

**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,†}, Lifan Jiang¹, Fengqi You⁵, Benjamin Houlton³, Ying Sun¹, Carla Gomes², and Yiqi Luo^{1*}

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

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

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

⁴Currently at: Department of Informatics and Intelligent Systems, Institute of Energy and the Environment, The Pennsylvania State University

⁵Department 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@psu.edu), Carla Gomes (gomes@cs.cornell.edu)

Abstract

The increasing availability of large-scale observational 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 these large-scale 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. Furthermore, by incorporating Monte Carlo (MC) dropout to generate posterior distributions, we demonstrate that BINN can effectively quantify uncertainty in estimated parameters. We use BINN to predict six major processes (or components in process-based models) regulating the soil carbon cycle from 25,925 observed SOC profiles across the contiguous US and compare them with the same processes previously retrieved by a Bayesian inference-based PROcess-guided deep learning and DATA-driven modeling (PRODA) approach. The good agreement between the spatial patterns retrieved by BINN and PRODA (average correlation coefficient = 0.86) suggests that BINN's ability of capturing mechanistic knowledge is consistent with the established Bayesian-based methods. 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 an efficient framework that harnesses the power of ~~both~~ AI, large-scale data, and process-based modeling to understand large scale soil carbon cycle. Applying BINN to other domains of Earth system science will facilitate new scientific discoveries while improving the predictive performance of Earth system models.

41 **Plain Language Summary**

42 Conventional artificial intelligence (AI) is great at predicting biogeochemical patterns (such as
43 the amount of organic carbon in soils) across space and time, when plenty of data are available.
44 However, it is difficult to use these AI tools to understand the underlying mechanisms that drive
45 these patterns, because many of these processes cannot be directly measured. In this study, we
46 introduce a new framework called Biogeochemistry-Informed Neural Network (BINN) that
47 integrates a process-based model into a multilayer neural network. By integrating large-scale
48 observational datasets, BINN optimizes process-based model predictions while retrieving critical
49 processes governing SOC storage within the framework of the chosen process-based model.
50 Critically, BINN is more than 50 times faster than existing approaches that use Bayesian
51 inference to align scientific models with observational data. This advance in understanding soil
52 carbon cycling and computational speed could help scientists better predict the soil carbon cycle
53 under different environments. BINN can be applied to retrieve mechanistic understanding of
54 other issues of Earth's biogeochemical cycles from large-scale data with process-based models.

1. Introduction

Artificial intelligence (AI) has revolutionized our ability to leverage big data to uncover relationships in complex systems such as the Earth system¹. Machine learning and deep learning, for example, 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 forests⁴, and estimating soil respiration⁵. 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 models⁶.

To address this challenge, various hybrid approaches have emerged, aiming to integrate scientific knowledge and reasoning with standard machine learning methods. This integration leverages the strengths of data-driven techniques alongside scientific theory^{7,8}. 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 predictions⁹, such as by adding loss terms that penalize violations of a known governing equation in physics-informed neural networks (PINN)¹⁰, or discrepancies between neural network and process-based model outputs¹¹. These methods do not guarantee that NN simulations will strictly obey physical principles; they only softly discourage violations.

78 While recent 'Hard-PINNs' developments have sought to enforce these constraints more strictly
79 to minimize the difference between NN predictions and numerical model simulations¹², this
80 requires a complex multi-level optimization procedure with many hyperparameters, and still does
81 not satisfy the constraint exactly. Another body of work thus tries to embed physical knowledge
82 inside the NN model itself, such as modifying an Long Short-Term Memory (LSTM) network to
83 guarantee mass conservation at each step¹³, or embedding hydrological models^{14,15}, terrestrial
84 models¹⁶, thermodynamic rules¹⁷, domain-specific objectives¹⁸, or Bragg's law for x-ray
85 diffraction¹⁹ into the NN model architecture. Recent studies have further advanced these end-to-
86 end setup to infer stomatal resistance and aerodynamic resistance in evapotranspiration
87 modeling²⁰ and to estimate global physical parameters for water cycle models²¹. Despite these
88 advances, previous efforts mostly used limited constraints in a system (e.g., a few empirical
89 relationships in photosynthesis²²) or simplified models that comprise latent variables and
90 conservation principles¹⁴. It remains challenging to integrate numerous differential equations into
91 a NN to study the dynamics of a complex system, such as SOC dynamics. This difficulty stems
92 from the tremendous computational cost required to execute complex models, and the associated
93 high-dimensional parameter spaces that need to be optimized for predictive model simulations.
94 In particular, the capacity of NN to efficiently optimize process-related parameter values with
95 faithful identifiability remains an under-explored area of research.

96 Meanwhile, a traditional method that uses scientific knowledge to guide inference of
97 biogeochemical processes from observational data is data assimilation. This method has played a
98 central role in advancing our understanding of many processes, such as terrestrial carbon flux
99 partitioning⁹, nutrient cycling²³, ecosystem respiration¹, plant phenology²⁴, and hydrology–
100 biogeochemistry couplings²⁵. However, the data assimilation method usually estimates a set of

101 fixed biogeochemical parameters without spatial or temporal variations, which has been argued
102 not represent variations in ecosystem properties over space and time²⁶. While statistical methods
103 such as Geographically-Weighted Regression allow spatially-varying regression coefficients,
104 they can only model linear relationships between inputs and outputs at each location^{27,28}. A
105 recently-developed approach, PROcess-guided deep learning and DATA driven modeling
106 (PRODA), first uses Bayesian inference-based data assimilation methods to find the optimal
107 biogeochemical parameters for each site, and then trains a deep learning model to infer
108 biogeochemical parameters at new sites, improving our understanding of the global soil carbon
109 cycle²⁹⁻³². This approach harnesses the strengths of both process-based modeling and deep
110 learning methods to improve SOC simulations by allowing biogeochemical parameters to vary
111 across space. However, the Bayesian inference-based, site-level data assimilation used in the
112 PRODA approach requires vast computational resources, making the method inefficient in both
113 time and energy and thus difficult to apply broadly.

114 In this study, we integrate a matrix form of a process-based model (i.e., Community Land
115 Model version 5, CLM5) that describes SOC dynamics into a neural network, thus developing a
116 Biogeochemistry-Informed Neural Network (BINN). BINN is an end-to-end framework that
117 combines data-driven machine learning with process-based modeling to facilitate the
118 interpretability of model-represented biogeochemical dynamics, such as SOC dynamics in this
119 study. Herein, we first introduce the structure of BINN, followed by a demonstration of BINN's
120 ability to recover biogeochemical parameters with high accuracy through a parameter recovery
121 experiment. We further implemented Monte Carlo (MC) dropout to generate posterior parameter
122 distributions to quantify uncertainties and identify equifinality issues. Lastly, we evaluate model
123 components from real-world SOC observations. The BINN-predicted biogeochemical parameters

124 accurately simulate SOC observations and are similar to those produced by PRODA, while being
125 much faster to compute. By combining data-driven learning with the mechanistic representations
126 of process-based models, BINN can accurately and quickly infer underlying physical processes
127 from SOC observations.

128

129 **2. Methods**

130 **2.1 Biogeochemistry-Informed Neural Network (BINN)**

131 BINN integrates a process-based model (CLM5, see Section 2.1.2 below) into a differentiable,
132 end-to-end framework (Fig. 1a) that maps environmental covariates to site-level biogeochemical
133 parameters, feeds those parameters into CLM5 for the forward simulation, and back-propagates
134 the loss through the whole framework during training (Fig. 1b).

135 We implemented BINN in Python using PyTorch and executed the experiments on the
136 supercomputer, Derecho, maintained by National Center for Atmospheric Research (NCAR).
137 The experiments were conducted using one compute node with 128 CPU cores, leveraging
138 PyTorch Distributed Data Parallel (DDP) for multi-CPU training. Meanwhile, BINN was also
139 tested with a multi-GPU compute node on a cluster in Cornell University's Center for Advanced
140 Computing, confirming its capability for training on GPU clusters when needed.

141

142 **2.1.1 Neural Network**

143 A fully-connected neural network learns the relationship between environmental covariates
144 (Table S1) and biogeochemical parameters (Table S2) over space. At each site, we have 60
145 environmental covariates (input features). While the numeric features can be directly passed into
146 the neural network, we use embedding layers to encode categorical (discrete) covariates and

147 spatial coordinates. For each categorical covariate, the embedding layer learns a vector
 148 embedding for each category (e.g. climate type). For spatial coordinates, we pass the latitude and
 149 longitude through a spatial positional encoder to obtain a location embedding vector,
 150 characterizing unobserved aspects of each location that are not captured in our covariates³³.

