Dear Editor,

We greatly appreciate the opportunity to revise our manuscript. We are also very grateful for the insightful and constructive feedback from both reviewers to improve our manuscript. We have thoroughly addressed the concerns and comments raised by the reviewers. Below we describe our major revisions.

In this revision, we restructured the manuscript to explicitly include separate Methods and Results sections. We expanded the discussion on the issue of equifinality, clarified the positioning of BINN as an end-to-end parameter learning framework (distinct from PINNs and other physics-informed AI approaches), and strengthened the side-by-side comparison with PRODA by using identical folds and reporting paired test Nash–Sutcliffe modelling efficiency coefficient (NSE) and spatial difference maps. We also expanded uncertainty quantification of BINN predictions and reported computational trade-offs by adding peak resident set size (RSS) alongside wall-clock time.

The major concerns and comments by the reviewers and our corresponding revisions are summarized as follows.

1. **Equifinality and Identifiability.** Both reviewers concerned about the equifinality issues when applied BINN across all the parameters training with real-world observations. In this revised version of manuscript, we extended the recovery from the top five to all 21 parameters, demonstrating that high-sensitivity parameters remain well recovered (Fig. S4). We then implemented Monte Carlo dropout to obtain posterior-like parameter distributions and site-level coverage, showing that prescribed values are frequently contained within BINN's uncertainty bands (Figs. S5–S6).

2. **Sensitivity Analysis.** Both reviewers asked for clearer documentation of our sensitivity analysis. In the revised manuscript, we moved the description of the first-order approximation from the Supplement to the Methods section. We also improved the figures to report mean ± standard deviation across sites (Figs. 2 and S3) and color-coded parameters according to their respective process component with clearer labels and legends, enhancing interpretability and justifying our selection of the most sensitive parameters. In addition, we corrected a bug in the sensitivity analysis, after which the parameter "beta" emerged as having the highest average sensitivity score across all depths. Accordingly, we updated the sensitivity results and reran the parameter recovery experiment using the five most sensitive parameters, now including "beta".

3. **Positioning & Terminology: BINN vs. PINN.** Reviewer 1 noted the name 'BINN' may cause potential confusion with PINNs. We revised the Introduction part of the manuscript to position BINN explicitly as an end-to-end local parameter learning framework that embeds a differentiable CLM5 operator. We also distinguished BINN from PINNs and other physics-informed AI approaches, referring to the references suggested by the reviewer.

4. **Comparison between BINN and PRODA.** Reviewer 1 suggested a clearer, common-ground evaluation. In addition to only comparing model components calculated from parameters

predicted by BINN and PRODA, in this revision, we added a paired 10-fold cross validation using identical folds for BINN and PRODA, and reported test NSE, normalized difference distributions, and spatial bias maps side-by-side (Fig. 6).

5. **Memory Trade-offs.** Reviewer 2 asked for memory usage comparison in addition to wall-clock time. In this revision, we report peak RSS alongside wall-clock time for MCMC, PRODA, and BINN (looped vs. vectorized CLM5). Our results show that BINN (vectorized) achieves the lowest RSS, in addition to more than 50 times faster than Bayesian-based methods (Fig. S8).

6. **Uncertainty Quantification in Spatial Maps.** Reviewer 2 noted that the normalized-difference map in the original manuscript did not convey spatial bias clearly. In this revision, we improved spatial interpretability by normalizing SOC difference maps using signed 99th-percentile scaling (rather than min–max) to avoid outlier-dominated color ranges (Figs. 5–6). We also added fold-to-fold coefficients of variation to map spatial uncertainty (Fig. 5b). These revisions clarify where predictions are stable vs. variable across the cross-validation folds.

For detailed responses, please refer to our point-by-point responses in the response letter to each reviewer. We hope that the revisions are satisfactory for both you and the reviewers.

Thanks for your time evaluating our revised manuscript.

Sincerely,

Haodi Xu
On behalf of all co-authors

Point-to-point responses to comments from reviewer #1
(Manuscript number: egusphere-2025-3282)

We thank Reviewer 1 for the thoughtful and constructive comments. In this response, we clarify BINN's scope and methods, explain PRODA and the comparison setup, distinguish BINN from related hybrid approaches, and expand the synthetic and uncertainty analyses. Reviewer comments appear in black, followed by our responses in blue.

Comment 1: In this article, the authors integrated a well-known process-based model CLM5 for simulating soil organic carbon states into a neural network framework which they call BiNN. They use a neural network to recover CLM5 parameters and compare this approach to Bayesian parameter estimation framework called PRODA. The strength of BiNNs is a significant speed up towards parameter calibration with PRODA and is supported by a preliminary experiment on synthetic data to showcase the approach. While this is an interesting and specific application for embedding neural networks and process-based models, as well as that the article builds on some interesting previous works, I'd recommend changes and extensions to the analysis, as well as restructuring the article before considering it for publication.

We thank the reviewer for the positive evaluation of our work and for providing constructive suggestions to further improve it. We have thoroughly addressed all the reviewer's concerns in a point-by-point manner. Our responses are marked with blue color in the following sections.

In general, I hesitate to consider this as a new model or method, but rather see it as an application of an already established method to an existing model. Further, to my knowledge, articles in GMD require a model version in the title, which in this case should contain at least CLM5, which is calibrated.

We appreciate the reviewer's thoughtful suggestion. The main contribution of our work is the development of BINN as a general framework for improving modeling and understanding in biogeochemistry. While we use the CLM5 model as a case study, BINN is a general approach that can be applied to other process-based models, as long as they contain unknown parameters. For example, we are working on extending BINN to model forest peak growth age and soil inorganic carbon. Thus, including a version name of BINN, i.e., "BINN v1.0", in the title would be more appropriate than a model name.

Comment 2: As the key idea of this article is that we should use BiNN because it speeds up parameter estimation at a similar level of accuracy as the second approach that estimates space-varying parameters, that is PRODA. However, the PRODA framework isn't properly introduced in the method section and evaluation of both frameworks in the reskults section is incomplete, so I suggest revisions of this part. As readers we don't know how good PRODA is performing. In section 4.2, only the correlations of PRODA and BiNN predictions are reported, but not the ones with the observations, they remain qualitatively described. Please add the according evaluation metrics here as well (is this what we see in figure 6? Then please refer to this figure). Accordingly, when comparing the two approaches against each other (Figure 5), it would be good to see both of them evaluated against a common ground truth, e.g. as is displayed in figure 4 for just the BiNNs. A good choice of experimental design would be to do this within the

synthetic experiment first, where the ground truth of the parameters is known. This means the results could compare 1) the correlation of recovery parameters from both approaches with the synthetic parameters and simulated SOC, and then 2) in applying this method after parameterisation compare the correlation of both predicted SOC (or other soil states) with the observed SOC. In the discussion, finally, PRODA should be contrasted with the parameter learning approach on a methodological level.

Thanks for highlighting the need to better introduce PRODA and to evaluate both approaches on a common basis. We've revised the introduction (Revised L95-L103) and the method sections (Revised Section 2.3.2) to add more details about PRODA's structure and performance. We've also included more comparisons on the performance between PRODA and BINN in the result section (Revised Fig. 6).

Following the reviewer's suggestion, in the revision, we've enriched Section 3.2 and Figure 6 to compare the behavior of BINN and PRODA in fitting the observations in addition to the present component retrieval comparison. To do this, we have extended the 10-fold cross-validation experiment to evaluate both BINN and PRODA side-by-side with real-world SOC observations (Figure R1). We found that while the testing accuracy between BINN and PRODA is similar (Fig. R1a), BINN exhibits lower spatial bias than PRODA (Fig. R1b). The spatial distributions of the differences in model simulations and observations further indicate that BINN (Fig. R1c) notably mitigates the underestimation of SOC in the central and eastern U.S. presented in the PRODA-optimized results (Fig. R1d).

We appreciate the reviewer's suggestion regarding the experimental design. The main objective of this study is to introduce BINN as an AI-assisted framework to accelerate data-model fusion in optimizing process-based models than conventional Bayesian-based optimization methods, i.e., PRODA in this case, while retaining consistent scientific interpretations. Moreover, the parameter recovery experiment using a Bayesian-based method like Monte-Carlo Markov Chain has been done previously (Tao et al. 2024). Thus, we believe that including the performance test of PRODA in a parameter recovery experiment may be out of the scope of this manuscript. In the revision, we've added a citation of Tao et al. (2024) where appropriate (Revised L99).

Tao F, Houlton BZ, Huang Y, Wang YP, Manzoni S, Ahrens B, Mishra U, Jiang L, Huang X, and Luo Y. 2024. Convergence in simulating global soil organic carbon by structurally different models after data assimilation. *Global Change Biology*, 30: e17297.

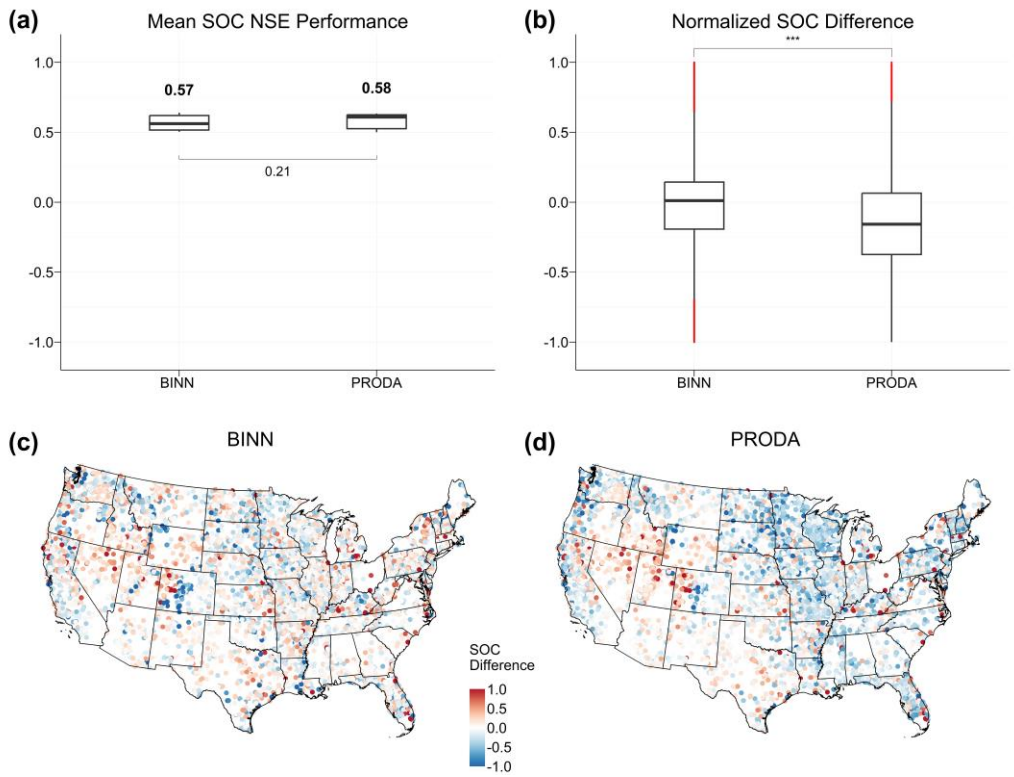

Figure R1. Side-by-side evaluation of BINN and PRODA on identical test sets across a 10-fold cross-validation. (a) Mean test Nash–Sutcliffe modelling efficiency coefficient (NSE) per fold. The *P* value above the bars donates the difference in test NSE between the two methods is insignificant. (b) Mean test-set differences after normalization: positive differences are scaled to [0,1] by the 99th percentile of positive deviations; negative differences are scaled to $[-1,0]$ by the 99th percentile of negative deviations. The red dots are outliers, and the horizonal significant bar shows that the mean differences in SOC between two methods are significantly different. (c–d) Spatial maps of the normalized differences for BINN (c) and PRODA (d); values near 0 indicate small bias, positive values indicate overestimation, and negative values indicate underestimation.

