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 will revise the introduction and the method sections to add more details about PRODA's structure and performance. We will also include more comparisons on the performance between PRODA and BINN.

Following the reviewer's suggestion, in the revision, we will enrich Section 4.2 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 will add a citation of Tao et al. (2024) where appropriate.

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.

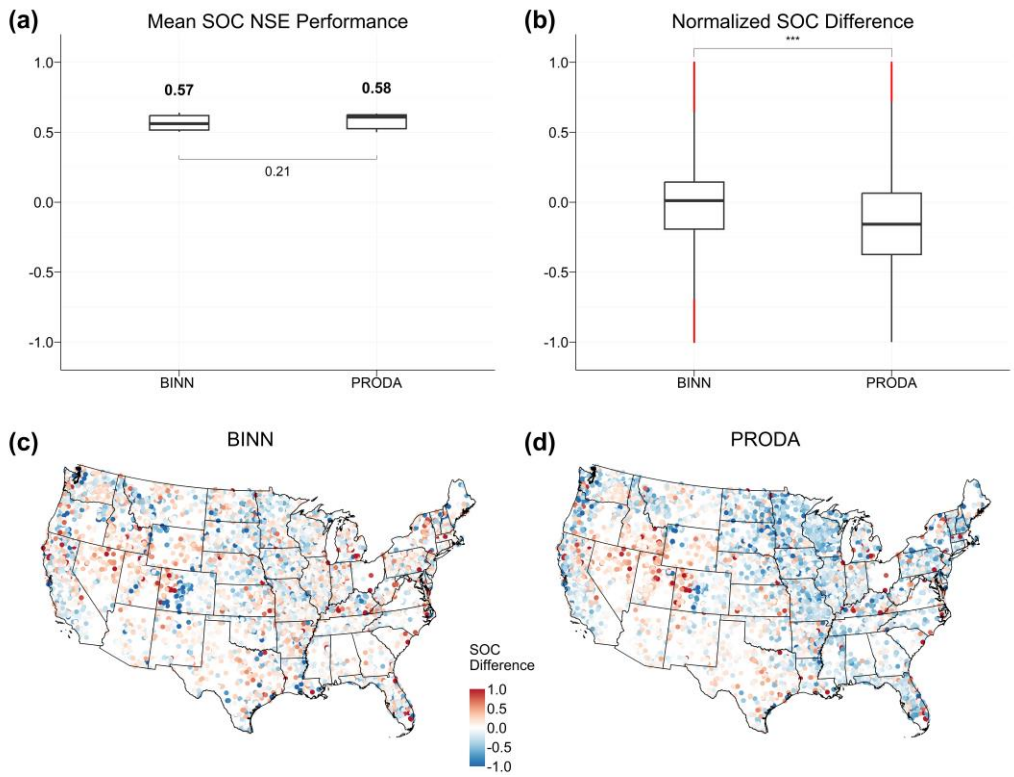

Figure R1. Side-by-side evaluation of BINN and PRODA on identical test sets across a 10-fold cross-validation. (a) Mean test Nash–Sutcliffe modelling efficiency coefficient (NSE) per fold. The *P* value above the bars donates the difference in test NSE between the two methods is insignificant. (b) Mean 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 the revised manuscript, we will clearly discuss this critical limitation of PRODA in the methods and discussion.

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 parameter 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 will extensively discuss the above results 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.