151 We then concatenate the categorical covariate embeddings, location embedding, and the
 152 remaining covariates into a vector $\mathbf{e}$, and pass this through a 4-layer neural network f_{NN}
 153 (Equation 1) with learnable weights/biases $\mathbf{w}$, which outputs a vector $\tilde{\mathbf{p}}$ with 21 values
 154 (representing *unconstrained* predictions for each parameter in the process-based model):

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

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

$$y = \sum w_i x_i + b \quad (2)$$

158 where y is an output from a neuron after the linear combination of its input, w_i is a learnable
 159 weight for x_i , x_i is the input from either the environmental covariates or the previous layer's
 160 outputs, and b is a learnable neuron-specific bias. After the linear transformation, a nonlinear
 161 activation function is applied to generate the eventual results at each neuron, so that the neural
 162 network can generate complex nonlinear relationships between the inputs (i.e., environmental
 163 covariates) and the outputs (i.e., the biogeochemical parameters in the process-based model).
 164 Note that the linear transformations and activation functions are both differentiable, so we can
 165 backpropagate gradients through the network. This allows us to calculate the gradient of the loss
 166 function with respect to all learnable weights w_i and biases b . To prevent overfitting and enable

Formatted Table

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

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

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

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

178 The final layer has 21 output neurons, with one corresponding to each biogeochemical
 179 parameter in the process-based model. We do not use *LeakyReLU* after the final linear
 180 transformation because it would bias the distribution of the predicted parameters. However, as
 181 we have prior knowledge about the plausible range of values for each biogeochemical parameter,
 182 the initial parameter predictions $\tilde{p}$ are first normalized to be in the range [0, 1] by element-wise
 183 *parameterized sigmoid* functions σ (Equation 4) and then further linearly scaled into the
 184 predicted parameters p (Equation 5). This ensures that each parameter p_i stays within their prior
 185 ranges:

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

Formatted Table

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

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 ³¹. 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)³⁴. 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, dropout 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 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'

207 outputs by re-centering and re-scaling, making training faster and more stable³⁵. We initialized γ
208 to 64.5 but allowed it to be further optimized throughout the training processes within the range
209 from 10 to 109. Specifically, we used a sigmoid to constrain the range:

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

210 where γ' is a learnable parameter that is initialized to 0. We set *negative_slope* to -0.3 in the
211 *LeakyReLU* by default. In total, our BINN architecture contains 74,793 trainable parameters.

212 We also performed a quantitative sensitivity analysis to assess how BINN performance
213 responds to variations in the loss and activation functions (Table S3). Specifically, we tested the
214 sensitivity of replacing the smooth L1 loss with L1 loss and L2 loss and substituting the
215 hyperbolic cosine parameter loss with a squared loss. Furthermore, we evaluated the sensitivity
216 to changes in activation functions by modifying the first three layers from leaky ReLU to ReLU
217 and the final layer from Sigmoid to hard Sigmoid.

218

219 2.1.2 Process-based Model

220 In this study, we use a matrix form of the soil carbon module of CLM5 to represent our
221 knowledge of SOC dynamics^{36,37} (Fig. S1). CLM5 has been continuously developed and refined
222 over the past decade for simulating biogeochemical cycles, including SOC dynamics³⁸; it
223 mathematically represents our knowledge of SOC dynamics with 140 partial differential
224 equations. We chose CLM5 to enable direct comparison with PRODA, which uses the same
225 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 $\times$ 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^{37,39} as:

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

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

In this study, we assume steady-state SOC storage for computational efficiency, 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

Formatted Table

thousands of years^{31,40}. 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:

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

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 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 cycle²⁶. Because those processes are highly variable depending on different climate conditions or soil properties, the values quantifying their properties (i.e., the parameter values) should differ ~~with locations~~ with different

environments²⁶. 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 S4) are used in Equation (8) to estimate steady-state SOC storage at sites across the Contiguous US.

To mitigate the impact of uncertainties in individual parameters and equifinality on our mechanistic analysis, we aggregated related parameters into functionally distinct model components. Following the method from Tao et al. (2024), these components, including carbon transfer efficiency, carbon input allocation, baseline decomposition, vertical transport rate, environmental modifiers, and plant carbon input, were constructed as weighted averages based on the relative contributions of constituent parameters to their respective carbon fluxes or pools.

2.1.3 Loss Function

Because we aim to accurately 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 :

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

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

Formatted Table

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 such as Mean Squared Error (MSE)⁴¹.

We also add an additional hyperbolic cosine loss (*cosh*) term that acts as a regulator, encouraging the neural network to predict biogeochemical parameters within reasonable bounds. Specifically, it penalizes parameter values that deviate substantially from the center of the prior distribution, thereby discouraging biogeochemically implausible extreme values. Eventually, the total loss, L_{batch} , is a linear combination of the two losses:

$$L_{batch} = \sum_{profile=1}^{batch\ size} \{Smooth\ L1\ Loss(\hat{y}_{profile}, y_{profile}) + w \sum_{j=1}^{21} cosh[\tau(p_j - 0.5)]\} \quad (10)$$

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 τ to 15 and loss weight w 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. Thus, in calculating the loss function value, 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 SOC concentration in deeper layers no longer changes much.

Formatted Table

309

310 **2.1.4 Backpropagation to Optimize Neural Network Parameters**

311 During training, BINN computes the loss function based on the current predicted SOC and
312 biogeochemical parameters; the loss quantifies how poorly its current predictions match the SOC
313 observations and prior knowledge. Through backpropagation, the loss signals propagate
314 backwards through the entire BINN structural chain: first through the CLM5 matrix equations
315 that generate modeled SOC, then through the biogeochemical parameters, and finally to the
316 neural network that predicts these biogeochemical parameters. At each step, PyTorch uses the
317 chain rule to automatically compute the gradient of the loss function with respect to each
318 learnable NN component (e.g. w in Equation 1 and γ' in Equation 26). These gradients indicate
319 how each component can be adjusted to increase or decrease the loss. The NN components are
320 adjusted slightly in the direction that decreases the loss, and then the above process is repeated.
321 The differentiability of the process-based model (CLM5 in this case) enables this continuous
322 gradient flow and thus allows the neural network to learn parameter values that produce better
323 SOC predictions.

324

325 **2.1.5 Monte Carlo dropout**

326 We used Monte Carlo (MC) dropout during BINN training to approximate the model's
327 predictive uncertainty⁴². MC dropout randomly disables a fraction of neurons (with probability
328 equals to the dropout rate) within the neural network during training. This process acts like
329 training an implicit ensemble of subnetworks and improves generalization and robustness of
330 BINN.

When the dropout rate 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 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 ability to recover the biogeochemical parameters of CLM5 from synthetic SOC data using a 10-fold cross-validation experiment⁴³. 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 approach²⁹; see Section 2.3.2), providing a controlled environment where true parameter values are known (Fig. 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.

2.2.1 First-Order Sobol Sensitivity Analysis

To decide which biogeochemical parameters to modify in the experiment of recovering the biogeochemical parameters of CLM5 from synthetic SOC, we conducted a sensitivity analysis of CLM5 using first-order Sobol method to identify the biogeochemical parameters that have the greatest influence on simulated SOC values.

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

363 For this analysis, we randomly selected 512 sites across the Contiguous US and
 364 employed the first-order approximation method. We first determined the unconditional variance
 365 $V(SOC)$ from the model output when all the 21 biogeochemical parameters (p) in CLM5 were
 366 allowed to vary freely within their initial ranges from their prior ranges. Specifically, we
 367 randomly sampled the biogeochemical parameter values 1000 times in their initial ranges at each
 368 site s and depth range d , ran the model, and calculated the variance of the simulations, which was
 369 considered the unconditional variance $V(SOC_{s,d})$.