Comment 3: Another larger issue I see with this article is that equifinality isn't addressed in a critical way, while the synthetic results from figure 4 indicate that it is already present when fitting just four parameters (see section 3, why else are correlations not stronger here?). For the application later on, the authors fit all 21 parameters of the model, and it is to be expected that this will not diminish their unidentifiability. If this work aims for methodological demonstration, it could be better to just remain with a smaller set of parameters, but discuss their estimation in more detail, or, if going for many parameters, also do the simulation on many. Generally, the strength of a traditional Bayesian calibration approach is that we get uncertainty with the posterior distributions, information which unfortunately, neither PRoda nor the BiNNs currently provide but which could greatly support interpretation of these findings. This is a limitation of PRODA, that is introduced as a Bayesian approach but doesn't leverage its capabilities. I think

the authors should openly discuss the limitations of PRODA in the methods section on PROda and in the discussion.

We thank the reviewer for pointing out the equifinality issue. We agree with the reviewer that while PRODA uses Bayesian-based data assimilation in its optimization algorithm, it does not fully utilize the uncertainty information of the posterior distributions because we only take the mean values of parameter posterior distributions to be predicted by the neural network in PRODA. Thus, PRODA cannot consider propagating the uncertainties of parameters' posterior distribution in the site-level data assimilation to larger scales.

In addition, to better address the equifinality issue in BINN, in the revision, we have implemented Monte Carlo dropout (Gal & Ghahramani 2016.) to obtain the posterior distributions of parameters at each site (Fig. R2). Although point estimates of parameters do not exactly match the prescribed values, the posterior distributions of parameter values by BINN frequently cover the prescribed values. We further quantify the site-level coverage (i.e., the fraction of prescribed values falling into the BINN-predicted intervals) in Fig. R3, which indicates the issue of equifinality can be controlled to a reasonable extent.

We fully agree with the reviewer that equifinality increases with increasing target parameter numbers during optimization. To address this concern, we conducted an additional parameter retrieval experiment using all the 21 parameters. The results show varying identifiability among different parameters (Fig. R4), depending on their deterministic influence on SOC storage (as also illustrated by Fig. 3 in the original manuscript). While parameters with smaller influence on SOC storage (e.g., cryo and taucwd) are less identifiable (i.e., low correlation coefficients between prescribed and retrieved values), parameters that are more deterministic to SOC storage (e.g., w-scaling and fs1s3) can be successfully retrieved with high correlations with their prescribed values (Fig. R4). We've extensively discussed the above results (Revised Section 4.1) in the revised manuscript as a supplementary to present Figure R4.

In addition, we rechecked the correlations and NSE in the parameter recovery experiment in Figure 4 of the original manuscript. We found some of the values were wrong and made corrections. The corrected analyses show that BINN recovers the four key parameters with high correlation (Fig. R5) and BINN-optimized CLM5 largely reproduces the synthetic SOC with high NSE.

Gal, Y. & Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, 2016.

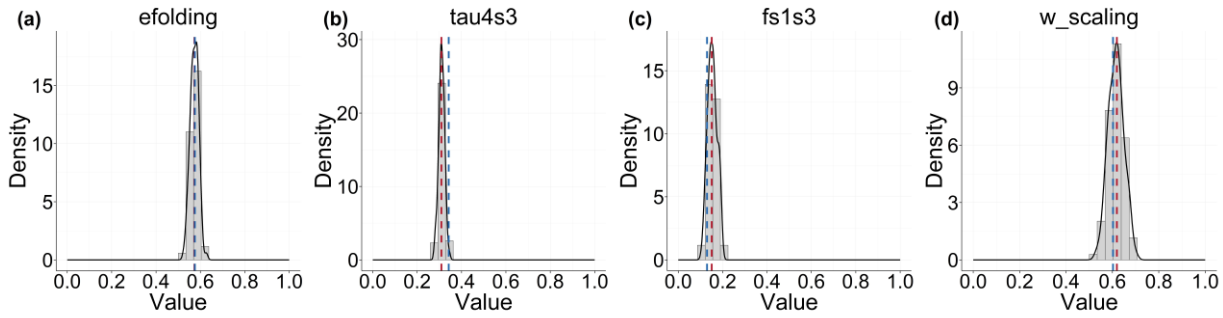

Figure R2. Uncertainty in posterior distributions by BINN parameters in relation to Monte Carlo dropout in the parameter recovery experiment at one site. For each parameter, the marginal posterior is shown, with the vertical blue and red dashed lines indicating the prescribed ("true") and BINN's point estimate, respectively.

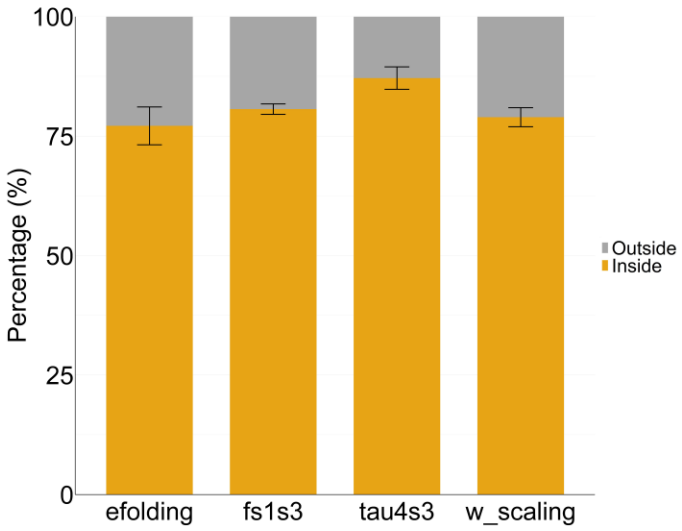

Figure R3: Coverage of parameters' posterior ranges of BINN. The coverage was quantified by the percentage of prescribed ("true") parameter values that fall into the Monte Carlo–dropout posterior range across test sites. Error bars show the mean ± SD across the 10 cross-validation folds of the synthetic parameter-recovery experiment.

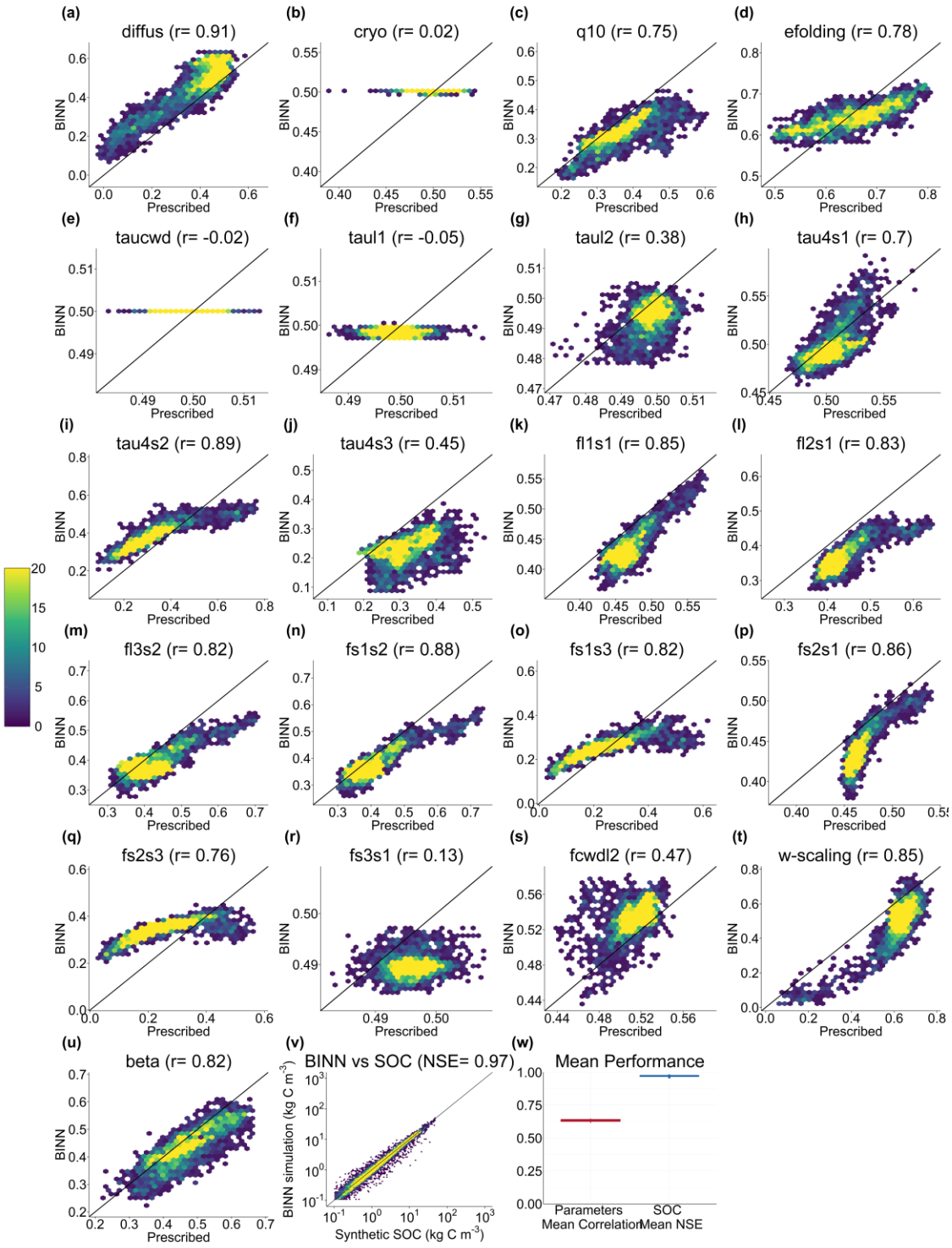

Figure R4: BINN parameter recovery for all CLM5 parameters. (a–u) Scatter plots of BINN-predicted versus prescribed values for each parameter. (v) Density scatter of simulated SOC versus synthetic SOC. (w) Summary across a 10-fold cross-validation: mean correlation for all parameters (predicted vs. prescribed), and NSE for SOC simulations.

