Reviewer 2

Comment 1: The revised manuscript shows clear and substantial improvement over the previous version, with several technical additions including the extension of the parameter recovery experiment, the incorporation of Monte Carlo dropout for uncertainty quantification, and improved clarity in the methodological description. The authors have also moderated some of the stronger claims and better positioned the work within the existing literature.

Response 1: We sincerely thank the reviewer for their positive evaluation and acknowledgement of our revision.

Comment 2: However, some key concerns remain only partially addressed. In particular, robustness to modelling assumptions (e.g., prior ranges and loss formulations) is discussed more clearly, but still relies primarily on conceptual justification rather than comprehensive empirical evaluation. While the sensitivity tests of prior ranges provide useful insight into interpretability, they do not fully assess the robustness of inferred parameters and resulting conclusions.

Response 2: We thank the reviewer for the constructive comment on the prior ranges for biogeochemical parameters. We agree that the selection of these prior ranges relies primarily on conceptual/quantitative justification rather than comprehensive empirical evaluation. Parameters in a model often cannot be directly measured by empirical observations or from manipulative experiments, and thus, have rarely been empirically evaluated. Accordingly, prior ranges of model parameters utilized in data assimilation are frequently selected to be in the vicinity of their values used in simulations. Indeed, these model parameter values usually have been quantitatively turned/calibrated to make simulations work successfully. Thus, the prior ranges of model parameters have been quantitatively but not empirically evaluated. As this study was built upon previous data assimilation studies, we fully adopted the prior ranges utilized in these previous studies to ensure they are as realistic as possible (please see the references in Table S2). In the revised manuscript, we highlight the need for comprehensive empirical syntheses of processes related to biogeochemical parameters used in process-based models for future data assimilation study (L724-726) as:

"In the future, more robust data assimilation will also rely on comprehensive empirical syntheses of processes related to the biogeochemical parameters used in process-based models."

Meanwhile, we are keen on learning new ideas from the reviewer on comprehensive empirical evaluation of model parameters.

Comment 3: In addition, the validation framework remains limited to in-domain cross-validation without independent or out-of-domain evaluation.

Furthermore, some interpretations—especially regarding mechanistic insights and biome-level patterns—still extend beyond what is strictly supported by the presented validation. While these optimized parameters provide a mechanically consistent representation of processes within the process-based model, the resulting mechanistic interpretations should be more explicitly framed as model-dependent and would benefit from clearer acknowledgement of the need for independent validation using complementary observational constraints (e.g., satellite observations, in situ sensor networks, and field or laboratory experiments). Incorporating relevant references in this context would strengthen the discussion (some are suggested attached).

Response 3: We appreciate the reviewer emphasizing the importance of out-of-domain evaluation. To address this limitation, we have revised the manuscript to state that our interpretations are strictly confined to the specific model and data domain. For example, we use terms such as "in domain predictive" to prevent overgeneralization.

We also thank the reviewer for highlighting the necessity of using complementary observational constraints to validate mechanistic discoveries. We fully agree that our findings are model-dependent. Accordingly, we incorporated sentences in the revised manuscript to acknowledge this limitation and the need for independent validation (L719-L728):

"However, it is important to recognize that the mechanistic understanding inferred through this framework is inherently tied to the CLM5 model structure and the BINN-optimized parameters. While these optimized parameters provide a mechanistically consistent representation of different processes within the process-based model, the resulting mechanistic understanding may still require independent validations from observation via satellites, sensors, and field and laboratory experiments. In the future, more robust data assimilation will also rely on comprehensive empirical syntheses of processes related to the biogeochemical parameters used in process-based models. Nonetheless, BINN-retrieved patterns represent a mechanistic understanding grounded in observation-model synthesis rather than that gained from process-based models alone."

Comment 4: Overall, the manuscript is significantly improved and technically sound, but would benefit from further strengthening of robustness analysis, clearer limitation of claims, and modest expansion of the discussion on validation. I leave it up to co-authors to polish the work further considering these aspects. I therefore recommend minor revision.

Response 4: We thank the reviewer for providing valuable suggestions that have significantly enhanced the scientific rigor of our manuscript. In this revision, we strengthened the robust analyses, systematically moderate our claims, and explicitly discussed the limitations of our current research. We hope that our revision has greatly improved the manuscript and fully addressed the reviewer's concerns.

