Biogeochemistry-Informed Neural Network (BINN v1.0) for Improving Accuracy of Model Prediction and Scientific Understanding of Soil Organic Carbon Storage

Haodi Xu[1, †], Joshua Fan[2, †], Feng Tao[3,4 †], Lifen Jiang[1], Fengqi ~~You~~[4]You[5], 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~~[4]Currently at: Department of Informatics and Intelligent Systems, Institute of Energy and the Environment, The Pennsylvania State University

[5]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~~feng.tao@psu.edu), Carla Gomes (gomes@cs.cornell.edu)

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

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) regulating the soil carbon cycle from 25,925 observed SOC profiles across the ~~conterminous~~contiguous US and ~~compared~~compare 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~~. The good agreement between the spatial patterns of the retrieved processes using the two approaches with an average correlation coefficient of 0.~~81~~77 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~~ to understand large scale soil carbon cycle. Applying BINN to other domains of earth system science will facilitate new scientific discoveries while improving ~~interpretability and accuracy~~the predictive performance of ~~Earth~~earth system ~~models.~~model simulations.

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

Plain Language Summary

~~The rapid development of~~ Conventional artificial intelligence (AI) ~~can uncover hidden~~ is great at predicting biogeochemical patterns (such as the amount of organic carbon in ~~biogeochemistry from big~~ soils) across space and time, when plenty of data~~, yet translating~~ are available. However, it is difficult to use these AI tools to understand the underlying mechanisms that drive these patterns ~~into scientific insights of underlying~~ , because many of these processes ~~is challenging~~.cannot be directly measured. In this study, we introduce a new framework called ~~the~~ Biogeochemistry-Informed Neural Network (BINN~~). BINN blends~~) that integrates a ~~traditional~~ process-based model ~~— built on established scientific theories — with the power of AI, enabling us to learn from large~~ into a multilayer neural network. By comparing the process-based simulations against big observational datasets, BINN optimizes process-based model simulations while ~~maintaining transparency in the underlying science of the model. Testing BINN on soil organic carbon (SOC) data, we found it can~~ accurately ~~identify~~identifying critical processes governing SOC storage~~, while working~~. Critically, BINN is more than 50 times faster than ~~previous methods.~~existing approaches that use Bayesian inference to align scientific models with observational data. This advance in ~~speed and~~ understanding soil carbon cycling and computational speed could help scientists better predict ~~how soils behave~~the soil carbon cycle under different environments. ~~We therefore believe~~ BINN can ~~partner with~~also be extended to use in incorporating other process-based models ~~to benefit~~, helping researchers ~~exploring many different aspects~~explore various issues of Earth's biogeochemical cycles.

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

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

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

~~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.~~

Artificial intelligence (AI) has revolutionized our ability to leverage big data to uncover relationships in complex systems such as the Earth system[1]. 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 (SOC) concentrations[2,3], predicting aboveground carbon accumulation rates in naturally regenerating forests[4], and estimating soil respiration[5]. 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[6].

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

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[7]. By introducing physical or knowledge-based constraints and reasoning, AI-based predictions can be further constrained by scientific knowledge in addition to being driven by observational data. More importantly, unlike the uninterpretable weights and biases in a conventional neural network (NN), 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. For example, we can use physical knowledge to regularize the model towards physically-consistent predictions[8], such as by adding loss terms that penalize violations of a known governing equation in physics-informed neural networks (PINN)[9], or discrepancies between neural network and process-based model outputs[10]. However, these methods do not guarantee that model simulations will strictly obey physical constraints; they only softly discourage violations of physical laws. Another body of work thus tries to embed physical knowledge inside the NN model itself, such as modifying an Long Short-Term Memory (LSTM) network to guarantee mass conservation at each step[11], or embedding hydrological models[12,13], terrestrial models[14], thermodynamic rules[15], domain-specific objectives[16], or Bragg's law for x-ray diffraction[17] into the NN model architecture. Despite these advances, previous efforts mostly used limited constraints in a system (e.g., a few empirical relationships in photosynthesis[18]) or simplified models that comprise latent variables and conservation principles[12]. It remains challenging to integrate numerous differential equations with free

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

parameters into a neural network to study the dynamics of a complex system, such as SOC dynamics.

Meanwhile, a traditional method that uses scientific knowledge to guide inference of biogeochemical processes from observational data is data assimilation. This method has played a central role in advancing our understanding of many processes, such as terrestrial carbon flux partitioning[8], nutrient cycling[19], ecosystem respiration[1], plant phenology[20], and hydrology–biogeochemistry couplings[21]. However, the data assimilation method usually estimates a set of fixed biogeochemical parameters without spatial or temporal variations, which has been argued not represent changes in ecosystem properties over space and time[22]. While statistical methods such as Geographically-Weighted Regression allow spatially-varying regression coefficients, they are not flexible enough to model arbitrary nonlinear relationships between environmental covariates and biogeochemical parameters[23]. A recently-developed approach, PROcess-guided deep learning and DAta driven modeling (PRODA), first uses Bayesian inference-based data assimilation methods to find the optimal biogeochemical parameters for each site, and then trains a deep learning model to infer biogeochemical parameters at new sites, improving our understanding of the global soil carbon cycle[24–27]. This approach harnesses the strengths of both process-based modeling and deep learning methods to improve SOC simulations by allowing biogeochemical parameters to vary across space. However, the Bayesian inference-based, site-level data assimilation used in the PRODA approach requires vast computational resources, making the method inefficient in both time and energy and thus difficult to apply broadly.

In this study, we integrate a matrix form of a process-based model (i.e., Community Land Model version 5, CLM5) that describes SOC dynamics into a neural network, thus developing a Biogeochemistry-Informed Neural Network (BINN). BINN is a novel framework that combines

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

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~~. Lastly, we evaluate model components from real-world SOC observations ~~(Figure 1b). Our~~ . The BINN-predicted biogeochemical parameters accurately simulate ~~real~~ SOC observations~~,~~ and are similar to those produced by PRODA, while being much faster to ~~estimate computationally.~~compute. By combining data-driven learning with reasoning about existing ~~geoscientific~~biogeochemical knowledge, BINN can accurately and quickly infer underlying physical processes from ~~only~~ SOC observations.

2. Methods

2.1 Biogeochemistry-Informed Neural Network (BINN)

BINN ~~incorporates~~integrates a process-based model (CLM5, see Section 2.1.2 below) into a ~~neural network~~ differentiable, end-to-end framework (Fig. 1a~~). First, a neural network learns the relationships between~~) that maps environmental covariates ~~and~~ to site-level biogeochemical parameters, ~~which quantify~~feeds those parameters into CLM5 for the forward simulation, and back-propagates the ~~strength of important processes in~~ loss through the ~~soil carbon cycle. We pass these parameters to a process-based model to simulate SOC storage, and compare~~whole framework during training (Fig. 1b).

We implemented BINN in Python using PyTorch and executed the experiments on the supercomputer, Derecho, maintained by National Center for Atmospheric Research (NCAR). The experiments were conducted using one compute node with ~~field observations.~~ 128 CPU

Formatted: Indent: First line: 0.5"

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

cores, leveraging PyTorch Distributed Data Parallel (DDP) for multi-CPU training. Meanwhile, BINN was also tested with 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.

Formatted: Font: Times New Roman, 12 pt

2.1.1 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~~A fully-connected neural network learns the relationship between environmental covariates (Table S1) and biogeochemical parameters (Table S2) over space. At each site, we have 60 environmental covariates (input features). While the numeric features can be directly passed into the neural network, we use embedding layers to encode categorical (discrete) covariates and spatial coordinates. For each categorical covariate, the embedding layer learns a vector embedding for each category (e.g. climate type). For spatial coordinates, we pass the latitude and longitude through a spatial positional encoder to obtain a location embedding vector, characterizing unobserved aspects of each location that are not captured in our covariates[28].

Formatted: Indent: First line: 0.5"

We then concatenate the categorical covariate embeddings, location embedding, and the remaining covariates into a vector $\boldsymbol{e}$, and pass this through a 4-layer neural network f_{NN}

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

(Equation 1) with learnable weights/biases $\boldsymbol{w}$, which outputs a vector ~~$\boldsymbol{z}$~~$\tilde{\boldsymbol{p}}$ with 21 values (~~one~~representing *unconstrained* predictions for each parameter in the process-based model):

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

Formatted Table

~~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:~~

Each layer of the neural network comprises prescribed numbers of neurons that receive information either from the environmental covariates (for the first layer) or outputs of the previous layer. It then computes linear transformations of the input:

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

Formatted Table

$$y = \sum w_i x_i + b$$

Formatted: Font color: Black

~~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~~where y is an output from a neuron after the linear combination of its input, w_i is a learnable weight for x_i, x_i is the input from either the environmental covariates or the previous layer's outputs, and b is a learnable neuron-specific bias. After the linear transformation, a nonlinear

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

activation function is applied to generate the eventual results at each neuron, so that the neural network can generate complex nonlinear relationships between the inputs (i.e., environmental covariates) and the outputs (i.e., the biogeochemical parameters in the process-based model). Note that the linear transformations and activation functions are both differentiable, so we can backpropagate gradients through the network. This allows us to calculate the gradient of the loss function with respect to all learnable weights w_i and biases b.