370 Next, we estimated the conditional expectation of the variable $SOC_{s,d}$ for each
 371 biogeochemical parameter p_i ($i = [0, 20]$) at each site s . We randomly selected a value (p_i^*) for
 372 each biogeochemical parameter p_i from a uniform distribution within its prior range. For the
 373 remaining biogeochemical parameters ($p_j: j \neq i$), we produced 1000 random settings from
 374 uniform distributions within their respective prior ranges. Using the sample of 1000
 375 biogeochemical parameter sets, we estimated the conditional expectation $E(SOC_{s,d}|p_i = p_i^*)$.
 376 We repeated this sampling process for 100 randomly selected values of p_i and used the results to
 377 estimate the variance $V(E(SOC_{s,d}|p_i))$. This value quantifies the variance in the output variable

378 $SOC_{s,d}$ as a result of modifying the biogeochemical parameter p_i . We discarded the simulations
 379 when NaN values appeared due to randomly sampled biogeochemical parameter sets. Finally, a
 380 first-order Sobol sensitivity index $S_{i,s,d}$ was calculated for each biogeochemical parameter p_i at
 381 each site s and depth d as:

$$S_{i,s,d} = \frac{V(E(SOC_{s,d}|p_i))}{V(SOC_{s,d})} \quad (11)$$

382 The final sensitivity value for each biogeochemical parameter was obtained by averaging the
 383 sensitivity values across all the randomly selected sites across all depths (Fig. 2) and individual
 384 depth ranges (Fig. S3). Error bars represent the mean sensitivity $\pm$ one standard deviation across
 385 sites.

386 Our analysis identified the five most sensitive parameters: the parameter "beta" governs
 387 the allocation of carbon inputs across soil depths; the parameter "w-scaling" represents the
 388 influence of soil water on SOC decomposition; the parameter "tau4s3" represents the
 389 decomposability of the passive SOC pool; the parameter "fs1s3" indicates the efficiency of
 390 carbon transforming from active SOC to passive SOC; and the parameter "efolding" quantifies
 391 the impacts of soil depth on SOC decomposition.

392

393 **2.2.2 Simulation Experiment to Recover Biogeochemical Parameters from Synthetic Data**

394 We selected the five most sensitive parameters identified from the sensitivity test to minimize
 395 equifinality, which is a phenomenon that a similar simulation output is generated with different
 396 combinations of biogeochemical parameter values⁴⁴, for the parameter recovery experiment.

397 To test the recovery efficiency of the biogeochemical parameters with BINN, we
 398 modified the final layer of the neural network by reducing the number of neurons from 21 to 5 to

399 predict these five biogeochemical parameters. Combining the five biogeochemical parameters
400 predicted by BINN and the remaining 16 biogeochemical parameters from the prescribed
401 parameter values, BINN was able to simulate SOC values and update itself through
402 backpropagation. After training BINN with the synthetic SOC dataset, we compared the five
403 parameters predicted by BINN with their prescribed parameter values in the testing dataset to
404 evaluate the accuracy of BINN in retrieving the prescribed biogeochemical parameters. To
405 examine the recovery efficiency beyond the top five parameters, we repeated the experiment
406 with all 21 parameters following the same processes.

407

408 2.2.3 10-fold cross-validation

409 To conduct 10-fold cross-validation on the simulations, the entire dataset was
410 randomly divided into ten equal-sized subsets. In each iteration, 8 subsets were used for training,
411 1 subset for validation, while the remaining subset served as the test set. Specifically, the training
412 subsets were utilized for gradient-based optimization, the validation subset served as an
413 independent check to guide hyperparameter tuning, best model selection, and early stopping to
414 prevent overfitting, and the test subset was reserved for an unbiased assessment of the BINN's
415 generalization capability. This process was repeated ten times, with each subset serving as the
416 test set once. The performance metrics, including Nash–Sutcliffe modelling efficiency
417 coefficient (NSE) and Pearson correlation coefficient (r), were calculated for each iteration. Final
418 performance evaluations were conducted by averaging the metrics across all iterations, and grid-
419 level predictions were averaged across the ten iterations. This cross-validation approach provides
420 a robust assessment of BINN's internal predictive stability by testing its performance on multiple

independent datasets, reducing the impact of data partitioning bias and thus enabling evaluation of model stability across different training-testing combinations.

2.3 BINN performance with real-world SOC Observations

We evaluated the performance of BINN by comparing BINN’s SOC predictions with observed SOC across the contiguous US (a total of 25,925 profiles) (Fig. 1b).

2.3.1 Data Preparation

We processed SOC observations from the World Soil Information Service (WoSIS) following Tao et al. (2020, 2023)^{29,31}. 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)³¹ 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 the previous year and current year, the number of soil layers reaching bedrock, the soil oxygen scalar for decomposition, the soil nitrogen scalar for decomposition, soil temperature, and soil water potential. These forcings were derived from 20 years of monthly CLM5 simulations at steady state using a preindustrial forcing (that is, I1850CIm50Bgc) 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.

445

446 2.3.2 PRODA

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

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

455 To generalize these site-level estimates to the contiguous US, we trained a fully-
456 connected multilayer neural network using the same 60 environmental variables as predictors.
457 The neural network used in the final training consisted of four hidden layers. The node numbers
458 for each hidden layer were 256, 512, 512 and 256, respectively, with *ReLU* as the activation
459 function and *adadelta* as the optimizer. PRODA was trained with 6000 epochs, utilizing an early
460 stopping patience of 1200 epochs, and validated with a 10-fold cross-validation. When
461 comparing with BINN, the test dataset for each fold was set the same as the corresponding fold
462 during BINN training.

463 To facilitate a direct comparison of model performance between BINN and PRODA (Fig.
464 4 and 6), we used the identical prior ranges of parameters governing substrate decomposability
465 as used in PRODA. Specifically, the prior ranges were for tau4s1 were [0.0001, 1]), tau4s2 [1,
466 50]), and tau4s3 ([200, 1000]). Aside from this specific comparative analysis, the values of prior
467 parameter ranges in Table S2 were used for all other analyses.

468 The robustness of the BINN framework to the variations in prior range selection was
 469 further evaluated by adjusting the prior ranges by 10 percent where the flux related parameters
 470 remained within their physical bounds. The results in Table S3 indicate that BINN is robust to
 471 the test of prior ranges as evidenced by the consistent NSE across different settings of modeling
 472 choices.

473

474 2.3.3 Summary Statistics

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

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

477 where obs_i is one of the SOC observations, $\overline{obs_i}$ is the mean of the SOC observations, and
 478 $simu_i$ is the simulated SOC by CLM5 embedded in BINN.

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

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

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

486

487 2.3.4 Normalized Difference

488 For each site, we averaged the simulated SOC across the 10 cross-validation folds and subtracted
489 the observed SOC to obtain a mean difference. To improve interpretability over space, we
490 applied a signed percentile normalization: positive differences were scaled to $[0,1]$ by dividing
491 by the 99th percentile of positive deviations, and negative differences were scaled to $[-1,0]$ by
492 dividing by the 99th percentile of negative deviations. This approach preserves the sign and
493 relative magnitude of biases while preventing a small number of extreme values from
494 dominating the colored scale.

495

496 **2.4 Computational Efficiency**

497 We compared computational efficiency between BINN and PRODA on the same dataset (2,000
498 soil profiles) and the same personal computer, reporting their wall-clock time and peak resident
499 set size (RSS). For BINN, we evaluated two implementations and recorded their respective
500 runtime and peak RSS: (i) a matrix version using the original CLM5 formulation and (ii) a
501 vectorized version that removes explicit for-loops. Each implementation was trained for 300
502 epochs (no early stopping) to standardize timing and ensure convergence on the training set. For
503 PRODA, the computational bottleneck is the site-level MCMC calibration of CLM5 parameters.
504 We measured MCMC runtime at 10 randomly selected sites (with 20,000 test iterations and
505 50,000 formal iterations per site) and extrapolated to 2,000 sites. Because per-site memory usage
506 is stable across sites, peak RSS was taken from these 10-site runs. PRODA's second stage using
507 a neural network that maps environmental covariates to the MCMC-optimized parameters was
508 trained for 6,000 epochs, with an early stopping patience of 1200 epochs. We recorded its peak
509 RSS and reported PRODA's overall peak RSS as the maximum of the MCMC stage or the NN
510 stage.

511

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-validation (Fig. 3g). Moreover, BINN achieved an average NSE of 0.99 on the test dataset when comparing simulated SOC with synthetic SOC (Fig. 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 (Fig. 3a). The parameter “tau4s3”, representing the baseline turnover time of the passive SOC pool, showed a correlation coefficient of 0.93 (Fig. 3b). The “fs1s3” parameter, indicating the transfer fraction from fast to passive SOC pools, achieved a correlation coefficient of 0.82 (Fig. 3c). The “w-scaling” parameter, which is the scaling factor of soil water, had a correlation coefficient of 0.9 (Fig. 3d). Lastly, the “beta” parameter that governs the vertical distribution of carbon inputs exhibited a correlation coefficient of 0.83 (Fig. 3e).

