Biogeochemistry-Informed Neural Network (BINN) for Improving Accuracy of Model Prediction and Scientific Understanding of Soil Organic Carbon

Haodi Xu[1, †], Joshua Fan[2, †], Feng Tao[3, †], Lifen Jiang[1], Fengqi You[4], Benjamin Houlton[3], Ying Sun[1], Carla Gomes[2], and Yiqi Luo[1*]

[1]Soil and Crop Sciences Section, School of Integrative Plant Science, Cornell University, Ithaca, New York, USA

[2]Department of Computer Science, Cornell University, Ithaca, New York, USA

[3]Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York, USA

[4]Department of Systems Engineering, Cornell University, Ithaca, New York, USA

[†]These authors contributed equally to this work.

[*]Correspondence to: Yiqi Luo (yiqi.luo@cornell.edu), Feng Tao (feng.tao@cornell.edu), Carla Gomes (gomes@cs.cornell.edu)

Abstract

Big data and the rapid development of artificial intelligence (AI) provide unprecedented opportunities to enhance our understanding of the global carbon cycle and other biogeochemical processes. However, retrieving mechanistic knowledge from big data remains a challenge. Here, we develop a Biogeochemistry-Informed Neural Network (BINN) that seamlessly integrates a vectorized process-based soil carbon cycle model (i.e., Community Land Model version 5, CLM5) into a neural network (NN) structure to examine mechanisms governing soil organic carbon (SOC) storage from big data. BINN demonstrates high accuracy in retrieving biogeochemical parameter values from synthetic data in a parameter recovery experiment. We use BINN to predict six major processes regulating the soil carbon cycle (or components in process-based models) from 25,925 observed SOC profiles across the conterminous US and compared them with the same processes previously retrieved by a Bayesian inference-based PROcess-guided deep learning and DAta-driven modeling (PRODA) approach [1,2]. The high agreement between the spatial patterns of the retrieved processes using the two approaches with an average correlation coefficient of 0.81 confirms BINN's ability in retrieving mechanistic knowledge from big data. Additionally, the integration of neural networks and process-based models in BINN improves computational efficiency by more than 50 times over PRODA. We conclude that BINN is a transformative tool that harnesses the power of both AI and process-based modeling, facilitating new scientific discoveries while improving interpretability and accuracy of Earth system models.

1 Introduction

Artificial intelligence (AI) has revolutionized our ability to leverage big data to uncover relationships in complex systems such as the Earth system [3]. Methods such as machine learning and deep learning have shown power in discovering key patterns from data in biogeochemistry, such as representing soil organic carbon concentrations [4,5], predicting aboveground carbon accumulation rates in naturally regenerating forests [6], and estimating soil respiration [7]. However, most AI-based approaches primarily learn black-box statistical correlations from the data rather than causality, making it challenging to translate the learned relationships and patterns into mechanisms and controls on processes. This lack of mechanistic insight is an inherent weakness of AI-based models [8].

To address this challenge, various hybrid approaches have emerged, aiming to integrate scientific knowledge and reasoning with standard machine learning methods. This powerful combination leverages the strengths of data-driven techniques along with scientific theories and reasoning to enhance the mechanistic understanding of the Earth system through big data [9]. By introducing physical or knowledge-based constraints and reasoning, such as mass or energy conservation [10–12], thermodynamic rules, interpretable latent spaces, and entropy-based reasoning constraints [13,14], Bragg's law for X-ray diffraction [15], empirical functional relationships [16,17], or partial differential equations [18], into standard ML approaches such as a neural networks, AI-based predictions can be further constrained by scientific knowledge in addition to being driven by observational data. More importantly, unlike the uninterpretable weights and biases in a conventional neural network, the latent physical parameters embedded in the knowledge-based constraints explicitly represent physical and biological processes, providing mechanistic interpretability to the neural network's predictions. However, previous efforts mostly used

limited constraints in a system (e.g., a few empirical relationships in photosynthesis [19]) or simplified models that comprise latent variables and conservation principles [12]. It remains challenging to integrate dozens of partial or ordinary differential equations with numerous free parameters into a neural network to study the dynamics of a complex system, such as soil organic carbon (SOC) dynamics.

A recently-developed approach, PROcess-guided deep learning and DAta driven modeling (PRODA), leverages a deep learning model to learn the site-by-site optimized biogeochemical parameters by Bayesian inference-based data assimilation, to improve our understanding of the global soil carbon cycle [1,2,20,21]. This approach successfully harnesses the strengths of both process-based modeling and deep learning methods to improve SOC simulations with spatially varying biogeochemical parameters. However, the Bayesian inference-based, site-level data assimilation embedded in the PRODA approach requires vast computational resources, making the method time- and energy- inefficient and difficult to apply broadly.

In this study, we integrate a matrix form of a process-based model that describes SOC dynamics into a neural network, thus developing a Biogeochemistry-Informed Neural Network (BINN). BINN is a novel framework that combines data-driven machine learning with process-based modeling to enable interpretability of biogeochemical dynamics, such as SOC dynamics in this study (Figure 1). Herein, we first introduce the structure of BINN (Figure 1a), followed by a demonstration of BINN's ability to recover biogeochemical parameters with high accuracy through a parameter recovery experiment (Figure 2) and estimation of model components from real-world SOC observations (Figure 1b). Our predicted biogeochemical parameters accurately simulate real SOC observations, and are similar to those produced by PRODA, while being much

faster to estimate computationally. By combining data-driven learning with reasoning about existing geoscientific knowledge, BINN can accurately and quickly infer underlying physical processes from only SOC observations.

2 Biogeochemistry-Informed Neural Network (BINN)

BINN incorporates a process-based model into a neural network to infer SOC concentrations and underlying processes from observational data (Figure 1a). First, a neural network learns the relationships between environmental covariates and biogeochemical parameters, which quantify the strength of important processes in the soil carbon cycle. We pass these parameters to a process-based model to simulate SOC storage, and compare with field observations.

2.1 Neural Network

We employ a fully-connected neural network to learn the relationship between environmental covariates (Table S1) and biogeochemical parameters over space. The neural network uses embedding layers to encode categorical covariates, and a spatial positional encoder to compute a vector embedding for each location. We combine these embeddings with the numeric covariates into a vector $\boldsymbol{e}$, and pass this through a 4-layer neural network f_{NN} (Equation 1) with learnable weights/biases $\boldsymbol{w}$, which outputs a vector $\boldsymbol{z}$ with 21 values (one for each parameter in the process-based model):

$$\boldsymbol{z} = f_{NN}(\boldsymbol{e}; \boldsymbol{w}) \quad (1)$$

Because we have prior knowledge about the plausible range of values for each biogeochemical parameter, $\boldsymbol{z}$ is further processed by element-wise sigmoid functions σ (Equation 2) into predicted parameters $\boldsymbol{p}$. This ensures each parameter $\boldsymbol{p_i}$ stays within these prior ranges:

$$\boldsymbol{p_i} = \sigma\left(y_i, \gamma, \theta_{i,max}, \theta_{i,min}\right) = \frac{1}{1+\exp\left(-\frac{z_i}{\gamma}\right)} * \left(\theta_{i,max} - \theta_{i,min}\right) + \theta_{i,min} \quad (2)$$

where z_i is the i-th output of the neural network, $\theta_{i,max}$ and $\theta_{i,min}$ are plausible limits for each biogeochemical parameter i, taken from previous literature [2], and γ is a learnable parameter that controls how fast the predictions converge to $\theta_{i,min}$ or $\theta_{i,max}$. Thus, the final output of the neural network is $\boldsymbol{p}$, a vector of 21 biogeochemical parameters for each location, where each parameter $\boldsymbol{p_i}$ is constrained to be in its prior range $(\theta_{i,min}, \theta_{i,max})$. We used a grid search to select hyperparameters; additional details about hyperparameters and model architecture can be found in Appendix 1.

2.2 Process-based Model

In this study, we used the soil carbon module of the Community Land Model version 5 (hereafter referred to as CLM5) to represent our knowledge of SOC dynamics (Figure S1). The CLM5 model has been continuously developed and refined over the past decade for simulating SOC dynamics [22]; it mathematically represents our knowledge of SOC dynamics with 140 partial differential equations. We chose CLM5 to enable direct comparison with PRODA, which uses the same model.

CLM5 simulates SOC dynamics across 20 soil layers to 8 m. Each layer contains 7 carbon pools, including one coarse woody debris pool, three litter pools corresponding to metabolic, cellulose, and lignin materials, and three SOC pools classified by different turnover

times into fast, slow, and passive pools. This structure results in a total of 140 carbon pools (7 pools × 20 layers).

A key innovation of our approach is that we incorporate into our neural network framework a differentiable CLM5 model, whose structure can be represented in a matrix form [23,24] as:

$$\frac{dX(t)}{dt} = B(t)I(t) - A\xi(t)KX(t) - V(t)X(t) \quad (3)$$

where $I(t)$ is the total carbon input from vegetation at time *t*, $B(t)$ (140×1) is the allocation of carbon input to different pools: A (140×140) is the carbon transfer matrix, quantifying horizontal carbon movement between pools in the same layer; K (140×140) is the intrinsic decomposition rate of each carbon pool, which is the same for each pool across 20 layers; ξ(t) (140×140) captures how the environment modifies the intrinsic decomposition rate in the K matrix by temperature (ξ_T), water (ξ_W), oxygen (ξ_O), and depth (ξ_D) scalars, V(t) (140×140) defines how SOC enters and leaves each layer; and $X(t)$ is carbon pool size. The term $B(t)I(t)$ represents the vegetation carbon input, $A\xi(t)KX(t)$ describes the SOC movements among the 7 pools within each layer, and V(t)X(t) indicates vertical SOC movements along the soil profile. The t in parentheses means that the corresponding process changes with time.

In this study, we assumed steady-state SOC dynamics for computational efficiency (Appendix 2), which is justified by previous research showing that recent disequilibrium effects from climate change and human activities are relatively minor compared to the SOC storage that has developed over thousands of years [2,25].

Equation (3) contains 21 biogeochemical parameters (Table S2) that quantify the strength and reflect properties of different processes (e.g., transformation and stabilization of SOC, temperature sensitivity of soil respiration, and substrate quality) in the soil carbon cycle. Because

those processes are highly variable depending on different climate conditions or soil properties, the values quantifying their strength or properties (i.e., the parameter values) should differ with changing environments [26]. Thus, in this study, the neural network embedded in BINN (Section 2.1) predicts these biogeochemical parameter values from environmental covariates. The predicted values of the 21 biogeochemical parameters and the environmental forcings (Table S3) are used in Equation (3) to estimate steady-state SOC storage at sites across the continental US.