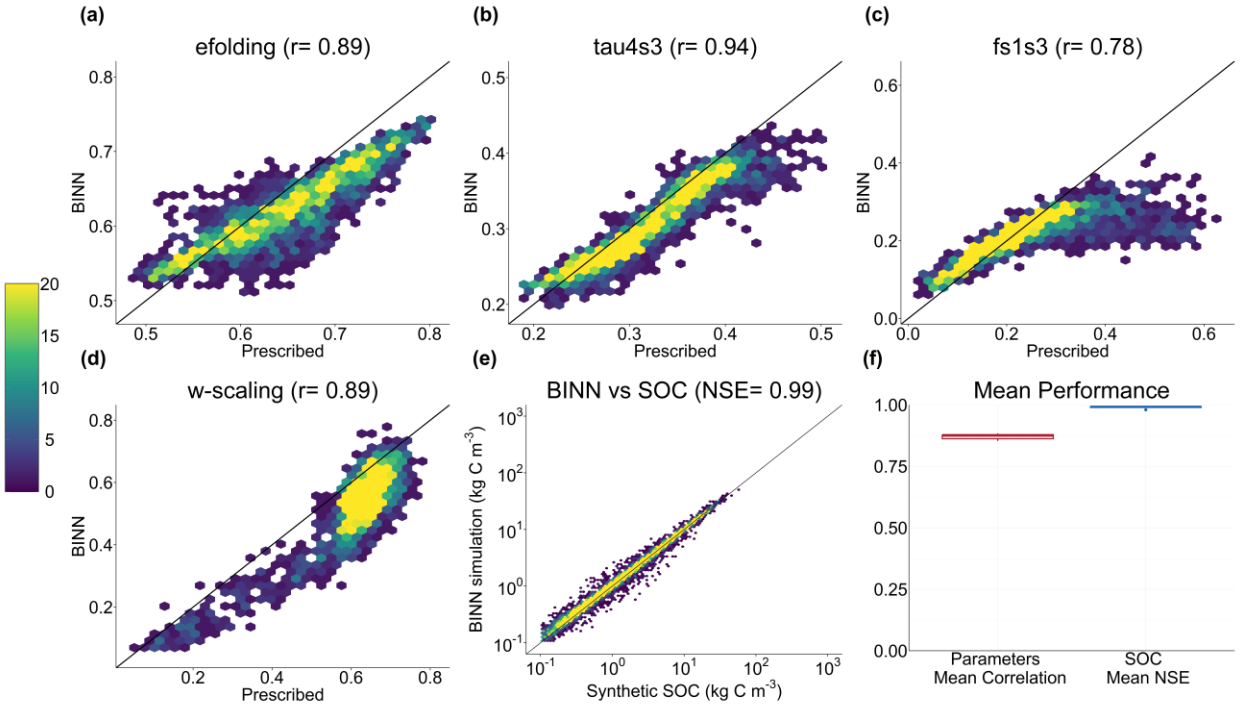

Figure R5: BINN's performance in retrieving the four most sensitive parameters and simulating SOC. Scatter plots with colors representing the density of points comparing the parameter values (a) "efolding", (b) "tau4s3", (c) "fs1s3", and (d) "w-scaling" predicted by BINN against the prescribed parameter values. (e) Comparison of the simulated SOC using BINN-optimized vs prescribed parameter values. (f) Mean performance of BINN in retrieving the 4 parameters, as measured by the correlation as well as NSE of the simulated SOC with BINN-optimized parameters compared to SOC simulated with prescribed parameter values.

Comment 4: As a last major point, I have doubts about the naming that may be misleading. While the terminology is certainly used broadly, more recently, PiNNs have been re-defined more specifically as incorporating the learned ode in a combined loss function as teacher forcing (https://doi.org/10.1007/s44379-025-00015-1). The approach introduced in this study falls into the general realm of physics-informed machine learning but differs in the methods significantly from this definition of PiNNs, hence I believe the article would profit from re-specifying their method more precisely in title and abstract to, e.g. an end-to-end local parameter learning/recovery approach (see e.g. https://doi.org/10.1038/s41467-021-26107-z). Further, this specific approach to model calibration should be reviewed in the introduction and the physics-informed ML approach distinguished from other process-informed ML approaches. This will help structuring the article and methods section (for an overview of process-informed ML approaches see e.g. https://doi.org/10.1111/ele.70012).

We thank the reviewer for flagging the potential confusion around terminology. We agree with the reviewer that previous usage of "PINN" typically refers to loss-minimizing neural networks that add a penalty for ODE/PDE violations in the loss. However, our approach deeply embeds the physical model within the neural network; a neural network maps environmental covariates to process-based model parameters, and then executes the forward simulation, so physical

consistency is enforced by the model itself rather than as soft residual terms in the loss. BINN ensures that the model's predictions strictly obey the physical constraints in the process-based model, while PINN does not enforce hard constraints and allows the governing equations to be violated. In the taxonomy of the paper you shared (Wesselkamp et al. 2024), PINN typically refers to "physics regularisation" methods, while BINN is similar to "physics embedding". Note that unlike Wesselkamp et al. 2024, our process-based model takes in forcing variables; we also add prior ranges on the parameters, which we enforce through sigmoid constraints (Revised Eq 4-5) and the hyperbolic cosine loss (Revised Eq 10). We've revised the introduction, and other related sections to clarify these distinctions (Revised L71-L85). We also thank the reviewer for providing the three insightful references. We will include them in the revised manuscript to better introduce the context of related research and methods.

Wesselkamp, M., Moser, N., Kalweit, M., Boedecker, J., & Dormann, C. F. (2024). Process-Informed Neural Networks: A Hybrid Modelling Approach to Improve Predictive Performance and Inference of Neural Networks in Ecology and Beyond. *Ecology Letters*, 27(11), e70012.

Comment 5: Structure: Section 5: These algorithmic details should come earlier and in a section on the fitting the network to CLM5 parameters (e.g. 2.1) . PRODA should totally not be introduced here but in an own subsection of section 2 after the NN and process model. Also, section 2 should be general methods. Why is there an own section on observational performance and computational efficiency? I suggest summarising results in a results section of their own. And move the data preparation methods section. And distribute the contents of section 4.2 to methods and results.

We thank the reviewer for the great suggestions on the structure. We've followed these suggestions to adjust the manuscript structure in the revision. We've also revised section 2 to be more general, including an overview of the PRODA approach. We've reorganized all the results to make them more straightforward.

Comment 6: **Section 1** (Introduction): 1) The beginning of the introduction lacks biogeochemical examples of parameter learning beyond soil, while mentioning a wide range of fields where hybrid models are applied. Preferably, mention other biogeochemical applications. 2) More importantly, hybrid approaches are introduced without any differentiation of how physical constraints are integrated with machine learning. See general comment on BiNNs, this would greatly help the reader at the beginning to locate the introduced approach (for an overview for e.g. for carbon flux with difference equations see https://doi.org/10.1111/ele.70012).
3) Further, if the goal is enabling interpretability of biogeochemical dynamics, as stated at the end of the introduction, I would like to know why and how BiNNs can lead to improvement here towards, e.g. traditional Bayesian approaches.
4) In contrast to hybrid and mechanistic approaches, there are also established statistical models that estimate spatially-varying parameters while maintaining direct interpretability of their coefficients, such as SVCMs or Geographically-weighted regression. See for example: (10.1186/s12862-024-02260-z). This should at least be mentioned.

We thank the reviewer's valuable suggestions for strengthening the introduction and have followed these suggestions to revise our manuscript as follows.

(1) We've expand our introduction to include more biogeochemical examples of parameter learning beyond soils (e.g., terrestrial carbon flux partitioning (Wesselkamp et al. 2024), nutrient cycling (Shi et al. 2016), ecosystem respiration (Reichstein et al. 2022), plant phenology (van Bree et al. 2025), and hydrology–biogeochemistry couplings (Liu et al. 2020)) to better situate our BINN within ongoing work in the related fields (Revised L86-L90).

Wesselkamp, M., Moser, N., Kalweit, M., Boedecker, J. & Dormann, C. F. Process-Informed Neural Networks: A Hybrid Modelling Approach to Improve Predictive Performance and Inference of Neural Networks in Ecology and Beyond. *Ecology Letters* **27**, e70012 (2024).

Shi, Z. *et al.* Inverse analysis of coupled carbon–nitrogen cycles against multiple datasets at ambient and elevated CO2. *J Plant Ecol* **9**, 285–295 (2016).

Liu, Y., Kumar, M., Katul, G. G., Feng, X. & Konings, A. G. Plant hydraulics accentuates the effect of atmospheric moisture stress on transpiration. *Nat. Clim. Chang.* **10**, 691–695 (2020).

Reichstein, M., Ahrens, B., Kraft, B., Camps-Valls, G., Carvalhais, N., Gans, F., ... & Winkler, A. J. (2022). Combining system modeling and machine learning into hybrid ecosystem modeling. In *Knowledge guided machine learning* (pp. 327-352). Chapman and Hall/CRC.

van Bree, R., Marcos, D., & Athanasiadis, I. N. (2025, April). Hybrid phenology modeling for predicting temperature effects on tree dormancy. *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 27, pp. 28458-28466).

(2) We've described the differences of various hybrid approaches that integrate physical knowledge by machine learning in previous studies and compare them with BINN (L71-L85).

(3) We've explained why and how BINN facilitates mechanistic understanding more efficiently than traditional Bayesian calibration in the introduction (Revised L86-L103). We've described how conventional Bayesian inference-based approaches usually cannot address the spatial heterogeneity of biogeochemical parameters, which has been proven critical to understanding key ecological processes in global carbon cycle (Luo & Schuur 2020). The PRODA approach combines traditional Bayesian calibration at site level with a neural network to retrieve spatial patterns of parameters at larger scales, yet its application has been hindered by its prohibitive computational cost for large datasets (particularly for Bayesian MCMC). BINN proposed in this manuscript, by seamlessly connecting varying environmental covariates, process-based models, and big soil data and by leveraging the backpropagation optimization algorithm in deep learning, presents an integrative algorithm considering the spatial patterns of those biogeochemically interpretable parameters while substantially improving the computational efficiency in model optimization.

Luo, Y., & Schuur, E. A. (2020). Model parameterization to represent processes at unresolved scales and changing properties of evolving systems. *Global Change Biology*, 26(3), 1109-1117.

(4) We've also acknowledged established statistical approaches that estimate spatially varying coefficients (e.g., SVCMs) to provide a more comprehensive context when introducing BINN (Revised L92-L95).

Comment 7: **Section 2:** General introduction – state more precisely if this end-to-end or a two-step procedure – from Figure 1 and description in 2.2 I expect a fully integrated hybrid model, but here it sounds as if its two fragmented steps – please clarify.
Equation (2): Please check for mistakes, y_i is not defined, could be z_i.
Equation (5): How do you choose tau, also in the HP search? And why p_j – 0.5, i.e. could you elaborate on why you chose 0.5 - from the description I would expect a parameter-specific value for each p_j, i.e. the center of the prior distribution, unless you scaled them. If so, please mention.
Section 2.4: Please give the training details you have in section 5 here.

We appreciate the reviewer for pointing out these ambiguities and typos in our manuscript.

(1) To address the reviewer's comments, we've clarified in Section 2 that BINN is an end-to-end, differentiable framework that maps environmental covariates to local parameters, which are then fed directly into the process-based model for the forward simulations, with gradients being back-propagated through this operator during training. We've revised our descriptions to avoid impressions that BINN is a fragmented two-step procedure.

(2) Yes, in Eq. 2, y_i should be z_i. Thank you for catching this; we've corrected it in the revision.

(3) For Eq. 5, τ is a hyperparameter that is selected via a grid search within the hyperparameter selection processes (Appendix 1). We apologize for the confusion about 0.5; there are some notational inconsistencies that we've fixed it in the revised manuscript. Specifically, we will split Eq 2. First, we use

$$p_{i,normalized} = \frac{1}{1 + \exp\left(-\frac{z_i}{\gamma}\right)}$$

to normalize the predicted parameters z_i to be in the range [0, 1]. Then, they are linearly scaled to the true prior range $(\theta_{i,min}, \theta_{i,max})$:

$$p_i = p_{i,normalized} * \left(\theta_{i,max} - \theta_{i,min}\right) + \theta_{i,min}$$

The loss (Eq. 5) operates on the normalized parameters (in the range [0, 1]), encouraging them to be close to the middle of the prior range (e.g. 0.5).

$$\sum_{j=1}^{21} cosh\left[\tau\left(p_{j,normalized} - 0.5\right)\right]$$

Comment 8: **Section 3**: If the goal of the paper is clearly stated in title and abstract, this section could simply be called simulation experiment.
What type of sensitivity analysis did you use for CLM5?

Given the equifinality and wide distributions in Figure 3, as this experiment was repeated in a CV, would it be possible to also report the standard deviation on the correlations?
Figure 3: See above. What sensitivity index was used and how was this done? Looks to me like feature importances.