Reviewer 4

Comment 1: It isn't explained why 10-fold cross validation was used.
Response 1: Thanks for the question. We employed 10-fold cross-validation to systematically generate independent test sets from the same dataset, thereby providing a more robust and unbiased assessment of model performance. We have explicitly described the cross-validation method in Section 2.2.3 of the manuscript as:

"To conduct 10-fold cross-validation on the simulations, the entire dataset was randomly divided into ten equal-sized subsets. In each iteration, 8 subsets were used for training, 1 subset for validation, while the remaining subset served as the test set. 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. This process was repeated ten times, with each subset serving as the test set once. The performance metrics, including Nash–Sutcliffe modelling efficiency coefficient (NSE) and Pearson correlation coefficient (r), were calculated for each iteration. Final performance evaluations were conducted by averaging the metrics across all iterations, and grid-level predictions were averaged across the ten iterations. This cross-validation approach provides a robust assessment of BINN's internal predictive stability by testing its performance on multiple independent datasets, reducing the impact of data partitioning bias and thus enabling evaluation of model stability across different training-testing combinations."

Furthermore, we have corrected inconsistent terminology from the previous version to use consistent terms such as "cross-validation" throughout the manuscript.

Comment 2: Lines 594: "Our study demonstrates that PRODA is a novel approach...", wording should be changed to something like "PRODA is a viable approach" or "efficient"...etc Novelty itself should come with much more comparison to other similar work.
Response 2: We thank the reviewer for this suggestion. We assume the reviewer intended to refer to "BINN" rather than "PRODA" in the quoted sentence. We have revised the sentence to describe BINN as an efficient approach rather than a novel one.

Comment 3: Figure 5: What is "SOC mean NSE" across folds?

Response 3: NSE (Nash-Sutcliffe Efficiency) is our main performance metric (defined in Section 2.3.3), and SOC means soil organic carbon. Using cross-validation, we trained the model 10 times with different test folds (splits). For each fold we computed the NSE using all soil organic carbon observations, then took the mean across folds. We have updated Figure 5d to display the mean test NSE values directly above the box within the box plot (Fig. R1).