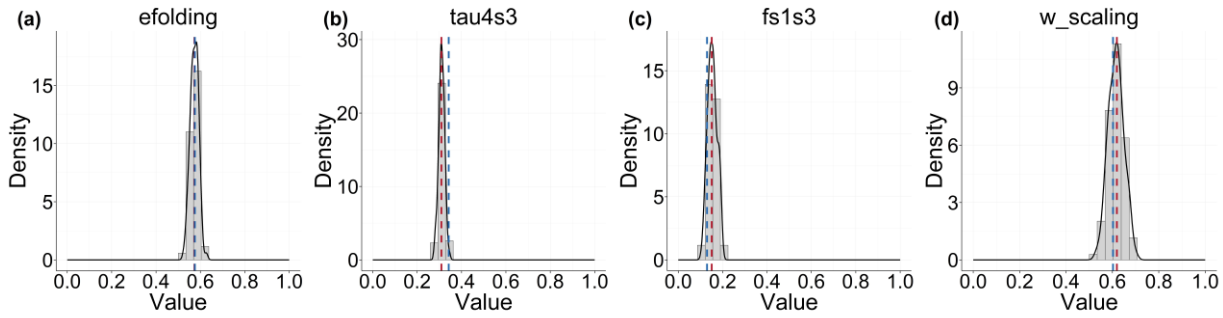

Figure R2. Uncertainty in posterior distributions by BINN parameters in relation to Monte Carlo dropout in the parameter recovery experiment at one site. For each parameter, the marginal posterior is shown, with the vertical blue and red dashed lines indicating the prescribed ("true") and BINN's point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

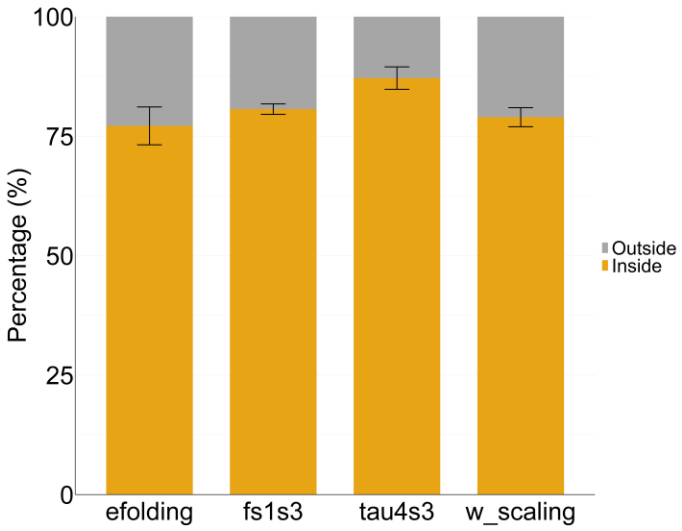

Figure R3: Coverage of parameters' posterior ranges of BINN. The coverage was quantified by the percentage of prescribed ("true") parameter values that fall into the Monte Carlo–dropout posterior range across test sites. Error bars show the mean ± SD across the 10 cross-validation folds of the synthetic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