In our study, for the first three layers of the neural network, we assign each of them to have 128 neurons to process information from the previous layer and use *LeakyReLU* as the activation function:

$$LeakyReLU(y) = max(0, y) + negative\_slope * min(0, y) \quad (3)$$

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* is chosen over traditional *ReLU* because it allows gradients to flow through the network even when the inputs to the activation are negative.

The final layer has 21 output neurons, with one corresponding to each biogeochemical parameter in the process-based model. We do not use *LeakyReLU* after the final linear transformation because it would bias the distribution of the predicted parameters. However, as we have prior knowledge about the plausible range of values for each biogeochemical parameter, the initial parameter predictions $\tilde{\boldsymbol{p}}$ are first normalized to be in the range [0, 1] by element-wise

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

sigmoid functions σ (Equation 4) and then further linearly scaled into the predicted parameters p (Equation 5). This ensures that each parameter p_i stays within their prior ranges:

$$p_{norm,i} = \sigma(\tilde{p}_i, \gamma) = \frac{1}{1 + exp\left(-\frac{\tilde{p}_i}{\gamma}\right)} \quad (4)$$

$$p_i = f_{scale}\left(p_{norm,i}, \theta_{i,max}, \theta_{i,min}\right) = p_{norm,i} * (\theta_{i,max} - \theta_{i,min}) + \theta_{i,min} \quad (5)$$

where $\tilde{p}_i$ is the i-th output of the neural network, γ is a learnable parameter, and $\theta_{i,max}$ and $\theta_{i,min}$ are the plausible upper and lower limits for each biogeochemical parameter p_i. Thus, the final output of the neural network is p, a vector of 21 biogeochemical parameters for each location, with each parameter p_i being constrained to be in its prior range $[\theta_{i,min}, \theta_{i,max}]$. $\theta_{i,min}$ and $\theta_{i,max}$ are values taken from previous literature to indicate plausible limits for processes quantified by each biogeochemical parameter p_i[26]. Note that we introduce a γ value in Equation 4 to control how fast the results after activation function can converge to 0 or 1 and to $\theta_{i,max}$ or $\theta_{i,min}$ after scaling. When the γ value is small, the predicted parameters may quickly get stuck at $\theta_{i,min}$ and $\theta_{i,max}$. If 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 2.1.4)[29]. Thus, we tuned γ to be a relatively large value to facilitate neural network optimization.

We conducted a hyperparameter search to determine suitable settings for BINN (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). Based on a grid search for these hyperparameters, we chose to train BINN for 300 epochs with a batch size of 32, using PyTorch DDP to distribute training across 128 CPUs, as this performed best on our validation dataset. The

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

optimized BINN model was recorded each time when the validation loss improved over the previous best model. We used the *AdamW* optimizer with a learning rate of 0.01. The embedding size was set as 64, and the dropout rate was set as 0.3. The network also used batch normalization after each *LeakyReLU* activation function to normalize the layers' outputs by re-centering and re-scaling, making training faster and more stable[30]. We initialized γ to 64.5 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:

$$\gamma = 10 + 109 \cdot \frac{1}{1 + exp(-\gamma')} \quad (6)$$

where γ' is a learnable parameter that is initialized to 0. We set *negative_slope* to -0.3 in the *LeakyReLU* by default.

2.1.2 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.~~In this study, we use a matrix form of the soil carbon module of CLM5 to

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

represent our knowledge of SOC dynamics[31,32] (Figure S2). CLM5 has been continuously developed and refined over the past decade for simulating biogeochemical cycles, including SOC dynamics[33]; 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 into fast, slow, and passive pools according to their turnover times. 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:~~

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[32,34] as:

$$\frac{dX(t)}{dt} = B(t)I(t) - A\xi(t)KX(t) - V(t)X(t) \quad (~~3~~ 7)$$

~~where~~where $dX(t)/dt$ is carbon pool change at time t; $I(t)$ 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

Formatted: Font: Not Italic
Formatted Table
Formatted: Font: Not Italic
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Formatted: Font: Times New Roman, 12 pt
Formatted: Font: Times New Roman, 12 pt

the same for each pool across 20 layers; ~~$\xi(t)$~~$\xi(t)$ (140×140) captures how the environment modifies the intrinsic decomposition rate in the ~~K~~K matrix by temperature (~~ξ_T~~,ξ_T), water (~~ξ_W~~,ξ_W), oxygen (~~ξ_O~~,ξ_O), and depth (~~ξ_D~~)ξ_D) scalars~~, V(t)~~; $V(t)$ (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)~~$V(t)X(t)$ indicates vertical SOC movements along the soil profile. The ~~t~~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~~ In this study, we assume steady-state SOC storage for computational efficiency, which is justified by previous research showing that recent disequilibrium effects from climate change

Formatted: Font: 12 pt
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Formatted: Font: 12 pt
Formatted: Font: Times New Roman, 12 pt
Formatted: Font: Times New Roman, 12 pt

and human activities are relatively minor compared to the SOC storage that has developed over thousands of years[26,35]. 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:

$$\hat{X}(t) = \left(A\xi(t)K + V(t)\right)^{-1} B(t)I(t) \quad (8)$$

The matrix representation of CLM5 is implemented in PyTorch utilizing vectorized functions to replace most of 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 vectorized functions. 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 $\boldsymbol{w}$ of the neural network, differentiating through all the operations in the process-based model (Section 2.1.4).

Equation (7) contains 21 biogeochemical parameters (Table S2) that represent properties of different processes (e.g., transformation and stabilization of SOC, temperature sensitivity of soil respiration, and substrate quality) in the soil carbon cycle [22]. Because those processes are highly variable depending on different climate conditions or soil properties, the values

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

quantifying their properties (i.e., the parameter values) should differ with locations with different environments[22]. Thus, in this study, the neural network embedded in BINN (Section 2.1.1) 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 (8) to estimate steady-state SOC storage at sites across the Contiguous US.

2.1.3 Loss Function

Because we ~~are interested in~~aim to accurately ~~simulating~~simulate 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~~.~~:

$$Smooth\ L1\ Loss\left(\hat{y}_{profile}, y_{profile}\right) = \begin{cases} \frac{0.5\left(\hat{y}_{profile} - y_{profile}\right)^2}{\beta} & if\ \mid \hat{y}_{profile} - y_{profile} \mid < \beta \\ \mid \hat{y}_{profile} - y_{profile} \mid - 0.5 \times \beta & o.w. \end{cases} \quad (9)$$

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]~~.~~such as Mean Squared Error (MSE)[36].

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.

Formatted Table
Formatted: Font: Not Italic
Formatted: Font: Not Italic
Formatted: Font: Times New Roman, 12 pt
Formatted: Font: Times New Roman, 12 pt

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, L_{batch}, 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~~τ to ~~1~~15 and loss weight of a 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~~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.

2.1.4 Backpropagation to Optimize Neural Network Parameters

Formatted Table
Formatted: Font: Not Italic
Formatted: Font: Not Italic
Formatted: Font color: Auto
Formatted: Font: Times New Roman, 12 pt
Formatted: Font: Times New Roman, 12 pt

During training, BINN computes the loss function based on the current predicted SOC and biogeochemical 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. ~~w~~ w in Equation 1 and ~~γ~~ 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.

Formatted: Font color: Black

Formatted: Font: Not Bold

~~3~~2.1.5 Monte Carlo dropout

We used Monte Carlo (MC) dropout during BINN training to approximate the model's predictive uncertainty[37]. MC dropout randomly disables a fraction of neurons (with probability *dropout_p*) within the neural network during training. This process acts like training an implicit ensemble of subnetworks and improves generalization and robustness of BINN.

When *dropout_p* is set larger than 0, the posterior distributions of the predicted biogeochemical parameters can be estimated by MC dropout at each site. To be specific, during forward simulation with the best-trained BINN, we kept dropout active and perform 100

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

stochastic forward passes with different dropout masks, resulting in an efficient approximation of the uncertainty in the model's parameter predictions.

2.2 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 S) [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[+]), 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.~~

We evaluated BINN's ability to recover the biogeochemical parameters of CLM5 from synthetic SOC data using a 10-fold cross-validation experiment[38]. Unlike real-world observations that contain measurement uncertainties and potentially unresolved processes, the synthetic SOC data across multiple depths used in this study were generated by running CLM5 with prescribed spatially-varying parameter values (obtained from a previous work with PRODA approach[24]; see Section 2.3.2), providing a controlled environment where true parameter values are known (Figure S2). This synthetic dataset allowed us to quantitatively assess BINN's parameter recovery accuracy by comparing predicted parameters with the known values used to generate the synthetic data. Parameter recovery was assessed under two configurations: one using a subset of the most sensitive parameters as targets, and another using all biogeochemical parameters as targets.

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

2.2.1 Sensitivity Analysis

To decide which biogeochemical parameters to modify in ~~this~~the experiment~~, we conduct~~ of recovering the biogeochemical parameters of CLM5 from synthetic SOC, we conducted a sensitivity analysis of CLM5 to identify the biogeochemical parameters that have the greatest influence on simulated SOC values ~~(Appendix 4). We~~ .

