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

We thank Reviewer 1 for the thoughtful and constructive comments. To resolve reviewer's concern, we have recalibrated the manuscript's framing to prioritize the functional utility of the BINN framework while removing broader conceptual overstatements. We also stated the analysis of equifinality and uncertainty quantification via MC dropout more prominently. Additionally, we clarified the BINN nomenclature to distinguish it from traditional PDE-emulating models.

Reviewer #1 from the previous comment-response round
Comment 1: The reviewer questioned the novelty of the approach. I share this concern, and the response of the authors could not fully resolve that. The possibility of including other models is also inherent to other code bases and not demonstrated within the manuscript, furthermore do the following two articles describe similar end-to-end setups (and further articles could be found):
1) Inferring dynamic parameters from dynamic covariates: ElGhawi et. al, Hybrid modeling of evapotranspiration: inferring stomatal and aerodynamic resistances using combined physics-based and machine learning, Environmental Research Letters, 2023.
2) Inferring static parameters from static covariates (in dynamic model): Baghirov et. al, H2MV (v1.0): Global Physically-Constrained Deep Learning Water Cycle Model with Vegetation, Geoscientific Model Development, 2024.
The novelty of doing similar but in a purely static or steady-state dynamic setup, using the matrix approach, uncertainty estimation or the application in a soil context are indeed given, but current phrasing overstates the innovation.

Response 1: We sincerely thank the reviewer for pointing out possibly overstated innovation in our manuscript. We have thoroughly revised text to avoid any possible overstatements in this version.

We also agree that the ability to incorporate additional models to such alike frameworks is not unique to BINN. Thanks to the reviewer for pointing us the two relevant studies, we have included them into our introduction section on line 85-87.

We are grateful to the reviewer to suggest that the novelty of our study lies in applying similar ideas in a purely static or steady-state framework, using a matrix-based approach, incorporating uncertainty estimation, and applying it specifically to soil systems. We have highlighted the novelty in the revised text on lines 594-601 as below:

"Our study demonstrates that BINN is a novel approach to retrieve spatially varying model parameters from big data and predict their spatial distributions for determining SOC dynamics over the contiguous US and thereby facilitate process understanding. By embedding a matrix form of process-based model into BINN, BINN can significantly enhance computational efficiency during parameter optimization compared to conventional data assimilation techniques. Additionally, the incorporation of MC dropout enables the estimation of posterior distribution for each parameter, facilitating rigorous uncertainty quantification within this hybrid modeling framework."

Comment 2: The reviewer asked for further introduction of PRODA (as the article focuses on comparison to it) and clarification on the PRODAs abilities. The extended introduction in response is sufficient to inform the readers. Also the added evaluation is completing the comparison of BINNs to PRODA. Nevertheless, are some conclusions, drawn from the correlation to PRODA, overstated.

1) 31-32: R of 0.77 with PRODA confirms correlation with PRODA, and the here made conclusion of that is not proven.

2) 46: Modeling assumptions are combined with a large dataset. As predictions of processes were not compared to independent data sources, identifying is too strong wording. This claim of identifying processes is made multiple times in the manuscript (e.g. 586-588).

Generally is PRODA presented as methodologically flawless. The fact that wrong models can also fit to data is not mentioned. The requested comparison on a methodological level is also not yet present.

Response 2: We thank the reviewer for raising concerns on the comparison of BINN with PRODA. We have replaced the term "identifying" with "retrieving" to emphasize the point of using BINN to infer processes from observations. We agree with the reviewer that "wrong models can also fit to data". We compared retrieved parameters between BINN and PRODA partly because PRODA uses Bayesian optimization, which is widely accepted in earth system modeling for parameter estimation through data assimilation. The comparison may not be perfect, but their agreement suggests that BINN can effectively capture spatial variations in key model components while maintaining physical interpretability. This point is stated in the paragraph on lines 640-649 as:

"We further tested BINN with real-world SOC observations across the contiguous US and quantified the six model components of CLM5. BINN-retrieved model components, calculated from the estimated biogeochemical parameters, showed good agreement with those generated by PRODA. Since Bayesian optimization as used in PRODA is widely accepted in earth system modeling for parameter optimization through data assimilation, the agreement between BINN and PRODA suggests that BINN can effectively capture spatial variations in key model components while maintaining physical interpretability. This capability enables BINN to evaluate the relative importance of different processes controlling SOC storage within the model and data domain, as done with PRODA, while offering computational advantages through its integrated neural network architecture."

Comment 3: The reviewer requested more analysis and discussion on equifinality. In response this issue is named in the text, but not prominently in abstract or conclusion. MC dropout is a interesting approach to identify equifinality, which could also be stated more prominently. But methodological I have following the following concern: 265-266: Implausible values should be prevented by the sigmoid already. Is this done to limit uncertainty of retrieved parameters? Please explain which effects the use of the hyperbolic cosine might have on the dropout-inferred uncertainty.