2.3 Loss Function

Because we are interested in accurately simulating SOC, our primary loss quantifies the discrepancy between simulated and observed SOC values. Specifically, we use a smooth L1 loss function, which transitions from quadratic behavior near zero to linear behavior beyond a specified threshold β.

$$Smooth\ L1\ Loss\left(\hat{y}_{profile}, y_{profile}\right) = \begin{cases} \frac{0.5\left(\hat{y}_{profile} - y_{profile}\right)^2}{\beta} & if\ \mid \hat{y}_{profile} - y_{profile} \mid < \beta \\ \mid \hat{y}_{profile} - y_{profile} \mid - 0.5 \times \beta & o.w. \end{cases} \quad (4)$$

where $\hat{y}_{profile}$ represents the simulated SOC profile at all observation depths for a single site by CLM5, $y_{profile}$ denotes the corresponding observed SOC profile at the same site, and β is a threshold hyperparameter that determines the transition point between quadratic and linear behaviors of the loss function. The smooth L1 loss function's linear asymptotic behavior makes it more robust to outliers compared to conventional loss functions [27].

We also add an additional hyperbolic cosine loss ($cosh$) term that acts as a regulator, encouraging the neural network to predict biogeochemical parameters within reasonable bounds.

Specifically, it penalizes parameter values that deviate substantially from the center of the prior distribution, thereby discouraging biogeochemically implausible extreme values. Eventually, the total loss is a linear combination of the two losses (Equation 5):

$$L_{batch} = \sum_{profile=1}^{batch\ size} \{Smooth\ L1\ Loss(\hat{y}_{profile}, y_{profile}) + w \sum_{j=1}^{21} cosh\left[\tau(p_j - 0.5)\right]\} \quad (5)$$

where $batch\ size$ is a hyperparameter describing the number of soil profiles processed in each training iteration before performing one backpropagation, w is a weighting hyperparameter that balances the two loss components, p_j represents the predicted biogeochemical parameter from the neural network, and τ is a scaling factor that controls the strength of regularization by the hyperbolic cosine function. From the hyperparameter grid search, we set beta to 1 and biogeochemical-parameter-loss weight to 100.

While CLM5 simulates SOC dynamics at 20 specific depths, SOC data collected from the field were not necessarily measured at the depth nodes set in CLM5 simulation. Thus, in calculating the loss function value, for observations at depths equaling CLM5 nodes, the simulated values were directly from CLM5 outputs. When observations occur at depths between two CLM5 nodes, we employed linear interpolation to estimate simulated SOC values at the observation depths. In cases where observations extend beyond 8 meters (i.e., the deepest node in CLM5 simulations), we used the values at 8 meters as simulated SOC as SOC concentration in deeper layers no longer changes much.

2.4 Backpropagation to Optimize Neural Network Parameters

During training, BINN computes the loss function based on the current predicted SOC and parameters; the loss quantifies how poorly its current predictions match the SOC observations and prior knowledge. Through backpropagation, the loss signals propagate backwards through the entire BINN structural chain: first through the CLM5 matrix equations that generate modeled SOC, then through the biogeochemical parameters, and finally to the neural network that predicts these biogeochemical parameters. At each step, PyTorch uses the chain rule to automatically compute the gradient of the loss function with respect to each learnable NN component (e.g. $\boldsymbol{w}$ in Equation 1 and γ in Equation 2). These gradients indicate how each component can be adjusted to increase or decrease the loss. The NN components are adjusted slightly in the direction that decreases the loss, and then the above process is repeated. The differentiability of the process-based model (CLM5 in this case) enables this continuous gradient flow and thus allows the neural network to learn parameter values that produce better SOC predictions.

3 Recovering Biogeochemical Parameters from Synthetic Data

We evaluated BINN's capability to recover the biogeochemical parameters of CLM5 from synthetic SOC data using a 10-fold cross-validation experiment (Appendix 5) [28]. Unlike real-world observations that contain measurement uncertainties and potentially unresolved processes, synthetic SOC data across multiple depths was generated by running CLM5 with prescribed spatially-varying parameter values (obtained from previous work[1]), providing a controlled environment where true parameter values are known (Figure 2). This synthetic dataset allows us to quantitatively assess BINN's parameter recovery accuracy by comparing predicted parameters with the known values used in data generation.

To decide which biogeochemical parameters to modify in this experiment, we conduct a sensitivity analysis of CLM5 to identify the biogeochemical parameters that have the greatest

influence on simulated SOC values (Appendix 4). We selected the four most sensitive parameters from this sensitivity test: the parameter "w-scaling" represents the influence of soil water on modifying SOC decomposition; the parameter "tau4s3" represents the decomposability of the passive SOC pool; the parameter "fs1s3" indicates the efficiency of carbon transforming from active SOC to passive SOC; and the parameter "efolding" quantifies the impacts of soil depth in SOC decomposition. While equifinality, by which different combinations of biogeochemical parameter values can lead to similar simulations, remains a challenge even with this reduced parameter set, focusing on these highly sensitive parameters allows us to evaluate BINN's parameter recovery capabilities with greater confidence. This approach minimizes confounding effects from less influential parameters while targeting the parameters that most strongly influence SOC dynamics in CLM5.

To test the recovery efficiency of the biogeochemical parameters with BINN, we modified the final layers of the neural network by reducing the number of neurons from 21 to 4 to predict these 4 biogeochemical parameters. Combining the 4 biogeochemical parameters predicted by BINN and the remaining 17 biogeochemical parameters from the prescribed parameter values, BINN was able to simulate SOC values and update itself through backpropagation. After training BINN with the synthetic SOC dataset, we compared the 4 parameters predicted by BINN with the prescribed parameter values in the testing dataset to evaluate the accuracy of BINN in retrieving the prescribed biogeochemical parameters.

When BINN predicted the top 4 most sensitive biogeochemical parameters, the recovered parameter values exhibited strong consistency with the prescribed biogeochemical parameters used during synthetic data generation, achieving an average correlation coefficient of 0.73 across 10 cross-validation iterations (Figure 4f). BINN achieved an average Nash–Sutcliffe modelling

efficiency coefficient (NSE) of 0.67 (Supplementary 6) on the test dataset when comparing simulated SOC with synthetic SOC (Figure 4f). In one of the cross-validation iterations with the median NSE of simulated SOC, the parameter “efolding” recovered by BINN, representing the depth scalar, had a correlation coefficient of 0.76 in comparison with the prescribed parameter values (Figure 4a). The parameter "tau4s3," representing the baseline turnover time of passive SOC pools, showed a correlation coefficient of 0.78 (Figure 4b). The "fs1s3" parameter, indicating the transfer fraction from fast SOC pool to passive SOC pool, achieved a correlation coefficient of 0.71 (Figure 4c). Lastly, "w-scaling," representing the scaling factor of soil water scalar, had a correlation coefficient of 0.70 (Figure 4d).

4. BINN performance with real-world SOC Observations

We then evaluated the performance of BINN by comparing BINN’s SOC predictions with observed SOC across the Conterminous United States (a total of 25,925 profiles) (Figure 1b).

4.1 Data Preparation

We processed SOC observations from the World Soil Information Service (WoSIS) following Tao et al. [1,2]. Each profile (site) may have SOC observations at multiple depths. Only profiles with at least three observations were kept, yielding 25,925 profiles (169,104 SOC measurements) across the conterminous US. We used 60 environmental covariates at each site from Tao et al. [2] as input to BINN (Table S1). To achieve better training effectiveness, we normalized all the environmental covariates to the interval [0, 1] according to their maximum and minimum values.

We applied eight types of forcing data to drive the simulations of SOC using CLM5, which are the mean annual net primary productivity (NPP), active soil layer depth from last year and current year, soil layer number that reaches the bedrock, soil oxygen scalar for

decomposition, soil nitrogen scalar for decomposition, soil temperature, and soil water potential. These forcings are from 20 years of monthly CLM5 simulations at the steady state using a preindustrial forcing (that is, I1850Clm50Bgc) at 0.5° resolution.

The 10-fold cross-validation divided the whole dataset randomly into 10 folds, and we took one-fold (i.e., 10%) data as the testing dataset in each iteration. The remaining data were further split into training (8/9) and validation (1/9) sets.

4.2 Real-world SOC Observations Analysis

After BINN optimization, we used its predicted biogeochemical parameters to calculate six model components that indicate different properties in soil carbon cycle over the US continent: carbon transfer efficiency, baseline decomposition, environmental modifier, carbon input allocation, vertical transport rate, and plant carbon inputs. The BINN-retrieved components were compared with the results generated from PRODA (Figure 5). We assessed the spatial patterns of six model components that emerged from both approaches. The plant carbon input component was identical between BINN and PRODA, due to the use of the same NPP forcing data (Figure 5p, 5q, 5r). Spatial distributions of the other five components were similar between BINN and PRODA with an average correlation of 0.81.

The carbon transfer efficiency predicted by BINN, which quantifies the weighted average ratio of decomposed carbon being transferred from one carbon pool to another relative to the total carbon decomposition, exhibited more spatial variation than PRODA's prediction. While both methods indicated higher carbon transfer efficiency in the northwestern region and lower efficiency in the middle west (Figure 5a, 5b), BINN predicted higher values in the northeastern and southeastern parts of the Conterminous United States compared to PRODA, resulting in a

relatively high average value. Even so, the correlation coefficient between the two approaches still reaches 0.86 (Figure 5c).

Both BINN and PRODA predicted the baseline decomposition, which describes the substrate decomposability of each soil pool, with similar spatial patterns across the Conterminous United States, showing higher values in the northwestern and eastern areas (Figure 5d, 5e). The average baseline decomposition values from BINN were relatively higher than those by PRODA (Figure 5f).

The environmental modifier predicted by BINN achieved a correlation coefficient of 0.82 with PRODA's results (Figure 5i). Their spatial patterns were nearly identical across the Conterminous United States, with lower values in the northwestern part and gradually increasing to the highest values in the southeastern part (Figure 5g, 5h).

The carbon input allocation predicted by both methods also displayed similar spatial patterns across the Conterminous United States with a correlation coefficient of 0.74 (Figure 5l). Both the methods predicted low carbon input allocation rates in the eastern and western US but high rates in the mid US (Figure 5j, 5k), though BINN predicted higher rates of carbon input allocation in the mid US than the PRODA's predictions.

BINN and PRODA predicted the vertical transport rate with nearly identical spatial distributions across the Conterminous United States (Figure 5m, 5n), with a high correlation coefficient of 0.91 (Figure 5o).

BINN demonstrated better accuracy than PRODA in predicting SOC across the Conterminous US as well (Figure 6) [1]. Fewer geographical biases were observed when comparing BINN's SOC predictions with observed SOC from the testing dataset (Figure 6a). The predicted and observed SOC values were highly correlated, with a NSE value of 0.66 (Figure

6b). The training and validation NSE values recorded throughout model training at each epoch showed the validation NSE reaching as high as 0.63 (Figure S3).

5. Computational Efficiency

We compared BINN's computational efficiency with PRODA using 2,000 soil profiles on a personal computer with two BINN versions: one using the original matrix equation of CLM5 and the other using the vectorized matrix equation by removing for-loops (Appendix 2). Both BINN implementations were trained for 300 epochs to ensure that BINN has finished learning from the soil profiles.