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 ~~the four~~512 sites across the Contiguous 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 their prior ranges. 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 p_i (i = [0, 20]) at each site. We randomly selected a value (p_i^*) for each biogeochemical parameter p_i from a uniform distribution within its prior range. For the

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

remaining biogeochemical parameters ($p_j: j \neq i$), we produced 1000 random settings from uniform distributions within their respective prior ranges. Using the sample of 1000 biogeochemical parameter sets, we estimated the conditional expectation $E(SOC|p_i = p_i^*)$. We repeated this sampling process for 100 randomly selected values of p_i and used the results to estimate the variance $V(E(SOC|p_i))$. This value quantifies the variance in the output variable *SOC* as a result of modifying the biogeochemical parameter p_i. We discarded the simulations when *NaN* values appeared due to randomly sampled biogeochemical parameter sets. Finally, we repeated this procedure for each biogeochemical parameter p_i, and a sensitivity index S_i was calculated for each biogeochemical parameter at each site as:

$$S_i = \frac{V(E(SOC|P_i))}{V(SOC)} \quad (11)$$

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 2) and individual depth ranges (Figure S3). Error bars represent the mean sensitivity ± one standard deviation across sites.

Our analysis identified the five most sensitive parameters ~~from this sensitivity test:~~; the parameter "beta" governs the allocation of carbon inputs across soil depths; 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~~

Formatted: Indent: First line: 0"

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

~~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~~on SOC decomposition.

2.2.2 Simulation Experiment to Recover Biogeochemical Parameters from Synthetic Data

We selected the five most sensitive parameters identified from the sensitivity test to minimize equifinality, which is a phenomenon that a similar simulation output is generated with different combinations of biogeochemical parameter values[39], for the parameter recovery experiment.

To test the recovery efficiency of the biogeochemical parameters with BINN, we modified the final ~~layers~~layer of the neural network by reducing the number of neurons from 21 to ~~4~~5 to predict these ~~4~~five biogeochemical parameters. Combining the ~~4~~five biogeochemical parameters predicted by BINN and the remaining ~~17~~16 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~~five parameters predicted by BINN with ~~the~~their prescribed parameter values in the testing dataset to evaluate the accuracy of BINN in retrieving the prescribed biogeochemical parameters. To examine the recovery efficiency beyond the top five parameters, we repeated the experiment with all 21 parameters following the same processes.

~~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~~

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

~~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.~~

2.2.3 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, 8 subsets were used for training, 1 subset for validation, 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 Nash–Sutcliffe modelling efficiency coefficient (NSE) and Pearson correlation coefficient (r), were calculated for each iteration. Final performance evaluations were conducted by averaging the metrics across all iterations, and grid-level predictions were averaged across the ten iterations. This cross-validation approach provides a robust assessment of BINN's 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.

2.3 BINN performance with real-world SOC Observations

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

We ~~then~~ evaluated the performance of BINN by comparing BINN's SOC predictions with observed SOC across the ~~Conterminous United States~~contiguous US (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~~**2.3.1 Data Preparation**

We processed SOC observations from the World Soil Information Service (WoSIS) following Tao et al. (2020, 2023)[24,26]. Each profile (i.e., each 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 contiguous US. We used 60 environmental covariates at each site from Tao et al. (2023)[26] 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, except the categorical covariates.

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

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

decomposition, the soil nitrogen scalar for decomposition, soil temperature, and soil water potential. These forcings ~~are~~were derived 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.

2.3.2 PRODA

We used PRODA as a baseline to compare SOC simulation skill and derived process components with BINN. PRODA couples site-level Bayesian calibration of CLM5 with a neural network that maps environmental covariates to the calibrated parameters.

The site-level data assimilation was conducted by an adaptive Metropolis Markov Chain Monte Carlo (MCMC) algorithm with 20,000 iterations for the test run and 50,000 iterations for the formal run. The posterior distribution for each parameter was generated during the second half of the formal run. The mean values of the posterior distributions of the parameters were calculated as the final point estimates.

To generalize these site-level estimates to the contiguous US, we trained a fully-connected multilayer neural network using the same 60 environmental variables as predictors. The neural network used in the final training consisted of four hidden layers. The node numbers for each hidden layer were 256, 512, 512 and 256, respectively, with *ReLU* as the activation function and *adadelta* as the optimizer. PRODA was trained with 6000 epochs and validated with a 10-fold cross-validation. When comparing with BINN, the test dataset for each fold was set the same as the corresponding fold during BINN training.

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

2.3.3 Summary Statistics

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

$$NSE = 1 - \frac{\sum(obs_i - simu_i)^2}{\sum(obs_i - \overline{obs_i})^2} \quad (12)$$

where obs_i is one of the SOC observations, $\overline{obs_i}$ is the mean of the SOC observations, and $simu_i$ is the simulated SOC by CLM5 embedded in BINN.

We used the Pearson correlation coefficient, r, between the predicted and prescribed biogeochemical parameters to evaluate the effectiveness of BINN in recovering each of the five biogeochemical parameters as follows:

$$r = \frac{\sum[(para_{BINN} - \overline{para_{BINN}}) \times (para_{true} - \overline{para_{true}})]}{\sqrt{\sum(para_{BINN} - \overline{para_{BINN}})^2 \times \sum(para_{true} - \overline{para_{true}})^2}} \quad (13)$$

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.

2.3.4 Normalized Difference

For each site, we averaged the simulated SOC across the 10 cross-validation folds and subtracted the observed SOC to obtain a mean difference. To improve interpretability over space, we applied a signed percentile normalization: positive differences were scaled to [0,1] by dividing by the 99th percentile of positive deviations, and negative differences were scaled to [−1,0] by dividing by the 99th percentile of negative deviations. This approach preserves the sign and

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

relative magnitude of biases while preventing a small number of extreme values from dominating the colored scale.

2.3.5 Traceability Analysis

To further demonstrate how BINN helps understand processes governing SOC dynamics, we conducted a traceability analysis[40]. The traceability analysis uses the process-based model to separate BINN-predicted SOC storage into carbon influx and ecosystem residence time, with the latter calculated by dividing carbon storage by carbon influx. The ecosystem residence time was jointly determined by baseline residence time and the environmental modifier. These decomposed components at each grid were then averaged across different biomes and visualized in scatter plots for comparison among the biomes. Biome types were assigned to each grid based on the dominant ecological region within that grid, using the Level 1 Ecological Regions of North America map provided by the US Environmental Protection Agency[41].

2.4 Computational Efficiency

We compared computational efficiency between BINN and PRODA on the same dataset (2,000 soil profiles) and the same personal computer, reporting their wall-clock time and peak resident set size (RSS). For BINN, we evaluated two implementations and recorded their respective runtime and peak RSS: (i) a matrix version using the original CLM5 formulation and (ii) a vectorized version that removes explicit for-loops. Each implementation was trained for 300 epochs (no early stopping) to standardize timing and ensure convergence on the training set. For PRODA, the computational bottleneck is the site-level MCMC calibration of CLM5 parameters. We measured MCMC runtime at 10 randomly selected sites (with 20,000 test iterations and 50,000 formal iterations per site) and extrapolated to 2,000 sites. Because per-site memory usage

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

is stable across sites, peak RSS was taken from these 10-site runs. PRODA's second stage using a neural network that maps environmental covariates to the MCMC-optimized parameters was trained for 6,000 epochs. We recorded its peak RSS and reported PRODA's overall peak RSS as the maximum of the MCMC stage or the NN stage.

3. Results

3.1 Parameter Recovery Accuracy

When BINN predicted the five most sensitive biogeochemical parameters, the recovered parameter values exhibited high consistency with the prescribed biogeochemical parameters used during synthetic data generation, achieving an average correlation coefficient of 0.88 across the 10-fold cross-validations (Figure 3g). Moreover, BINN achieved an average NSE of 0.99 on the test dataset when comparing simulated SOC with synthetic SOC (Figure 3g). In the cross-validation iteration with the median NSE of simulated SOC, the BINN-recovered parameter "efolding", representing the depth scalar, had a correlation coefficient of 0.92 in comparison with the prescribed parameter values (Figure 3a). The parameter "tau4s3", representing the baseline turnover time of the passive SOC pool, showed a correlation coefficient of 0.93 (Figure 3b). The "fs1s3" parameter, indicating the transfer fraction from fast to passive SOC pools, achieved a correlation coefficient of 0.82 (Figure 3c). The "w-scaling" parameter, which is the scaling factor of soil water, had a correlation coefficient of 0.9 (Figure 3d). Lastly, the "beta" parameter that governs the vertical distribution of carbon inputs exhibited a correlation coefficient of 0.83 (Figure 3e).

Extending recovery to all 21 parameters revealed heterogeneous identifiability that is consistent with their influence on SOC storage (Fig. S4). ~~2~~Parameters with low sensitivity (e.g.,

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

cryo, taucwd) showed weak correlations between retrieved and prescribed values, whereas high-influence terms (e.g., w-scaling, fs1s3) remained well recovered.

In addition, to better address the equifinality issue in BINN, we implemented MC dropout[37] to obtain the posterior distributions of parameters at each site (Fig. S5). Although point estimates of parameters may not exactly match the prescribed values in parameter recovery experiment, the posterior distributions of parameter values by BINN frequently cover the prescribed values. We further quantified the site-level coverage (i.e., the fraction of prescribed values falling into the BINN-predicted intervals) in Fig. S6, which indicated the issue of equifinality can be further controlled to a reasonable extent in BINN.