Response 3: We thank the reviewer for recognizing our effort to address the issue of equifinality and our use of MC dropout. In this revision, we have explicitly stated this issue in the abstract and discussion sections. Specifically, MC dropout is introduced in the abstract L26-L28 as below:

“Furthermore, by incorporating Monte Carlo (MC) dropout to generate posterior distributions, we demonstrate that BINN can effectively quantify uncertainty in estimated parameters.”

We also have expanded discussion on how MC dropout evaluates equifinality explicitly in the discussion L633-639 as below:

“To further characterize and mitigate equifinality, we applied MC dropout to approximate the posterior distributions of site-level parameters. Although point estimates did not necessarily match the prescribed values exactly, the posterior distributions frequently cover the ground truth (i.e., the prescribed parameters in this case). This demonstrates BINN’s ability to identify equifinality issues. By generating a distribution of potential parameter values instead of a single point estimate, the framework provides a robust quantification of uncertainty that carries through from the biogeochemical parameters to the final model simulations.”

Regarding the use of the hyperbolic cosine loss term, this was implemented to improve training stability and prevent vanishing gradient issues that can occur when parameters approach the extreme limits of the sigmoid function. While the sigmoid ensures that outputs remain within a [0,1] range, the additional loss term prevents parameters from saturating at these extremes, which generally represent rare ecological circumstances.

Because the weight of this loss term is relatively small, its influence on dropout-inferred uncertainty is primarily concentrated on weakly constrained parameters. In these instances where the signal from SOC data is less informative, the hyperbolic cosine loss provides a supplementary constraint that helps shape the posterior distribution.

Comment 4: The reviewer identified possible confusion given the naming of the framework. The authors extended the introduction in response to position the framework as not being within the scope of PINNs. As conclusion I see the potential confusion even more obvious. Unfortunately is the almost identical naming to a (within the hybrid modeling world) very different method only resolvable by different naming.
In my understanding are PINNs emulating Differential equation (systems), while the here presented work learns parameters from covariates and thereby extends rather the given system to include unknown variables (covariates):
1) Raissi, Perdicaris, Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, Journal of Computational Physics, 2019.
2) Karpatne, Jia, Kumar, Knowledge-guided Machine Learning: Current Trends and Future Prospects, 2024

Response 4: We thank the reviewer for raising the point of potential confusion about the naming of BINN. We agree that PINN and BINN represent different frameworks, as PINNs typically emulate differential equations while BINN retrieves parameters for defined equations. However, our framework was inspired by the broader development of physics-informed neural networks and has a similar motivation of using physical knowledge (process-based models) to inform and constrain neural networks. For that reason, we have chosen to retain the name BINN.

Comment 5: Furthermore do I see following issues in the positioning of the method (apart from points stated to Comment 1):
75: Hard-PINNs would do that. (Le et. al, Physics-Informed Neural Networks with Hard Constraints for Inverse Design, Computational Methods in Science and Engineering, 2021)

Response 5: We thank the reviewer for the suggestions regarding hard physical constraints in PINN and we have incorporated a discussion of Hard-PINNs into our introduction in L78-81 as follows:

“While recent 'Hard-PINNs' developments have sought to enforce these constraints more strictly to minimize the difference between NN predictions and numerical model simulations[12], this requires a complex multi-level optimization procedure with many hyperparameters and still does not satisfy the constraint exactly.”

Comment 6: 83-85: What is challenging about it? The impact of the paper depends on the challenges which were overcome here.

Response 6: We have expanded description on the specific challenges hybrid models face when applied to complex systems like soil organic carbon dynamics, particularly the high computational costs and the equifinality issues arising from high-dimensional parameter spaces. Specifically, in L90-95, the challenges of current hybrid models are stated as:

“It remains challenging to integrate numerous differential equations into a NN to study the dynamics of a complex system, such as SOC dynamics. This difficulty stems from the tremendous computational cost required to execute complex models, and the associated high-dimensional parameter spaces that need to be optimized for predictive model simulations. In particular, the capacity of NN to efficiently optimize process-related parameter values with faithful identifiability remains an under-explored area of research.”

Comment 7: 95: Does reference 23 really say, that GWR has this limit?

Response 7: Regarding the GWR citation, we have included the additional citation to the original Geographically Weighted Regression paper (Brunsdon et al. 1996 [1]) and explain that:

“While statistical methods such as Geographically-Weighted Regression allow spatially-varying regression coefficients, they can only model linear relationships between inputs and outputs at each location[27,28]” in L102-104.

[1] Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. Geographical analysis, 28(4), 281-298.

Comment 8: While the mentioned issues were resolved, I had following additional concerns while reading this section.
325-326: Why separation of depth here? Especially as results are later summarized again.

Response 8: We greatly appreciate the reviewer for the comments. Regarding the separation of depth, we retain results across different depth ranges because sensitivity scores vary significantly for the same parameters across depths, which provides critical information for studies focusing on specific soil layers.

Comment 9: 400: Only positive inputs means, that LeakyReLU uses only positive linear part. No activation happens and effect of nodes collapses to a linear combination equally to Linear regression. Further should the sigmoid function only get positive inputs and only the upper half of the parameter range is effectively used. Problem might be fixed by the embeddings. But please clarify.