PRODA is a two-step approach that combines Bayesian-inference approaches with neural networks. Its computational bottleneck lies in its first step: site-level Markov Chain Monte Carlo (MCMC) optimization of CLM5 parameters. To quantify this, we measured MCMC runtime for 10 profiles (20,000 test iterations and 50,000 formal iterations per site) and extrapolated to 2,000 profiles. PRODA's second step, which uses a neural network to learn relationships between environmental covariates and MCMC-optimized parameters, required 6,000 epochs of training.

On a single CPU, BINN with for-loops required 52.5 hours (Figure 8), whereas the vectorized version took 10.5 hours (5× faster). In contrast, MCMC alone took 574.69 hours, leading to 577.69 hours for PRODA. Thus, BINN is 57 times faster than PRODA and can be further accelerated via PyTorch's Distributed Data Parallel (DDP), enabling parallel training across multiple CPUs. While both approaches achieve parameter interpretability (Section 4.2), BINN does so more efficiently by directly integrating the process-based model into the neural network architecture, eliminating the need for computationally intensive site-by-site MCMC optimization.

6. Determination of SOC over the conterminous US

To further demonstrate how BINN helps understand processes governing SOC dynamics, we conducted a traceability analysis [29]. The traceability analysis separates BINN-predicted SOC storage from section 4.2 into carbon influx and ecosystem residence time, with the latter being calculated by dividing carbon storage by carbon influx. The ecosystem residence time is jointly determined by baseline residence time and the environmental modifier. The decomposition at each grid was then averaged across different biomes and visualized in a scatter plot for comparison among the biomes. Biome types were assigned to each grid based on the dominant ecological region within the grid, utilizing the Level 1 Ecological Regions of North America map provided by the US Environmental Protection Agency [30].

The traceability analysis revealed that ecosystems with similar average carbon storage, which is illustrated by the close proximity of the dots to the contour line, can result from distinct underlying processes in different biomes (Figure 7a). For instance, North American Deserts and Mediterranean California exhibited similar carbon storage, despite contrasting underlying mechanisms, with North American Deserts having longer ecosystem carbon residence times coupled with smaller carbon inputs than Mediterranean California. Additionally, Mediterranean California and Northwestern Forested Mountains showed similar SOC residence time (Figure 7b), which is attributed to similar baseline carbon residence times and environmental scalars.

7. Discussion

This paper introduces BINN, a novel approach for retrieving model parameters from big data and predicting their spatial distributions over the globe, accelerating computational efficiency, and facilitating process understanding.

7.1. BINN's ability to retrieve and predict biogeochemical parameters

In this study, BINN's ability of retrieving biogeochemical parameters was first validated through a parameter recovery experiment, which used synthetic SOC data generated by CLM5 with prescribed parameter values across the Conterminous US. To minimize the effects of equifinality, we performed a sensitivity analysis to select the four most sensitive biogeochemical parameters to be retrieved and predicted by BINN. These highly sensitive parameters are often well-constrained in the Bayesian-inference approach. Therefore, we expected BINN to perform similarly to the Bayesian approach in retrieving these parameters. The retrieval results showed that BINN could accurately recover the prescribed values using only environmental data and synthetic SOC data at each site without the Bayesian method.

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 [26,31].

We further tested BINN with real-world SOC observations across the Conterminous US and quantified the six model components of CLM5. BINN-predicted model components, which are calculated from estimated 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 estimation through data assimilation, the agreement between BINN and PRODA predictions suggests that BINN can effectively capture spatial variations in critical model components while maintaining physical interpretability. This enables BINN to evaluate the relative importance of different processes controlling SOC storage, similar to PRODA, while offering computational advantages through its integrated neural network architecture.

7.2. BINN's Computational Efficiency

BINN shows significant improvement in computational efficiency compared to PRODA while performing similar functionality in terms of retrieving parameters and predicting spatial distributions of SOC storage and their components from big data. Compared to PRODA, BINN reduces computational time by more than 50-fold in a test with 2,000 profiles. PRODA requires running a Bayesian optimization algorithm for each site independently; it does not use gradients to optimize parameter values but only perturbs parameters randomly and checks if the accuracy improved. By contrast, BINN reimplements the CLM5 process-based model in a differentiable way using PyTorch, and leverages this differentiability to rapidly find parameters (for all sites simultaneously) that accurately simulate SOC observations (Table S4). Additionally, we used vectorized functions to replace for-loops in the old CLM5 model, which further enhances computational efficiency. Finally, BINN can utilize PyTorch's Distributed Data Parallel (DDP) to parallelize computations, saving real physical time and allowing researchers to iterate more quickly on improving the model. High computational efficiency is also more environmentally-friendly, saving energy when dealing with large datasets.

7.3. BINN's facilitation of mechanistic understanding

BINN's aim is to integrate machine learning and process-based modeling to assist in identifying controls over biogeochemical systems by leveraging the power of big data. A process-based model is an abstraction of a real-world system and represents processes that govern the system, yet such model-based predictions generally fit poorly with empirical observations [31]. This discrepancy arises because complex systems, such as the terrestrial ecosystems, contain

numerous mechanisms regulating carbon cycling; although some of these mechanisms are well-understood, many remain unresolved. Process-based models may explicitly represent the well-understood processes in its structure while using parameters to represent unresolved processes [26]. Without taking advantage of the extensive information present in observations, model parameters are usually not well constrained.

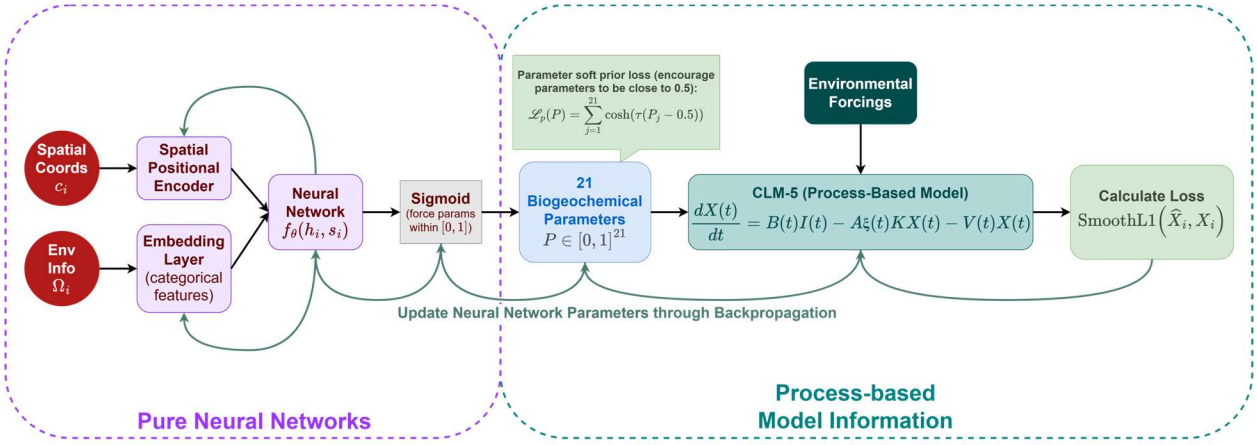
When these parameters are properly constrained by empirical data, they can more accurately reflect the unresolved biogeochemical processes, enabling models to better simulate ecosystem behavior. For example, Liu et al. [32] demonstrated that optimized parameters representing xylem water potential in an eco-hydraulic model aligned well with measured values for dominant species across different sites. To effectively learn about unresolved processes through parameter optimization, model simulations must closely match real-world observations. BINN achieves this by optimizing the relationships between environmental covariates and model parameters through its neural network component. This optimization process enables BINN-trained CLM5 to simulate soil carbon dynamics more accurately and efficiently than the original CLM5, thereby allowing parameters to better represent unresolved processes.

While optimized parameters can represent unresolved biogeochemical processes in well-performing models, deeper scientific analysis is needed to fully understand these processes and their roles in ecosystems. As demonstrated by traceability analysis, process-based models after parameters are constrained can be used to examine how various processes influence SOC storage across space, revealing spatial patterns in mechanisms like carbon residence time as shown in the study by Tao et al. [2]. Furthermore, such a model can help evaluate how newly incorporated mechanisms affect existing processes, as demonstrated by Xia et al. [29] in their study of nitrogen processes' impact on carbon storage capacity.

BINN's potential for advancing scientific understanding can be extended well beyond SOC dynamics to various fields of research in biogeochemistry and ecology. Whenever current scientific understanding of a biogeochemical system can be mathematically formulated in a process-based model, BINN can help uncover mechanisms from big data. This framework is particularly valuable for studying complex biogeochemical cycles, such as nutrient cycles, where some processes are well-understood and explicitly represented in models, while others remain unresolved. By combining process-based representations of known mechanisms with big data, BINN could help identify previously unknown mechanisms governing biogeochemical cycling.

Furthermore, BINN's flexible architecture allows integration of diverse data sources. This capability is particularly valuable for incorporating limited but important datasets, such as isotope measurements, which can reveal spatial and temporal mechanisms in soil carbon dynamics despite their scarcity. Even if these measurements are only available at a few sites, BINN can still learn from them by incorporating them in the loss function where they are available. By leveraging multiple data sources, BINN maximizes the potential to facilitate our scientific understanding while maintaining biogeochemical consistency.

(a)



(b)

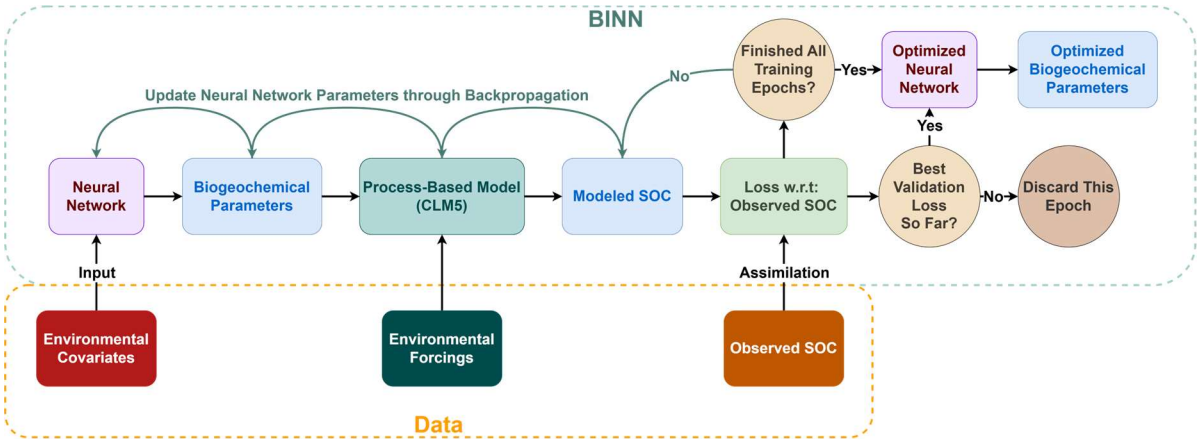


Figure 1. Schematic diagram of BINN architecture and training process. (a) Detailed BINN structure showing the integration of neural networks with CLM5. The neural network component processes spatial coordinates through a positional encoder and categorical environmental covariates through an embedding layer. The network outputs are transformed via a sigmoid activation to generate 21 biogeochemical parameters. These parameters, constrained by a soft prior loss, are input to CLM5 along with environmental forcings to simulate SOC dynamics. The

model's performance is evaluated using a smooth L1 loss function. The entire framework is differentiable, enabling end-to-end training through backpropagation (teal arrow).