3.2 Performance in Predicting Real-world SOC Observations ~~Analysis~~

After BINN optimization, we used its predicted biogeochemical parameters to calculate six model components ~~that indicate~~representing different ~~properties~~processes in soil carbon cycle over the contiguous US ~~continent~~: carbon transfer efficiency, baseline decomposition, environmental modifier, carbon input allocation, vertical transport rate, and plant carbon inputs. The comparisons between BINN-retrieved six model components ~~were compared with the results~~and those generated from PRODA ~~(Figure 5). We assessed the~~, along with their spatial patterns ~~of six model components~~, are shown in Figure 4. Note that ~~emerged from both approaches. The~~the plant carbon input component was identical between BINN and PRODA~~,~~ (Figure 4p-r), due to the use of the same NPP forcing data ~~(Figure 5p, 5q, 5r). Spatial~~. The spatial distributions of the other five components were similar between BINN and PRODA, with an average correlation coefficient of 0.~~81~~77.

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

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

total carbon decomposition, exhibited more spatial variation than PRODA's ~~prediction.~~predictions (Figure 4a–b). While both methods ~~indicated~~simulated higher carbon transfer efficiency in the northwestern region and lower efficiency in the ~~middle west (Figure 5a, 5b),~~midwestern region, BINN predicted slightly higher values in the northeastern and southeastern parts of the ~~Conterminous United States~~contiguous US 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~~reached 0.77 (Figure 4c). Baseline decomposition, which describes the substrate decomposability of each soil pool, showed similar spatial patterns across the contiguous US for both BINN and PRODA (Figure 4d–e), although the average baseline decomposition values from BINN were somewhat higher than those from PRODA (Figure 4f).

~~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~~78 with PRODA's results (Figure ~~5i~~4i). Their spatial patterns were nearly identical across the ~~Conterminous United States~~contiguous US, with lower values in the northwestern ~~part~~region and values gradually increasing ~~to the highest values in~~toward the southeastern ~~part~~region (Figure ~~5g, 5h).~~

~~The carbon~~4g–h). Carbon input allocation predicted by both methods also displayed similar spatial patterns ~~across the Conterminous United States~~ with a correlation coefficient of 0.~~74~~73 (Figure ~~5l~~4l). Both ~~the~~ methods predicted low carbon input allocation rates in the eastern

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

and western US but high rates in the ~~mid~~central US (Figure ~~5j, 5k), though~~4j–k), although BINN predicted higher rates of carbon input allocation in the ~~mid~~central US than ~~the PRODA's predictions.~~

~~———~~PRODA. BINN and PRODA also predicted ~~the~~ vertical transport rate with nearly identical spatial distributions ~~across the Conterminous United States~~ (Figure ~~5m, 5n~~4m–n), with a high correlation coefficient of 0.91 (Figure ~~5o~~4o).

~~——— BINN demonstrated better accuracy than PRODA in predicting SOC across the Conterminous US as well (Figure 6)[+]. 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.~~ When applying the predicted parameters to simulate SOC, BINN showed an average NSE of 0.71 in predicting the observed SOC across the 10-fold cross validations (Figure 5c), with an average NSE of about 0.57 on the test datasets (Figure 5d). The training and validation NSE values recorded throughout model training at each epoch indicated a validation NSE of 0.63 (Figure S6). The map of the normalized difference between BINN predicted and observed SOC shows that BINN captured SOC dynamics in most parts of the contiguous US, except in the north-central and western regions (Figure 5a), which may be due to the large uncertainty across the 10-fold cross validations (Figure 5b).

Overall, BINN demonstrated accuracy comparable to PRODA in predicting SOC across the contiguous US (Figure 6a). However, PRODA showed a tendency toward negative bias

Formatted: Indent: First line: 0.5"

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

(underestimation) over broad areas, whereas BINN's normalized difference was closer to zero (Fig. 6b). In the north-central and eastern US, where PRODA mainly underestimated observed SOC, BINN exhibited fewer geographical biases (Figure 6c–d).

3.3 Spatial Patterns of the Determinant Processes of SOC

Traceability analysis revealed that ecosystems with similar average soil carbon storage among different biomes, which is illustrated by the proximity of the points to a given contour line, may result from distinct underlying processes (Figure 7a). For instance, North American Deserts and Mediterranean California exhibited similar carbon storage, despite contrasting underlying mechanisms, with North American Deserts had 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 can be attributed to similar baseline carbon residence times and environmental scalars.

3.4 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~~

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

~~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~~),~~) to complete 2,000 profiles, whereas the vectorized version took only 10.5 hours (5× faster). ~~In contrast~~Moreover, MCMC alone took 574.~~69~~7 hours, leading to a total of 577.~~69~~7 hours for PRODA~~.~~ when including ~3 hours of neural network training. Thus, BINN is approximately 57 times faster than PRODA and can be further accelerated via ~~PyTorch's Distributed Data Parallel (~~PyTorch DDP~~),~~ ~~enabling~~, which enables parallel training across multiple CPUs. While both approaches achieve parameter interpretability (Section ~~4~~3.2), 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. In addition, BINN with vectorized CLM5 yields the lowest peak memory usage among all the methods (Figure S7).

Formatted: Indent: First line: 0"

~~**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].~~

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

~~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,~~**4. Discussion**

Our study demonstrates that BINN is a novel approach ~~for retrieving~~to retrieve spatially varying model parameters from big data and ~~predicting~~predict their spatial distributions for determining SOC dynamics over the ~~globe,~~contiguous US, while greatly 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~~

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

~~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~~**4.1. 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 contiguous US. To minimize the effects of equifinality, we conducted a sensitivity analysis and focused on the five most sensitive biogeochemical parameters, which are also those typically well constrained by Bayesian calibration. As expected, BINN recovered these target parameters with high correlations with their respective prescribed values. Moreover, BINN reproduced the synthetic SOC patterns well using only environmental covariates and synthetic SOC, without using a Bayesian method. These results suggest that BINN learns causal relationships between environmental covariates and SOC simulations consistent with the underlying dynamics, rather than simply exploiting spurious correlations[22].

We then extended recovery to all 21 parameters; the results showed that highly sensitive parameters remained well recovered. Although parameters with low influence on SOC showed weak recovery by BINN, this is a common limitation also observed in Bayesian approaches when data are weakly informative.

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

To further characterize and mitigate equifinality, we applied MC dropout at inference to obtain posterior-like parameter distributions at each site. Although point estimates did not necessarily match the prescribed values exactly, the BINN uncertainty intervals frequently covered the ground truth (i.e., the prescribed parameters in this case). Reporting these NN-induced uncertainty ranges (for both parameters and resulting SOC simulations) makes the sources of indeterminacy explicit and helps keep the practical impact of equifinality under control.

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

~~7~~4.2. 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 the spatial distributions of SOC storage and ~~their~~its components from big data. Compared to PRODA, BINN reduces computational time by more than 50-fold in a test with 2,000 profiles~~.~~, while

Formatted: English (United Kingdom)

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

keeping memory usage the lowest. 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~~. Instead, it only perturbs parameters randomly and checks whether the accuracy has been improved or not ~~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].~~

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

~~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~~

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

~~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.~~By contrast, BINN reimplements the process-based model, CLM5, in a differentiable way using PyTorch, and leverages this differentiability to rapidly find the optimal parameters (for all sites simultaneously) that accurately simulate SOC observations (Table S4). Additionally, we used vectorized functions in PyTorch to replace for-loops in the original matrix form of CLM5 model, which further enhances computational efficiency. BINN can also utilize PyTorch DDP to parallelize computations, allowing much faster iterations and thus greatly saving real physical time. Meanwhile, the memory usage of BINN can be kept bounded via smaller batches and gradient checkpointing, whereas site-by-site Bayesian samplers must maintain full state per chain, which tends to increase peak RSS despite longer running time. High computational efficiency is also more environmentally friendly, saving energy when dealing with large datasets.

4.3. BINN's Facilitation of Mechanistic Understanding

A process-based model is an abstraction of a real-world system and represents the processes that govern the system's dynamics, yet such model-based predictions often fit with empirical observations poorly[42]. 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 those well-understood processes in their structure while using parameters to represent

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

unresolved processes[22]. Without taking advantage of the extensive information present in observations, these model parameters are usually not well constrained.

BINN learns site-specific parameters from environmental covariates while the process-based model executes the forward simulation. This end-to-end setup improves model fit to observations and yields spatially varying parameters. 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. After training, we further transform the learned parameters into model components (as in Figure 5) and apply traceability analysis to quantify each component's contribution to SOC storage across space, thereby linking learned parameters to individual processes and, ultimately, to SOC outcomes. The model components from BINN reveal interpretable patterns (e.g., baseline decomposition and carbon transfer efficiency) consistent with prior work using Bayesian-based methods[26]. Traceability analysis results can further quantify the contribution of each model component to variations in SOC storage across ecosystems. Similar logic underlies evaluations of how newly added processes shift system behavior (e.g., nitrogen effects on carbon storage capacity[40]).

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 unknown mechanisms governing biogeochemical cycling.