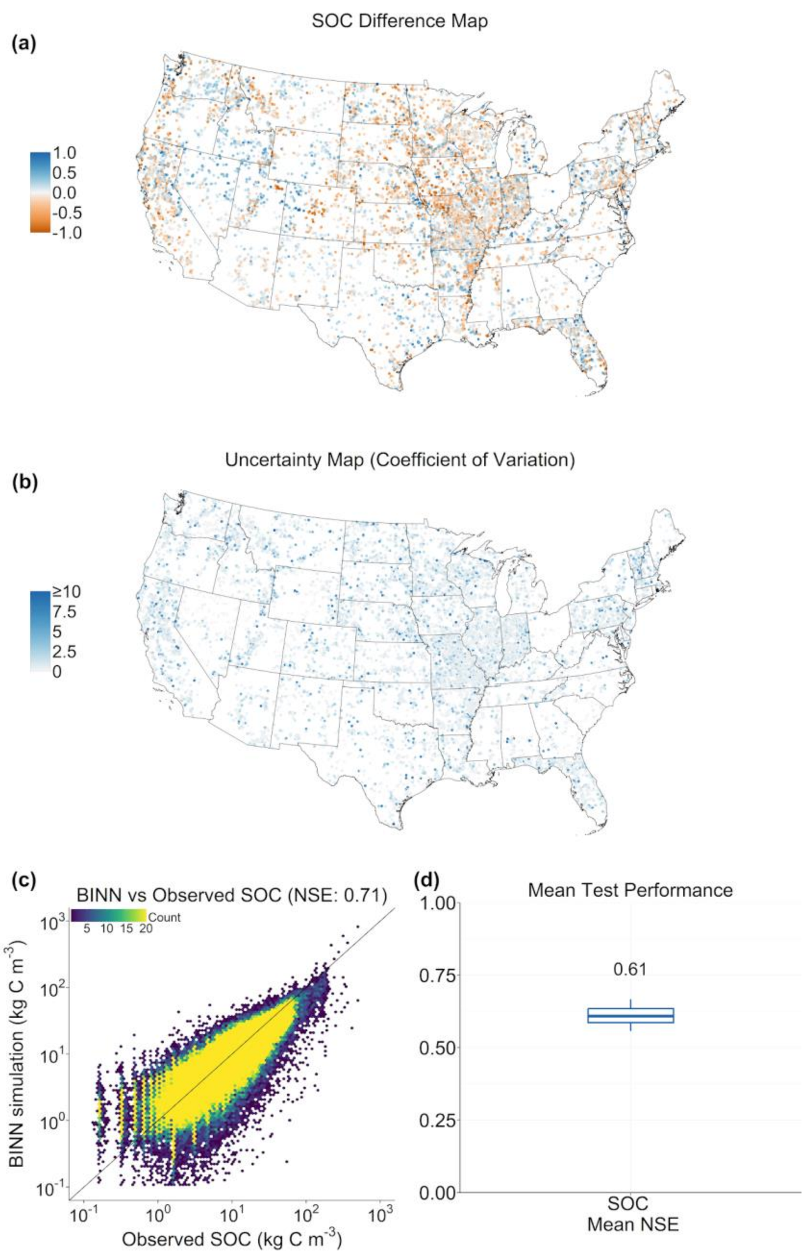

Figure R1: Comparison of simulated and observed SOC storage using BINN across 10-fold cross-validation training. (a) Map of normalized difference across all 10 folds. (b) Coefficient of variation of difference across 10 folds, indicating spatial uncertainty. (c) Observed vs. fold-average simulated SOC across all depths and sites. (d) Box plot of test NSE across 10 folds.

Comment 4: The total number of parameters of the neural network should be given, at least approximately. As well as the detailed architecture for reproducibility in supplementary material.

Response 4: Thank the reviewer for the suggestions. We've detailed the neural network architecture previously in Section 2.1.1. To be clearer as the reviewer suggested, we have now included the exact number of trainable parameters in this section on L211: "In total, our BINN architecture contains 74,793 trainable parameters." Furthermore, to ensure complete reproducibility, we updated the code to strictly align with the hyperparameters described in the manuscript.

Comment 5: In figure S5: What is "Extremely Para"?

Response 5: Thank the reviewer for asking. "Extreme Para" denotes instances where extreme parameter values, specifically those less than 0.00001 or greater than 0.99999 prior to the sigmoid transformation, are predicted by the model during training, which results in instability. We have added an explanatory note below the table title of Table S3 to clarify it.

Comment 6: The use of Latitude and Longitude as inputs through the positional encoder is highly questionable, if the NN is large enough (which this one appears to be), it can simply memorize its inputs, similar to assigning a unique ID to every (latitude, longitude) combination.

Response 6: We appreciate the reviewer raising this concern regarding the positional encoder. To address this, we conducted an ablation study where we tried removing the positional encoder from the network architecture. The modified BINN achieved similar predictive accuracy (Fig. R2), which demonstrates that the model is learning generalizable relationships from the environmental covariates rather than simply memorizing specific latitude and longitude coordinates. While this indicates that the positional encoder is not strictly necessary for this specific large dataset, we describe that the positional encoder is an architectural option within the framework, as it can learn valuable information about each datapoint's geospatial context beyond what is given in the environmental covariates (Klemmer et al. 2023).

Klemmer, K., Safir, N. S., & Neill, D. B. (2023, April). Positional encoder graph neural networks for geographic data. In *International conference on artificial intelligence and statistics* (pp. 1379-1389). PMLR.