(b) Overview of BINN training workflow. Environmental covariates at each site serve as input to the neural network to predict biogeochemical parameters. These parameters, along with environmental forcings, drive the process-based model (CLM5) to simulate SOC. The difference between modeled and observed SOC is used to compute the loss function, which guides neural network parameter updates through backpropagation (teal arrow). This training process continues until reaching the maximum number of epochs or achieving optimal validation performance.

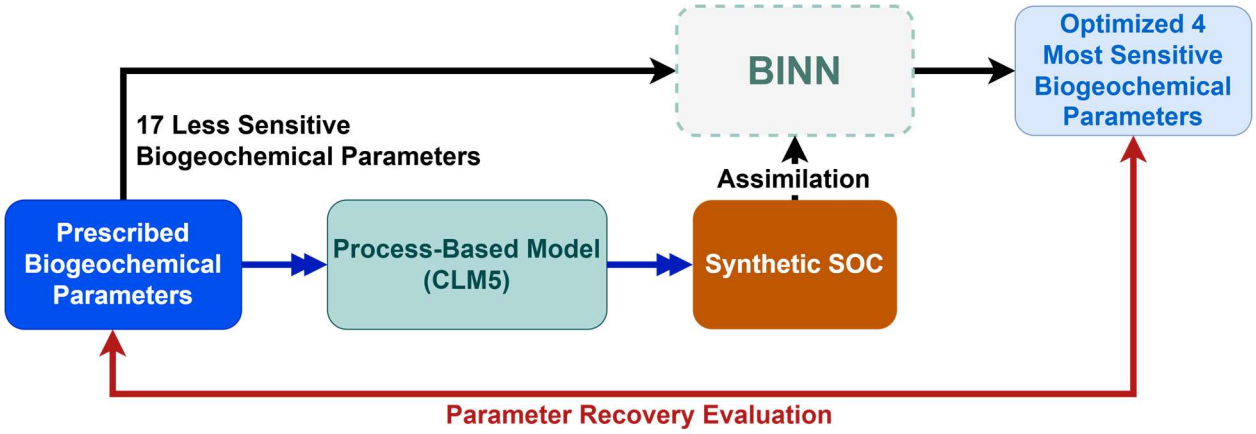


Figure 2: Schematic of the parameter recovery experiment to evaluate BINN's ability to retrieve the model processes regulating SOC. The parameter recovery experiment involves three main steps. 1). Blue two-headed arrows: Synthesizing a SOC dataset using CLM5 with prescribed parameter values (21 parameters). 2). Single-headed arrows: Using the synthetic SOC dataset to train BINN to predict the 4 most sensitive parameters. 3). Red double arrow: By comparing the BINN-predicted parameters with the prescribed parameters used to generate the

synthetic dataset, we can assess BINN's effectiveness in retrieving the processes regulating SOC from observational data.

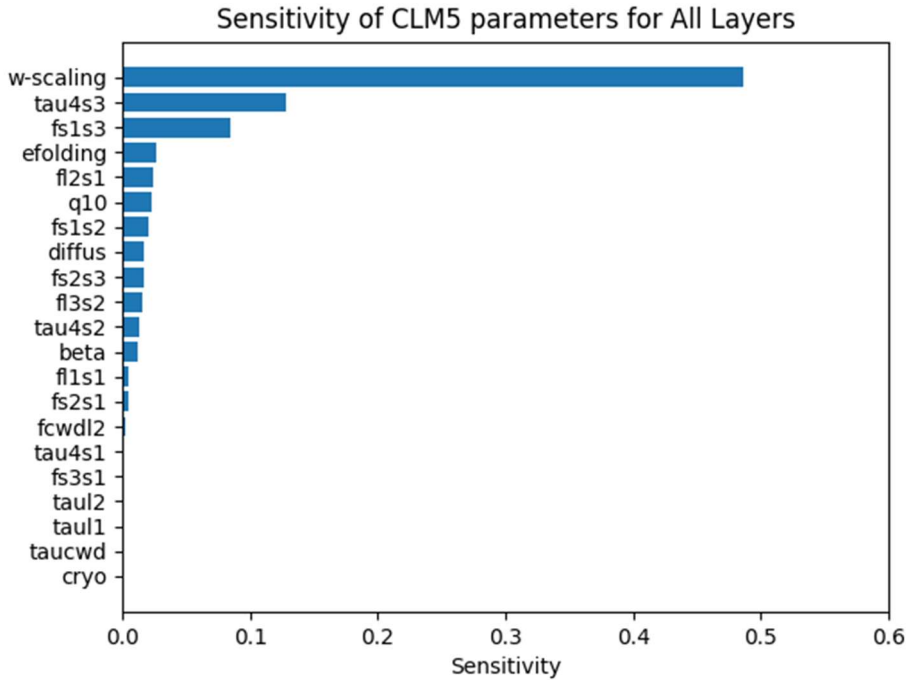


Figure 3: Sensitivity Indices for CLM5 Biogeochemical Parameters Across All Soil Depths. The bar plot illustrates the sensitivity of CLM5 to each parameter across all soil layers. Parameters are listed on the y-axis in descending order based on their sensitivity scores. The x-axis represents the sensitivity scores, indicating how changes in each parameter influence the model's performance.

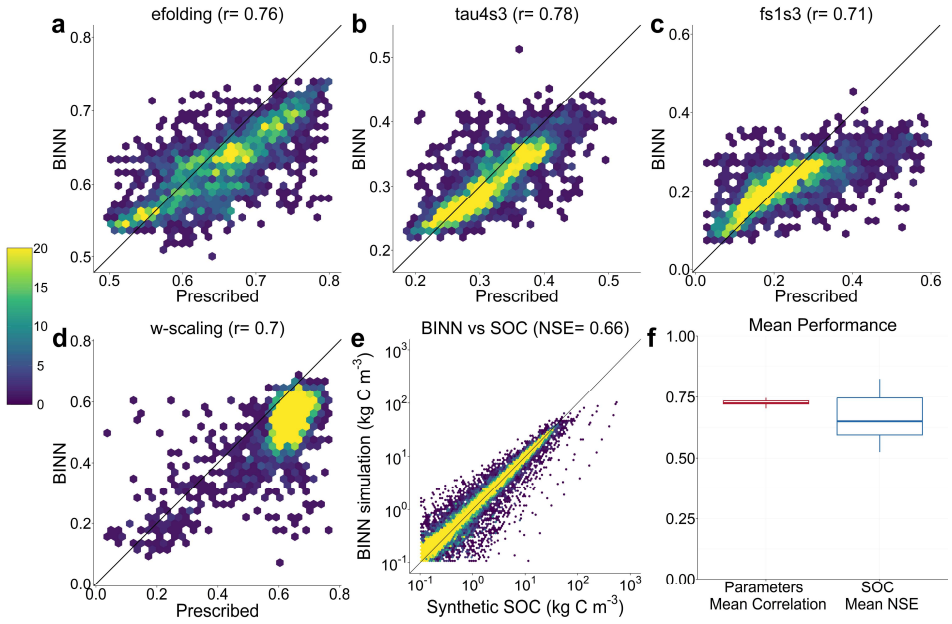


Figure 4: Evaluation of BINN's performance in retrieving the four most sensitive parameters from the synthetic SOC dataset. Scatter plots comparing the parameter values (a) "efolding", (b) "tau4s3", (c) "fs1s3", and (d) "w-scaling" predicted by BINN (BINN) against the prescribed parameter values. The color of each point in the scatter plots represents the number of data points within each hexagonal bin. The correlation coefficient between the predicted and prescribed parameter values is shown in the title of each plot. (e) Comparison of the simulated and synthetic SOC values, with colors representing the density of points. (f) Mean performance of BINN in retrieving the 4 parameters, as measured by the correlation coefficient and NSE between the predicted and prescribed parameter values as well as NSE of the simulated SOC values compared to the synthetic SOC data.

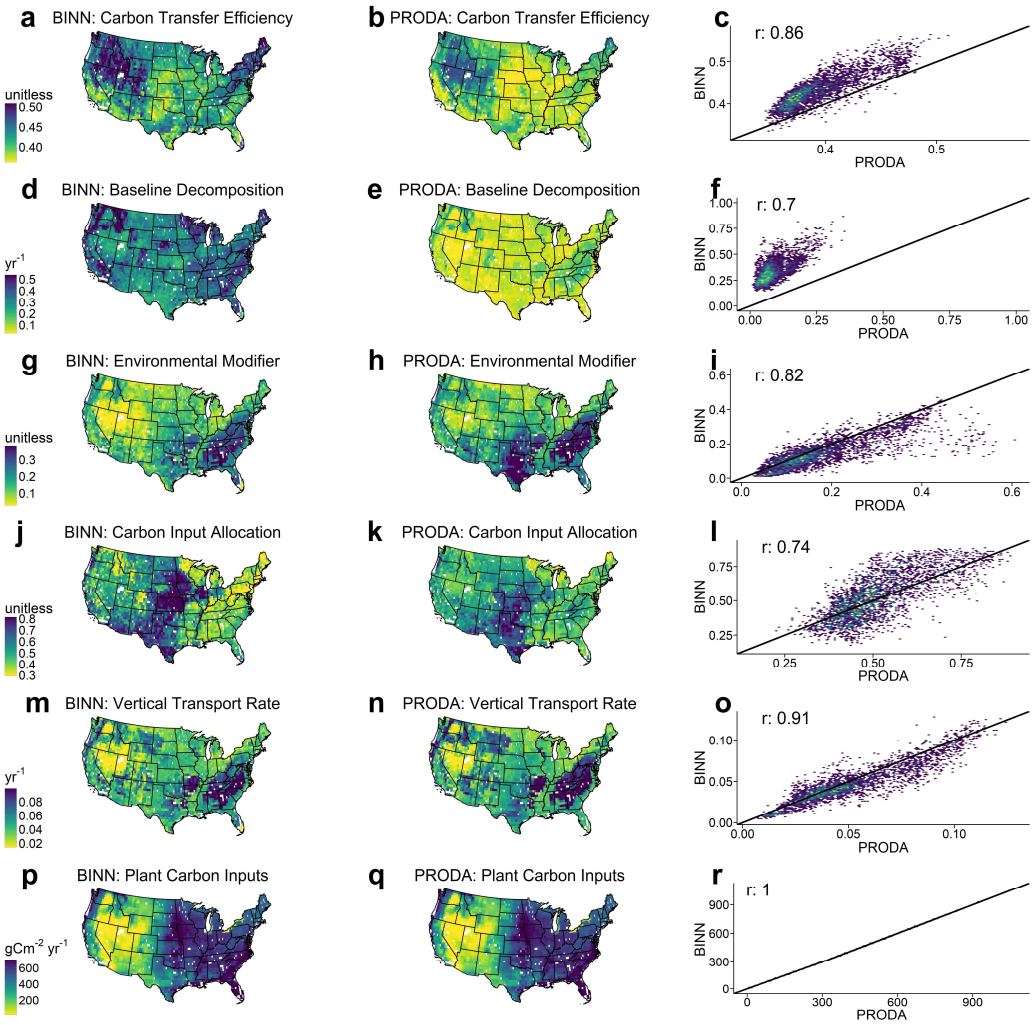


Figure 5: Comparison of the spatial patterns of model components retrieved by BINN and PRODA across the Conterminous United States. The model components include carbon transfer efficiency (a, b, c), baseline decomposition (d, e, f), environmental modifier (g, h, i), carbon input allocation (j, k, l), vertical transport rate (m, n, o), and plant carbon inputs (p, q, r). The left column (a, d, g, j, m, p) shows the model components retrieved by BINN, while the middle column (b, e, h, k, n, q) displays the model components retrieved by PRODA. The scatter plots in the right column (c, f, i, l, o, r) compare the values of each model component retrieved by BINN (y-axis) against those retrieved by PRODA (x-axis). The correlation coefficient between the BINN and PRODA values for each model component is shown in the top left corner

of the corresponding scatter plot. The plant carbon inputs (p, q, r) are identical for both methods due to the use of the same input forcing data.