We appreciate the reviewer's valuable suggestions on framing and reporting. We've changed the heading of section 2.2.2 to "Simulation Experiment to Recover Biogeochemical Parameters from Synthetic Data" to be clearer.

Regarding sensitivity analysis, we used a first-order approximation method following Gao et al. (2011). We perform the sensitivity randomly at 512 sites across the Conterminous U.S. In the revision, we include the standard deviation of the sensitivity in addition to the mean values in the sensitivity test results (Fig. R6). The detailed processes in performing the sensitivity analysis have been specified in the supplementary material in the original manuscript, but we will move the core descriptions about sensitivity analysis to the main text in the revision. Figure 2 in the revised manuscript has been updated to Fig. R6 with the first-order sensitivity index in the title to avoid confusion with feature importance.

During the revision, we've discovered and corrected a bug in our original sensitivity-analysis code. After this correction, the parameter "beta" emerged as having the highest average sensitivity across all depths. Accordingly, we updated the sensitivity results (Revised Fig. 2; Fig. R6) and reran the parameter recovery experiment using the five most sensitive parameters, now including "beta" (Revised Fig. 3). The main conclusions are unchanged, but the updated analysis more accurately reflects the relative importance of the key biogeochemical parameters.

Gao, C. *et al.* Assimilation of multiple data sets with the ensemble Kalman filter to improve forecasts of forest carbon dynamics. *Ecological Applications* **21**, 1461–1473 (2011).

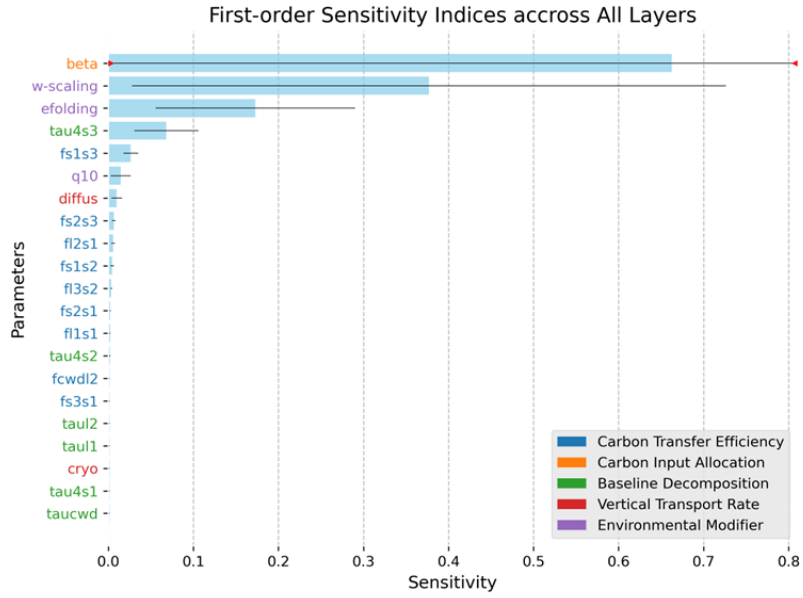

Figure R6. First-order sensitivity of CLM5 parameters across all soil depths. Bars show the first-order sensitivity index for each biogeochemical parameter over all soil layers. Red triangles indicate error bars clipped by the x-axis (uncertainty extends beyond the displayed range). Parameters are ranked (y-axis) by decreasing sensitivity and color-coded by their associated process component (e.g., K, A, B, ξ, V). The x-axis reports sensitivity scores, quantifying the influence of small parameter perturbations on model outputs, with larger values indicating greater influence.

Comment 9: **Section 7:** General: Quickly introduce BiNNs and the Bayesian approach at the beginning.
"High correlations between BINN-retrieved and prescribed biogeochemical parameter values in a controlled parameter recovery experiment demonstrate BINN's ability to recover causal relationships between covariates and SOC dynamics. Faithful retrieval of biogeochemical parameters from data substantially reduces uncertainty in SOC model predictions"
These two sentences come across quite lonely as they are. Please link back to your findings: Where do we see this?

We thank the reviewer for the suggestions. We've revised section 7 to start with a brief recap of BINN and the Bayesian-based approach (PRODA) to orient readers (Revised Section 4.1). We've also revised the sentences that the reviewer mentioned to tie them directly to the reported results and related surroundings, ensuring a smoother, continuous flow rather than isolated statement. We've also direct our arguments to specific results in the revision to be clearer.

Comment 10: **Section 7.3:** I agree that this approach may provide a new tool to model unresolved processes, it is not very clear on how it can help towards better mechanistic understanding with e.g. tracebility analysis mentioned. Could you explain this better? Also, there's a lot of repetition in this paragraph.

We thank the reviewer for the comments. We've streamlined this paragraph and make the logic more straightforward (Revised Section 4.3): BINN learns site-specific parameters that map directly onto CLM5 process components (e.g., baseline decomposition K). From these parameters we compute "model components" in Fig. 5, where each component has a clear biogeochemical meaning—for example, baseline decomposition Krepresents the intrinsic SOC decay rate absent environmental modulation by ξ. We then quantify each component's contribution to spatial variations in SOC storage in a traceability analysis, thereby linking learned parameters to processes and further to SOC outcomes. We've revise this section accordingly, removing repetition and including a concise example to illustrate how BINN helps better mechanistic understanding.

Comment 11: The positional encoder that was used in the NN that informs the networks about the location. This design decision may blur the biogeochemical interpretation of parameter estimates if that is the goal. Here, a post-hoc check of sensitivity to location could be useful and if sensitive, run the analysis without the positional encoder.

We thank the reviewer for raising this important point. To assess whether spatial features blur mechanistic interpretation, in the revision, we conducted an experiment, i.e., BINN trained

without the positional encoder (Fig. R7). The resulting SOC predicting skill (NSE) and spatial residual patterns are comparable to those reported in Fig. 6 of the original manuscript. Given the negligible impact, we retain the positional encoder but will provide a setting in our released code to disable it if users prefer a purely covariate-driven NN.

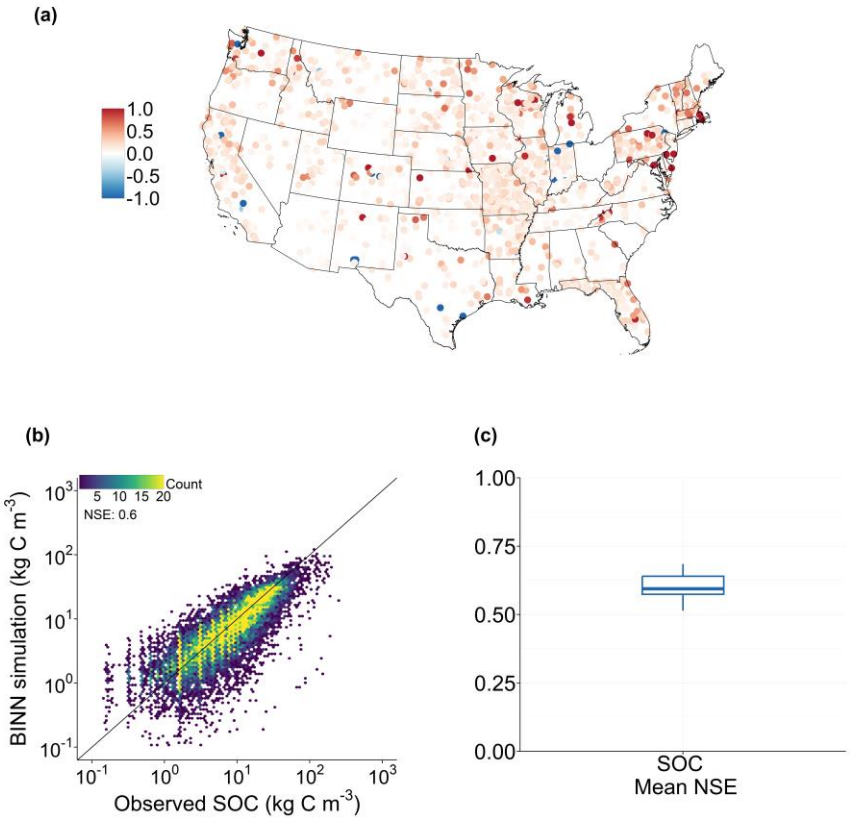

Figure R7. Comparison of observed and simulated SOC storage using BINN without positional encoder. (a) Spatial distribution of normalized differences for one representative fold (median NSE). (b) The scatter plot presents the NSE between observed SOC and simulated SOC

derived from the testing dataset at varied soil depths, with the NSE values shown in the plot. (c) The box plot shows the mean performance of testing NSE in the 10-fold cross validation test.

Point-to-point responses to comments from reviewer #2
(Manuscript number: egusphere-2025-3282)

We thank Reviewer 2 for the careful and constructive feedback. In this response, we clarify computational trade-offs (runtime and memory), expand validation and uncertainty analyses (CV diagnostics, MC-dropout), discuss choices on parameter prior ranges and activation functions, and streamline figures and captions to improve interpretability. Reviewer comments appear in black, followed by our responses in blue.

Comment1: The main contribution of this work is integration of CLM5 into a BINN framework. The paper walks the reader through BINN's design, validation, benchmarking, and performance: Figure 1 lays out the architecture—neural network parameters bounded by sigmoids flow into a differentiable CLM5 SOC module, trained end-to-end against Smooth-L1 losses with soft priors. Figures 2–4 show synthetic tests: CLM5-generated data used to recover the most sensitive parameters, with moderate success ($r \approx 0.7$) and acceptable SOC skill, supported by sensitivity analysis but limited by equifinality and assumptions. Figures 5–6 compare BINN outputs to PRODA and observations, demonstrating high spatial correlation and $NSE \approx 0.66$, though validation may be optimistic. Figure 7 uses traceability analysis to illustrate how different biomes balance inputs vs residence time, offering mechanistic interpretation, and Figure 8 highlights computational efficiency, with BINN running over 50× faster than PRODA by virtue of vectorization and gradient-based learning.

The language of the article reads clear most of the times except for using buzzwords that exaggerates and overstretch the results/claims, e.g. "transformative", "harness the power of AI". The work feels methodologically rigor and advances engineering problem with good computational efficiency. The manuscript presents a technically solid and well-executed methodological advance, and the figures clearly illustrate the architecture, parameter recovery, benchmarking, and efficiency of the BINN framework. However, the work in its current form suffers from a gap between claims and evidence: while the method is convincingly demonstrated in synthetic tests and with observational SOC profiles, the validation strategy (random ross-validation, reliance on shared forcing, limited exploration of equifinality, absence of independent benchmarks or uncertainty quantification) does not fully support the breadth of the conclusions drawn. Strengthening the validation would likely require substantial additional work, which may not be feasible in a short revision cycle. Therefore, I recommend one of two paths forward: (i) reframe the manuscript more modestly, toning down broad claims of ecological insight and general applicability, focusing instead on the clear computational and methodological contributions; or (ii) extend the analysis with additional validation and uncertainty assessment to bring the evidence base up to the level of the claims. Either path would improve the agreement between the strong methodological innovation and the scientific narrative presented.

We appreciate the reviewer's comprehensive summary and positive evaluation of our work. The reviewer's constructive comments and suggestions will help us further improve the robustness and rigorousness of this work. Following the reviewer's suggestions, we've revised related sentences in our manuscript (Revised L33-37) to remain neutral in descriptions. We've reframed the manuscript to focus more on the computational and methodological contributions of BINN.