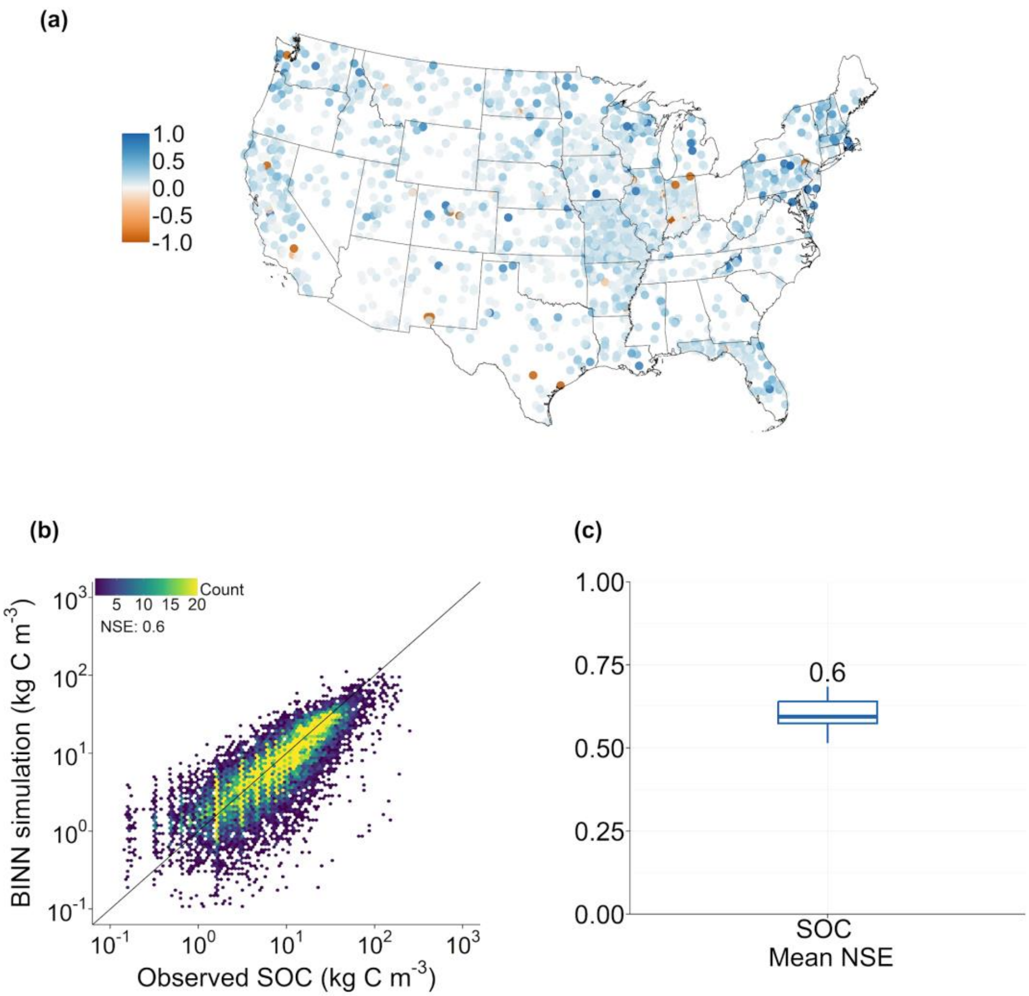

Figure R2. Comparison of observed and simulated SOC storage using BINN without positional encoder. (a) Spatial distribution of normalized differences for one representative fold (median NSE). (b) The scatter plot presents the NSE between observed SOC and simulated SOC derived from the testing dataset at varied soil depths. (c) The box plot shows the mean performance of testing NSE in the 10-fold cross validation test.

Comment 7: The synthetic dataset does not really represent a valid experiment; It depends heavily on how the function that maps covariates to parameters is, the neural network in this case is only trying to map that function. If the function is simple then the neural network will of course have no issues obtaining it. Thus the conclusions obtained from the synthetic data experiment are really non-decisive.

Response 7: We appreciate the reviewer's concern that an overly simplistic mapping function behind the synthetic dataset could yield inconclusive results. However, the synthetic parameters utilized in our experiment were not generated by a simplistic or predefined mathematical function. Instead, these parameter values were derived directly from Bayesian data assimilation conducted at individual sites in a previous study (Tao et al. 2020). Therefore, the dataset realistically simulates the true complexity of the relationship between environmental covariates and biogeochemical parameters rather than relying on artificial simplifications. Thus, we believe that the conclusions drawn from our synthetic data experiment are both rigorous and valid.

Tao, Feng, Zhenghu Zhou, Yuanyuan Huang, et al. 2020. "Deep Learning Optimizes Data-Driven Representation of Soil Organic Carbon in Earth System Model Over the Conterminous United States." *Frontiers in Big Data* 3 (June): 17.

Comment 8: The scripts in the provided code all have n_epochs=50 (or 100), meanwhile the paper mentions that the BINN was trained for 300 epochs (Line 200), impacting the reproducibility of the work. Please upload the correct scripts used, as well as an explicit list of obtained hyperparameters from hyperparameter optimization (and preferably the hyperparameter optimization script).

Response 8: We appreciate the reviewer for pointing out the discrepancy between the hyperparameters in the provided scripts and those described in Section 2.1.1. We have updated the code repository to strictly align with the manuscript, ensuring reproducibility of our results.