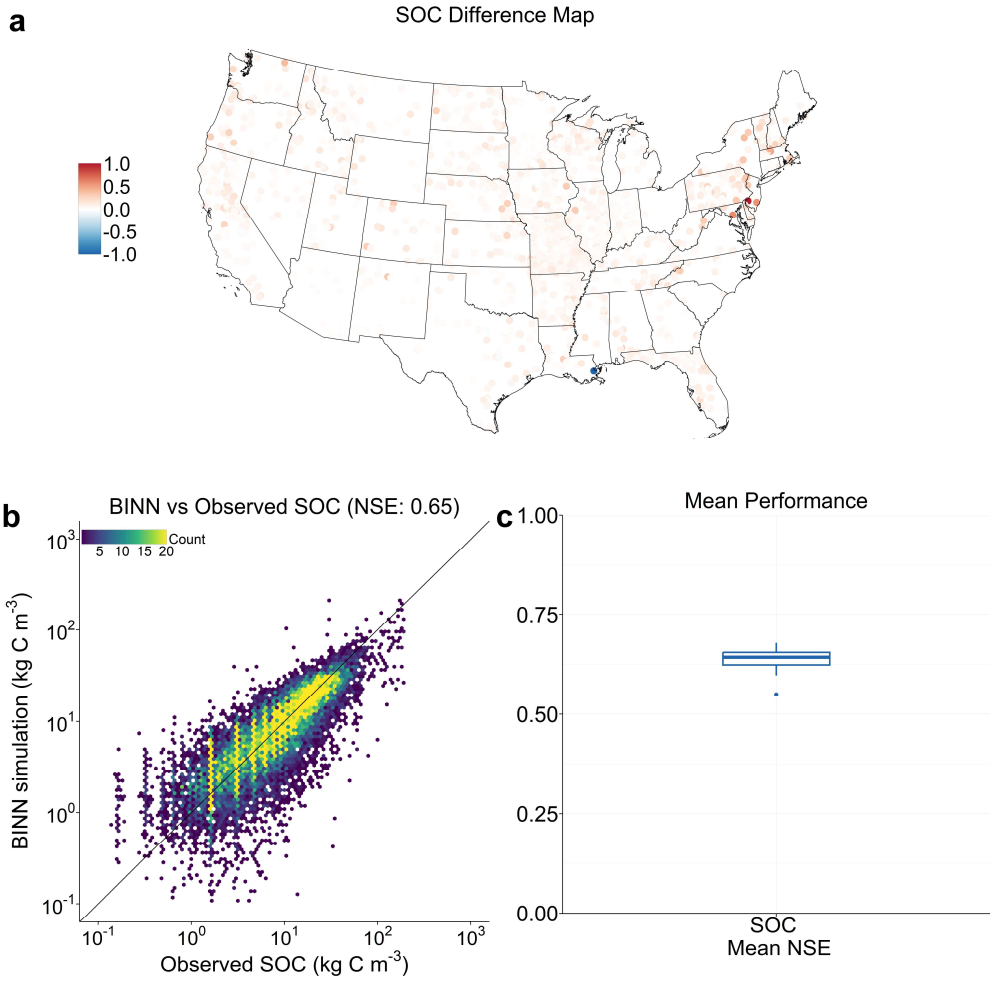


Figure 6: Comparison of observed and simulated Soil Organic Carbon (SOC) Storage Using BINN. (a) A spatial deviation mapping showcases the difference in simulated SOC storage by BINN relative to real-world observations across the soil profile for each test location. The test data comes from one cross-validation fold with median NSE values. The map normalizes positive discrepancies (to 0~1) and negative discrepancies (to 0~-1) against the

maximum positive deviation and the minimum negative deviation, respectively, to enhance the visualization of model performance. (b) The scatter plot presents the SOC from data points derived from the testing dataset between observed and simulated SOC storage at various soil depths, with the correlation coefficient values shown in the title. (c) The box plot shows the mean performance of testing NSE in the 10-fold cross validation test.

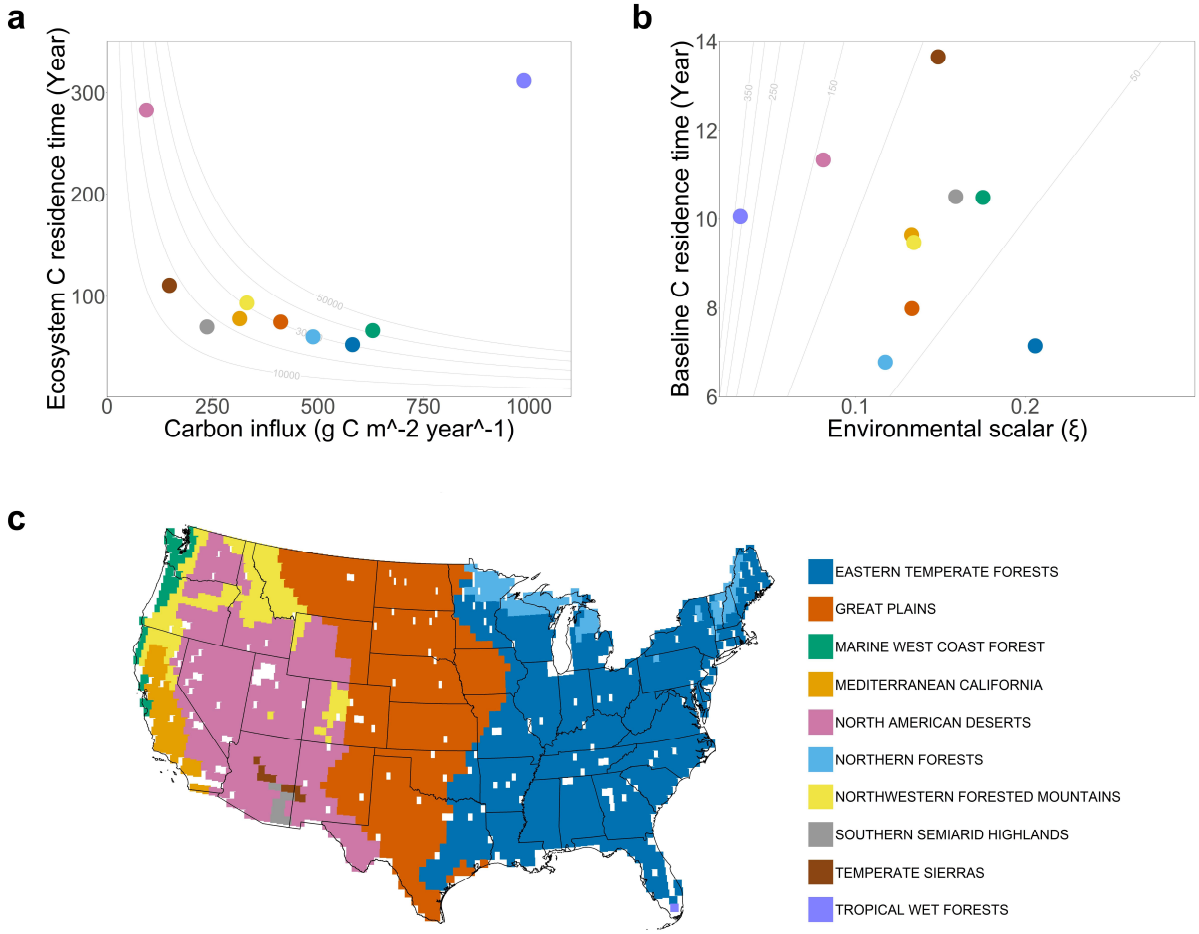


Figure 7: Traceability analysis on (a) how influx and ecosystem carbon residence time determine SOC storage (contour lines) and (b) how environmental scalar (ξ) and baseline carbon

residence time determines ecosystem carbon residence time (contour lines) in different biomes. The color-coded points represent the average values in different biomes.

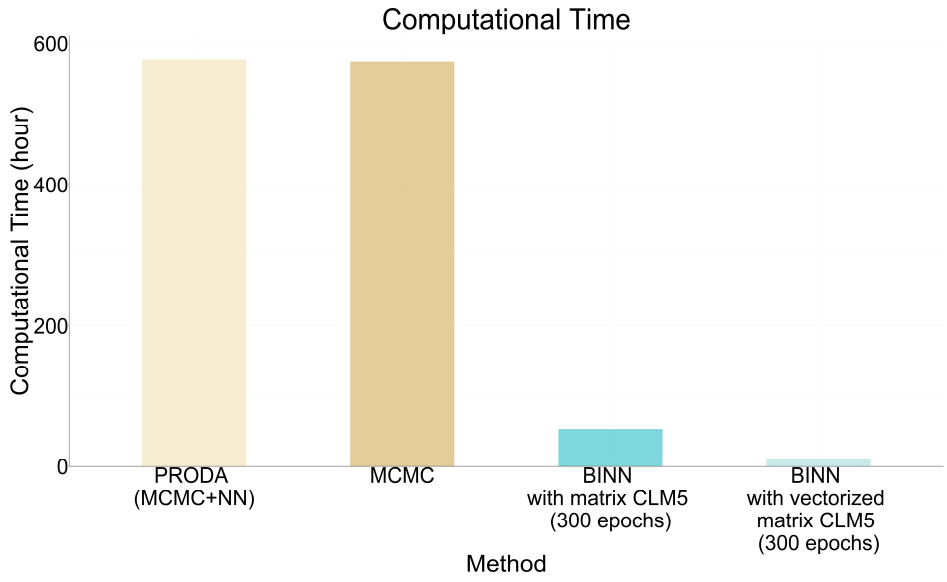


Figure 8: Comparative Analysis of Computational Time Required for Integrating 2000 Soil Profiles into Process-Based Models (CLM5). The figure shows the computational time (in hours) for PRODA (MCMC+NN), which uses a Bayesian-inference approach (MCMC) combined with a neural network (NN), and for BINN with the matrix form of CLM5 before and after vectorization. BINN with the vectorized matrix form of CLM5 achieves the highest computational efficiency, reducing the computational time by more than 50-fold compared to PRODA (MCMC+NN) and by approximately 5-fold compared to BINN with the non-vectorized matrix form of CLM5. The computational time is based on running each method for 300 epochs to ensure that the models have finished learning from the 2,000 soil profiles.