Response 9: For the neural network architecture, we clarify that even if all the inputs to the neural network are positive, if we pass them through a single linear layer (with both positive and negative weights/biases), the result will contain both positive and negative values. Thus, the LeakyReLU activation should receive both positive and negative inputs.

Comment 10: 434: Why is NSE used for accuracy and r for reporting parameter identification?

Response 10: We utilize NSE to assess the accuracy of SOC simulations because it evaluates how well the model replicates physical observations. For parameter comparisons, however, the correlation coefficient (r) is more appropriate because we are evaluating the spatial agreement between two different inference methods rather than a comparison against direct field measurements.

Comment 11: 443: This approach doesn't distinguish between train and test samples. More informative would be to take only the test samples per fold.

Response 11: We appreciate the reviewer's request for clarification regarding the data subsets utilized in our visualizations. Figures 6c and 6d represent the aggregated results of the independent test partitions across the 10-fold cross-validation. We maintained identical test sets for both BINN and PRODA to ensure a rigorous and standardized benchmark comparison. While Figure 5a incorporates the training, validation, and testing subsets to provide a comprehensive diagnostic of simulation uncertainty across the entire 10-fold cross-validation ensemble. We have updated the caption for Figure 5 to reflect this distinction: "Figure 5: Comparison of simulated and observed SOC storage using BINN across 10-fold cross-validation training."

Comment 12: 467-468: Realistic time would be with early stopping (evaluation is provided everywhere)
for each method. One converging faster than the other is an essential part of efficiency.

Response 12: In this study, PRODA utilized early stopping with patience of 1200 epochs while BINN did not, while both methods were trained until convergence to ensure a consistent comparison of computational efficiency. We have updated the Method section in L459-460 with "PRODA was trained with 6000 epochs, utilizing an early stopping patience of 1200 epochs, and validated with a 10-fold cross-validation."

Comment 13: 168: This is called a parameterized sigmoid.

Response 13: We updated the manuscript to use the term "parameterized sigmoid" as suggested in L183.

Comment 14: 181: CPU number is a hardware setting rather than a hyperparameter

Response 14: While the CPU number is a hardware setting, it is also a hyperparameter because it impacts the effective batch size and optimization dynamics via PyTorch DistributedDataParallel.

Comment 15: 183: What are the loss function hyperparameters?

Response 15: The loss function incorporates a scale factor (τ = 15) for the hyperbolic cosine loss term, with a balancing weight (w = 100). Detailed loss function hyperparameters are now provided in section 2.1.3.

Comment 16: 188: Dropout rate was not mentioned earlier. Generally should the hyperparameter optimization be described after the framework, so terms are clear to the reader.

Response 16: The dropout rate is now introduced earlier in L166-170 to improve clarity:

"To prevent overfitting and enable uncertainty quantification through MC dropout, we incorporated a dropout layer after each hidden layer (Equation 2). During BINN training process, neurons are randomly deactivated with a prescribed dropout rate (see below for hyperparameter selections), allowing the network to explore a range of model configurations."

Comment 17: Eq. 11: Why is a capital p used here? Additional S indices for site and depth could clarify
further what was done. Generally could paragraph 2.2.1 profit from rephrasing. Name it "first-order Sobol index".

Response 17: We appreciate the reviewer's rigorous review of our mathematical notation. To improve clarity and maintain consistency, we replaced the capital P with a lowercase p. Additionally, we have updated the sensitivity notation to $S_{i,s,d}$ to represent indices specific to each site and depth range. Following the reviewer's suggestion, Section 2.2.1 has been renamed to "First Order Sobol Sensitivity Analysis".

Comment 18: 380-381: What was validation used for?

Response 18: The validation set was used during the training phase as an independent dataset to guide hyperparameter tuning, best model selection, and early stopping to prevent overfitting. We added this explanation on line 411-415. "Specifically, the training subsets were utilized for gradient-based optimization, the validation subset served as an independent check to guide hyperparameter tuning, best model selection, and early stopping to prevent overfitting, and the test subset was reserved for an unbiased assessment of the BINN's generalization capability."

Comment 19: 388: By averaging, potential model instability is rather hidden.

Response 19: Thanks for the reviewer for raising this issue. Potential model instability can be directly detected from the boxplot of the test accuracy (Figure 5d). In our 10-fold cross validation results, we could check the standard deviations throughout the boxplot results and did not observe model instability during our training process.

Comment 20: 456-458: I am not sure to which results this sentence hints, please clarify!

Response 20: Thanks for the comments. We have deleted the traceability analysis section and therefore the specific sentence and the associated results referenced in this comment are no longer present in the revised version.

Comment 21: Why was spatial encoding used if it bares potential of data leakage and did not improve predictions?

Response 21: Spatial encoding is used that if there are unobserved, spatially-dependent variables that impact the biogeochemical parameters, the spatial encoding could be able to learn this "spatially-dependent error" as a function of (latitude, longitude). While it is not strictly needed here, it could be useful for other applications where there are missing environmental covariates.