Formatted: Font: Times New Roman, 12 pt

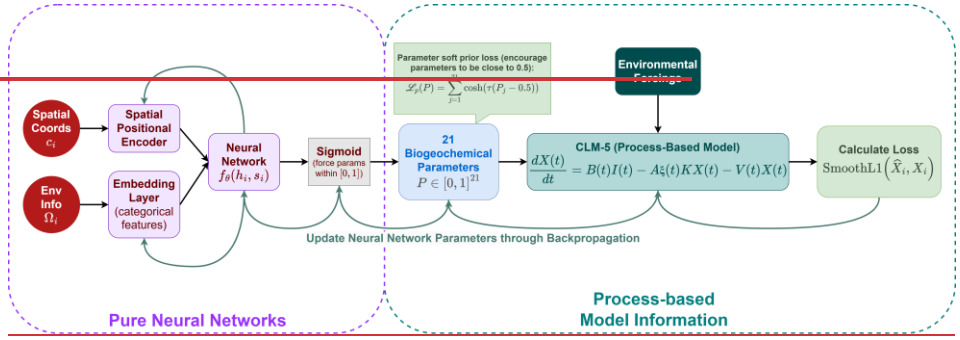
Formatted: Font: Times New Roman, 12 pt

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.

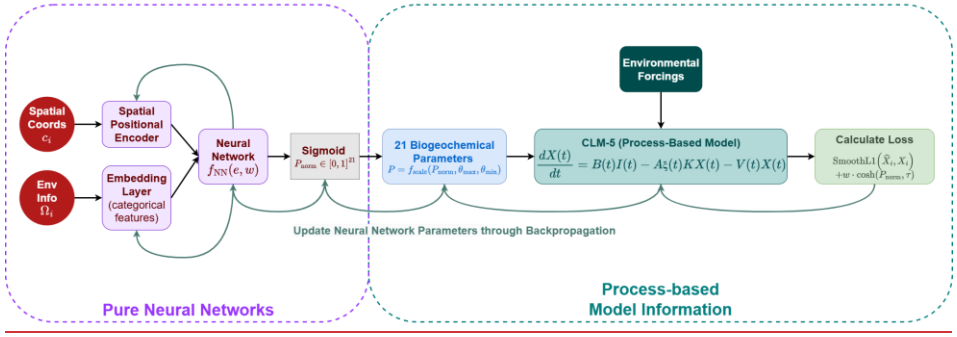
Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

(a)



BINN v1.0 with CLM5



(b)

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

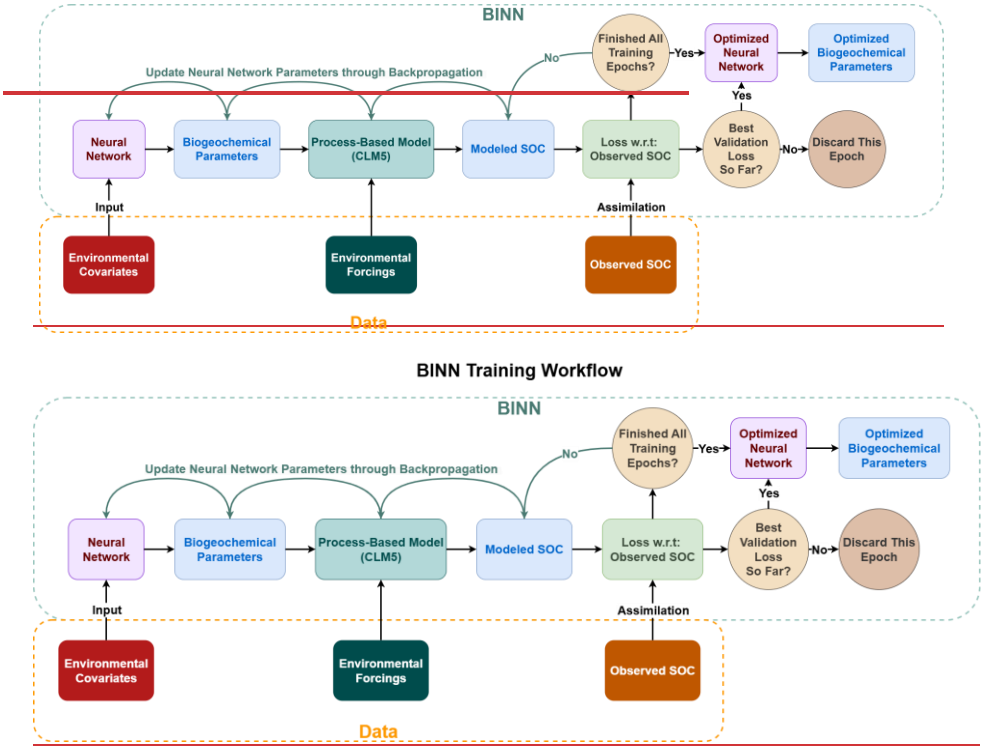


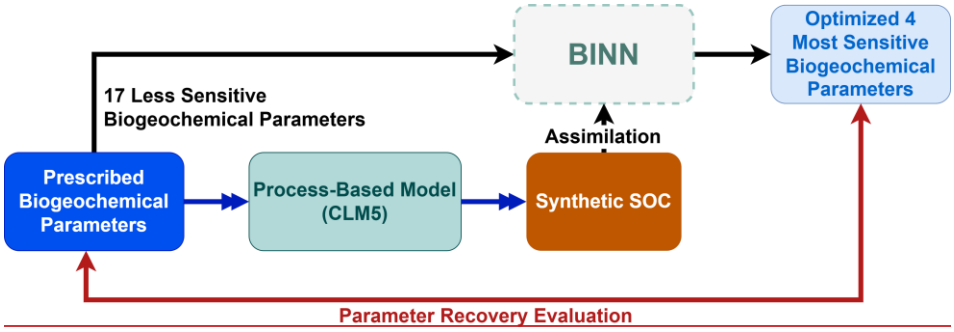
Figure 1. Schematic diagram of BINN architecture and training ~~process~~workflow. (a) Detailed BINN structure showing the integration of neural networks with CLM5. ~~The~~A neural network ~~component~~(purple) processes spatial coordinates (red) through a positional encoder and categorical environmental covariates through an embedding layer. The network outputs are transformed via a sigmoid activation function to generate 21 ~~biogeochemical~~normalized parameters~~.~~ (grey). These parameters~~, constrained by a soft prior loss,~~ are ~~input to~~then scaled by their corresponding prior range (blue) and pass into CLM5 along with environmental forcings to simulate SOC ~~dynamics.~~(teal). The model's performance is evaluated using a smooth L1 loss

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

function- and a soft prior penalty on biogeochemical parameters (green). 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.



Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

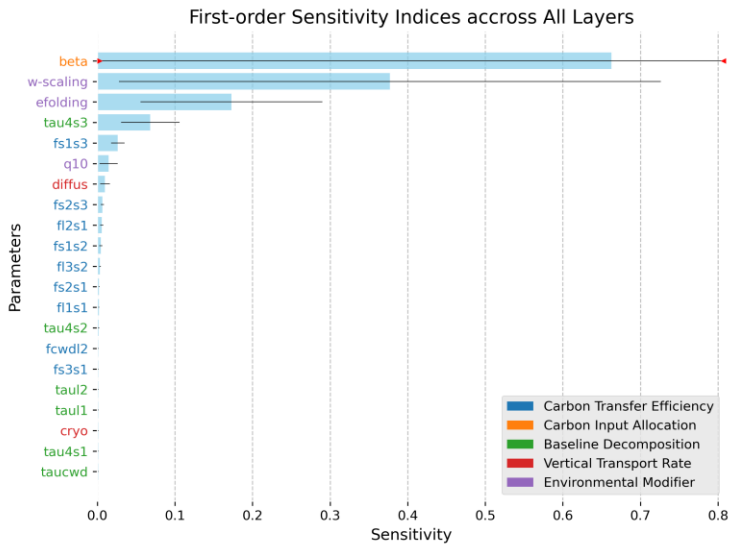
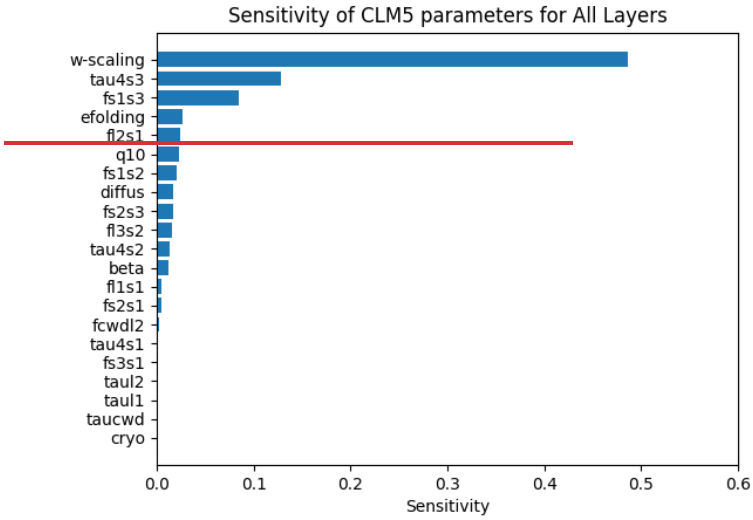