Moreover, we've strengthened validation and uncertainty reporting in the revised manuscript. First, beyond reporting test NSE across the 10-fold cross validation (Fig. 6c in the original manuscript), we now provide the spatial distribution of mean residual (with sign indicating over or underestimation) and coefficients across all 10 folds (Fig. R1a–b), together NSE of 10 folds including all sites (Fig. R1c) and report the mean test NSE (Fig. R1d). This makes it clear where the uncertainty is data-driven versus model-driven and avoids over-interpreting any single evaluation. Second, in the revision, we quantify parameter uncertainty from BINN with Monte Carlo dropout, showing site-level posterior distributions (Fig. R2) and coverage in the posterior distributions of BINN on the prescribed parameter values (Fig. R3). The new results indicate that equifinality is manageable, in particularly for key, sensitive parameters.

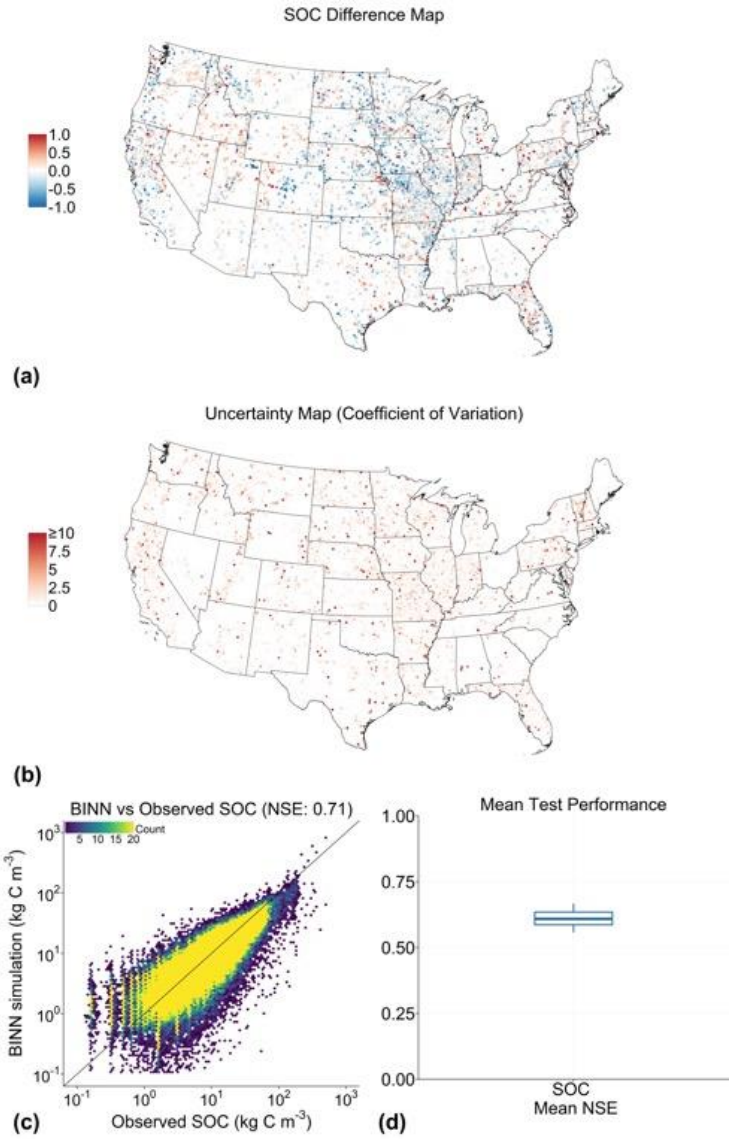

Figure R1: Comparison of observed and simulated SOC storage using BINN across 10 fold cross-validation test. (a) Spatial map of normalized difference across all 10 folds. The residual

is calculated by averaging the simulated SOC at each site across all 10 folds. Mean test-set differences are plotted after normalization: positive differences are scaled to [0,1] by the 99th percentile of positive deviations; negative differences are scaled to $[-1,0]$ by the 99th percentile of negative deviations. (b) Spatial map of the coefficients of variation across all 10 folds. (c) The scatter plot presents the SOC from data points derived from the mean values across all 10 folds between observed and simulated SOC storage at various soil depths, with the Nash–Sutcliffe modelling efficiency coefficient (NSE) shown in the title. (d) The box plot shows the mean performance of test NSE accross the 10-fold cross validation test.

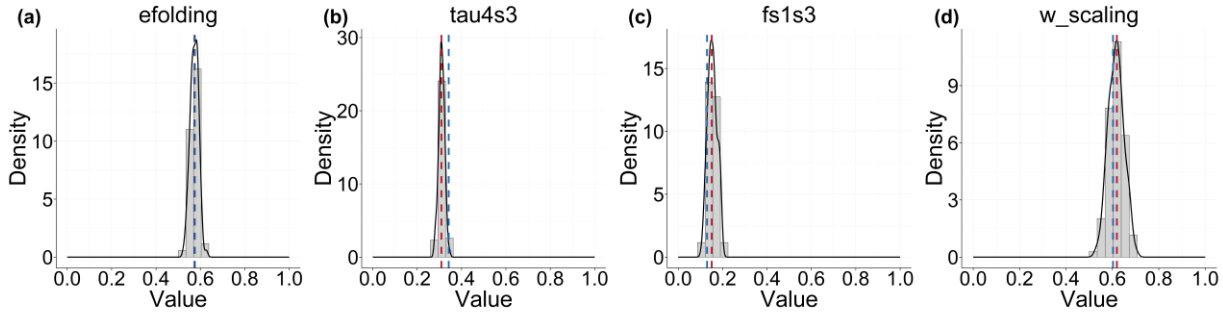

Figure R2. Uncertainty in posterior distributions by BINN parameters in relation to Monte Carlo dropout in the parameter recovery experiment at one site. For each parameter, the marginal posterior is shown, with the vertical blue and red dashed lines indicating the prescribed ("true") and BINN's point estimate, respectively.

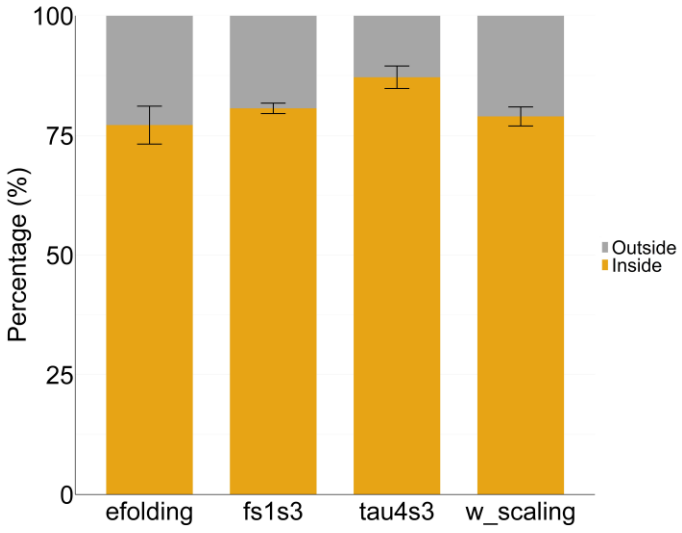

Figure R3: Coverage of parameters' posterior ranges of BINN. The coverage was quantified by the percentage of prescribed ("true") parameter values that fall into the Monte Carlo–dropout posterior range across test sites. Error bars show the mean ± SD across the 10 cross-validation folds of the synthetic parameter-recovery experiment.

Comment 2: **Sensitivity to Design Choices**: The sensitivity to certain study design choices that may have affected the entire article are not investigated: (1) Choice of a sigmoids activation

function per-parameter risks stabilizes the training, and is a regularization mechanism [1] but may cause the output parameters to be biased rather than interpretable results especially in the presence of noisy and/or sparse data [2]. (2) While I understand the prior is chosen based on literature, for certain parameters it doesn't seem certain other literature values. Most importantly: **tau4s3** (*Turnover time of passive SOC*) is set to 20-400 years seems to be short, **tau4s1** (*Turnover time of fast SOC*) minimum rate is set to 0.8 hr, and **w_folding** (*the influence of soil water on modifying SOC decomposition*) seems too wide, allowing **0.0001** may nullify water limitation and **5** being a large amplification.

1. Please, include references for each choice of the priors' range.
2. Please, investigate, if the choice of activation function and priors has biased the results.
3. Please, show more evidence as why you are convinced the model results indicate interpretability rather than bias.
4. Please, also explain why other model choices (like loss function) are made

We thank the reviewer for the valuable comments and suggestions. We address the concerns point-by-point as below.

(1) Parameter prior range. We've added Table R1 and related references for parameter prior ranges in in the supplementary materials of the revised manuscript (please see Table R1 and references at the end of this response letter). Specifically, we determine the prior ranges of parameters based on a synthesis of previous modeling work (e.g., CLM5 Technote), meta-analyses (e.g., Xu et al. 2016), and data assimilation studies (e.g., Shi et al. 2018). Following a common practice in ecosystem modeling and data assimilation, we select the minimum and maximum possible values reported in literature as the lower and upper ends, respectively in our prior range. This ensures the interpretability of parameter values that are within the range while also maintaining flexibility in searching the optimal values. Regarding the three parameters, i.e., tau4s3, tau4s1, and w_folding, that the reviewer pointed out, we set tau4s3 to be within 20 and 400 years, mainly considering the default value of about 230 years in CLM5 for tau4s3. Mechanistically, the respective passive SOC pool of tau4s3 represents SOC that is chemically and/or physically protected. Thus, we choose the lower end of 20 years and upper end of 400 years to represent least and strongest physiochemical protection, respectively. Similar rationale also applies to the prior range for tau4s1. While the lower end represents the turnover of the simplest, most decomposable organic matter, such as glucose, the upper end indicates slower turnover of more complex organic matters. For w_folding, the lower end reflects strongest soil water stress that depresses SOC decomposition, such as in the desert regions, whereas the upper end indicates most suitable soil water conditions that favor SOC decomposition. In the revised manuscript, we will describe how we set the prior ranges in the main text with more details in the supplementary materials.

(2) Influence of the choice of activation function on training results. Our use of the sigmoid function is to purely constrain the neural network outputs to be within [0, 1] and then mapped to each parameter's prior range. To address the reviewer's concern, in the revision, we conducted an additional experiment to examine the potential bias introduced by the choice of activation functions. Specifically, we retrained BINN with a hard-sigmoid (same bounding, but different gradient shape from the sigmoid function used in the original manuscript). We found that the

selection of activation functions has minimal influence on our optimization results and there is no significant change in NSE values when using different activation functions (Fig. R4b, c). Meanwhile, the spatial residual patterns between BINN-optimized simulation and observations (Fig. R4a) are nearly unchanged using the two different activation functions. All these results suggest that it is unlikely that the choice of activation functions will introduce systematic bias in model optimization (Fig. R4a–c).

(3) Evidence for interpretability. We've explicitly included three lines of evidence in the revised manuscript to demonstrate that BINN-optimized results offer high interpretability (Revised Section 4). First, in the synthetic experiment, BINN well recovers prescribed parameters and reproduces synthetic SOC, demonstrating that BINN learns parameters matching the known "ground truth", i.e., the prescribed parameters, rather than exploiting spurious correlations. Second, each of the predicted parameters is directly mapped to a CLM5 process component with clear biogeochemical meaning, for example, q10 describes how SOC decomposition may change under increasing temperature. When applying BINN to simulate SOC, the process components derived from BINN closely match those derived by PRODA, supporting a common mechanistic signal rather than activation function or prior-induced artifacts, as PRODA uses Bayesian-based calibration which is the current gold-standard data assimilation method for ecological modeling. Third, we apply traceability analysis to quantify the contribution of each component to variations in SOC storage across ecosystems, thereby linking learned parameters to process components to SOC outcomes. Taking them together, these results indicate that BINN provides mechanistic insight, not sorely statistical fit.