Reviewer #2 from the previous comment-response round

Comment 22: The reviewer requests more modest wording. Although this was taken care of mostly, following points still show problematic wording:

Response 22: We sincerely thank the reviewer for the follow-up assessment. We acknowledge that maintaining a modest and precise technical tone is essential for the clarity of this manuscript.

Comment 23: 18: Big data is a, not defined, buzzword and should be avoided.

Response 23: We have updated the terminology throughout the manuscript to ensure more precise and modest wording. Specifically, in abstract, we replace "big data" to "large-scale observational data" or "large-scale data" to provide a clearer technical definition.

Comment 24: 34: "transformative tool that harnesses the power of both AI and process-based modeling" was specifically mentioned by the reviewer and still remains in the text.

Response 24: As suggested, we also changed "transformative tool" to "efficient framework" in L37 to more accurately represent BINN's contribution.

Comment 25: 53-55: Starting from AI is too broad here. Machine learning is the discipline, deep learning is one method of machine learning.

Response 25: Thank the reviewer for clarifying the differences among these terms. Regarding the word choices of AI, we have retained the broader term in the opening to maintain accessibility for an interdisciplinary audience before detailing the specific machine learning and deep learning methodologies in the subsequent text.

Comment 26: 65-67: Redundant.

Response 26: We revised L66 to 69 to remove redundant phrasing and ensure the content is concise and distinct:

"This integration leverages the strengths of data-driven techniques alongside scientific theory. By introducing physical or knowledge-based constraints and reasoning, AI-based predictions can be further constrained by scientific knowledge in addition to being driven by observational data."

Comment 27:
1) Resolved.
2) Vide notes to Reviewer #1, Comment 3.
3) Vide notes to Reviewer #1, Comment 2.
4) Vide notes to Reviewer #1, Comment 3.

Response 27: Thank the reviewer for acknowledging that we have addressed these issues. These comments, i.e., Comments 2 and 3 by Reviewer #1, are addressed above in our responses #2 and #3.

Comment 28:
1) Resolved.
2) This could partly support the claims on causality. It is fine to leave that, but sticking to modest claims it necessary then.

Response 28: Thank the reviewer for acknowledging that we have resolved these issues.

Comment 29: Apart from concerns mentioned earlier, I consider these as resolved.

Response 29: Thanks for the acknowledgement.

Comment 30: Fig. 5: I recommend to use a different color bar (partly white color code on white background).

Response 30: Thanks for pointing out the color scheme issue, we have replaced white with grey for better visualization (Figs. 5, 6).

Comment 31: This could partly support the claims on causality. It is fine to leave that, but sticking to modest claims it necessary then.

Response 31: Thanks for reviewing the comment.

Additional remarks
Comment 32: 498: approximate posterior distribution

Response 32: We have updated the word "obtain" to "approximate" in L326.

Comment 33: 502: Which interval is used here?

Response 33: We thank reviewer for pointing out the potential confusions, we have changed "intervals" to "posterior distribution" in L331.

Comment 34: 503: Can be, could be or should be? What is the general take home message of this paragraph? That equifinality is controlled well? If not, what would be the implications for results?

Response 34: We have added a sentence explaining that the equifinality can be effectively identified through MC dropout in L536-540: " We further quantified the site-level coverage (i.e., the fraction of prescribed values falling into the BINN-predicted posterior distributions) in Fig. S6, which indicated the issue of equifinality is effectively identified and accounted for within the BINN framework, as the uncertainty intervals consistently encompass the ground-truth parameter values at most sites."

Comment 35: 506-507: Maybe I oversaw a reference on how this is done. Please call it 'Traceability analysis' throughout the text to help the reader.

Response 35: We added a detailed explanation in the methods section (L274-279) to distinguish our model components from standard traceability analysis. In addition, we have removed the traceability analysis section from this revision, and thus we believe model components will no longer be messed up with 'Traceability analysis' in our manuscript.

Comment 36: 540: I think that is not Figure S6.

Response 36: Thanks for pointing out the typo. It's now fixed.

Comment 37: 553: This would be the definition of equifinality.

Response 37: The sentence on that line explains how similar SOC levels can be attributed to fundamentally different underlying mechanisms (i.e., traceability analysis). However, since the traceability analysis section has been removed from the current revision, the specific discussion referenced in this comment has been omitted.

Comment 38: 570: S7 should be S8.

Response 38: Thanks for pointing out the typo. It's now fixed.

Comment 39: 4.1 and 4.3 could be one paragraph.

Response 39: We chose to maintain the separation of sections 4.1 and 4.3 because they serve distinct purposes. Section 4.1 focuses on the current results and performance of BINN while section 4.3 discusses the future potential for using this framework to guide independent validation and new mechanistic studies. Nevertheless, we greatly appreciate the reviewer's suggestion.

Comment 40: 635-637: Also well understood processes are parameterized.