Extending recovery to all 21 parameters revealed heterogeneous identifiability that is consistent with their influence on SOC storage (Fig. S4). Parameters with low sensitivity (e.g., cryo, taucwd) showed weak correlations between retrieved and prescribed values and were typically distributed around the midpoints of their prior ranges, 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 dropout⁴² to approximate 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 posterior distributions) in Fig. S6, which indicated the issue of equifinality is effectively identified and accounted for within the BINN framework, as the uncertainty intervals consistently encompass the ground-truth parameter values at most sites.

541

542 **3.2 Performance in Predicting Real-world SOC Observations**

After BINN optimization, we used its predicted biogeochemical parameters to calculate six model components representing different processes in soil carbon cycle over the contiguous US: 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 and those generated from PRODA, along with their spatial patterns, are shown in Fig. 4. Note that the plant carbon input component was identical between BINN and PRODA (Fig. 4p-r), due to the use of the same NPP forcing data. The spatial distributions of the other five components were similar between BINN and PRODA, with an average correlation coefficient of 0.86.

The carbon transfer efficiency predicted by BINN, which quantifies the weighted average ratio of decomposed carbon being transferred from one carbon pool to another relative to the total carbon decomposition, displayed similar spatial patterns as PRODA's predictions (Fig. 4a–b), with a correlation coefficient of 0.92 (Fig. 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 as well (Fig. 4d–e), with the correlation coefficient of 0.91 (Fig. 4f).

559 The environmental modifier predicted by BINN achieved a correlation coefficient of 0.82
560 with PRODA's results (Fig. 4i). Their spatial patterns were nearly identical across the contiguous
561 US, with lower values in the northwestern region and values gradually increasing toward the
562 southeastern region (Fig. 4g–h). While both methods simulated higher carbon input allocation in
563 the central US, BINN predicted slightly higher values in the western and central parts of the
564 contiguous US compared to PRODA (Fig. 4j–k), resulting in a relatively high average value.
565 Even so, the correlation coefficient between the two approaches still reached 0.71 (Fig. 4l).
566 BINN and PRODA also predicted vertical transport rate with nearly identical spatial distributions
567 (Fig. 4m–n), with a high correlation coefficient of 0.92 (Fig. 4o).

568 When applying the predicted parameters to simulate SOC, BINN showed an average
569 NSE of 0.71 in predicting the observed SOC across the 10-fold cross-validation (Fig. 5c), with
570 an average NSE of ~~about 0.5761~~ on the test datasets (Fig. 5d). The training and validation NSE
571 values recorded throughout model training at each epoch indicated a validation NSE of 0.63 (Fig.
572 S7). The map of the normalized difference (Fig. 5a) identifies geographic regions where BINN's
573 predictive performance is limited, showing systematic underestimation in the North-Central
574 region and overestimation within the Rocky Mountains., which persisted across the 10-fold
575 cross-validation ensemble (Fig. 5b).

576 Overall, BINN demonstrated in-domain predictive accuracy comparable to PRODA in
577 predicting SOC across the contiguous US (Fig. 6a). However, PRODA showed a tendency
578 toward negative bias (underestimation) over broad areas, whereas BINN's normalized difference
579 was closer to zero (Fig. 6b). In the north-central and eastern US, where PRODA mainly
580 underestimated observed SOC, BINN exhibited fewer geographical biases (Fig. 6c–d).

581

3.3 Computational Efficiency

On a single CPU, BINN with for-loops required 52.5 hours (Fig. 7) to complete 2,000 profiles, whereas the vectorized version took only 10.5 hours ($5\times$ faster). Moreover, MCMC alone took 574.7 hours, leading to a total of 577.7 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 DDP, which enables parallel training across multiple CPUs. While both approaches achieve parameter interpretability (Section 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 (Fig. S8).

4. Discussion

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

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

604 In this study, BINN's ability of retrieving biogeochemical parameters was first validated through
605 a parameter recovery experiment, which used synthetic SOC data generated by CLM5 with
606 prescribed parameter values across the contiguous US. To minimize the effects of equifinality,
607 we conducted a sensitivity analysis using a first-order Sobol index and focused on the five most
608 sensitive biogeochemical parameters, which are also those typically well constrained by
609 Bayesian calibration. It is important to note that while the first-order Sobol index identifies the
610 direct contribution of individual parameters to model variance, it may not fully account for all
611 nonlinear interactions within the system. Therefore, future research involving higher-order
612 sensitivity analysis would be beneficial to further explore complex parameter relationships.
613 As expected, BINN recovered these target parameters with high correlations with their respective
614 prescribed values. Moreover, BINN reproduced the synthetic SOC patterns well using only
615 environmental covariates and synthetic SOC, without using a Bayesian method. These results
616 suggest that BINN learns causal relationships between environmental covariates and SOC
617 simulations consistent with the underlying dynamics, rather than simply exploiting spurious
618 correlations²⁶.

619 We then extended recovery to all 21 parameters; the results showed that highly sensitive
620 parameters remained well recovered. Although parameters with low influence on SOC showed
621 weak recovery by BINN, this is a common limitation also observed in Bayesian approaches
622 when data are weakly informative. To overcome this limitation, future applications could
623 integrate complementary data streams beyond SOC, which might provide additional constraints
624 necessary to inform those less-sensitive parameters. Alternatively, the inherent uncertainty of
625 high-dimensional parameter spaces can be mitigated by aggregating individual parameters into
626 functionally related model components. As demonstrated in this study, grouping parameters with

627 similar process-based meanings and weighing them by their relative contributions to total SOC
628 help characterize the underlying biogeochemical processes more powerfully (Fig. 4). This
629 approach effectively prioritizes the most influential parameters; as weakly recovered parameters
630 generally exert minimal impact on total SOC storage. Such parameters are assigned lower
631 weights within the calculated model components, thereby mitigating the impact of poorly
632 constrained parameters on our mechanistic interpretations.

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

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

650 While both models show comparable NSE values for SOC simulations (Fig. 6a), their
651 spatial error structures differ significantly. PRODA demonstrates a widespread tendency to
652 underestimate SOC stocks, whereas BINN's normalized differences are centered nearer to zero,
653 indicating reduced systematic bias (Fig. 6b). This difference likely stems from the architectural
654 distinction between the two frameworks: while the neural network in PRODA is trained to
655 replicate site-level parameter distributions derived from a separate site-level data assimilation
656 step, BINN's end-to-end framework directly optimizes the neural network to minimize the
657 deficit in SOC simulations across the US continent. By coupling the parameter inference and the
658 physical simulation within a single objective function, BINN can more effectively mitigate
659 spatial biases that might otherwise arise from the multi-step training process used in PRODA.

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

668 However, even though BINN shows less underestimation than PRODA, spatial ~~bias~~biases
669 still persist within the North-Central and Rocky Mountain domains (Fig. 5a). These biases likely
670 stem from a fundamental mismatch between the spatial resolution of the environmental
671 covariates that derived from global gridded products (Table S1), and the local-scale drivers of
672 each SOC observation. In the North-Central region, the high spatial density of SOC observations

reveals sub-grid variability that coarse gridded covariates cannot adequately resolve. Conversely, the overestimation observed in the Rocky Mountains highlights a scale-dependent limitation where gridded inputs fail to represent the complex topographic heterogeneity inherent to mountainous terrain. Consequently, while the BINN framework effectively captures broad environmental gradients, its localized precision remains constrained by the granularity of the available geospatial drivers.

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 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 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. Instead, it only perturbs parameters randomly and checks whether the accuracy has been improved or not.

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 S5). 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

696 chain, which tends to increase peak RSS despite longer running time. High computational
697 efficiency is also more environmentally friendly, saving energy when dealing with large datasets.
698

699 **4.3. BINN's Facilitation of Mechanistic Understanding**

700 A process-based model is an abstraction of a real-world system and represents the processes that
701 govern the system's dynamics, yet such model-based predictions often fit with empirical
702 observations poorly⁴⁵. This discrepancy arises because complex systems, such as the terrestrial
703 ecosystems, contain numerous mechanisms regulating carbon cycling. Although some of these
704 mechanisms are well-understood, many remain unresolved. Process-based models may explicitly
705 represent those well-understood processes in their structure while these processes need to be
706 specified via parameterization at given locations under given conditions to represent the
707 heterogeneity of environmental conditions²⁶. Without taking advantage of the extensive
708 information present in observations, these model parameters are usually not well constrained.