(4) Choice of loss function. We choose the Smooth-L1 loss to measure the SOC loss because it is robust in excluding outliers yet retains stable gradients near zero. Moreover, the soft-prior penalty (i.e., centered cosh function) encourages parameters toward middle of the ranges while allowing data-driven deviation. We've improved our descriptions about specific design of the BINN framework and the rationale for the selected loss function in section 2 to make our justifications clearer.

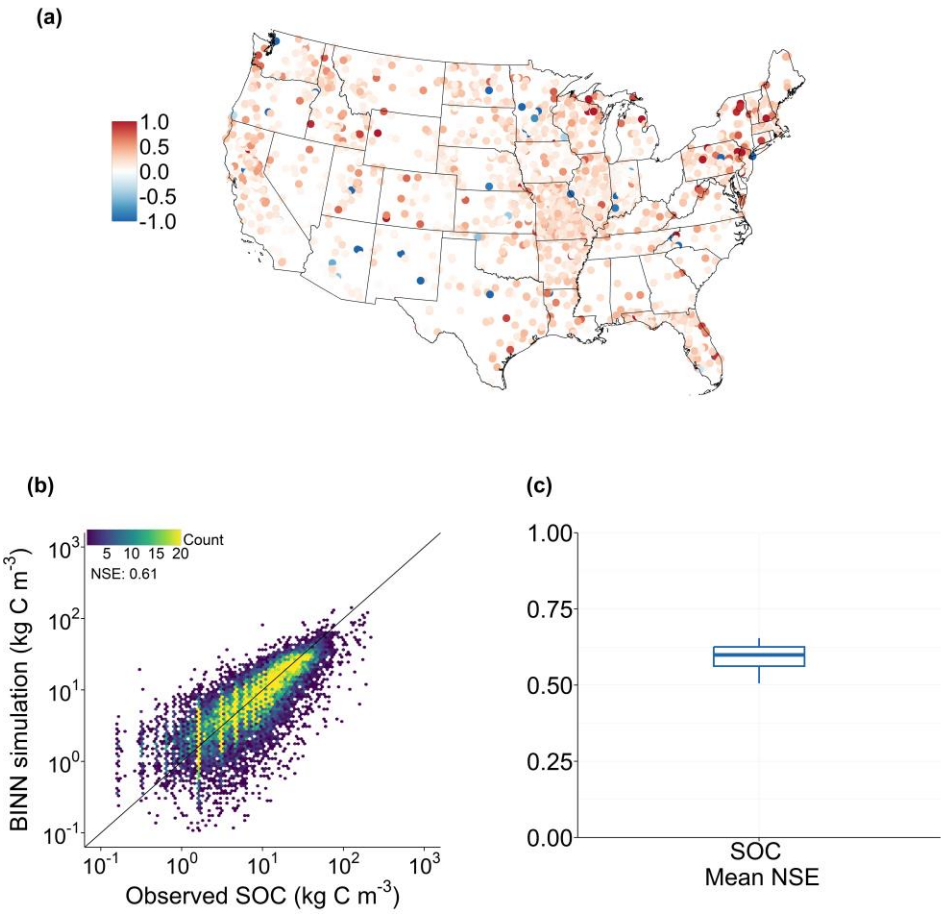

Figure R4: Comparison of observed and simulated SOC storage using BINN with Hardsigmoid activation function. (a) Spatial map of the normalized differences for one representative fold (median NSE). (b) Observed vs simulated SOC storage from the testing dataset. (c) The box plot shows the mean performance of testing NSE in the 10-fold cross validation test.

Comment 3: **Computational trade-offs:** The main difference between PRODA and this model is that this model considers all sites data across space at the same time. It is correct that reduction in computational time is expected, but the computational memory cost is expected to increase, with increased chance of data leakage in space. This model is claimed to offer a computational advantage to PRODA, but it is not made clear how much computational time is saved, how much memory load is increased. Please, consider acknowledging the trade-offs made in the BINN modeling framework to save computational costs compared to PRODA. Some are summarized as follows.

1. To claim the potential to extend to other regions or spatial generalizability of the framework, you would need to use leave-one-biome-out would test for the special generalizations (and if same NSE and correlation coefficients will be achieved)
2. Uncertainty quantification is a strong point in PRODA which also enhances model robustness and interpretability.

We thank the reviewer for bringing up the concerns about memory usage in BINN training. To address this concern, in the revision, we report the peak memory usage (RSS) for MCMC, PRODA with MCMC and BINN (with and without vectorized CLM5) on the same PC in Fig. R5. Our measurements indicate that BINN with vectorized CLM5 yields the least peak memory usage among all the methods, while also substantially reducing computational time (Fig. 8). Conceptually, BINN's memory usage can be kept bounded via smaller batches and gradient checkpointing, whereas the site-by-site Bayesian samplers maintain full state per chain, which tends to increase peak RSS despite longer running time. Thus, memory is not a key limiting factor in BINN training. We've included Figs. R5 and 8 in the revised section 3.4 and discussed this point in the revised manuscript.

Thanks for the wonderful suggestion on the leave-one-biome-out tests, which can be done in our future work. Our current scope is to infer parameters within the training data domain that involves various biomes, and we are not trying out-of-domain transfer to the biomes without training data. Thus, we do not include biome as a categorical environmental covariate. To clarify, our intent is to show that, even without explicitly encoding biome labels, BINN can recover processes that regulate SOC storage across biomes within the training data domain. Accordingly, we claim that BINN can be applied in regions where both models and observations are available to infer interpretable parameters to improve local SOC simulations. At present, BINN requires data from all target regions during training to support mechanism discovery and site-specific inference accuracy, whose inference cannot be generalized to those biomes without observations. We will clarify this and discuss this limitation in the revised manuscript. We note that PRODA also has this limitation; after it performs data assimilation on observation sites, it relies on a neural network to extrapolate to new sites.

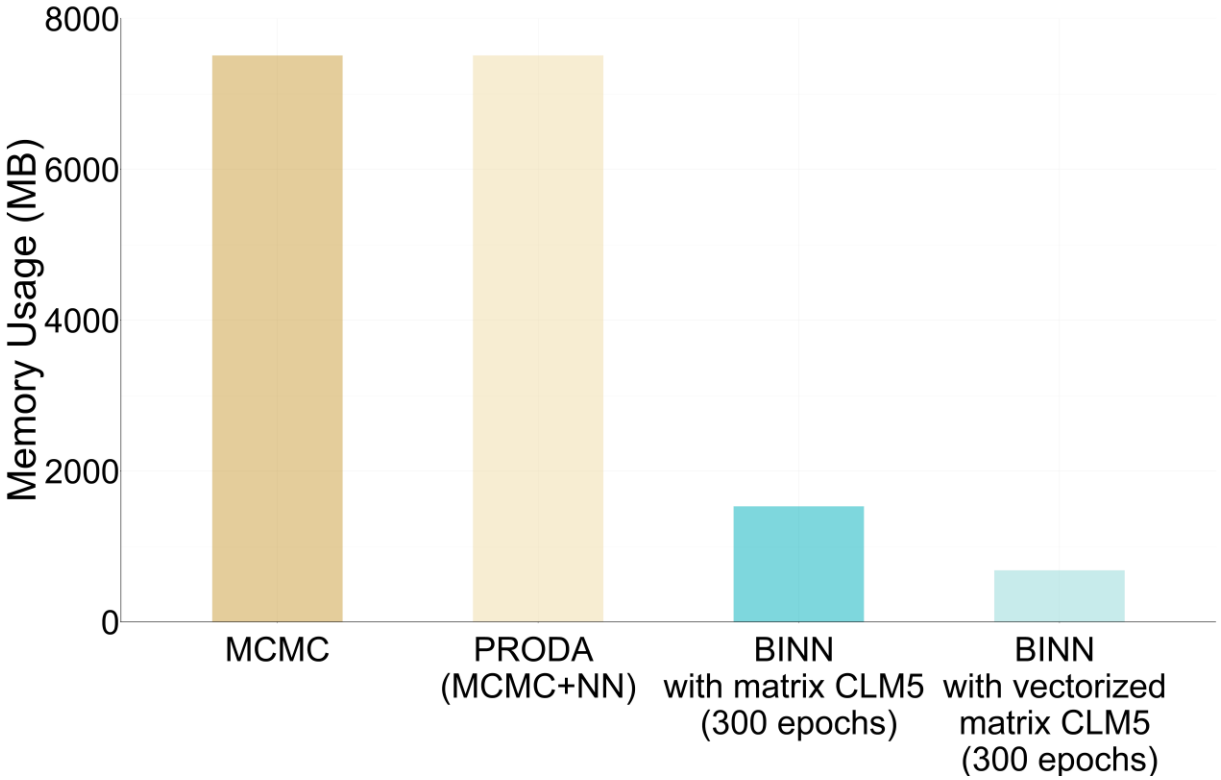

Figure R5. Comparative analysis of the peak memory (RSS) required for integrating 2000 soil profiles into process-based model (CLM5). The figure shows the maximum memory usage (in MB) for MCMC, for PRODA (MCMC+NN), which uses a Bayesian-inference approach (MCMC) combined with a neural network (NN), and for BINN with the matrix form of CLM5 before and after vectorization. The memory requirement is based on running each method for 300 epochs to ensure that the models have finished learning from the 2,000 soil profiles.

Comment 4: **Limited parameter testing:** Limited test cases and benchmarking to only a few of the 21 parameters. Only 4 out of 21 CLM5 parameters were actually recovered and validated. The sensitivity analysis (Fig. 3) justifies focusing on these, but ignores interactions and leaves 17 parameters untested. If the framework is claimed to be generalizable and "interpretable," but only 4 parameters were realistically tested, then the claims exceed the demonstrated evidence.

We fully understand the reviewer's concern. Initially, we focused on the four most sensitive parameters to show that BINN achieves comparable recovery where identifiability is strongest because these four parameters are typically well-constrained in Bayesian-based calibration. To address the reviewer's concern, in the revision, we have extended the synthetic recovery to all 21 biogeochemical parameters and now report correlations for each parameter (Fig. R5). As anticipated, highly sensitive parameters are recovered with strong correlations, whereas insensitive parameters exhibit weak identifiability, consistent with the sensitivity indices in Fig. 3 of the original manuscript. As our mechanistic inferences are based on combinations of

parameters rather than any single parameter, the low identifiability of some of the parameters may not hinder interpretability in downstream analyses.