Acknowledgements

This research is supported by AI-CLIMATE: "AI Institute for Climate-Land Interactions, Mitigation, Adaptation, Tradeoffs and Economy", funded by the USDA National Institute of Food and Agriculture (NIFA) and the NSF National AI Research Institutes Competitive Award (No. 2023-67021-39829). This research is also partially supported by Schmidt Sciences programs, an AI2050 Senior Fellowship and an Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship; NSF grants (DEB 2242034, DEB 2406930, and DEB 2425290); the US Department of Energy's Terrestrial Ecosystem Sciences Grant DE-SC0023514, subcontract CW55561 from Oak Ridge National Laboratory to Cornell University; the CALS Moonshot Seed Grant program; the "NYS Connects: Climate Smart Farms & Forestry" project, funded by the USDA, the New York State Department of Environmental Conservation, and the New York State Department of Agriculture and Markets; an NSF Research Traineeship (NRT) fellowship in Digital Plant Science (DGE-1922551), and the Air Force Office of Scientific Research (AFOSR, grants FA9550-23-1-0322, FA9550-23-1-0569, FA9550-21-1-0316).

Data and Code Availability

The code for BINN is available at phxtao/BINNS at BINN_clean.

Conflicts of Interest

The authors declare no competing interests.

References

1. Tao, F. *et al.* Deep Learning Optimizes Data-Driven Representation of Soil Organic Carbon in Earth System Model Over the Conterminous United States. *Front. Big Data* **3**, (2020).
2. Tao, F. *et al.* Microbial carbon use efficiency promotes global soil carbon storage. *Nature* **618**, 981–985 (2023).
3. Reichstein, M. *et al.* Deep learning and process understanding for data-driven Earth system science. *Nature* **566**, 195–204 (2019).
4. Tan, T. *et al.* Importance of Terrain and Climate for Predicting Soil Organic Carbon Is Highly Variable across Local to Continental Scales. *Environ. Sci. Technol.* **58**, 11492–11503 (2024).
5. Hengl, T. *et al.* SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE* **12**, e0169748 (2017).
6. Cook-Patton, S. C. *et al.* Mapping carbon accumulation potential from global natural forest regrowth. *Nature* **585**, 545–550 (2020).
7. Jian, J., Steele, M. K., Thomas, R. Q., Day, S. D. & Hodges, S. C. Constraining estimates of global soil respiration by quantifying sources of variability. *Glob. Change Biol.* **24**, 4143–4159 (2018).
8. Willard, J., Jia, X., Xu, S., Steinbach, M. & Kumar, V. Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems. *ACM Comput Surv* **55**, 66:1-66:37 (2022).
9. Karpatne, A., Jia, X. & Kumar, V. Knowledge-guided Machine Learning: Current Trends and Future Prospects. (2024) doi:10.48550/ARXIV.2403.15989.
10. Hoedt, P.-J. *et al.* MC-LSTM: Mass-Conserving LSTM. in *Proceedings of the 38th International Conference on Machine Learning* 4275–4286 (PMLR, 2021).
11. Liu, L. *et al.* Knowledge-guided machine learning can improve carbon cycle quantification in agroecosystems. *Nat. Commun.* **15**, 357 (2024).

12. Kraft, B., Jung, M., Körner, M., Koirala, S. & Reichstein, M. Towards hybrid modeling of the global hydrological cycle. *Hydrol. Earth Syst. Sci.* **26**, 1579–1614 (2022).
13. Chen, D. *et al.* Automating Crystal-Structure Phase Mapping: Combining Deep Learning with Constraint Reasoning. Preprint at https://doi.org/10.48550/arXiv.2108.09523 (2021).
14. Chen, D., Zhu, Y., Cui, X. & Gomes, C. Task-Based Learning via Task-Oriented Prediction Network with Applications in Finance. in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* 4476–4482 (International Joint Conferences on Artificial Intelligence Organization, Yokohama, Japan, 2020). doi:10.24963/ijcai.2020/617.
15. Min, Y. *et al.* Physically Informed Graph-Based Deep Reasoning Net for Efficient Combinatorial Phase Mapping. in *2023 International Conference on Machine Learning and Applications (ICMLA)* 392–399 (2023). doi:10.1109/ICMLA58977.2023.00061.
16. Shen, C. *et al.* Differentiable modelling to unify machine learning and physical models for geosciences. *Nat. Rev. Earth Environ.* **4**, 552–567 (2023).
17. Fang, J., Bowman, K., Zhao, W., Lian, X. & Gentine, P. Differentiable Land Model Reveals Global Environmental Controls on Ecological Parameters. Preprint at https://doi.org/10.48550/arXiv.2411.09654 (2024).
18. Raissi, M., Perdikaris, P. & Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **378**, 686–707 (2019).
19. Tramontana, G. *et al.* Partitioning net carbon dioxide fluxes into photosynthesis and respiration using neural networks. *Glob. Change Biol.* **26**, 5235–5253 (2020).
20. Tao, F. & Luo, Y. PROcess-Guided Deep Learning and DAta-Driven Modelling (PRODA). in *Land Carbon Cycle Modeling* (CRC Press, 2022).

21. Tao, F. *et al.* Convergence in simulating global soil organic carbon by structurally different models after data assimilation. *Glob. Change Biol.* **30**, e17297 (2024).
22. Lawrence, D. M. *et al.* The Community Land Model Version 5: Description of New Features, Benchmarking, and Impact of Forcing Uncertainty. *J. Adv. Model. Earth Syst.* **11**, 4245–4287 (2019).
23. Luo, Y. *et al.* Matrix Approach to Land Carbon Cycle Modeling. *J. Adv. Model. Earth Syst.* **14**, e2022MS003008 (2022).
24. Huang, Y. *et al.* Matrix approach to land carbon cycle modeling: A case study with the Community Land Model. *Glob. Change Biol.* **24**, 1394–1404 (2018).
25. Lu, X., Wang, Y.-P., Luo, Y. & Jiang, L. Ecosystem carbon transit versus turnover times in response to climate warming and rising atmospheric CO_2 concentration. *Biogeosciences* **15**, 6559–6572 (2018).
26. Luo, Y. & Schuur, E. A. G. Model parameterization to represent processes at unresolved scales and changing properties of evolving systems. *Glob. Change Biol.* **26**, 1109–1117 (2020).
27. Girshick, R. Fast R-CNN. Preprint at https://doi.org/10.48550/arXiv.1504.08083 (2015).
28. Arnold, C. P. & Dey, C. H. Observing-Systems Simulation Experiments: Past, Present, and Future. *Bull. Am. Meteorol. Soc.* **67**, 687–695 (1986).
29. Xia, J., Luo, Y., Wang, Y.-P. & Hararuk, O. Traceable components of terrestrial carbon storage capacity in biogeochemical models. *Glob. Change Biol.* **19**, 2104–2116 (2013).
30. U.S. Environmental Protection Agency. NA_Eco_Level1. (2010).
31. Luo, Y. *et al.* Toward more realistic projections of soil carbon dynamics by Earth system models. *Glob. Biogeochem. Cycles* **30**, 40–56 (2016).
32. 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. Change* **10**, 691–695 (2020).

Supplementary Information

Appendix 1: Neural Network Architecture Details

Here we provide more details on our neural network architecture. Our network architecture includes an embedding layer, a spatial positional encoder, and a 4-layer fully connected network (Figure 1a).

Embedding categorical and spatial data. First, we convert categorical covariates in the environmental dataset via embedding layers into numerical vectors. The spatial location of each site (longitude and latitude) are passed through a spatial positional encoder [1] to obtain a location embedding vector, characterizing unobserved aspects of each location that are not captured in our covariates. We then concatenate the categorical covariate embeddings, location embedding, and the remaining covariates, and pass them through a 4-layer fully-connected neural network.

Fully-connected layers. Each layer of the neural network comprises prescribed numbers of neurons that receive information either from the environmental covariates (for the first layer) or the previous layer. It then computes a linear transformation of the inputs:

$$y = \sum w_i x_i + b \tag{S1}$$

where x_i is an input from either the environmental covariates or the previous layer's outputs, w_i is a learnable weight for x_i, b is a learnable neuron-specific bias, and y is an output from a neuron after the linear combination of its input. After the linear transformation, a nonlinear activation function is applied to generate the eventual results at each neuron, such that the neural network can generate complex nonlinear relationships between the input (i.e., environmental covariates) and the outputs (i.e., the biogeochemical parameters in CLM5). Meanwhile, the application of a linear transformation and activation function also ensures that the nonlinear relationships explored by the neural network are differentiable, such that we can calculate the gradient of the cost function with respect to the learnable weights w_i and biases b (Section 2.4). In our study, for the first three layers of the neural network, we assigned each of them to have 128 neurons to process information from the previous layer and used LeakyReLU as the activation function:

$$LeakyReLU(y) = max(0, y) + negative\_slope * min(0, y) \quad \text{(S2)}$$

where *max(0, y)* is a function that returns the larger value of *0* or *y*, while *min(0, y)* returns the smaller value of *0* or *y*. The *negative_slope* is a hyperparameter that determines how "leaky" the function is for negative y. LeakyReLU was chosen over traditional ReLU because it allows gradients to flow through the network even when the inputs to the activation are negative.

Final layer and parameter constraints. The final layer only has 21 output neurons, one corresponding to each biogeochemical parameter in CLM5. We do not use a leaky ReLU after the final linear transformation, as it would bias the distribution of the predicted parameters. Instead, we have prior knowledge about the plausible range of values for each biogeochemical parameter. Thus, we pass the final layer's output $\boldsymbol{z}$ through a sigmoid (σ) function to ensure that the parameter predictions fall within these prior ranges (Equation 2 in the main body).

After the activation, the final outputs of the neural network will be 21 values falling in the range between $\theta_{i,min}$ and $\theta_{i,max}$, each corresponding to the investigated biogeochemical parameters. $\theta_{i,min}$ and $\theta_{i,max}$ are values taken from previous literature to indicate plausible limits for processes quantified by each biogeochemical parameter i [2]. Note that we introduced a γ value in the activation function to control how fast the results after activation can converge to $\theta_{i,max}$ or $\theta_{i,min}$. When the γ value is small, the predicted parameters may quickly get stuck at $\theta_{i,min}$ and $\theta_{i,max}$; when this happens, the derivative of the activation approaches zero and it will be difficult to further optimize the predictions via gradient-based optimization (see Section 2.4) [3]. Thus, we tuned γ to be a relatively large value to facilitate neural network optimization.