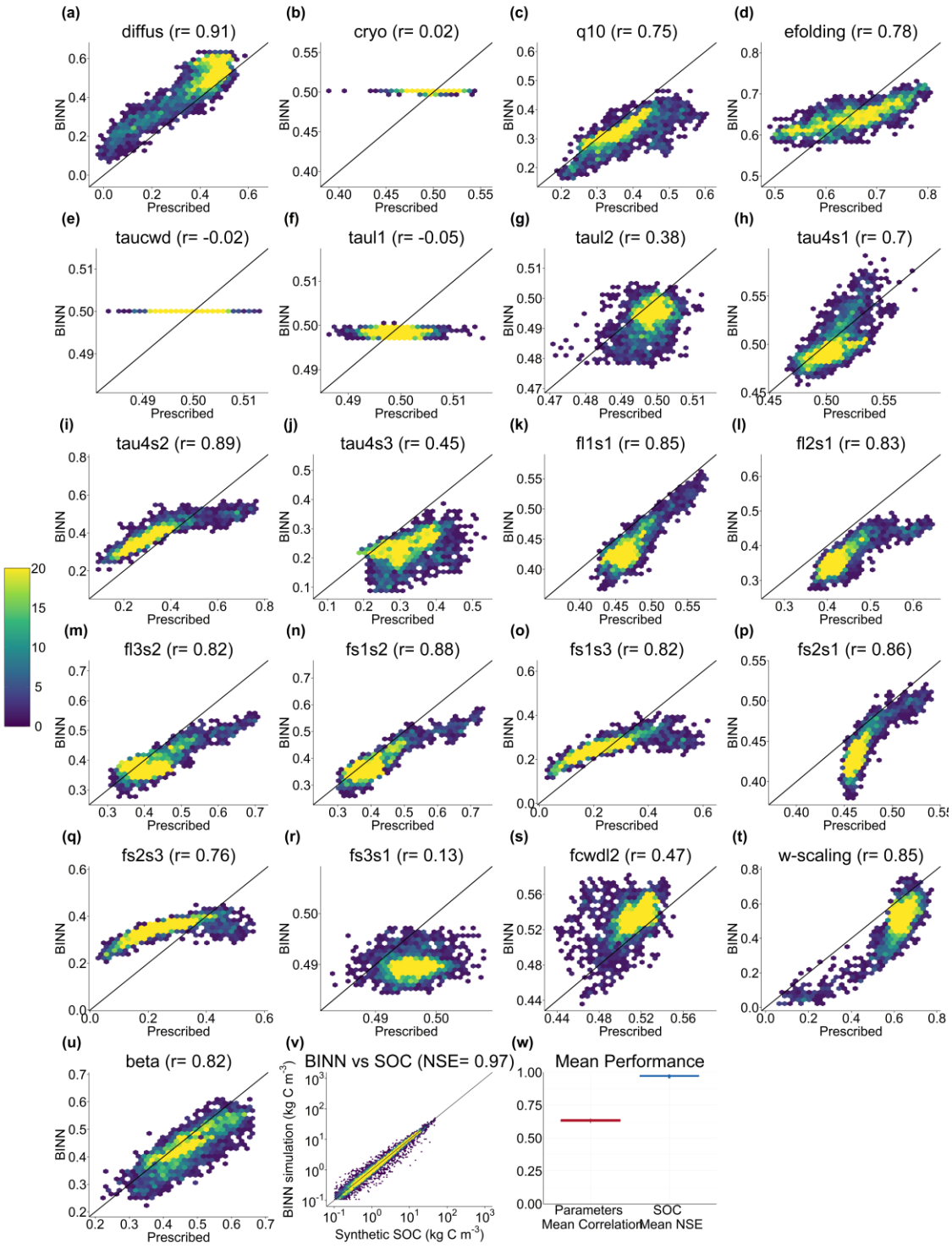

Figure R4: BINN parameter recovery for all CLM5 parameters. (a–u) Scatter plots of BINN-predicted versus prescribed values for each parameter. (v) Density scatter of simulated SOC versus synthetic SOC. (w) Summary across a 10-fold cross-validation: mean correlation for all parameters (predicted vs. prescribed),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