Response 40: We believe that even well-understood processes require parameterization within a model structure. In soil systems, significant spatial heterogeneity necessitates the processes be parameterized differently across various locations and environmental. To reflect this, we have revised the text in L704–707:

"Process-based models may explicitly represent those well-understood processes in their structure while these processes need to be specified via parameterization at given locations under given conditions to represent the heterogeneity of environmental conditions"

Comment 41: 640: Fig 6a shows that there is no improved model fit.

Response 41: We thank the reviewer for the good observation. Yes, BINN does not achieve a better model fit than PRODA. Rather, we developed BINN to improve computational efficiency and reduce spatial biases. To address the ambiguity in the referenced sentence, we have refined the text (Lines 710-711) to clarify that the improvement in model fit refers to BINN's performance relative to the unconstrained model, rather than a comparison between two methods.

Comment 42: 652-666: This is an outlook and should be formulated throughout in the conjunctive.

Response 42: These two paragraphs are now formulated in the conjunctive mood to reflect their hypothetical nature.

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

We thank Reviewer 2 for the thoughtful and constructive comments on our revised manuscript. In this revision, we have performed a sensitivity analysis, which confirms the framework's robustness to choices of prior ranges, activation functions, and loss functions. We have strictly recalibrated the manuscript's scope to define the retrieved process-related parameters as model-derived and within training-data domain, and also removed the traceability analysis section to avoid potential confusion. We further discuss how we use model components in our downstream analyses to address the issue of heterogeneous identifiability. Furthermore, we resolved previous systematic biases between BINN- and PRODA-predicted model components by harmonizing prior ranges for parameters.

General Assessment

Comment 1: While the manuscript now clearly defines and justifies the prior ranges and the role of the sigmoid mapping and regularization terms, the evaluation still relies on the same bounded priors. No quantitative sensitivity analysis is provided to assess how dependent the results are on the choice of prior bounds. Thus, my original concern regarding robustness to modelling choices (activation function, priors, and loss formulation) is addressed mainly at the level of justification rather than empirical testing.

Response 1: We appreciate the reviewer's emphasis on empirical testing. In this revision, we have performed a quantitative sensitivity analysis to assess how BINN performance responds to variations in the loss function, activation function, and prior ranges. For example, we adjusted the prior ranges by 10 percent where the flux related parameters remained within their physical bounds. We also tested the sensitivity of replacing the smooth L1 loss with L1 loss and L2 loss and substituting the hyperbolic cosine parameter loss with a squared loss. Furthermore, we evaluated the sensitivity to changes in activation functions by modifying the first three layers from leaky ReLU to ReLU and the final layer from Sigmoid to hard Sigmoid.

The results in Table R1 indicate that BINN is robust to the test of prior ranges as evidenced by the consistent NSE across different settings. It has also been documented in the literature that changes in prior ranges within physical bounds usually do not result in much alteration in estimated parameters (Xu et al. 2006). We noted that substituting our preferred loss terms with L2 or squared difference functions led to some changes in estimated parameters, which approach their extreme limits during training, negatively impacting training stability.

We have added the new sensitivity analysis results in the Supplementary Materials (Table S3).

Table R1: Sensitivity Analysis of BINN Modeling Choices

Modelling Part	Test Choices	Test NSE	Training Stability
Original Settings	Smooth L1 + Leaky ReLU + Sigmoid	0.611 ± 0.036	-

Loss Function	L1 Loss	0.589 ± 0.045	-
	L2 Loss	0.615 ± 0.025	Extreme Para
	Para Squared Diff	0.616 ± 0.042	Extreme Para
Activation	ReLU	0.608 ± 0.038	-
	Hard Sigmoid	0.59 ± 0.047	-
Prior Range	Prior Range Increased by 10%	0.6 ± 0.04	-
	Prior Range Decreased by 10%	0.602 ± 0.039	-

Xu, Tao, Luther White, Dafeng Hui, and Yiqi Luo. 2006. "Probabilistic Inversion of a Terrestrial Ecosystem Model: Analysis of Uncertainty in Parameter Estimation and Model Prediction." *Global Biogeochemical Cycles* 20 (2). https://doi.org/10.1029/2005GB002468.

Comment 2: Similarly, although uncertainty maps and normalised bias visualisations have been added, the biome-level traceability analysis remains unsupported by independent validation and therefore still risks being over-interpreted as mechanistic insight rather than model-derived structure.

Response 2: We appreciate the reviewer's comments on traceability analysis. Our original idea was to show that the CLM5 model that was trained with large-scale observations can offer insights into biome-level variations in SOC based on the knowledge gained about model components. The information of the traceable components of biome-level SOC storage was contained in the observations while BINN retrieved such information from the observations in this study. Conceptually, it was necessary to validate the retrieved traceable components again.

It was our apology to cause the confusion on the traceability analysis. Since it is not the essential part of this study, we have decided to delete the part on traceability analysis from the manuscript to avoid any potential confusion by readers.

