non-uniform priors in a Bayesian setting [2]. AIC is more readily deployed in frequentist model comparisons in which normally distributed prior information is not used. DIC can be computed under Bayesian settings, but is an approximation of WAIC that is calculated from posterior means, whereas WAIC involves integration over the posterior distribution sample [1]. Since WAIC can already be readily calculated, has been demonstrated to be more stable and accurate than DIC in much statistics literature, and is itself an approximation of LOO [2], we felt that calculating and plotting WAIC, LPML, and LOO was sufficient for illustrating the influence of the pre-warming steady state ratios on goodness-of-fit. We initially felt that some of this information was too technical and distracting to discuss in a paper not being submitted to a statistics journal, but we now agree that we should summarize this information with citations in the introduction or methods section to justify our use of WAIC, LPML, and LOO.

> L 255-259: which Figures showing this? Also, the sentences are too complex. Please re-write simpler sentences

There was an oversight here, so we will add a figure citation in this paragraph referring to Supp. Fig. 3a, b which supports the assertions made in that paragraph. The sentences will be rewritten to be less convoluted.

> L 273: is 50 mg SOC g-1 soil same in line 218 (SOC =50)? Please use consistent unit

The line here was not referring to the specific SOC50 simulation and HMC run, but to the observation that we were unable to initialize simulations with pre-warming SOC below 50 mg C g^{-1}, so we felt that it was appropriate to write the full units out rather than abbreviate here.

> L 275: Supplemental Table 3?

Will be corrected.

> L 316: Supplemental Fig. 5

Will be corrected.

> L 322: Supplemental Fig. 8a, b

Will be corrected.

> Typing error
>
> 1. Please check upper case expression; r-squared is R2 (uppercase)
> 2. L 156: mg C g-1 soil
> 3. Put period after abbreviation of figure, e.g., Fig. xx
> 4. Please double-check figures and table numbers
> 5. Supplemental or supplementary?

We will fix upper case, exponent, abbreviation, and numbering issues. We will replace isntances of "supplementary" with "supplemental." We appreciate the close reading, fixes, and suggestions, and will address them in our revision.

References

[1] M. Betancourt. A unified treatment of predictive model comparison, 2015. arXiv:1506.02273.

[2] A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014.