Hyperparameters. We conducted an experiment to determine the best hyperparameters (i.e., epochs of training, batch size, CPU number, optimizer, learning rate, embedding size of the embedding layer, whether to use batch normalization, the initialization of γ, *negative_slope* in the LeakyReLU and the loss function hyperparameters) for BINN. By performing a grid search for these hyperparameters, we

chose to train BINN for 300 epochs with a batch size of 32, using PyTorch Distributed Data Parallel (DDP) to distribute training across 128 CPUs. The optimized BINN model was recorded each time the validation loss improved over the previous best model. We used the AdamW optimizer with a learning rate of 0.01. The embedding size is 64. The network also uses batch normalization after each ReLU activation function to normalize the layers' outputs by re-centering and re-scaling, making training faster and more stable [4]. We initialized γ to 59.5 but allowed it to be further optimized throughout the training processes within the range from 10 to 109. Specifically, we used a sigmoid to constrain the range:

$$\gamma = 10 + 99 \cdot \frac{1}{1 + exp(-\gamma')} \quad \text{(S3)}$$

where γ' is a learnable parameter that is initialized to 0 (making the initial $\gamma = 59.5$). We set *negative_slope* to -0.3 in the LeakyReLU by default. **Appendix 2:** Steady-state SOC Simulations

The steady-state SOC storage $\hat{X}(t)$ can be obtained by letting dX(t)/dt on the left-hand side of equation (3) equal 0. Solving for $\hat{X}(t)$ we obtained:

$$\hat{X}(t) = \left(A\xi(t)K + V(t)\right)^{-1} B(t)I(t) \quad \textit{(S4)}$$

The matrix representation of CLM5 is implemented in PyTorch utilizing vectorized functions to replace all the for-loops in the original code. Vectorized functions are designed to operate on entire arrays of data simultaneously, rather than processing elements one by one. This enables more efficient computation of SOC predictions in response to changes in parameters. For example, we constructed two vectors for each carbon pool: (1) an environmental scalar vector containing temperature, moisture, oxygen, and depth modifiers that affect decomposition rates, and (2) a decomposition vector containing pool-specific baseline decomposition rates and carbon transfer coefficients. By constructing these vectors for all pools simultaneously (2 vectors for each pool, 7 carbon pools each layer, and 20 layers in total in CLM5), we can directly construct a matrix using the vectorized function. By implementing all mathematical operations (such as addition, matrix multiplication, and matrix inverse) using PyTorch functions, PyTorch can track the gradient of each operation. Using backpropagation, PyTorch can then

automatically compute the gradient of the loss function with respect to the learnable weights/biases of the neural network, differentiating through all the operations in the process-based model. Since the goal of training BINN is to minimize a loss function that quantifies the difference between simulated and observed SOC, a fully differentiable CLM5 allows BINN to trace differences in loss function values back to changes in biogeochemical parameters and eventually environmental covariates (via backpropagation), enabling gradient-based optimization to let CLM5 best simulate SOC observations.

Appendix 3: Computational Software and Hardware

We implemented BINN in Python using PyTorch and executed the experiments on the NCAR Derecho supercomputer. The experiments were conducted using one compute node with 128 CPU cores, leveraging PyTorch's DDP for multi-CPU training. However, BINN was also tested on a multi-GPU compute node on a cluster in Cornell University's Center for Advanced Computing, confirming its capability for training on GPU clusters when needed.

Appendix 4: Sensitivity Analysis

The sensitivity analysis was conducted on SOC simulations at various soil-depth ranges, including 0-0.3 m, 0.3-1 m, >1 m, and the entire soil profile (0-8 m). SOC simulations at each layer by CLM5 were aggregated based on the node depths falling into the above-mentioned depth ranges. Specifically, layers 1-6 were used to calculate SOC between 0-0.3 m, layers 7-9 for SOC between 0.3-1 m, and layers 10-20 for SOC greater than 1 m. Simulations from all 20 layers were summed up to calculate SOC across the whole soil profile. The variance and sensitivity for each depth range were calculated based on SOC values derived from the individual layers mentioned above.

For this analysis, we randomly selected 512 sites across the Conterminous US and employed the first-order approximation method. We first determined the unconditional variance $V(SOC)$ from the model output when all the 21 biogeochemical parameters (P) in CLM5 were allowed to vary freely within their initial ranges from Tao et al.[2]. Specifically, we randomly sampled the biogeochemical parameter values 1000 times in their initial ranges at each site, ran the model, and calculated the variance of the simulations, which was considered the unconditional variance V(SOC).

Next, we estimated the conditional expectation of the variable SOC for each biogeochemical parameter Pi (i = [0, 20]) at each site. We randomly selected a value (Pi*) for each biogeochemical parameter Pi from a uniform distribution within its prior range, as specified by Tao et al. (2023). For the remaining biogeochemical parameters (Pj: $j \neq i$), we randomly selected 1000 values from uniform distributions within their respective prior ranges. Using the sample of 1000 biogeochemical parameter sets, we estimated the conditional expectation E(SOC | Pi = Pi*). We repeated this sampling process for 100 randomly selected values of Pi and used the results to estimate the variance V(E(SOC | Pi)). This quantifies the variance in the output variable C as a result of modifying the biogeochemical parameter Pi. We discarded the simulations when NaN values appeared due to randomly sampled biogeochemical parameter sets. Finally, we repeated this procedure for each biogeochemical parameter Pi (i = [0, 20]), and a sensitivity index Si was calculated for each biogeochemical parameter at each site as:

$$S_i = \frac{V(E(SOC|P_i))}{V(SOC)} \quad (S5)$$

The final sensitivity value for each biogeochemical parameter was obtained by averaging the sensitivity values across all the randomly selected sites across all depths (Figure 3), and individual depth ranges (Figure S2).

Appendix 5: 10-Fold Cross-Validation

To conduct 10-fold cross-validations on the simulations, the entire dataset was randomly divided into ten equal-sized subsets. In each iteration, nine subsets were used for training, while the remaining subset served as the test set. This process was repeated ten times, with each subset serving as the test set once. The performance metrics, including NSE and r, were calculated for each iteration. Final performance evaluations were determined by averaging 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 generalizability 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.

Appendix 6: Summary Statistics

We calculated the Nash–Sutcliffe modelling efficiency coefficient (NSE) of simulated SOC (Equation 10) to evaluate the effectiveness of SOC predictions by BINN following the equation:

$$NSE = 1 - \frac{\sum(obs_i - simu_i)^2}{\sum(obs_i - \overline{obs_i})^2} \quad (S6)$$

where *obs* is the SOC observation, $\overline{obs}$ is the mean of the SOC observations, and *simu* is the simulated SOC by CLM5 embedded in BINN.

We used the Pearson correlation coefficient (r) between the predicted and prescribed biogeochemical parameters (Equation 11) to evaluate the effectiveness of BINN in recovering each of the 4 biogeochemical parameters:

$$r = \frac{\sum[(para_{BINN} - \overline{para_{BINN}}) \times (para_{true} - \overline{para_{true}})]}{\sqrt{\sum(para_{BINN} - \overline{para_{BINN}})^2 \times \sum(para_{true} - \overline{para_{true}})^2}} \quad (S7)$$

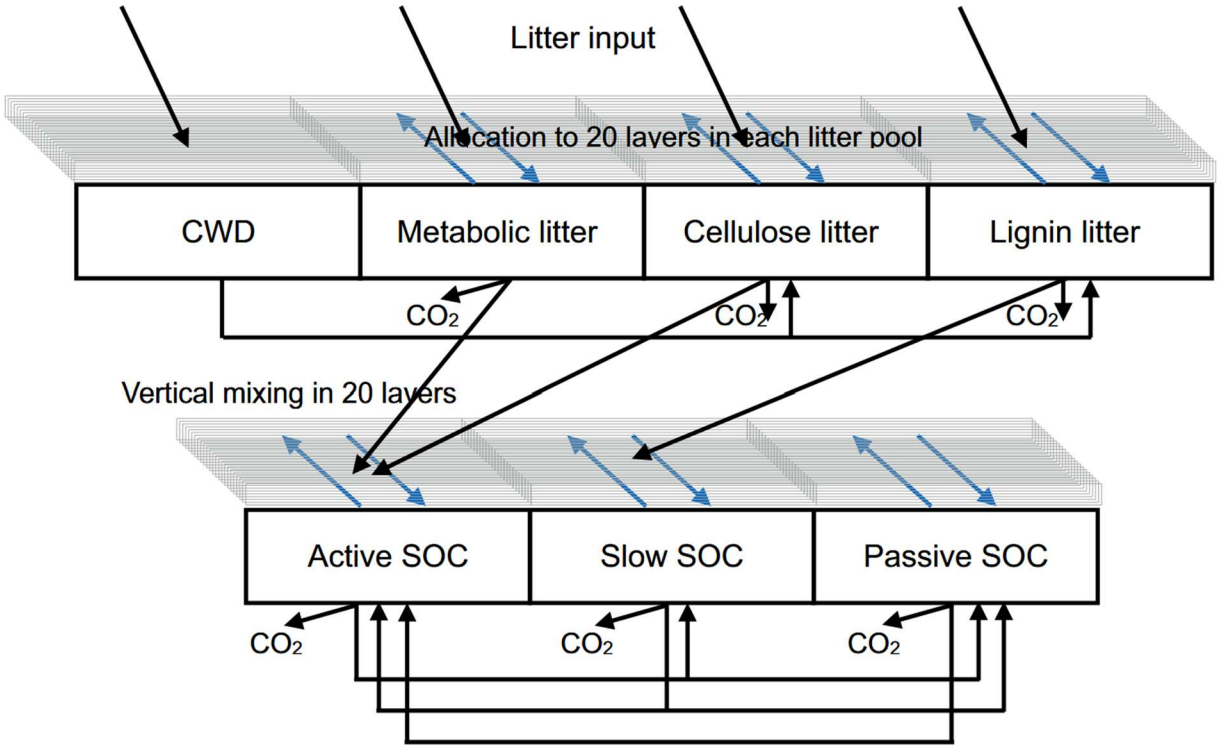
where $para_{BINN}$ is the biogeochemical parameter predicted by BINN, $para_{true}$ is the biogeochemical parameter previously prescribed at the same site, $\overline{para_{BINN}}$ is the mean of this biogeochemical parameter predicted by BINN, and $\overline{para_{true}}$ is the mean of this prescribed biogeochemical parameter.