Comments remained from previous response letter
Comment 3: While the authors have added uncertainty diagnostics and softened some language, the validation remains restricted to random cross-validation within the same data domain. The manuscript still lacks any independent or out-of-domain validation to support broader claims of ecological insight or general applicability.

Response 3: We thank the reviewer for raising this concern. We have thoroughly revised the manuscript, e.g., deleted the traceability analysis and ensured that all results are characterized as being derived from the model or within the current domain, to avoid any unnecessary broader claims on ecological insight or general applicability.

Comment 4: Although the authors now document and justify their choices of activation function, priors, and loss formulation, the evaluation does not quantify sensitivity to the prior ranges themselves. Robustness to modelling choices is therefore argued conceptually rather than demonstrated empirically.

Response 4: Thank the reviewer for raising those issues again. We did more sensitivity analysis to quantify changes in estimated parameters with different prior ranges as described in our responses to Comment 1 above. It is also well documented in the literature that changes in prior ranges within physical bounds usually do not result in much alteration in estimated parameters (Xu et al. 2006).

Xu, Tao, Luther White, Dafeng Hui, and Yiqi Luo. 2006. "Probabilistic Inversion of a Terrestrial Ecosystem Model: Analysis of Uncertainty in Parameter Estimation and Model Prediction." *Global Biogeochemical Cycles* 20 (2). https://doi.org/10.1029/2005GB002468.

Comment 5: The added memory benchmarks clarify computational trade-offs, but the manuscript still does not assess spatial or biome-level generalisation (e.g., leave-one-biome-out tests). As a result, the scope of applicability remains limited to the training domain.

Response 5: Thanks for pointing out this. This comment is primarily related to the traceability analysis, which was deleted from the manuscript as described in our responses to Comment 2 in general assessment above. We hope the reviewer is satisfied with our revision.

Comment 6: Extending the recovery experiment to all 21 parameters reveals heterogeneous identifiability, but interactions among parameters and their implications for interpretability remain unexplored. Weakly identifiable parameters are still used in downstream analyses without clear justification.

Response 6: We appreciate the reviewer's concern on the heterogeneous identifiability of individual parameters. By the definition of equifinality (or the lack of identifiability), any value within the posterior parameter range is equally valid for describing systems behavior given the information contained in the observations used in the training. Any interactions among these weakly identifiable parameters would be already reflected in their estimated parameters as the observations used in the training contain no or little information on these parameters. No information means that we have no way to understand them and their interactions. To fully resolve this issue, we need more observations that specifically contain information for these parameters. At this stage of scientific research, we have no way to obtain observations that can provide enough information to constrain all the parameters in any complex model, such as CLM5. It is important to remember that a model trained with large-scale observations via BINN, despite many of the weakly identifiable parameters, still can simulate ecosystem dynamics much more realistically than the model alone without any data constraints (Luo et al. 2016).

We used these weakly identifiable parameters in downstream analysis because our study is designed to understand SOC (i.e., system-level) dynamics instead of individual parameter values.

Luo, Yiqi, Anders Ahlström, Steven D. Allison, et al. 2016. "Toward More Realistic Projections of Soil Carbon Dynamics by Earth System Models." *Global Biogeochemical Cycles* 30 (1): 40–56. https://doi.org/10.1002/2015GB005239.

Comment 7: The revised sensitivity analysis is now clearly described and corrected, but it remains a first-order, local analysis. Potential nonlinear interactions among parameters and higher-order sensitivities are not assessed.

Response 7: We thank the reviewer for the insightful comment regarding nonlinear interactions and higher-order sensitivities. While these questions are hypothetical, do we have any practical examples to demonstrate that these nonlinear interactions, higher-order sensitivities are important in controlling SOC dynamics? The complex process-based models, such as CLM5, themselves were developed to account nonlinear interactions among processes. Parameters are just values that were assigned to specify processes at a given condition. BINN advances the model specification via parameter estimation. Any potential nonlinear interactions among parameters with higher-order sensitivity are supposed to be reflected in the estimated parameters.

Moreover, the first-order Sobol index can effectively capture the direct, non-linear contribution of individual parameters to the variance of the model output, it does not explicitly account for higher-order interactions between multiple parameters. Because the primary objective of the sensitivity analysis in this study was to identify a subset of influential parameters for the recovery experiments, we believe that the first-order indices provide a sufficiently robust basis for this selection.

We have revised the discussion in L625-627 to acknowledge the potential for complex parameter interactions:

"It is important to note that while the first-order Sobol index identifies the direct contribution of individual parameters to model variance, it may not fully account for all nonlinear interactions within the system. Therefore, future research involving higher-order sensitivity analysis would be beneficial to further explore complex parameter relationships."

Comment 8: The explanation of systematic bias between BINN and PRODA is plausible, but remains hypothetical. No targeted experiment is provided to demonstrate that the observed bias is indeed caused by prior-centering of weakly identifiable parameters.