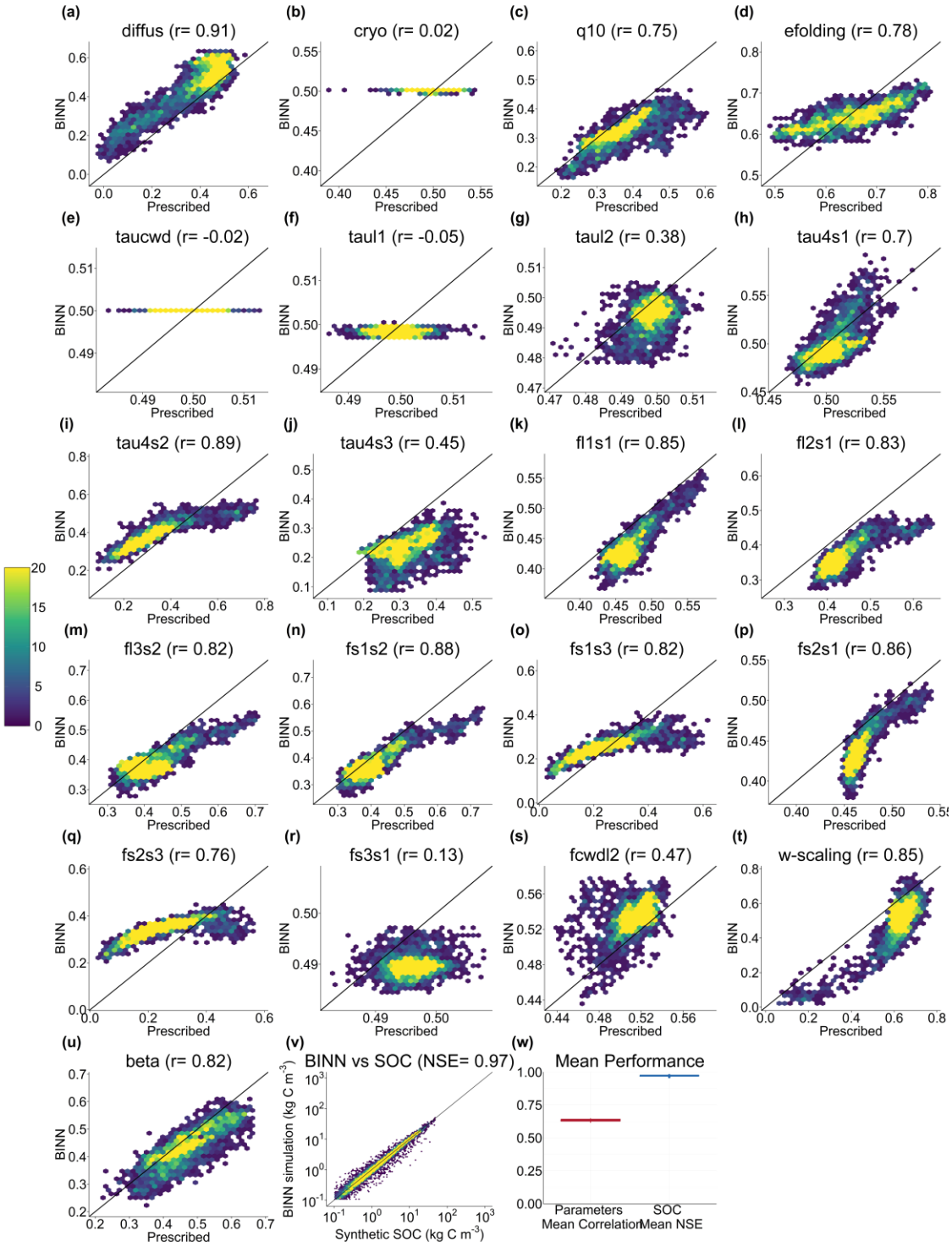

Figure R6: BINN parameter recovery for all CLM5 parameters. (a–u) Scatter plots of BINN-predicted versus prescribed values for each parameter. (v) Density scatter of simulated SOC versus synthetic SOC. (w) Summary across a 10-fold cross-validation: mean correlation for all parameters (predicted vs. prescribed), and NSE for SOC simulations.

Comment 5: The manuscript doesn't include line number, so I cannot unfortunately provide line-by-line comments. Please, consider this in the resubmission

We apologize for this inconvenience. We've included line numbers in the revised manuscript.

Comment 6: Figure 1, panel (a) and (b) need titles in the figure for better read. If colors contain information, please, be specific. "Priors" vs "Sigmoid activation" can be more explicitly separated.

Thanks for the suggestion. We've added concise titles to panels (a) and (b) in Figure 1 during the manuscript revision.

Comment 7: Figure 2, seems redundant.

Thank you for the suggestion. We agree that the main text can make the workflow clear enough without Figure 2. To keep the main text more streamlined, we've moved Figure 2 to Supplementary Information, as it may be useful for readers who are interested in the detailed parameter recovery workflow and its difference from the real-world application.

Comment 8: Figure 3, caption needs to explain better what kind of sensitivity test was carried out. The parameter labels need to be more intuitive. You can also consider giving colors to parameters that fall within one of your five broad categories (environmental modifier, CUE, substrate decomposability, ...). Include appropriate legends as needed.

We thank the reviewer for the suggestions. Regarding sensitivity analysis, we used a first-order approximation method following Gao et al. (2011) and will state the first-order sensitivity index in the title in the caption for Fig. 3. We performed the sensitivity analyses randomly at 512 sites across the Contiguous U.S., and in the revision, we will include the standard deviation of the sensitivity analysis in addition to the mean values in the sensitivity test results in Fig. R7, with parameters indicating individual process components (e.g., K, A, B, ξ, V) being distinguished by different colors.

During the revision, we've discovered and corrected a bug in our original sensitivity-analysis code. After this correction, the parameter "beta" emerged as having the highest average sensitivity across all depths. Accordingly, we updated the sensitivity results (Revised Fig. 2; Fig. R7) and reran the parameter recovery experiment using the five most sensitive parameters, now including "beta" (Revised Fig. 3). The main conclusions are unchanged, but the updated analysis more accurately reflects the relative importance of the key biogeochemical parameters.

Gao, C. *et al.* Assimilation of multiple data sets with the ensemble Kalman filter to improve forecasts of forest carbon dynamics. *Ecological Applications* **21**, 1461–1473 (2011).

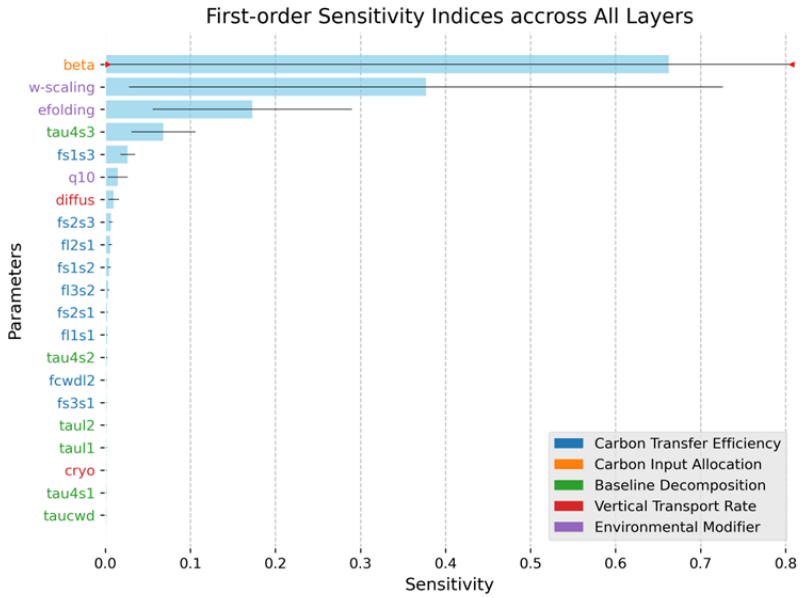

Figure R7. First-order sensitivity of CLM5 parameters across all soil depths. Bars show the first-order sensitivity index for each biogeochemical parameter over all soil layers. Red triangles indicate error bars clipped by the x-axis (uncertainty extends beyond the displayed range). Parameters are ranked (y-axis) by decreasing sensitivity and color-coded by their associated process component (e.g., K, A, B, ξ, V). The x-axis reports sensitivity scores, quantifying the influence of small parameter perturbations on model outputs, with larger values indicating greater influence.

Comment 9: Figure 4, please, use clear and sharp caption as these are your main contributions, such as "Fit of BINN to SOC data across CONUS in depth"

We've revised the figure 4 caption to be sharp and clear for readers to understand as the reviewer suggested (Fig. 3 in the revised manuscript).

Comment 10: Figure 5, please, make color scales consistent across all panels and legends be readable. Please, explain why in panel c and f BINN is constantly overestimating the PRADO, and why correlation collapses to 1 in r.

We thank the reviewer for the comments and suggestions. In the revision, we enlarge the legends for better readability. If panels share comparable ranges, we harmonize color scales. However, for components with distinct ranges with other components (e.g., plant carbon input), we will keep panel-specific scales (Fig. R8).

The reason why correlation between BINN and PRODA in the panel of 'Plant Carbon Inputs' collapses to 1 is because this component is NPP-driven and identical in both frameworks (both use the same CLM5 NPP simulations), so a collapse to r=1 is expected (Fig. R8r). We will explicitly explain this point in the revised manuscript.

The apparent overestimation in carbon transfer efficiency and baseline decomposition stems from the difference in the way sampling insensitive parameters under the two methods (Fig. R8c and f). In BINN, gradient-based training with a soft prior penalty tends to keep weakly identifiable parameters close to center of their prior range, which may shift composite components upward. In PRODA, the posterior distribution of the insensitive parameters is sampled uniformly from the prior value, thus the estimated parameters are not pulled toward the center as in BINN (Tao et al. 2024). Because these components aggregate both sensitive and insensitive parameters, BINN's mild central tendency can manifest as systematic positive bias relative to PRODA.

Tao F, Houlton BZ, Huang Y, Wang YP, Manzoni S, Ahrens B, Mishra U, Jiang L, Huang X, and Luo Y. 2024. Convergence in simulating global soil organic carbon by structurally different models after data assimilation. *Global Change Biology*, 30: e17297.

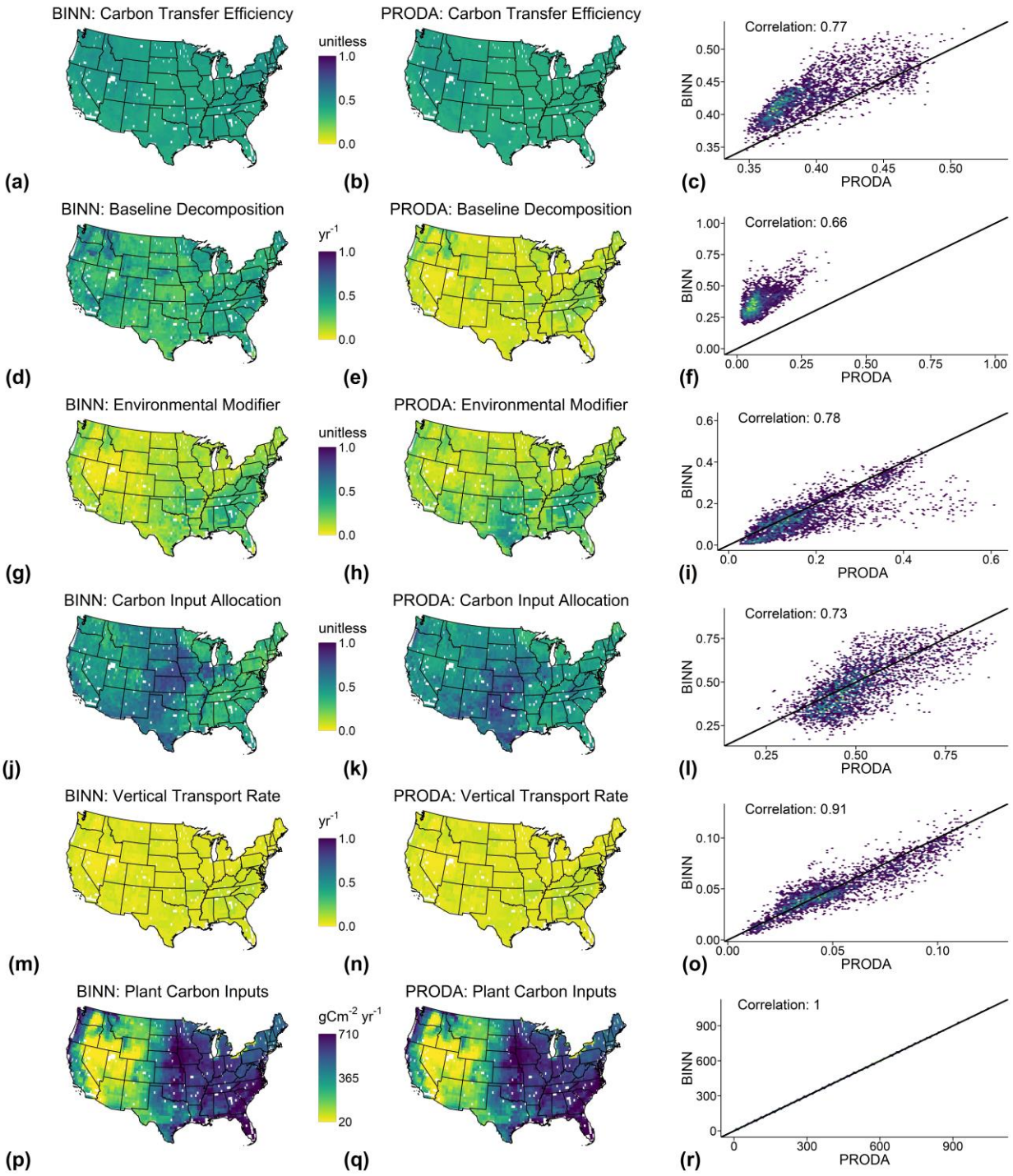

Figure R8. Comparison of the spatial patterns of model components retrieved by BINN and PRODA. The model components include carbon transfer efficiency (a, b, c), baseline decomposition (d, e, f), environmental modifier (g, h, i), carbon input allocation (j, k, l), vertical transport rate (m, n, o), and plant carbon inputs (p, q, r). The left column (a, d, g, j, m, p) shows the model components retrieved by BINN, while the middle column (b, e, h, k, n, q) displays the model components retrieved by PRODA. The scatter plots in the right column (c, f, i, l, o, r) compare the values of each model component retrieved by BINN (y-axis) against those retrieved