Figure 2: ~~**Schematic of the parameter recovery experiment to evaluate BINN's ability to retrieve the model processes regulating SOC.** The parameter recovery experiment involves three main steps. 1). Blue two-headed arrows: Synthesizing a SOC dataset using CLM5 with prescribed parameter values (21 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.~~

Formatted: Font: Not Bold

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt



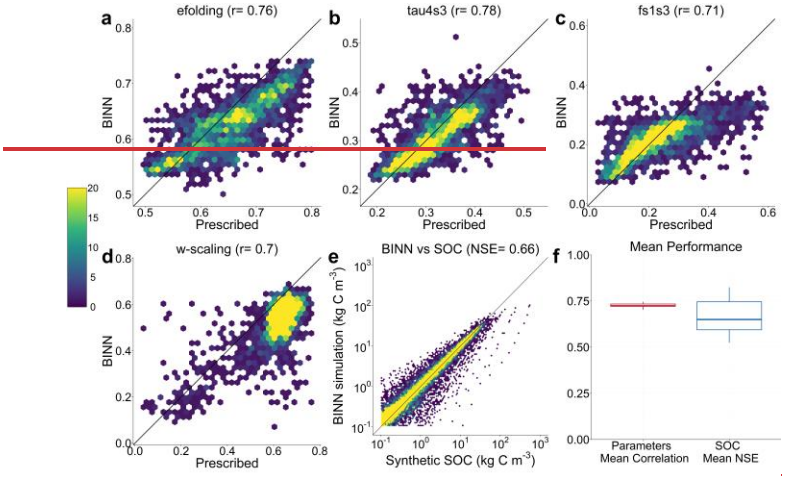
~~**Figure 3: Sensitivity Indices for CLM5 Biogeochemical Parameters Across All Soil Depths.**~~ ~~The bar plot illustrates the~~ **First-order sensitivity** ~~of CLM5 to each parameter~~**indices of biogeochemical parameters aggregated across all soil layers.** Bars show mean sensitivity, with horizontal error bars representing ±1 standard deviation. Red triangles indicate error bars clipped by the x-axis (uncertainty extends beyond the displayed range). Parameters are ~~listed on the~~ ranked (y-axis ~~in descending order based on their sensitivity scores.~~) by decreasing sensitivity and color-coded by their associated process component (e.g., K, A, B, ξ, V in Equation 7). The x-axis represents ~~the~~ sensitivity scores, indicating how changes in each parameter influence the model's performance.

Formatted: Font: Bold

Formatted: Font: Bold

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt



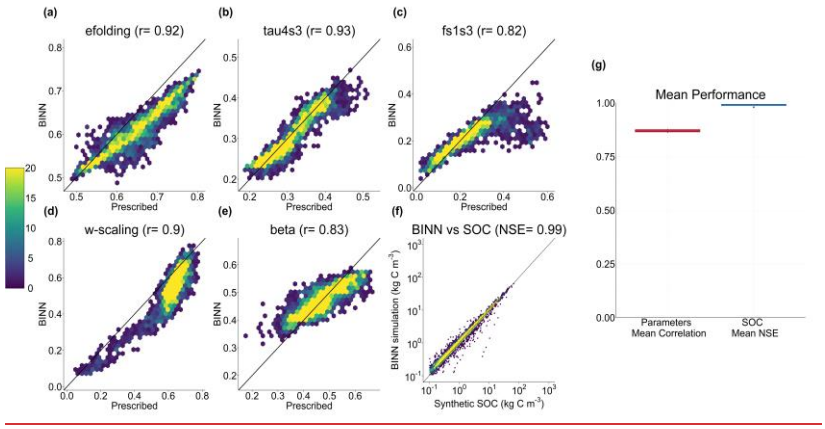


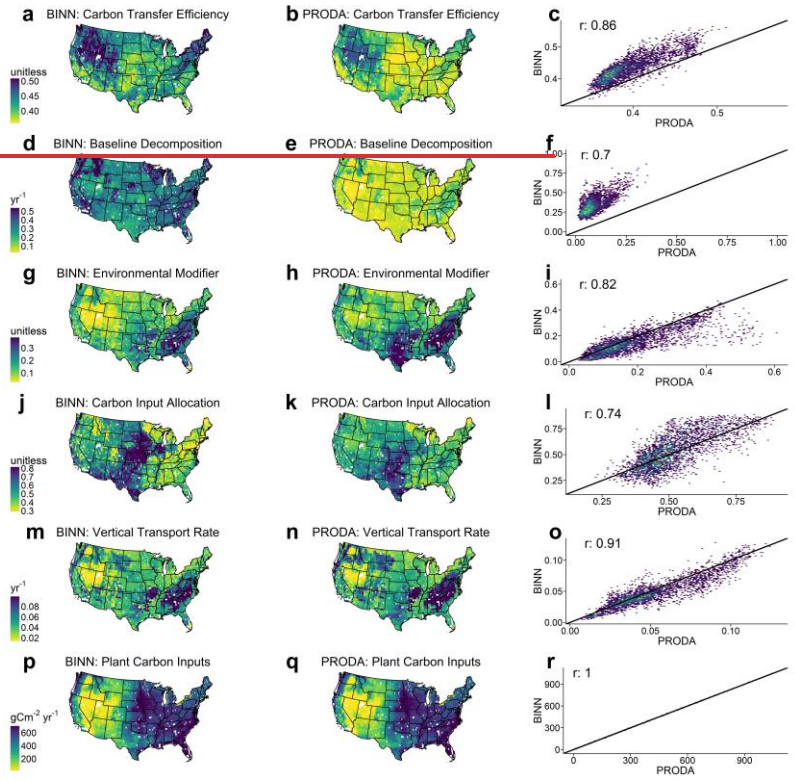
Figure ~~4: Evaluation of BINN's~~3: BINN's performance in ~~retrieving the four most sensitive parameters from the synthetic SOC dataset.~~the parameter recovery experiment. Scatter plots comparing the parameter values (a) "efolding", (b) "tau4s3", (c) "fs1s3", ~~and~~ (d) "w-scaling", and (e) "beta" predicted by BINN ~~(BINN)~~ against the prescribed parameter values. The

Formatted: Font: Bold

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

color of each point in the scatter ~~plots~~plot 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~~f) Comparison of the BINN-simulated and synthetic SOC values, with colors representing the density of points. (~~f~~g) Mean performance of BINN in retrieving the ~~4~~five parameters across 10-fold cross validations, 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.



Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

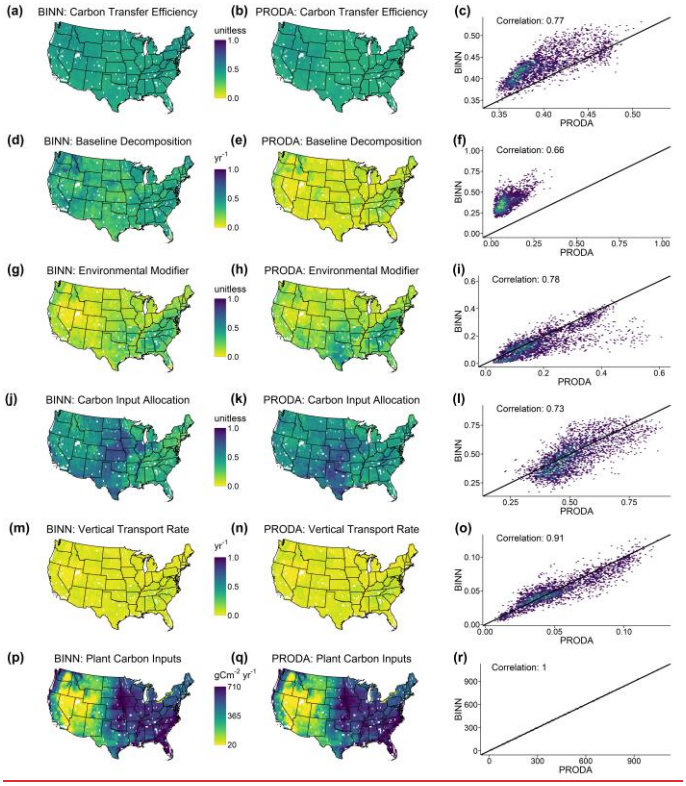
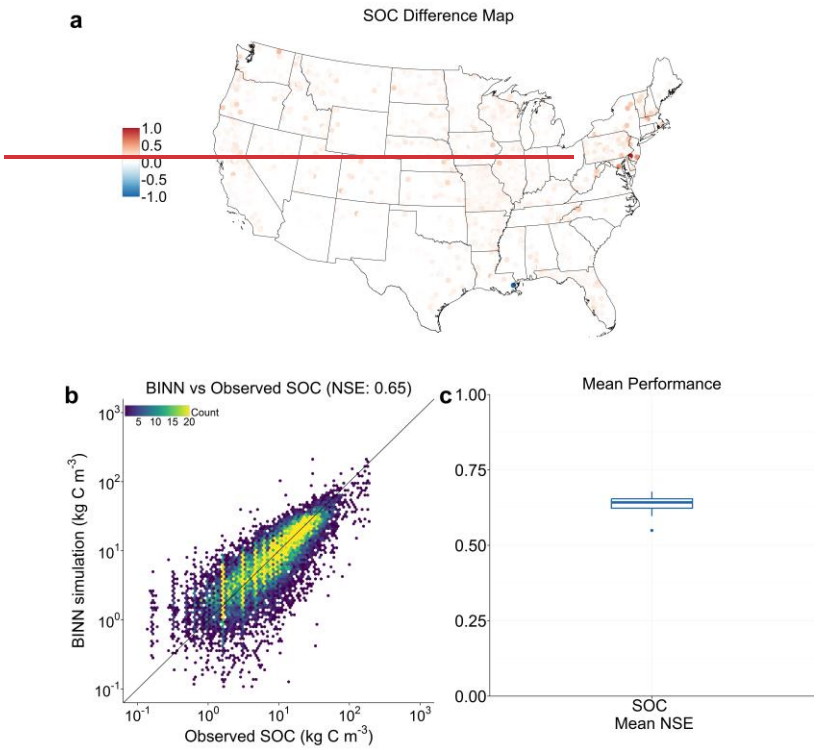