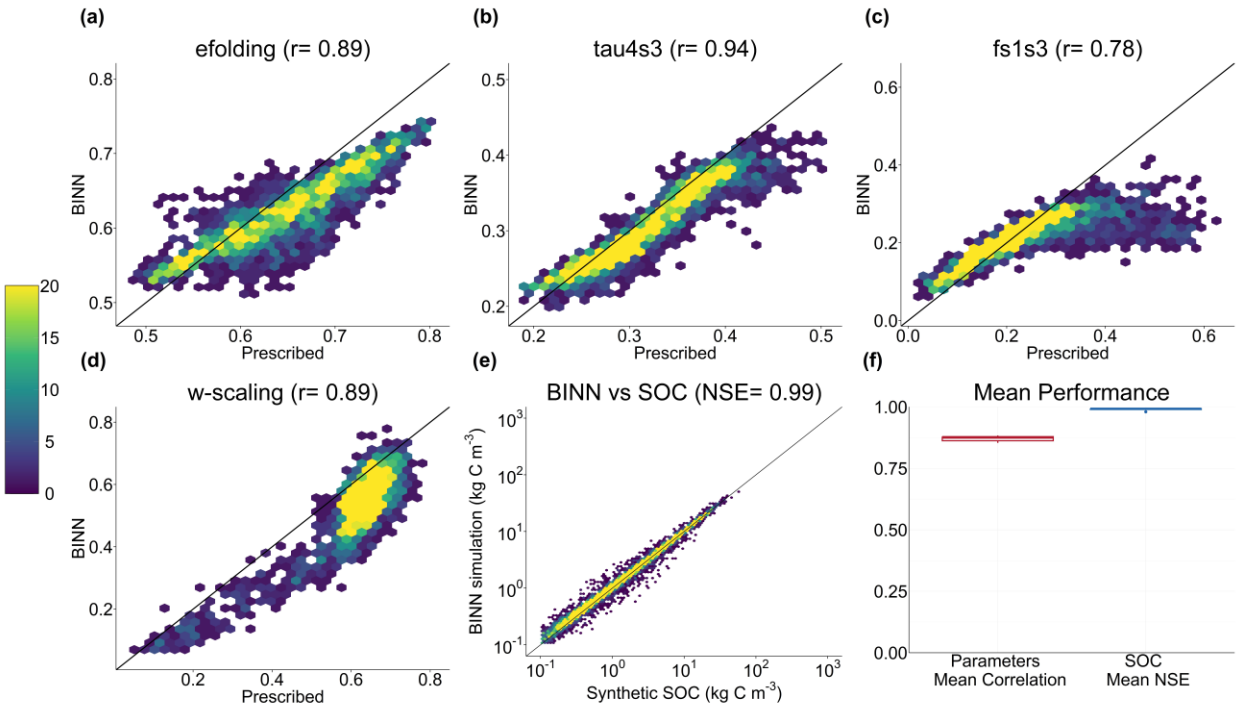

Figure R5: BINN's performance in retrieving the four most sensitive parameters and simulating SOC. Scatter plots with colors representing the density of points comparing the parameter values (a) "efolding", (b) "tau4s3", (c) "fs1s3", and (d) "w-scaling" predicted by BINN against the prescribed parameter 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 (Eq 2) and the hyperbolic cosine loss (Eq 5). We will revise the abstract, introduction, and other related sections to clarify these distinctions. 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 will follow these suggestions to adjust the manuscript structure in the revision. We will revise section 2 to be more general, including an overview of the PRODA approach. We will reorganize 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 will follow these suggestions to revise our manuscript as follows.

(1) We will 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.

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 will describe the differences of various hybrid approaches that integrate physical knowledge by machine learning in previous studies and compare them with BINN.

(3) We will explain why and how BINN facilitates mechanistic understanding more efficiently than traditional Bayesian calibration in the introduction. We will describe 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 will also acknowledge established statistical approaches that estimate spatially varying coefficients (e.g., SVCMs) to provide a more comprehensive context when introducing BINN.

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 will clarify 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 will revise 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 will correct 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 will fix. 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 will change the heading of section 3 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 3 in the original manuscript will be updated to Fig. R6 with the first-order sensitivity index in the title to avoid confusion with feature importance.

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).