709 BINN learns site-specific parameters from environmental covariates while the process-
710 based model executes the forward simulation. This end-to-end setup improves model fit to
711 observations by yielding spatially varying parameters. When these parameters are properly
712 constrained by empirical data, they can more accurately reflect the unresolved biogeochemical
713 processes, enabling models to better simulate ecosystem behavior. After training, we further
714 transform the learned parameters into model components (as in Fig. 4), thereby linking learned
715 parameters to individual processes. The model components from BINN reveal interpretable
716 patterns (e.g., baseline decomposition and carbon transfer efficiency) consistent with prior work
717 using Bayesian-based methods³¹.

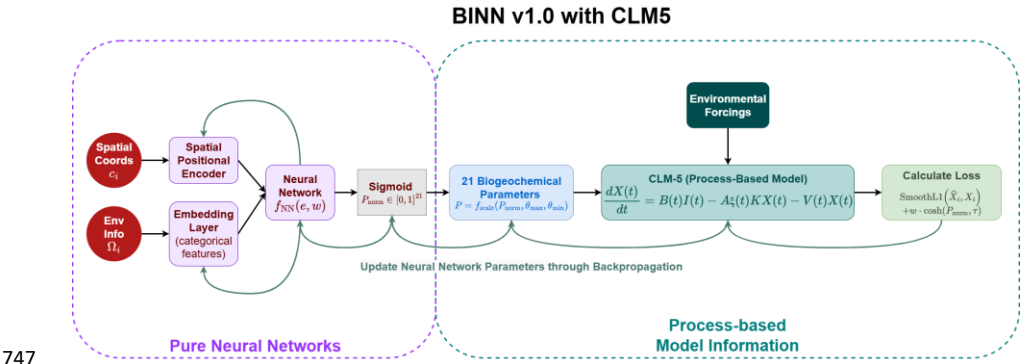
718 However, it is important to recognize that the ~~meehanism~~mechanistic understanding
719 inferred through this framework is inherently tied to the CLM5 model structure and the BINN-
720 optimized parameters. While these optimized parameters provide a ~~mechanically~~mechanistically
721 consistent representation of different processes within the process-based model, the resulting
722 mechanistic ~~discovery~~understanding may still require ~~further~~ independent validations, ~~such as~~
723 ~~observational data~~ from observation via satellites, sensors, and field and laboratory experiments.
724 In the future, more robust data assimilation will also rely on comprehensive empirical syntheses
725 of processes related to the biogeochemical parameters used in process-based models.
726 Nonetheless, BINN-retrieved ~~meehanism~~patterns represent a mechanistic understanding ~~is much~~
727 ~~better~~ grounded in ~~observations~~observation-model synthesis rather than these ~~that~~ gained by
728 ~~the~~from process-based models ~~or data~~ alone.

729 BINN's potential for advancing scientific understanding could be extended well beyond
730 SOC dynamics to various fields of research in biogeochemistry and ecology. Whenever current
731 scientific understanding of a biogeochemical system can be mathematically formulated in a
732 process-based model, BINN could help uncover mechanisms from big data when relevant
733 observational data are presented. This framework would be particularly valuable for studying
734 complex biogeochemical cycles, such as nutrient cycles, where some processes are well-
735 understood and explicitly represented in models, while others remain unresolved. By combining
736 process-based representations of known mechanisms with large-scale data, BINN could help
737 uncover unknown mechanisms governing biogeochemical cycling, providing a foundation for
738 future empirical studies to validate these model-derived insights.

739 Furthermore, BINN's flexible architecture might allow integration of diverse data
740 sources. This capability would be particularly valuable for incorporating limited but important

741 datasets, such as isotope measurements, which could reveal spatial and temporal mechanisms in
742 soil carbon dynamics despite their scarcity. Even if these measurements are only available at a
743 few sites, BINN could still learn from them by incorporating them in the loss function where
744 they are available. By leveraging multiple data sources, BINN would maximize the potential to
745 facilitate our scientific understanding while maintaining biogeochemical consistency.

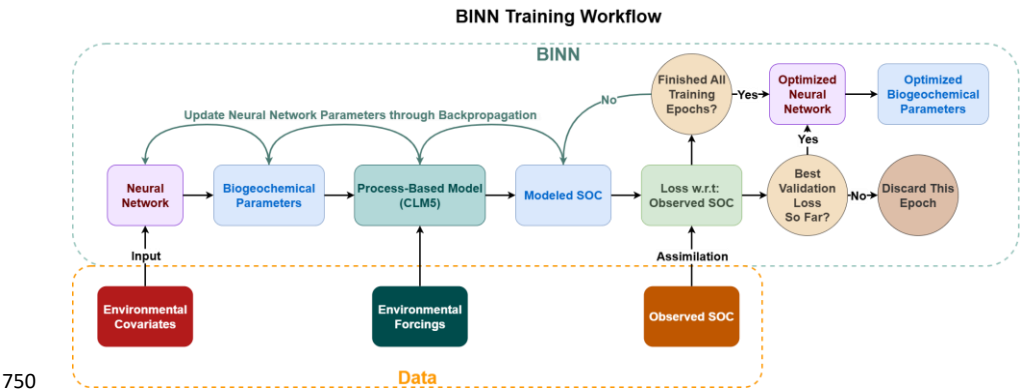
746 (a)



747

748

749 (b)



750

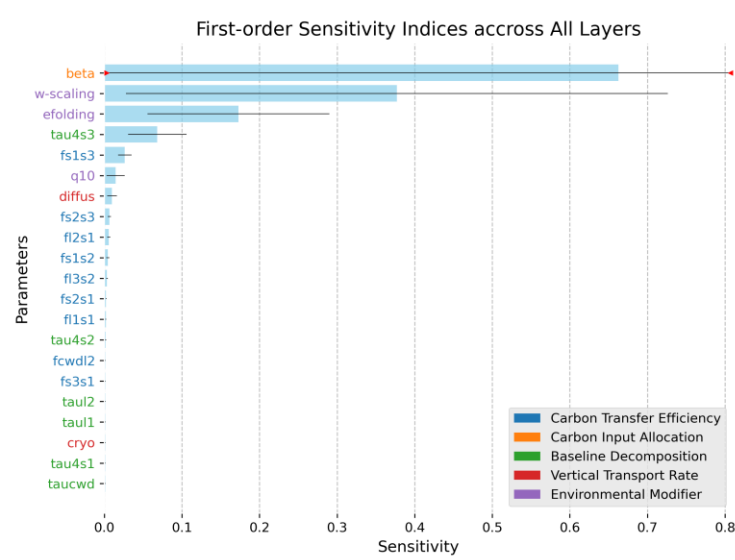
751

752 **Figure 1. Schematic diagram of BINN architecture and training workflow. (a) Detailed**

753 BINN structure showing the integration of neural networks with CLM5. A neural network

754 (purple) processes spatial coordinates (red) through a positional encoder and categorical

755 environmental covariates through an embedding layer. The network outputs are transformed via
756 a sigmoid activation function to generate 21 normalized parameters (grey). These parameters are
757 then scaled by their corresponding prior range (blue) and pass into CLM5 along with
758 environmental forcings to simulate SOC (teal). The model's performance is evaluated using a
759 smooth L1 loss function and a soft prior penalty on biogeochemical parameters (green). The
760 entire framework is differentiable, enabling end-to-end training through backpropagation (teal
761 arrow). (b) Overview of BINN training workflow. Environmental covariates at each site serve as
762 input to the neural network to predict biogeochemical parameters. These parameters, along with
763 environmental forcings, drive the process-based model (CLM5) to simulate SOC. The difference
764 between modeled and observed SOC is used to compute the loss function, which guides neural
765 network parameter updates through backpropagation (teal arrow). This training process
766 continues until reaching the maximum number of epochs or achieving optimal validation
767 performance.



