Supplement of Geosci. Model Dev., 19, 6777–6795, 2026
https://doi.org/10.5194/gmd-19-6777-2026-supplement

Geoscientific Model Development
Open Access

Supplement of

Biogeochemistry-Informed Neural Network (BINN v1.0) for improving accuracy of model prediction and scientific understanding of soil organic carbon storage

Haodi Xu et al.

Correspondence to: Feng Tao (feng.tao@ntu.edu.sg), Carla P. Gomes (gomes@cs.cornell.edu), and Yiqi Luo (yiqi.luo@cornell.edu)

Supplementary Tables and Figures

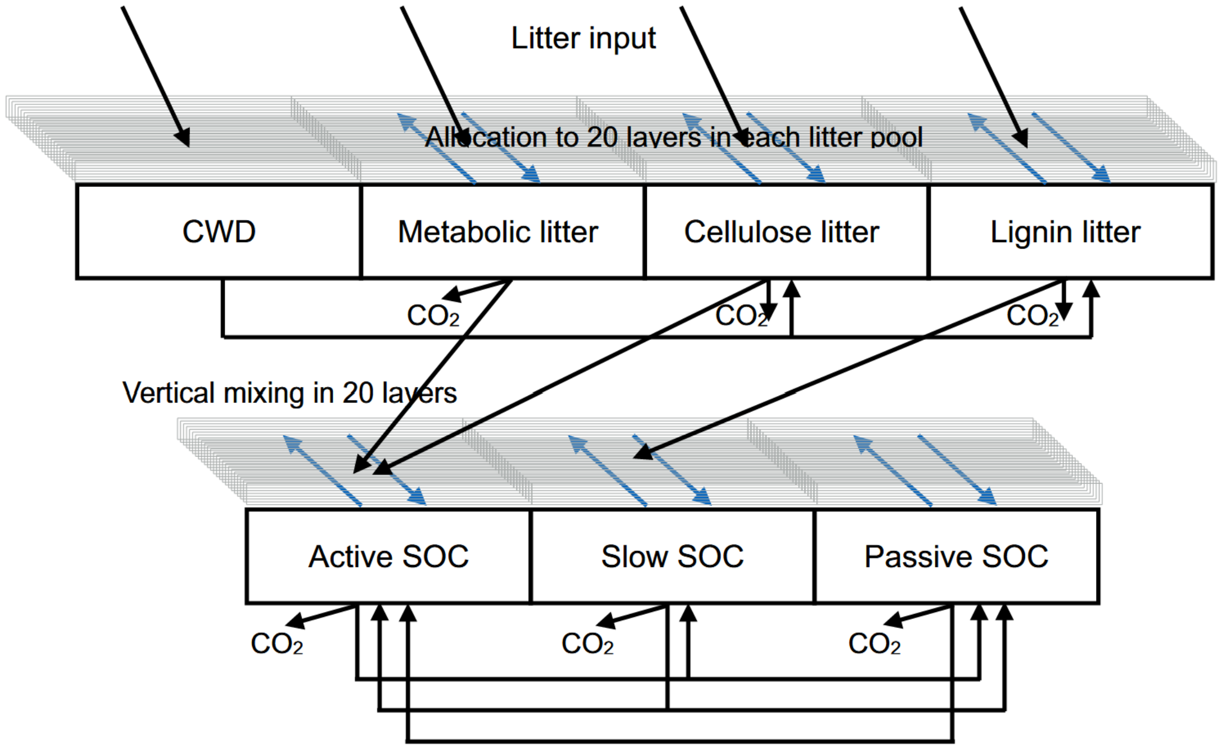


Figure S1: Model structures of CLM5 (Tao et al., 2024).