Figure ~~5~~4: Comparison of the spatial patterns of model components retrieved by BINN and PRODA ~~across the Conterminous United States.~~ The. Maps show BINN (left column panels) and PRODA (middle column panels) estimates for six model components ~~include:~~ carbon transfer efficiency (a~~, b,~~ 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~~ Right-column panels (c, f, i, l, o, r) compare ~~the values of each~~ model ~~component retrieved by~~components between BINN (y-axis) ~~against those retrieved by~~ and

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

PRODA (x-axis~~). The~~) with 1:1 lines; panel insets report Pearson correlation ~~coefficient between the~~ coefficients. Color scales are harmonized across 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~~; otherwise, panel-specific scales are noted in the legend.

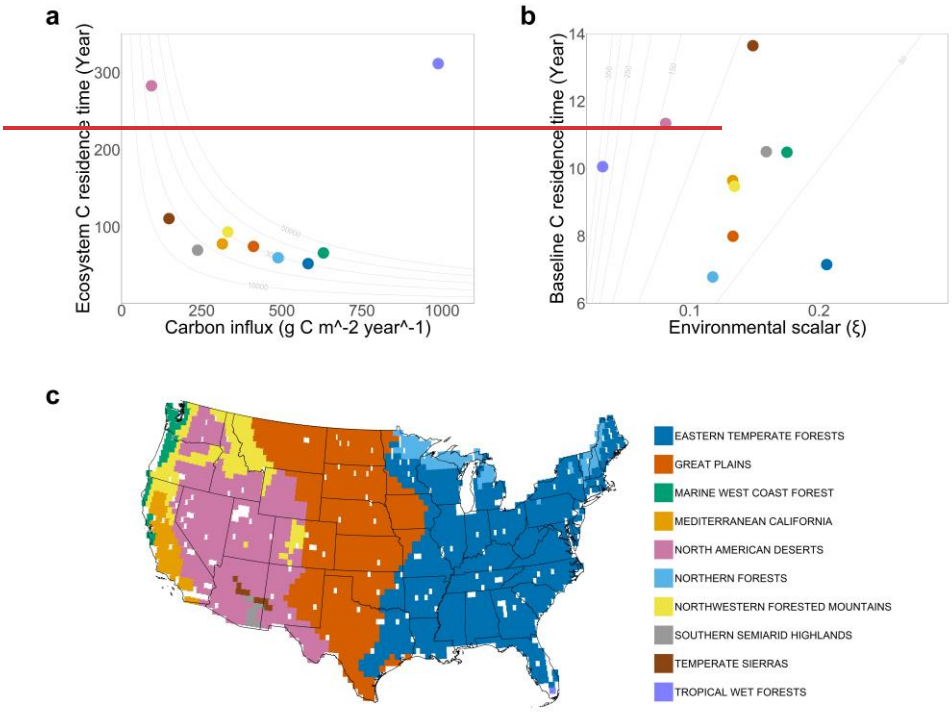


~~**Figure 6: Comparison of observed and simulated Soil Organic Carbon (SOC) Storage Using BINN.** (a) A spatial deviation mapping showcases the difference in simulated SOC~~

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

~~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.~~



Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

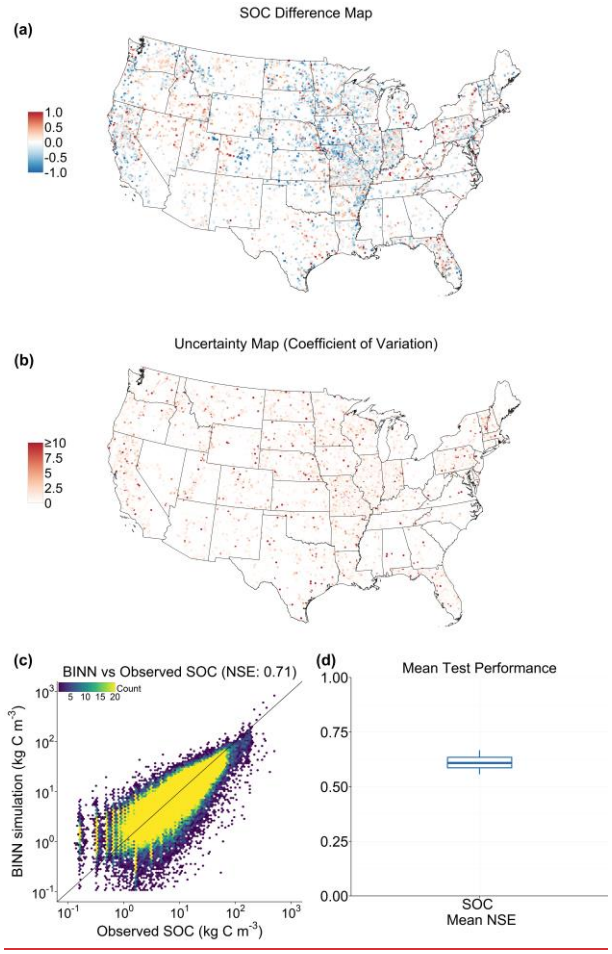


Figure 5: Comparison of simulated and observed SOC storage using BINN across 10-fold cross-validation test. (a) Map of normalized difference across all 10 folds. (b) Coefficient of variation of difference across 10 folds, indicating spatial uncertainty. (c) Observed vs. fold-average simulated SOC across all depths and sites. (d) Box plot of test NSE across 10 folds.

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

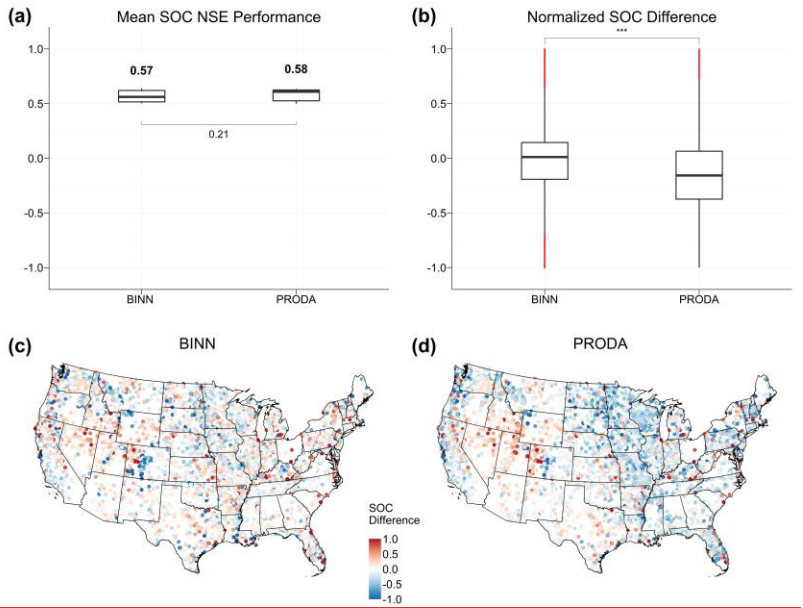


Figure 6. Side-by-side evaluation of BINN and PRODA across identical 10-fold test sets against the observed SOC. (a) Mean test NSE per fold for BINN and PRODA. NSE values are shown above the bars; the P value below the bars donates the difference in test NSE between the two methods is insignificant. (b) Mean test-set differences in simulated SOC between the two methods after normalization. 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 (c) BINN and (d) PRODA: values near 0 indicate small bias; positive and negative values indicate overestimation and underestimation, respectively.