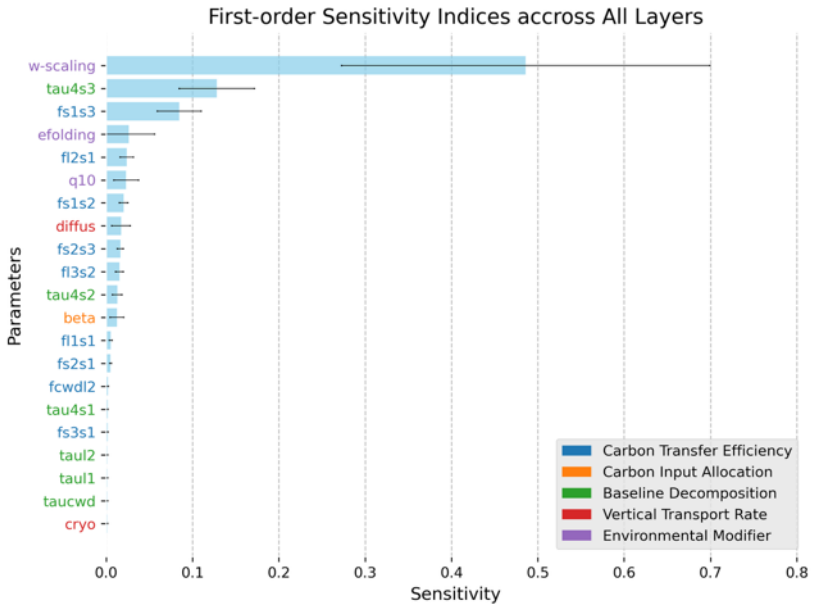

Figure R6. First-order sensitivity of CLM5 parameters across all soil depths. Bars show the first-order sensitivity index for each biogeochemical parameter over all soil layers. Parameters are ranked (y-axis) by decreasing sensitivity and color-coded by their associated process component (e.g., K, A, 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 will revise section 7 to start with a brief recap of BINN and the Bayesian-based approach (PRODA) to orient readers. We will also revise 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 will also direct our arguments to specific results in the revision to be clearer (Figs. 4 and 6)

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 will streamline this paragraph and make the logic more straightforward: 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 will revise section 7.3 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.

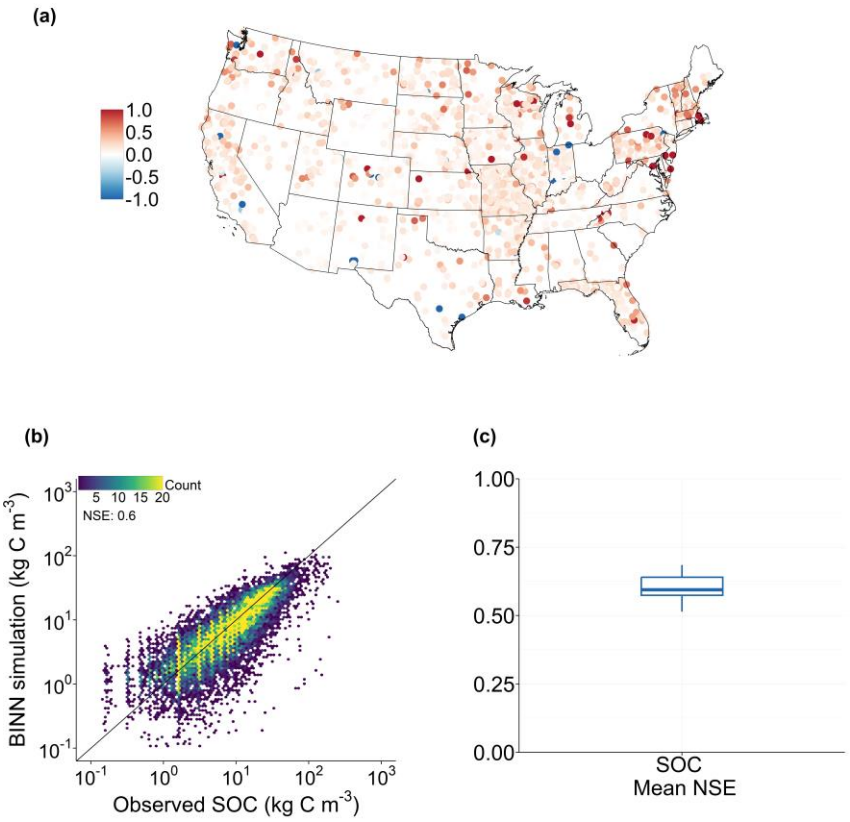

Figure R7. Comparison of observed and simulated SOC storage using BINN without positional encoder. (a) Spatial distribution of normalized differences for one representative fold (median NSE). (b) The scatter plot presents the NSE between observed SOC and simulated SOC derived from the testing dataset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