Response 8: We appreciate the reviewer's pointing out the systematic bias between BINN and PRODA. In this revision, we used the same set of prior ranges between BINN and PRODA to rerun the experiment. The updated results (Figure 4) show that most of the systematic bias is gone. The estimated carbon transfer efficiency by BINN is still slightly lower than in PRODA, mostly due to higher carbon input allocation to deeper soil layers in BINN. Since shallow layers typically maintain larger carbon stocks than deeper layers, the weighted carbon transfer efficiency in BINN is lower to compensate for the relatively less input in those shallow layers.

We have corrected most of the biases in the previous version. The remaining minor bias in carbon transfer efficiency is mostly due to carbon input allocation. We added this explanation on L660-667 as below:

"In addition, all model components predicted by BINN show high correlations with those predicted by PRODA. However, the estimated carbon transfer efficiency in BINN is slightly lower than in PRODA, primarily due to higher carbon input allocation to deeper soil layers in BINN. Since shallow layers typically maintain larger carbon stocks than deeper layers, the weighted carbon transfer efficiency in BINN is lower to compensate for the reduced input allocation in those shallow layers. This minor bias can also be attributed to the distinct architectural designs between BINN and PRODA and the resulting differences in optimized parameters."

Comment 9: Although normalization and uncertainty maps improve visual interpretability, the ecological meaning of the spatial bias patterns is still not analysed. The manuscript does not explain why specific regions exhibit persistent over- or underestimation.

Response 9: Thanks for the comment on the spatial bias of BINN simulation. We have added the discussions on L668-678 discussion the potential reasons for the spatial bias:

"However, even though BINN shows less underestimation than PRODA, spatial bias still persist within the North-Central and Rocky Mountain domains (Fig. 5a). These biases likely stem from a fundamental mismatch between the spatial resolution of the environmental covariates that derived from global gridded products (Table S1), and the local-scale drivers of each SOC observation. In the North-Central region, the high spatial density of SOC observations reveals sub-grid variability that coarse gridded covariates cannot adequately resolve. Conversely, the overestimation observed in the Rocky Mountains highlights a scale-dependent limitation where gridded inputs fail to represent the complex topographic heterogeneity inherent to mountainous terrain. Consequently, while the BINN framework effectively captures broad environmental gradients, its localized precision remains constrained by the granularity of the available geospatial drivers."

Comment 10: The decomposition of SOC into carbon inputs versus residence time across biomes remains unsupported by independent or analytical validation. Without such validation, the biome-level trade-off analysis risks being over-interpreted as mechanistic insight rather than a model-derived diagnostic.

Response 10: Thanks for the comments. We have deleted the traceability analysis as explained in response 2.

Line-by-Line Comment (focused on comments remained)

Comment 11: Validation scope & claims

Response 11: We thank the reviewer for these constructive suggestions regarding the validation scope and the framing of our claims.

Comment 12: Lines 20–21 (Abstract): Please qualify the claim 'retrieving mechanistic knowledge from big data'. Validation is limited to random cross-validation within the same dataset.

Response 12: We use the word "retrieve" to describe BINN's performance in parameter recovery experiments where it successfully recovers known, prescribed values. However, we acknowledge the reviewer's concern that our current validation is limited to the existing dataset and may not yet generalize to unobserved regions. To reflect this scope, we have replaced "identify" when describing parameters and model components inferred from real world data.

Comment 13: Lines 31–33 (Abstract): Agreement with PRODA does not constitute independent validation. Pleases often or restrict the claim.

Response 13: We have softened our claims on L32-34 to the following:

"The good agreement between the spatial patterns retrieved by BINN and PRODA (average correlation coefficient = 0.86) suggests that BINN's ability of capturing mechanistic knowledge is consistent with the established Bayesian-based methods."

Comment 14: Lines 34–36 (Abstract): The phrase 'transformative tool' should be tempered or restricted to the current dataset.

Response 14: On L37, we have modified "transformative tool" to "efficient framework"

Comment 15: Lines 107–115 (Introduction): Clarify that inferred processes are model-derived and not independently observed.

Response 15: On L116-L119, we have included the following words to clarify the inferred processes are model-derived:

"BINN is an end-to-end framework that combines data-driven machine learning with process-based modeling to facilitate the interpretability of model-represented biogeochemical dynamics, such as SOC dynamics in this study."

Comment 16: Lines 608–609 (Discussion): Acknowledge that process importance is conditional on the model structure and training domain.

Response 16: We have included the following in L646-649:

"This capability enables BINN to evaluate the relative importance of different processes controlling SOC storage within the model and data domain, as done with PRODA, while offering computational advantages through its integrated neural network architecture.".

Generalisation

Comment 17: Lines 386–388 (Methods): Random k-fold CV does not test biome-level transferability. Please clarify.

Response 17: We have changed the word "internal predictive stability" instead of "transferability" for clarification in L420

Comment 18: Lines 544–545 (Results): Clarify that comparable accuracy does not imply out-of-domain generalization.

Response 18: We added the word "in-domain predictive" to be clear in L576

Comment 19: Lines 652–655 (Discussion): Restrict claims about uncovering mechanisms to the training domain.

• Suggestion: Either add leave-one-biome-out testing or explicitly limit scope to in-domain performance.