768

769

770 **Figure 2: First-order sensitivity indices of biogeochemical parameters aggregated across all**

771 **soil layers.** Bars show mean sensitivity, with horizontal error bars representing ± 1 standard

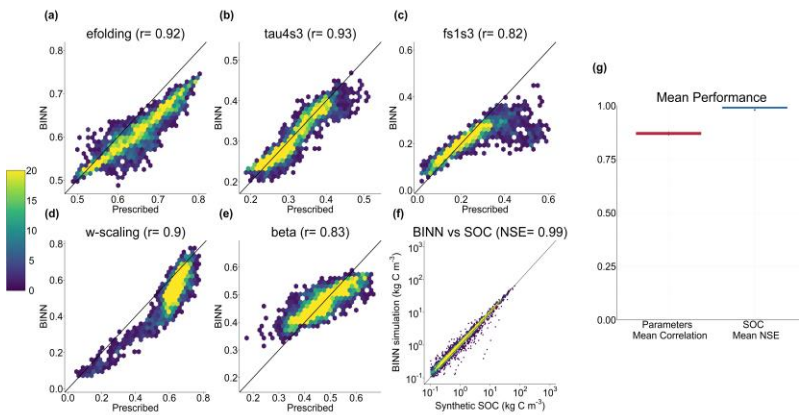
772 deviation. Red triangles indicate error bars clipped by the x-axis (uncertainty extends beyond the

773 displayed range). Parameters are ranked (y-axis) by decreasing sensitivity and color-coded by

774 their associated process component (e.g., K , A , B , ξ , V in Equation 7). The x-axis represents

775 sensitivity scores, indicating how changes in each parameter influence the model's performance.

776



777

778

779 **Figure 3: BINN's performance in the parameter recovery experiment.** Scatter plots

780 comparing the parameter values (a) "efolding", (b) "tau4s3", (c) "fs1s3", (d) "w-scaling", and (e)

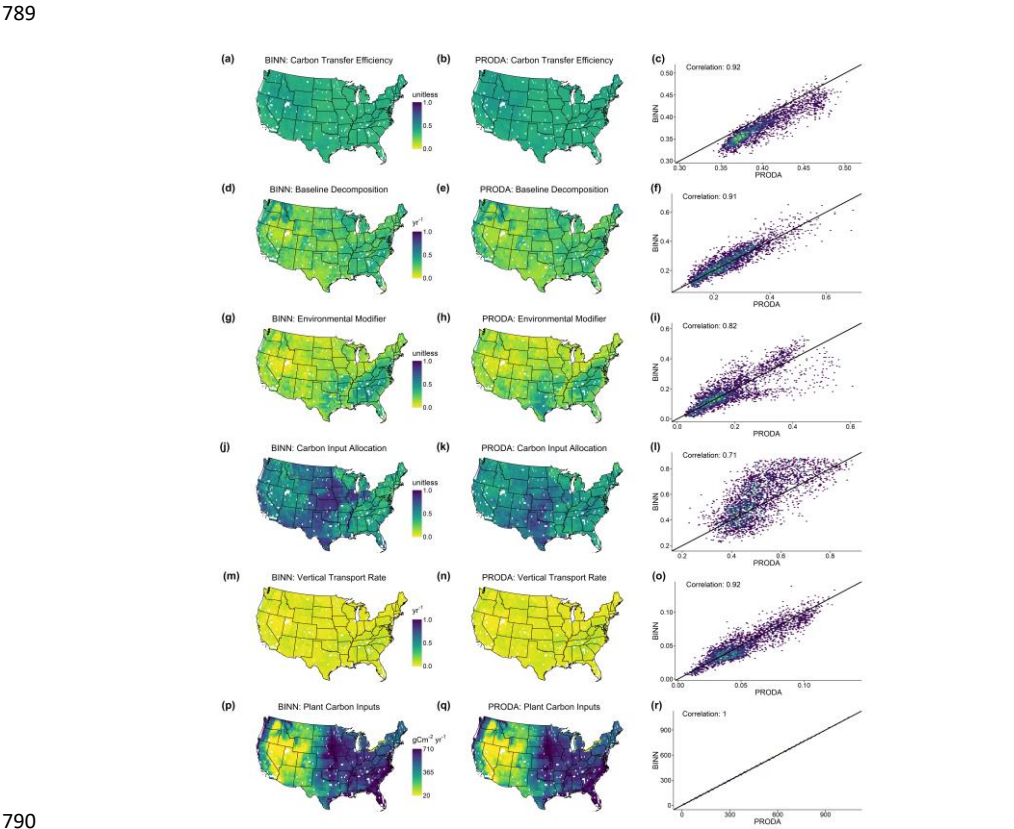
781 "beta" predicted by BINN against the prescribed parameter values. The color of each point in the

782 scatter plot represents the number of data points within each hexagonal bin. The correlation

783 coefficient between the predicted and prescribed parameter values is shown in the title of each

784 plot. (f) Comparison of the BINN-simulated and synthetic SOC values, with colors representing

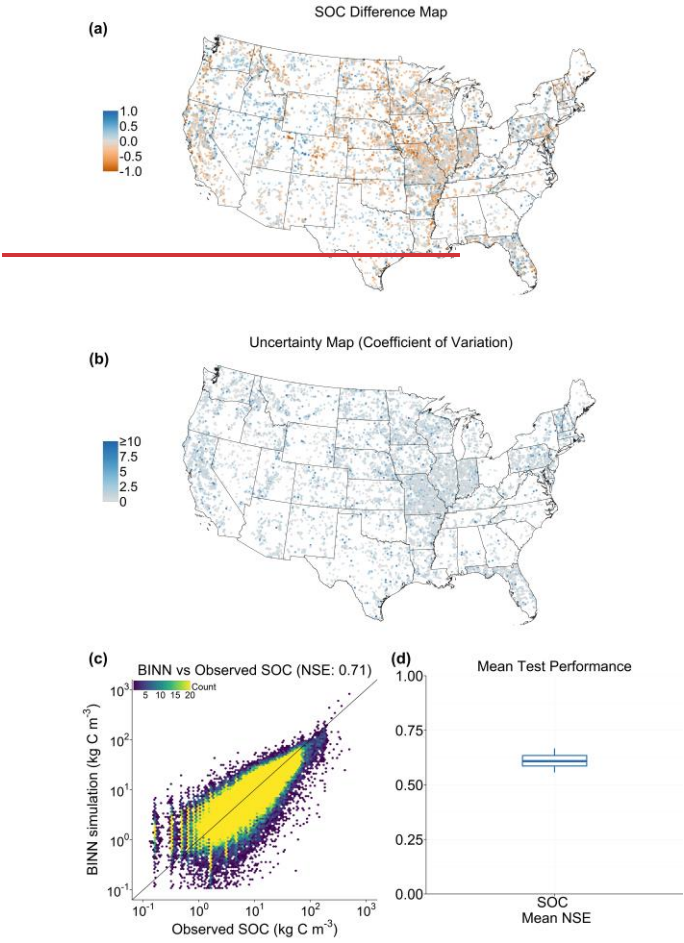
785 the density of points. (g) Mean performance of BINN in retrieving the five parameters across 10-
786 fold cross-validations, as measured by the correlation coefficient between the predicted and
787 prescribed parameter values as well as NSE of the simulated SOC values compared to the
788 synthetic SOC data.



791 **Figure 4: Comparison of the spatial patterns of model components retrieved by BINN and**
792 **PRODA.** Maps show BINN (left column panels) and PRODA (middle column panels) estimates
793 for six model components: carbon transfer efficiency (a–c), baseline decomposition (d–f),
794 environmental modifier (g–i), carbon input allocation (j–l), vertical transport rate (m–o), and

795 plant carbon inputs (p-r). Right-column panels (c, f, i, l, o, r) compare model components
796 between BINN (y-axis) and PRODA (x-axis) with 1:1 lines; panel insets report Pearson
797 correlation coefficients. Color scales are harmonized across BINN and PRODA; otherwise,
798 panel-specific scales are noted in the legend.