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

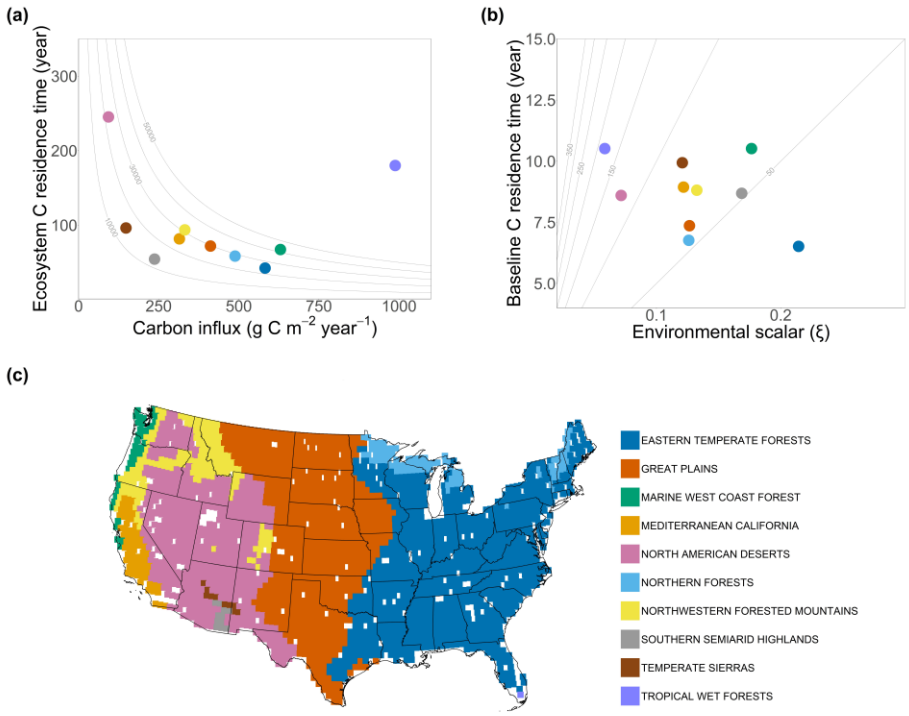


Figure 7: Traceability analysis on how model components determine SOC storage in different biomes. (a) ~~how~~How influx and ecosystem carbon residence time determine SOC storage (contour lines~~) and~~). (b) ~~how~~How environmental scalar (ξ) and baseline carbon residence time ~~determines~~determine ecosystem carbon residence time (contour lines~~) in different biomes.~~). The color-coded points in (a) and (b) represent the average values in different biomes. (c) Biome map for the contiguous US used to aggregate the variables in panels (a) and (b).

Formatted: Font: Not Bold

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

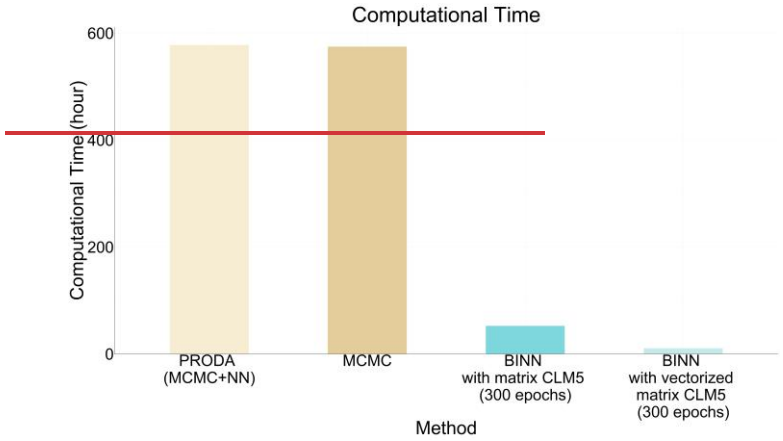


Figure 8: Comparative Analysis of Computational Time Required for Integrating 2000 Soil Profiles into Process-Based Models (CLM5). The figure shows the computational time (in hours) for PRODA (MCMC+NN), which uses a Bayesian inference approach (MCMC) combined with a neural network (NN), and for BINN 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.

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

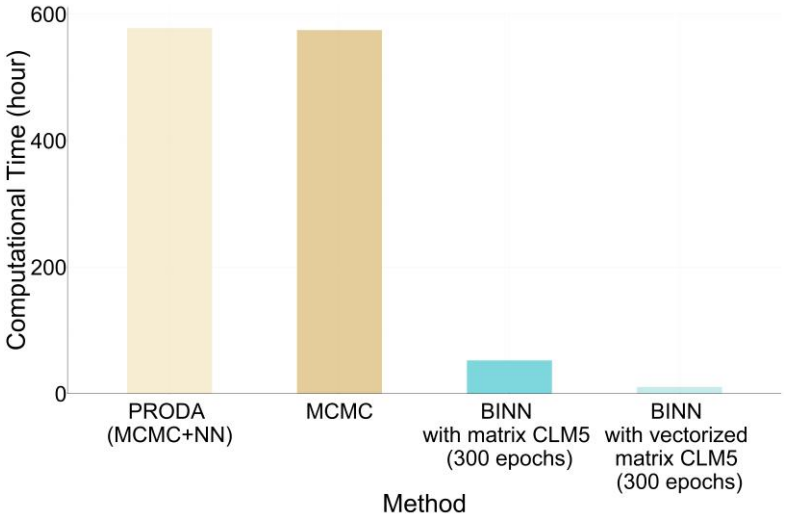


Figure 8: Computational time to process 2,000 soil profiles with CLM5. Wall-clock runtime (hours) for PRODA (MCMC + NN), MCMC only, and for BINN using the original matrix CLM5 and a vectorized CLM5 implementation. All runs were executed on the same platform.

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

Acknowledgements

This research is supported by AI-~~CLIMATE~~LEAF: "AI Institute for ~~Climate~~ Land ~~Interactions, Mitigation, Adaptation, Tradeoffs and,~~ Economy, Agriculture & Forestry", 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).

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

Data and Code Availability

The code and data for BINN is available at ~~https://github.com/Hardyxu8067/BINN.~~:

https://zenodo.org/records/17746275.

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

Conflicts of Interest

The authors declare no competing interests.

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

References

~~1.~~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~~2. 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~~3. Hengl, T. *et al.* SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE* **12**, e0169748 (2017).

~~6~~4. Cook-Patton, S. C. *et al.* Mapping carbon accumulation potential from global natural forest regrowth. *Nature* **585**, 545–550 (2020).

~~7~~5. 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~~6. 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~~7. Karpatne, A., Jia, X. & Kumar, V. Knowledge-guided Machine Learning: Current Trends and Future Prospects. https://doi.org/10.48550/ARXIV.2403.15989 (2024) doi:10.48550/ARXIV.2403.15989.

Formatted: Left

Formatted: Font: 12 pt

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

~~10.~~8. 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. *Ecol. Lett.* **27**, e70012 (2024).

9. 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).

10. Liu, L. *et al.* Knowledge-guided machine learning can improve carbon cycle quantification in agroecosystems. *Nat. Commun.* **15**, 357 (2024).

11. 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.~~13. Shen, C. *et al.* Differentiable modelling to unify machine learning and physical models for geosciences. *Nat. Rev. Earth Environ.* **4**, 552–567 (2023).

14. 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).

15. 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).

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

~~14~~16. 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~~17. 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.~~18. Tramontana, G. *et al.* Partitioning net carbon dioxide fluxes into photosynthesis and respiration using neural networks. *Glob. Change Biol.* **26**, 5235–5253 (2020).

19. 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).

20. Bree, R. van, Marcos, D. & Athanasiadis, I. Hybrid Phenology Modeling for Predicting Temperature Effects on Tree Dormancy. Preprint at https://doi.org/10.48550/arXiv.2501.16848 (2025).

21. 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).

22. 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).

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

23. Wesselkamp, M., Roberts, D. R. & Dormann, C. F. Identifying potential provenances for climate-change adaptation using spatially variable coefficient models. *BMC Ecol. Evol.* **24**, 70 (2024).

24. 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).

~~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~~25. Tao, F. & Luo, Y. PROcess-Guided Deep Learning and DAta-Driven Modelling (PRODA). in *Land Carbon Cycle Modeling* (CRC Press, 2022).

26. Tao, F. *et al.* Microbial carbon use efficiency promotes global soil carbon storage. *Nature* **618**, 981–985 (2023).

~~21~~27. Tao, F. *et al.* Convergence in simulating global soil organic carbon by structurally different models after data assimilation. *Glob. Change Biol.* **30**, e17297 (2024).

28. 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).

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

29. 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).

30. 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).

31. Lu, X. *et al.* Full Implementation of Matrix Approach to Biogeochemistry Module of CLM5. *J. Adv. Model. Earth Syst.* **12**, e2020MS002105 (2020).

32. 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).

~~22~~33. 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~~34. Luo, Y. *et al.* Matrix Approach to Land Carbon Cycle Modeling. *J. Adv. Model. Earth Syst.* **14**, e2022MS003008 (2022).

~~24.~~35 ~~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).~~

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt

~~27~~36. Girshick, R. Fast R-CNN. Preprint at https://doi.org/10.48550/arXiv.1504.08083 (2015).

~~28~~37. Gal, Y. & Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning.

38. Arnold, C. P. & Dey, C. H. Observing-Systems Simulation Experiments: Past, Present, and Future. *Bull. Am. Meteorol. Soc.* **67**, 687–695 (1986).

~~29~~39. Luo, Y. *et al.* Parameter identifiability, constraint, and equifinality in data assimilation with ecosystem models. *Ecol. Appl.* **19**, 571–574 (2009).

40. 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~~41. U.S. Environmental Protection Agency. NA_Eco_Level1. (2010).

~~31~~42. 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).~~

Formatted: Font: Times New Roman, 12 pt

Formatted: Font: Times New Roman, 12 pt