Response 19: We have now included the sentence "uncover mechanisms from big data when relevant observational data are presented" for clarification in L729-730.

Comment 20: Parameter identifiability & interpretation

Response 20: We thank the reviewer for raising these important questions regarding parameter identifiability and interpretation.

Comment 21: Lines 493–496 (Results): Clarify how heterogeneous identifiability affects interpretation of weak parameters.

Response 21: We've clarified on L528-530 that "Parameters with low sensitivity (e.g., cryo, taucwd) showed weak correlations between retrieved and prescribed values and were typically distributed around the midpoints of their prior ranges". However, the reason these parameters are not identified well and why we still include these parameters in the downstream analysis are justified in Response #6.

Comment 22: Lines 589–592 (Discussion): Explain whether weakly identifiable parameters are interpreted or excluded.

Response 22: We've added explanations on how we incorporate weakly identifiable parameters in the downstream analysis on L620-632:

"Although parameters with low influence on SOC showed weak recovery by BINN, this is a common limitation also observed in Bayesian approaches when data are weakly informative. To overcome this limitation, future applications could integrate complementary data streams beyond SOC, which might provide additional constraints necessary to inform those less-sensitive parameters. Alternatively, the inherent uncertainty of high-dimensional parameter spaces can be mitigated by aggregating individual parameters into functionally related model components. As demonstrated in this study, grouping parameters with similar process-based meanings and weighing them by their relative contributions to total SOC help characterize the underlying biogeochemical processes more powerfully (Fig. 4). This approach effectively prioritizes the most influential parameters; as weakly recovered parameters generally exert minimal impact on total SOC storage. Such parameters are assigned lower weights within the calculated model components, thereby mitigating the impact of poorly constrained parameters on our mechanistic interpretations."

Comment 23: (Discussion): Restrict mechanistic interpretation to parameters with demonstrated sensitivity and recovery skill.

Response 23: As noted in our previous responses, our mechanistic interpretations are focused on bulk model components that aggregate multiple individual parameters. For instance, the 'Baseline Decomposition' component illustrated in Figure 4 is comprised of all substrate decomposition-related parameters in CLM5. By interpreting these aggregated components, we ensure that the sensitivity of individual parameters and the potentially weak recovery of non-sensitive terms do not compromise the validity of our downstream interpretations.

Comment 24: Spatial bias interpretation

Response 24: We thank the reviewer for providing the comments regarding the spatial interpretation of BINN-predicted parameters

Comment 25: Lines 541–543 (Results): Provide interpretation or explicitly state that spatial bias is diagnostic only.

Response 25: On L572-573, we have included "The map of the normalized difference (Fig. 5a) identifies geographic regions where BINN's predictive performance is limited " to clarify the spatial bias is diagnostic only.

Comment 26: Lines 546–548 (Results): Clarify what data or model limitations may explain spatial bias.

Response 26: We believe that it is the architectural distinction causing the spatial bias between PRODA and BINN, specifically in L651-659: " PRODA demonstrates a widespread tendency to underestimate SOC stocks, whereas BINN's normalized differences are centered nearer to zero, indicating reduced systematic bias (Fig. 6b). This difference likely stems from the architectural distinction between the two frameworks: while the neural network in PRODA is trained to replicate site-level parameter distributions derived from a separate site-level data assimilation step, BINN's end-to-end framework directly optimizes the neural network to minimize the deficit in SOC simulations across the US continent. By coupling the parameter inference and the physical simulation within a single objective function, BINN can more effectively mitigate spatial biases that might otherwise arise from the multi-step training process used in PRODA."

Comment 27: Lines 552–559 (Results, Figure 7 description): The text interprets biome-level differences in SOC residence time and carbon inputs as reflecting distinct underlying mechanisms. Given that these patterns are derived entirely from model-based traceability and not independently validated, please clarify that Figure 7 provides a diagnostic decomposition of model behavior rather than evidence of true ecological mechanisms.

• Figure 7 caption (Lines ~739–741): Please explicitly state that the biome-level trade-offs shown in Figure 7 are conditional on the BINN–CLM5 structure and should not be interpreted as empirical validation of process differences among biomes.

Response 27: Since this comment is primarily related to the traceability analysis, which was deleted from the manuscript as described in our responses to Comment 2 in general assessment above. We hope the reviewer is satisfied with our revision.

Comment 28: Lines 630–646 (Discussion): Avoid implying ecological causality from bias patterns without independent support.

Response 28: We have added a paragraph to discuss that further independent support is needed for mechanistic understanding on L718-725: " However, it is important to recognize that the mechanism understanding inferred through this framework is inherently tied to the CLM5 model structure and the BINN-optimized parameters. While these optimized parameters provide a mechanically consistent representation of different processes within the process-based model, the resulting mechanistic discovery may still require further independent validations, such as observational data from satellites, sensors, and field and laboratory experiments. Nonetheless, BINN-retrieved mechanism understanding is much better grounded in observations than these gained by the process-based models or data alone."