799



800

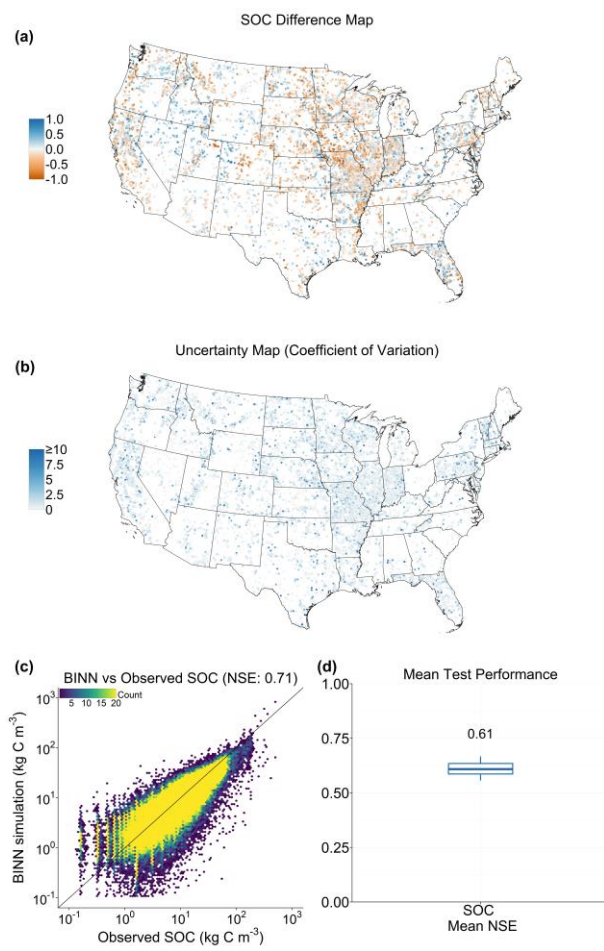


Figure 5: Comparison of simulated and observed SOC storage using BINN across 10-fold cross-validation training. (a) Map of normalized difference across all 10 folds. (b) Coefficient of variation of difference across 10 folds, indicating spatial uncertainty. (c) Observed vs. fold-average simulated SOC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

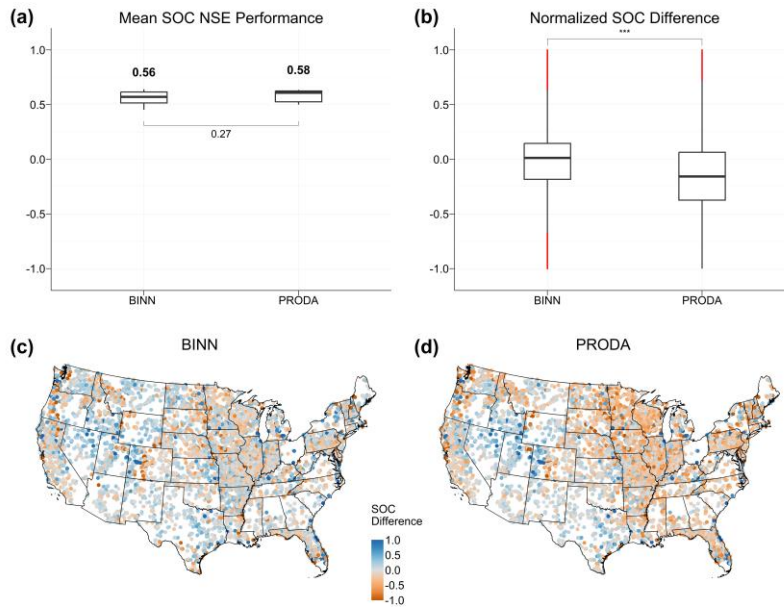
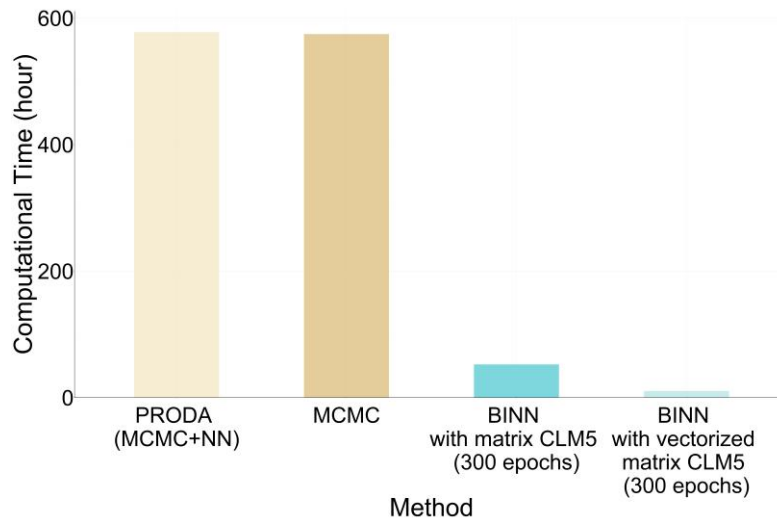


Figure 6. Side-by-side evaluation of BINN and PRODA across identical 10-fold test sets against the observed SOC. (a) Mean test NSE per fold for BINN and PRODA. NSE values are shown above the bars; the P value below the bars donates the difference in test NSE between the two methods is insignificant. (b) Mean test-set differences in simulated SOC between the two methods after normalization. The red dots are outliers, and the horizontal 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.



817
818 **Figure 7: Computational time to process 2,000 soil profiles with CLM5.** Wall-clock runtime
819 (hours) for PRODA (MCMC + NN), MCMC only, and for BINN using the original matrix
820 CLM5 and a vectorized CLM5 implementation. All runs were executed on the same platform.

821 **Acknowledgements**

822 This research is supported by AI-LEAF: “AI Institute for Land, Economy, Agriculture &
823 Forestry”, funded by the USDA National Institute of Food and Agriculture (NIFA) and the NSF
824 National AI Research Institutes Competitive Award (No. 2023-67021-39829). This research is
825 also partially supported by Schmidt Sciences programs, an AI2050 Senior Fellowship and an
826 Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship; NSF grants (DEB 2242034,
827 DEB 2406930, and DEB 2425290); the US Department of Energy’s Terrestrial Ecosystem
828 Sciences Grant DE-SC0023514, subcontract CW55561 from Oak Ridge National Laboratory to
829 Cornell University; the CALS Moonshot Seed Grant program; the “NYS Connects: Advancing
830 Markets for Producers” project, funded by the USDA and the New York State Department of
831 Environmental Conservation; an NSF Research Traineeship (NRT) fellowship in Digital Plant
832 Science (DGE-1922551), and the Air Force Office of Scientific Research (AFOSR, grants
833 FA9550-23-1-0322, FA9550-23-1-0569, FA9550-21-1-0316).

834 **Data and Code Availability**

835 The code and data for the BINN framework are available in the Zenodo repository at:

836 <https://doi.org/10.5281/zenodo.19237379.16541441> (~~Xu and Fan, 2025~~⁴⁶ (Xu and Fan, 2025⁴⁶))

837 **Author Contribution**

838 HX, YL, FT, and JF: Conceived this study; HX and JF: Developed the methods and performed
839 the analyses; HX: Prepared the first draft of the manuscript; HX, JF, FT, LJ, FY, BH, YS, CG,
840 and YL: Revised the manuscript.

841 **Conflicts of Interest**

842 The authors declare no competing interests.

843 **References**

- 844 1. Reichstein, M. *et al.* Deep learning and process understanding for data-driven Earth system science.
845 *Nature* **566**, 195–204 (2019).
- 846 2. Tan, T. *et al.* Importance of Terrain and Climate for Predicting Soil Organic Carbon Is Highly Variable
847 across Local to Continental Scales. *Environ. Sci. Technol.* **58**, 11492–11503 (2024).
- 848 3. Hengl, T. *et al.* SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE*
849 **12**, e0169748 (2017).
- 850 4. Cook-Patton, S. C. *et al.* Mapping carbon accumulation potential from global natural forest
851 regrowth. *Nature* **585**, 545–550 (2020).
- 852 5. Jian, J., Steele, M. K., Thomas, R. Q., Day, S. D. & Hodges, S. C. Constraining estimates of global soil
853 respiration by quantifying sources of variability. *Glob. Change Biol.* **24**, 4143–4159 (2018).
- 854 6. Willard, J., Jia, X., Xu, S., Steinbach, M. & Kumar, V. Integrating Scientific Knowledge with Machine
855 Learning for Engineering and Environmental Systems. *ACM Comput Surv* **55**, 66:1-66:37 (2022).
- 856 7. Karpatne, A., Jia, X. & Kumar, V. Knowledge-guided Machine Learning: Current Trends and Future
857 Prospects. <https://doi.org/10.48550/ARXIV.2403.15989> (2024) doi:10.48550/ARXIV.2403.15989.
- 858 8. Jin, Z. *et al.* Knowledge-Guided Machine Learning for Global Change Ecology Research. *Glob. Change*
859 *Biol.* **32**, e70742 (2026).
- 860 9. Wesselkamp, M., Moser, N., Kalweit, M., Boedecker, J. & Dormann, C. F. Process-Informed Neural
861 Networks: A Hybrid Modelling Approach to Improve Predictive Performance and Inference of Neural
862 Networks in Ecology and Beyond. *Ecol. Lett.* **27**, e70012 (2024).
- 863 10. Raissi, M., Perdikaris, P. & Karniadakis, G. E. Physics-informed neural networks: A deep learning
864 framework for solving forward and inverse problems involving nonlinear partial differential
865 equations. *J. Comput. Phys.* **378**, 686–707 (2019).