by PRODA (x-axis). The correlation coefficient between the BINN and PRODA values for each model component is shown in the top left corner of the corresponding scatter plot. The plant carbon inputs (p, q, r) are identical for both methods due to the use of the same input forcing data.

Comment 11: Figure 6, bias and error distributions. Hard to interpret geographically. Captions don't explain ecological meaning of bias hotspots. Needs to be restructured in agreement with how you would handle the major revisions.

We fully understand the reviewer's comment that it is challenging to interpret the current SOC residual map. To address this concern, we normalized the map with 99 percentiles instead of the maximum and minimum values for better interpretability. The revised figure now shows the major bias hotspots located in the center of U.S. (Fig. R9). During the revision, we've added a map with normalized differences across all cross-validations to show the uncertainty in data (Revised Fig. 5).

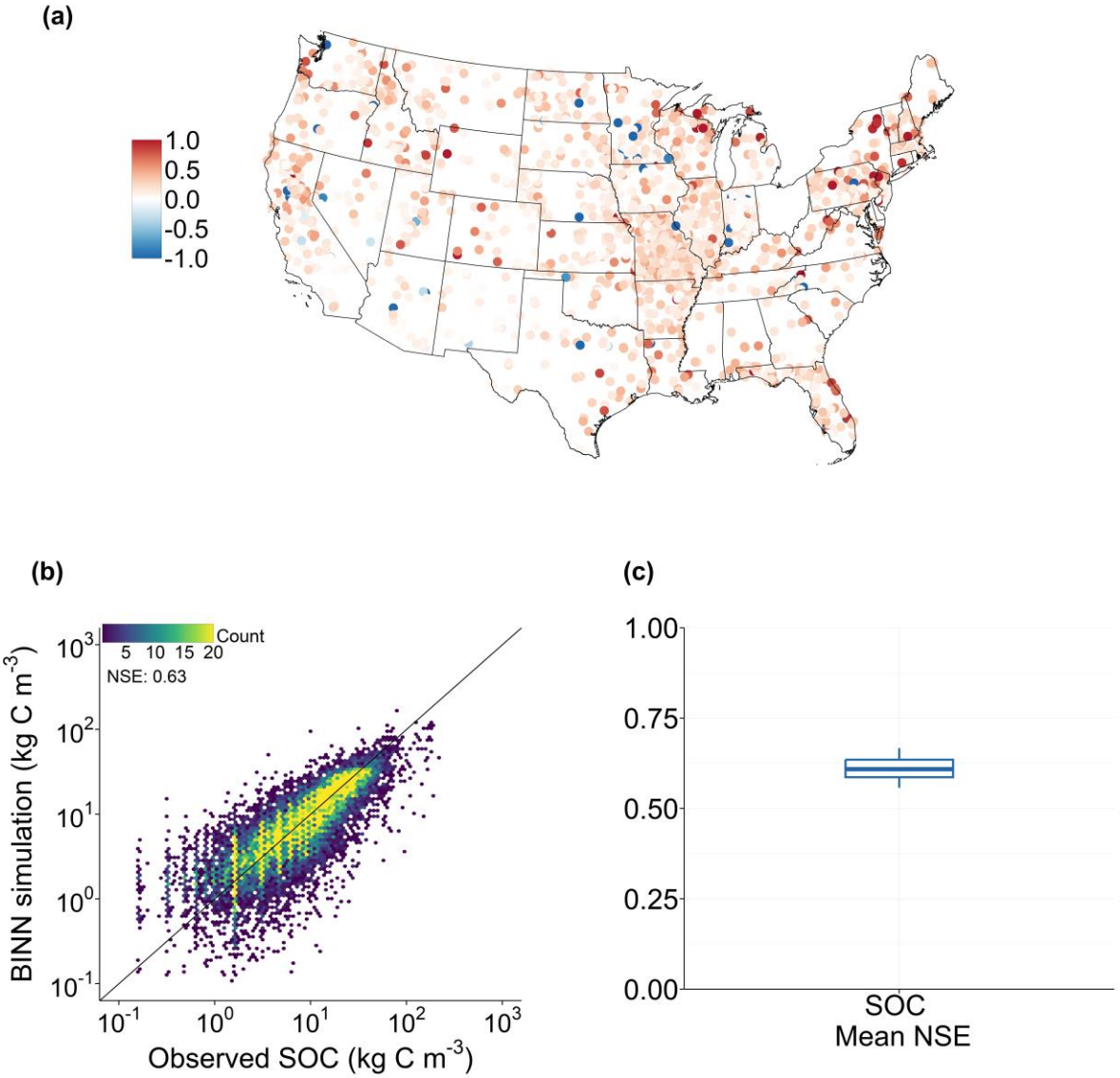

Figure R9. Comparison of observed and simulated SOC storage using BINN. (a) Spatial map of residuals for one representative fold (median NSE). (b) Observed vs simulated SOC storage from the testing dataset. (c) The box plot shows the mean performance of testing NSE in the 10-fold cross validation test.

Comment 12: Figure 7 decomposes SOC into carbon input vs residence time across biomes without any independent data or synthesis data to validate biome-level trade-offs. A needed step before presenting this figure is testing BINN on analytical or the reference [3] can be used to create a test case for validity of residence times derived from the BINN, before its application to CONUS.

We thank the reviewer for the helpful suggestion. We agree that validating biome-level trade-offs with independent benchmarks would strengthen the claim. As the primary objective of our

current traceability analysis is to help mechanistic interpretation from SOC data within the training data domain, out-of-domain or biome-level validation are outside the scope of this study. Please refer to our previous responses to Comment 3. However, we are grateful for the wonderful suggestion, which can help guide future work.

Comment 13: Figure 8, please, restructure and consider the points in the major revision

As suggested, we have added analysis on memory usage and included the new Fig. R5 in Supplementary Information and described the memory usage in the main text (Revised Section 3.4). Therefore, it would be appropriate if we keep Figure 8 to highlight the computational efficiency in the main text, as it is the most remarkable operational benefit of our newly developed BINN.

References in Table R1:

Lawrence, D. M. *et al.* The Community Land Model Version 5: Description of New Features, Benchmarking, and Impact of Forcing Uncertainty. *Journal of Advances in Modeling Earth Systems* **11**, 4245–4287 (2019).
Shi, Z., Crowell, S., Luo, Y. & Moore, B. Model structures amplify uncertainty in predicted soil carbon responses to climate change. *Nat Commun* **9**, 2171 (2018).
Zhang, D., Hui, D., Luo, Y. & Zhou, G. Rates of litter decomposition in terrestrial ecosystems: global patterns and controlling factors. *Journal of Plant Ecology* **1**, 85–93 (2008).
Xu, X. *et al.* Soil properties control decomposition of soil organic carbon: Results from data-assimilation analysis. *Geoderma* **262**, 235–242 (2016).
Davidson, E. A. & Janssens, I. A. Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature* **440**, 165–173 (2006).
Davidson, E. A., Janssens, I. A. & Luo, Y. On the variability of respiration in terrestrial ecosystems: moving beyond Q10. *Global Change Biology* **12**, 154–164 (2006).
Koven, C. D., Lawrence, D. M. & Riley, W. J. Permafrost carbon−climate feedback is sensitive to deep soil carbon decomposability but not deep soil nitrogen dynamics. *Proceedings of the National Academy of Sciences* **112**, 3752–3757 (2015).

Table R1: 21 biogeochemical parameters in CLM5

No.	Name	Matrix Term	Corresponding Mechanism	Description	Unit	Prior Range	Reference
1	fl1s1	A	Microbial carbon use efficiency (CUE)	Transfer fraction, from metabolic litter to fast SOC	unitless	[0.1, 0.8]	Lawrence et al. 2019; Shi et al. 2018
2	fl2s1			Transfer fraction, from cellulose litter to fast SOC	unitless	[0.2, 0.8]	
3	fl3s2			Transfer fraction, from lignin litter to slow SOC	unitless	[0.2, 0.8]	
4	fs1s2			Transfer fraction, from fast SOC to slow SOC	unitless	[0.0001, 0.4]	
5	fs1s3			Transfer fraction, from fast SOC to passive SOC	unitless	[0.0001, 0.1]	
6	fs2s1			Transfer fraction, from slow SOC to fast SOC	unitless	[0.1, 0.74]	
7	fs2s3			Transfer fraction, from slow SOC to passive SOC	unitless	[0.0001, 0.1]	
8	fs3s1			Transfer fraction, from passive SOC to fast SOC	unitless	[0.0001, 0.9]	
9	fcwdl2			Transfer fraction, from coarse woody debris to cellulose litter	unitless	[0.5, 1]	
10	tau4cwd	K	Substrate decomposability	Turnover time of coarse woody debris	year	[1, 6]	Zhang et al. 2008
11	tau4l1			Turnover time of metabolic litter	year	[0.0001, 0.11]	
12	tau4l2			Turnover time of cellulose litter	year	[0.1, 0.3]	
13	tau4s1			Turnover time of fast SOC	year	[0.0001, 0.5]	Xu et al. 2016
14	tau4s2			Turnover time of slow SOC	year	[1, 10]	
15	tau4s3			Turnover time of passive SOC	year	[20, 400]	
16	q10	ξ	Environmental modifiers	Temperature sensitivity	unitless	[1.2, 3]	Davidson et al. 2005; Davidson et al. 2006
17	efolding			E-folding parameter to calculate depth scalar	meter	[0.1, 1]	Lawrence et al. 2019; Shi et al. 2018
18	w_scaling			Scaling factor to soil water scalar	unitless	[0.0001, 5]	
19	bio	V	Vertical transport	Bioturbation rate	m2/yr	$[3\times10^{-5}\ 16\times10^{-4}]$	Koven et al. 2015
20	cryo			Cryoturbation rate	m2/yr	$[3\times10^{-5}\ 5\times10^{-4}]$	
21	beta	I	Carbon input	Vertical distribution of carbon input	unitless	[0.5, 0.9999]	Lawrence et al. 2019; Shi et al. 2018