Supplementary Tables and Figures

Supplementary Table 1: Environmental Covariate Data as BINN input

No.	Variable Name	Data Source	Category	Description
1	Lon	WoSIS	Geography	Longitude
2	Lat			Latitude
3	Elevation	NOAA		Elevation
4	Abs_Depth_to_Bedrock	(Hengl et al. 2017)		Soil layer depth that reaches the bedrock
5	Occurrence_R_Horizon	WoSIS		Probability of occurrence of R horizon
6	nbedrock	CLM5 simulation		Soil layer number that reaches the bedrock
7	Koppen_Climate_2018	(Beck et al. 2018)	Climate	Koppen Climate Classification
8	BIO1	(Fick and Hijmans 2017)		Annual Mean Temperature
9	BIO2			Mean Diurnal Range
10	BIO3			Isothermality
11	BIO4			Temperature Seasonality
12	BIO5			Max Temperature of Warmest Month
13	BIO6			Min Temperature of Coldest Month
14	BIO7			Temperature Annual Range
15	BIO8			Mean Temperature of Wettest Quarter
16	BIO9			Mean Temperature of Driest Quarter
17	BIO10			Mean Temperature of Warmest Quarter
18	BIO11			Mean Temperature of Coldest Quarter
19	BIO12			Annual Precipitation
20	BIO13			Precipitation of Wettest Month
21	BIO14			Precipitation of Driest Month
22	BIO15			Precipitation Seasonality
23	BIO16			Precipitation of Wettest Quarter
24	BIO17			Precipitation of Driest Quarter

25	BIO18			Precipitation of Warmest Quarter
26	BIO19			Precipitation of Coldest Quarter
27	USDA_Suborder			USDA 2014 Suborder Classes
28	WRB_Subgroup			WRB 2006 Subgroup Classes
29	Coarse_Fragments_v_0cm			Coarse Fragments Volumetric
30	Coarse_Fragments_v_30cm			
31	Coarse_Fragments_v_100cm			
32	Clay_Content_0cm			Clay Content
33	Clay_Content_30cm			
34	Clay_Content_100cm			
35	Silt_Content_0cm	(Hengl et al. 2017)	Soil Texture	Silt Content
36	Silt_Content_30cm			
37	Silt_Content_100cm			
38	Texture_USDA_0cm			Texture Classes
39	Texture_USDA_30cm			
40	Texture_USDA_100cm			
41	Sand_Content_0cm			Sand Content
42	Sand_Content_30cm			
43	Sand_Content_100cm			
44	Bulk_Density_0cm			Bulk Density
45	Bulk_Density_30cm			
46	Bulk_Density_100cm			
47	SWC_v_Wilting_Point_0cm			Soil Water Capacity
48	SWC_v_Wilting_Point_30cm			
49	SWC_v_Wilting_Point_100cm			
50	pH_Water_0cm	(Hengl et al. 2017)	Soil Chemical Properties	Soil pH in H2O
51	pH_Water_30cm			
52	pH_Water_100cm			
53	CEC_0cm			Cation Exchange Capacity
54	CEC_30cm			
55	CEC_100cm			
56	Garde_Acid			Grade of a Sub-Soil Being Acid
57	ESA_Land_Cover	ESA. Land Cover CCI Product User Guide Version 2. Tech. Rep. (2017)	Vegetation	ESA Land Cover
58	cesm2_npp	CLM5 simulation		NPP
59	cesm2_npp_std			Standard deviation of NPP
60	cesm2_vegc			Vegetation Carbon Stock



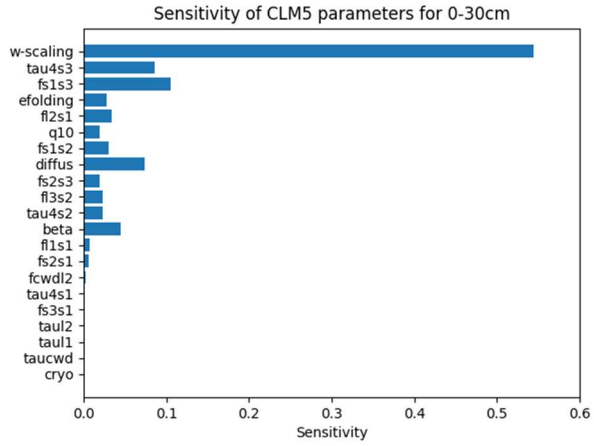
Supplementary Figure 1: Model structures of CLM5 [5].

Supplementary Table 2: 21 biogeochemical parameters in CLM5

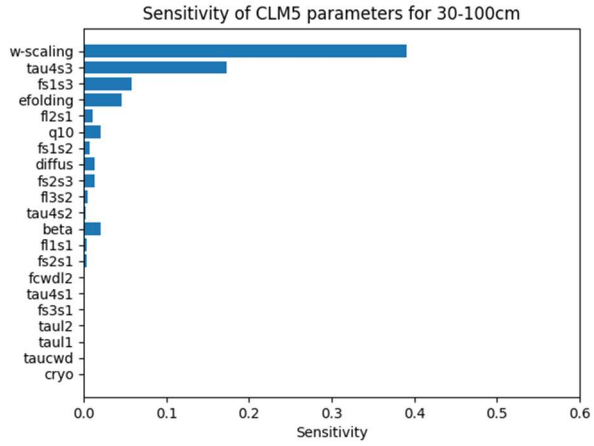
No.	Name	Matrix Term	Corresponding Mechanism	Description	Unit	Prior Range
1	fl1s1	A	Microbial carbon use efficiency (CUE)	Transfer fraction, from metabolic litter to fast SOC	unitless	[0.1, 0.8]
2	fl2s1			Transfer fraction, from cellulose litter to fast SOC	unitless	[0.2, 0.8]
3	fl3s2			Transfer fraction,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

Supplementary Table 3: 8 environmental forcings for CLM5

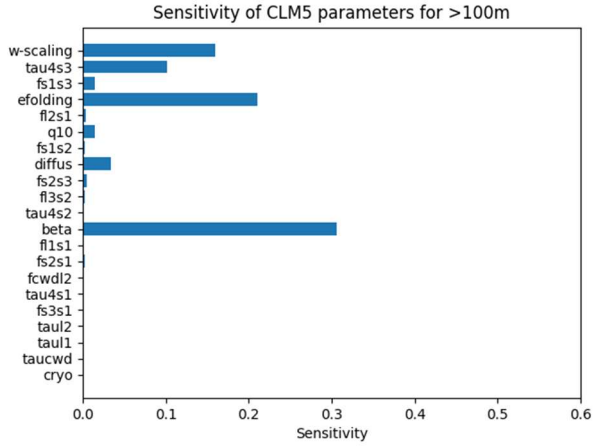
Variable Names	Description	Resolution
nbedrock	Soil layer number that reaches the bedrock	0.5 degree, average monthly values from the monthly record of 20-year simulation after the system reaches the steady state
ALTMAX	Maximum active layer depth of current year	
ALTMAX_LASTYEAR	Maximum active layer depth of last year	
CELLSAND	Sand content	
NPP	Net primary productivity	
SOILPSI	Soil water potential	
TSOI	Soil temperature	
O_SCALAR	Oxygen scalar for decomposition	
FPI_vr	Nitrogen scalar for decomposition	



(a)

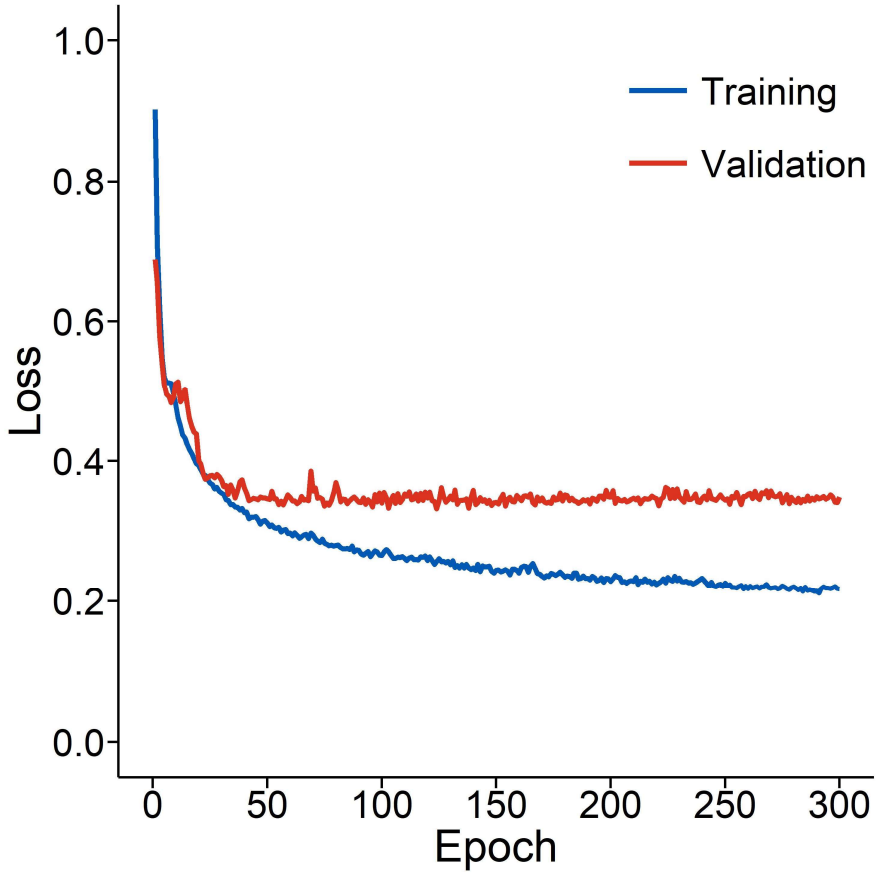


(b)



(c)

Supplementary Figure 2: Sensitivity indices for CLM5 biogeochemical parameters across: (a) 0-30cm, (b) 30-100 cm, (c) >100 cm.



Supplementary Figure 3: Training and validation modelling inefficiency history for one cross-validation fold with median NSE values.

Supplementary Table 4: Comparison between different data assimilation methods

Criteria	**BINN**	**PRODA**	**MCMC**	**Kalman Filter**	**Genetic Algorithm**
Optimization Speed	Fast	Slow	Slow	Fast	Slow
Optimization Target	Parameters	Parameters	Parameters	States	Parameters
Recognition of Spatial/Temporal Heterogeneity	Yes	Yes	No	No	No
Uncertainty Assessment	No	No	Yes	Yes	No
Multisource Data	Yes	Yes	Yes	Yes	Yes
Key Refs	This study	Feng Tao 2020 Frontiers	Oleksandra Hararuk 2014 JGR	MATHEW WILLIAMS 2005 GCB	Damian J. Barrett 2002 AGU ESS

Reference for Supplementary

1. Klemmer, K., Safir, N. & Neill, D. B. Positional Encoder Graph Neural Networks for Geographic Data. Preprint at https://doi.org/10.48550/arXiv.2111.10144 (2023).
2. Tao, F. *et al.* Microbial carbon use efficiency promotes global soil carbon storage. *Nature* **618**, 981–985 (2023).
3. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* 249–256 (JMLR Workshop and Conference Proceedings, 2010).
4. Bjorck, N., Gomes, C. P., Selman, B. & Weinberger, K. Q. Understanding Batch Normalization. in *Advances in Neural Information Processing Systems* vol. 31 (Curran Associates, Inc., 2018).
5. Tao, F. *et al.* Convergence in simulating global soil organic carbon by structurally different models after data assimilation. *Global Change Biology* **30**, e17297 (2024).