- 866 11. Liu, L. *et al.* Knowledge-guided machine learning can improve carbon cycle quantification in
867 agroecosystems. *Nat. Commun.* **15**, 357 (2024).
- 868 12. Lu, L. *et al.* Physics-informed neural networks with hard constraints for inverse design. Preprint at
869 <https://doi.org/10.48550/arXiv.2102.04626> (2021).
- 870 13. Hoedt, P.-J. *et al.* MC-LSTM: Mass-Conserving LSTM. in *Proceedings of the 38th International*
871 *Conference on Machine Learning* 4275–4286 (PMLR, 2021).
- 872 14. Kraft, B., Jung, M., Körner, M., Koirala, S. & Reichstein, M. Towards hybrid modeling of the global
873 hydrological cycle. *Hydrol. Earth Syst. Sci.* **26**, 1579–1614 (2022).
- 874 15. Shen, C. *et al.* Differentiable modelling to unify machine learning and physical models for
875 geosciences. *Nat. Rev. Earth Environ.* **4**, 552–567 (2023).
- 876 16. Fang, J., Bowman, K., Zhao, W., Lian, X. & Gentine, P. Differentiable Land Model Reveals Global
877 Environmental Controls on Ecological Parameters. Preprint at
878 <https://doi.org/10.48550/arXiv.2411.09654> (2024).
- 879 17. Chen, D. *et al.* Automating Crystal-Structure Phase Mapping: Combining Deep Learning with
880 Constraint Reasoning. Preprint at <https://doi.org/10.48550/arXiv.2108.09523> (2021).
- 881 18. Chen, D., Zhu, Y., Cui, X. & Gomes, C. Task-Based Learning via Task-Oriented Prediction Network
882 with Applications in Finance. in *Proceedings of the Twenty-Ninth International Joint Conference on*
883 *Artificial Intelligence* 4476–4482 (International Joint Conferences on Artificial Intelligence
884 Organization, Yokohama, Japan, 2020). doi:10.24963/ijcai.2020/617.
- 885 19. Min, Y. *et al.* Physically Informed Graph-Based Deep Reasoning Net for Efficient Combinatorial
886 Phase Mapping. in *2023 International Conference on Machine Learning and Applications (ICMLA)*
887 392–399 (2023). doi:10.1109/ICMLA58977.2023.00061.

- 888 20. ElGhawri, R. *et al.* Hybrid modeling of evapotranspiration: inferring stomatal and aerodynamic
889 resistances using combined physics-based and machine learning. *Environ. Res. Lett.* **18**, 034039
890 (2023).
- 891 21. Baghirov, Z., Jung, M., Reichstein, M., Körner, M. & Kraft, B. H2MV (v1.0): global physically
892 constrained deep learning water cycle model with vegetation. *Geosci. Model Dev.* **18**, 2921–2943
893 (2025).
- 894 22. Tramontana, G. *et al.* Partitioning net carbon dioxide fluxes into photosynthesis and respiration
895 using neural networks. *Glob. Change Biol.* **26**, 5235–5253 (2020).
- 896 23. Shi, Z. *et al.* Inverse analysis of coupled carbon–nitrogen cycles against multiple datasets at ambient
897 and elevated CO₂. *J. Plant Ecol.* **9**, 285–295 (2016).
- 898 24. Bree, R. van, Marcos, D. & Athanasiadis, I. Hybrid Phenology Modeling for Predicting Temperature
899 Effects on Tree Dormancy. Preprint at <https://doi.org/10.48550/arXiv.2501.16848> (2025).
- 900 25. Liu, Y., Kumar, M., Katul, G. G., Feng, X. & Konings, A. G. Plant hydraulics accentuates the effect of
901 atmospheric moisture stress on transpiration. *Nat. Clim. Change* **10**, 691–695 (2020).
- 902 26. Luo, Y. & Schuur, E. A. G. Model parameterization to represent processes at unresolved scales and
903 changing properties of evolving systems. *Glob. Change Biol.* **26**, 1109–1117 (2020).
- 904 27. Wesselkamp, M., Roberts, D. R. & Dormann, C. F. Identifying potential provenances for climate-
905 change adaptation using spatially variable coefficient models. *BMC Ecol. Evol.* **24**, 70 (2024).
- 906 28. Brunson, C., Fotheringham, A. S. & Charlton, M. E. Geographically Weighted Regression: A Method
907 for Exploring Spatial Nonstationarity. *Geogr. Anal.* **28**, 281–298 (1996).
- 908 29. Tao, F. *et al.* Deep Learning Optimizes Data-Driven Representation of Soil Organic Carbon in Earth
909 System Model Over the Conterminous United States. *Front. Big Data* **3**, (2020).
- 910 30. Tao, F. & Luo, Y. PROcess-Guided Deep Learning and DAta-Driven Modelling (PRODA). in *Land*
911 *Carbon Cycle Modeling* (CRC Press, 2022).

- 912 31. Tao, F. *et al.* Microbial carbon use efficiency promotes global soil carbon storage. *Nature* **618**, 981–
913 985 (2023).
- 914 32. Tao, F. *et al.* Convergence in simulating global soil organic carbon by structurally different models
915 after data assimilation. *Glob. Change Biol.* **30**, e17297 (2024).
- 916 33. Klemmer, K., Safir, N. & Neill, D. B. Positional Encoder Graph Neural Networks for Geographic Data.
917 Preprint at <https://doi.org/10.48550/arXiv.2111.10144> (2023).
- 918 34. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. in
919 *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* 249–
920 256 (JMLR Workshop and Conference Proceedings, 2010).
- 921 35. Bjorck, N., Gomes, C. P., Selman, B. & Weinberger, K. Q. Understanding Batch Normalization. in
922 *Advances in Neural Information Processing Systems* vol. 31 (Curran Associates, Inc., 2018).
- 923 36. Lu, X. *et al.* Full Implementation of Matrix Approach to Biogeochemistry Module of CLM5. *J. Adv.*
924 *Model. Earth Syst.* **12**, e2020MS002105 (2020).
- 925 37. Huang, Y. *et al.* Matrix approach to land carbon cycle modeling: A case study with the Community
926 Land Model. *Glob. Change Biol.* **24**, 1394–1404 (2018).
- 927 38. Lawrence, D. M. *et al.* The Community Land Model Version 5: Description of New Features,
928 Benchmarking, and Impact of Forcing Uncertainty. *J. Adv. Model. Earth Syst.* **11**, 4245–4287 (2019).
- 929 39. Luo, Y. *et al.* Matrix Approach to Land Carbon Cycle Modeling. *J. Adv. Model. Earth Syst.* **14**,
930 e2022MS003008 (2022).
- 931 40. Lu, X., Wang, Y.-P., Luo, Y. & Jiang, L. Ecosystem carbon transit versus turnover times in response to
932 climate warming and rising atmospheric CO₂ concentration. *Biogeosciences* **15**, 6559–6572 (2018).
- 933 41. Girshick, R. Fast R-CNN. Preprint at <https://doi.org/10.48550/arXiv.1504.08083> (2015).
- 934 42. Gal, Y. & Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in
935 Deep Learning.

- 936 43. Arnold, C. P. & Dey, C. H. Observing-Systems Simulation Experiments: Past, Present, and Future.
937 *Bull. Am. Meteorol. Soc.* **67**, 687–695 (1986).
- 938 44. Luo, Y. *et al.* Parameter identifiability, constraint, and equifinality in data assimilation with
939 ecosystem models. *Ecol. Appl.* **19**, 571–574 (2009).
- 940 45. Luo, Y. *et al.* Toward more realistic projections of soil carbon dynamics by Earth system models.
941 *Glob. Biogeochem. Cycles* **30**, 40–56 (2016).
- 942 46. ~~Xu, H. & Fan, J.~~[Hardyxu8067 & joshuafan](#). Hardyxu8067/BINN: v1.0.~~01~~. Zenodo
943 <https://doi.org/10.5281/zenodo.1923737920405976> (2026).
944