We are grateful to the reviewer for the thoughtful and constructive feedback. We address the specific reviewer comments below.

> Why was it that CON generally performed better at most SOC densities? What are the estimated parameter ranges compared to other models/literature values? For example the activation energy for SOM decomposition seems to be higher than the activation energy for other processes like uptake which may have some interesting implications.

We feel that the chief reason that CON performed relatively better than AWB at most SOC densities by information criteria and cross-validation was because of overfitting and parameter count. By R2, AWB could fit the data slightly better by absolute residual sums than CON at most pre-warming SOC (Fig. 5), but AWB has more parameters and was penalized for that, and we were reluctant to over-reach on conclusions. However, you make an excellent point about the need for more discussion about the biological implications of the parameter posteriors and fitting results. Discussion of the biological realism of the models we used is more limited to the extent of SOC loss over time in the current iteration of the manuscript. Thus, in our revision, we will discuss more how our posteriors from each model compare with empirical results from literature. We will also add our explanation for why we think the mean posterior SOM decomposition activation energy ended up higher than the mean posterior activation energies of other processes. We believe that was the case because SOC decomposition is generally the the rate-limiting step in C-cycling systems that represent microbial activity. If SOM decomposition Ea were too low, the soil C would cycle too fast and result in a poorer fit to the data set.

> L62-66: Citation for this discussion of R2? Also there are other metrics for evaluating Bayesian models that are not discussed here you dont need an exhaustive review but ROC/AUC and BIC seem common.

For the discussion of R2, we will cite Kvålseth 1985[3], Spiess and Neumeyer 2010[6], and Gelman et al. 2019[1].

We initially did not discuss BIC due to its similarity to AIC. BIC is closely related to AIC and its computation is dependent on the pointwise maximum likelihood estimate from frequentist methodology[2]. Hence, BIC, contrary to its name, is not a fully Bayesian metric calculated from the posterior distribution. However, as BIC is indeed used to compare out-of-sample predictive accuracy of groups of models, we agree that it should mentioned and will revise our manuscript to include BIC in the introduction.

We would like to avoid discussing ROC/AUC because they are indicators of the prediction accuracy of binary classifier models trained and tested on categorical data[5]. In our case, we were specifically looking at metrics that estimated out-of-sample prediction accuracy or goodness-of-fit for models conditional on ordinal data with elements in $\mathbb{R}$. LOO/WAIC and ROC/AUC correspond to fundamentally different model and data types, so we feel that ROC/AUC would be off-topic.

> L125: 0.9995 and 0.001 seems like extreme adaptation and step sizes to me, causing the model to take many small steps. If this is a supported strategy, can you provide a citation or justify further?

To start, it is worth clarifying the differences between traditional MCMC and the Hamiltonian Monte Carlo algorithm that Stan uses. The HMC is not a random walk algorithm, and each of its trajectories are deterministically calculated via Hamiltonian dynamics. An MCMC step size parameter is fixed through the duration of the sampling. In contrast, the step size is tuned at each step based on the adapt delta and calculated Hamiltonian trajectory. I should have clarified in the manuscript (and will do so in the revision) that the step size used was the initial step size. Consequently, starting with a small step size does not mean that the HMC algorithm takes fixed steps of the same size for the rest of the chain (refer to this page from the Stan documentation for more detail).

The adapt delta and initial step sizes were set as such in an effort to reduce the number of divergent transitions during the HMC sampling, which was an issue for the AWB model. As can be seen in the supplement, divergent transitions were still detected for AWB following the implementation of the strategy. This indicates the future need to re-parameterize and re-formulate AWB to obtain smoother and more stable parameter space geometries to be explored. However, the divergences occurred at a reduced rate compared to using Stan's default adapt delta of 0.8, so the strategy proved helpful for obtaining a suitable number of posterior samples.

We did not encounter any divergent transitions with CON, but mirrored the HMC parameters for both models since the HMC parameters ultimately do not alter the overall exploration of the parameter space. Increasing adapt delta results in less sensitive tuning of the step size per iteration. In our case, this does mean our step size will be smaller on average as our initial step size will be less responsive to tuning and slower to increase, but the algorithm will be better at navigating geometrically trickier parameter space to generate fewer trajectory divergences. This trade-off comes at the cost of more computationally expensive and less efficient calculated Hamiltonian trajectories. The smaller steps correspond to a drive for more changes in trajectory direction to cover the same amount of ground. With sufficient chain length and samples (which our Bayesian diagnostics indicates that we obtained), the parameter space should still have been adequately explored [4].

The strategy of using higher adapt delta and lower initial step sizes is less documented in formal literature, but has been previously used by other Stan users to mitigate divergent transitions (see this link and this link for further discussion). It was important for us to maximize the amount of samples we obtained for AWB, so the increased computational time per iteration was a worthwhile tradeoff for us.

> L191: As written, implies that AWB performs better because it has a higher RR in subsequent years after the first year, but the data show that the first year has the highest RR, so CON seems to correspond more closely to this. In the discussion, you can bring up the potential realism of oscillations given the Harvard Forest long term warming experiment.

We will do this.

> L325-333: This discussion of R2 and other cost metrics seems repetitive to the introduction.

We will prune redundant information in this part of the discussion.

> It seems like the performance metrics would be yet better with a lower SOC density (¡50 mg SOC/g soil), if it were possible to achieve them without the AWB instability. I think you could fix the instability by changing your decomposition/uptake kinetics. Right now in the uptake equation DOC is in the denominator but its initial concentration is much smaller than MIC. So you can either flip to Reverse M-M for uptake or use ECA where both quantities (SOC and ENZ or MIC and DOC) are in the denominator <- this may be harder to fit because it will be more constrained, but it is also harder to break.

These are perceptive and insightful modeling suggestions that we greatly appreciate. In the discussion section of this manuscript, we will add some sentences to describe the importance of exploring the effect of changes to microbial-explicit model structures including the ones you proposed on data fitting and posterior sampling in subsequent work. Then, we will apply those AWB model modifications to an in-progress follow-up model comparison project that fits models to a different, larger data set from Harvard Forest. We feel that these reparameterization suggestions could help reduce or eliminate the number of divergent transitions generated during AWB HMC posterior sampling.

References

[1] A. Gelman, B. Goodrich, J. Gabry, and A. Vehtari. R-squared for Bayesian Regression Models. *The American Statistician*, 73(3):307–309, jul 2019.

[2] A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24, jul 2013.

[3] T. O. Kvålseth. Cautionary Note about R2. *The American Statistician*, 39(4):279–285, nov 1985.

[4] S. Livingstone, M. Betancourt, S. Byrne, and M. Girolami. On the Geometric Ergodicity of Hamiltonian Monte Carlo, Jan 2016. arXiv:1601.08057.

[5] T. Saito and M. Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432–e0118432, mar 2015.

[6] A. N. Spiess and N. Neumeyer. An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: A Monte Carlo approach. *BMC Pharmacology*, 10(1):6, dec 2010.