Table S1: Environmental Covariate Data as BINN input

No.	Variable Name	Data Source	Category	Description
1	Lon	WoSIS	Geography	Longitude
2	Lat			Latitude
3	Elevation	NOAA		Elevation
4	Abs_Depth_to_Bedrock	(Hengl et al. 2017)		Soil layer depth that reaches the bedrock
5	Occurrence_R_Horizon	WoSIS		Probability of occurrence of R horizon
6	nbedrock	CLM5 simulation		Soil layer number that reaches the bedrock
7	Koppen_Climate_2018	(Beck et al. 2018)	Climate	Koppen Climate Classification
8	BIO1	(Fick and Hijmans 2017)		Annual Mean Temperature
9	BIO2			Mean Diurnal Range
10	BIO3			Isothermality
11	BIO4			Temperature Seasonality
12	BIO5			Max Temperature of Warmest Month
13	BIO6			Min Temperature of Coldest Month
14	BIO7			Temperature Annual Range
15	BIO8			Mean Temperature of Wettest Quarter
16	BIO9			Mean Temperature of Driest Quarter
17	BIO10			Mean Temperature of Warmest Quarter
18	BIO11			Mean Temperature of Coldest Quarter
19	BIO12			Annual Precipitation
20	BIO13			Precipitation of Wettest Month
21	BIO14			Precipitation of Driest Month
22	BIO15			Precipitation Seasonality
23	BIO16			Precipitation of Wettest Quarter
24	BIO17			Precipitation of Driest Quarter
25	BIO18			Precipitation of Warmest Quarter
26	BIO19			Precipitation of Coldest Quarter

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

Table S2: 21 biogeochemical parameters in CLM5

No.	Name	Matrix Term	Corresponding Mechanism	Description	Unit	Prior Range	Reference
1	fl1s1	A	Microbial carbon use efficiency (CUE)	Transfer fraction, from metabolic litter to fast SOC	unitless	[0.1, 0.8]	Lawrence et al. 2019; Shi et al. 2018
2	fl2s1			Transfer fraction, from cellulose litter to fast SOC	unitless	[0.2, 0.8]	
3	fl3s2			Transfer fraction, from lignin litter to slow SOC	unitless	[0.2, 0.8]	
4	fs1s2			Transfer fraction, from fast SOC to slow SOC	unitless	[0.0001, 0.4]	
5	fs1s3			Transfer fraction, from fast SOC to passive SOC	unitless	[0.0001, 0.1]	
6	fs2s1			Transfer fraction, from slow SOC to fast SOC	unitless	[0.1, 0.74]	
7	fs2s3			Transfer fraction, from slow SOC to passive SOC	unitless	[0.0001, 0.1]	
8	fs3s1			Transfer fraction, from passive SOC to fast SOC	unitless	[0.0001, 0.9]	
9	fcwdl2			Transfer fraction, from coarse woody debris to cellulose litter	unitless	[0.5, 1]	
10	tau4cwd	K	Substrate decomposability	Turnover time of coarse woody debris	year	[1, 6]	Zhang et al. 2008
11	tau4l1			Turnover time of metabolic litter	year	[0.0001, 0.11]	
12	tau4l2			Turnover time of cellulose litter	year	[0.1, 0.3]	
13	tau4s1			Turnover time of fast SOC	year	[0.0001, 0.5]	Xu et al. 2016
14	tau4s2			Turnover time of slow SOC	year	[1, 10]	
15	tau4s3			Turnover time of passive SOC	year	[20, 400]	
16	q10	ξ	Environmental modifiers	Temperature sensitivity	unitless	[1.2, 3]	Davidson et al. 2005; Davidson et al. 2006
17	efolding			E-folding parameter to calculate depth scalar	meter	[0.1, 1]	Lawrence et al. 2019; Shi et al. 2018
18	w_scaling			Scaling factor to soil water scalar	unitless	[0.0001, 5]	
19	bio	V	Vertical transport	Bioturbation rate	m2/yr	[3×10^{-5} 16×10^{-4}]	Koven et al. 2015
20	cryo			Cryoturbation rate	m2/yr	[3×10^{-5} 5×10^{-4}]	
21	beta	I	Carbon input	Vertical distribution of carbon input	unitless	[0.5, 0.9999]	Lawrence et al. 2019; Shi et al. 2018

Table S3: Sensitivity Analysis of BINN Modeling Choices. "Extreme Para" denotes instances where extreme parameter values, specifically those less than 0.00001 or greater than 0.99999 prior to the sigmoid transformation, are predicted by the model during training, which results in instability.

Modelling Part	Test Choices	Test NSE	Training Stability
Original Settings	Smooth L1 + Leaky ReLU + Sigmoid	0.611 ± 0.036	-
Loss Function	L1 Loss	0.589 ± 0.045	-
	L2 Loss	0.615 ± 0.025	Extreme Para
	Para Squared Diff	0.616 ± 0.042	Extreme Para
Activation	ReLU	0.608 ± 0.038	-
	Hard Sigmoid	0.59 ± 0.047	-
Prior Range	Prior Range Increased by 10%	0.6 ± 0.04	-
	Prior Range Decreased by 10%	0.602 ± 0.039	-

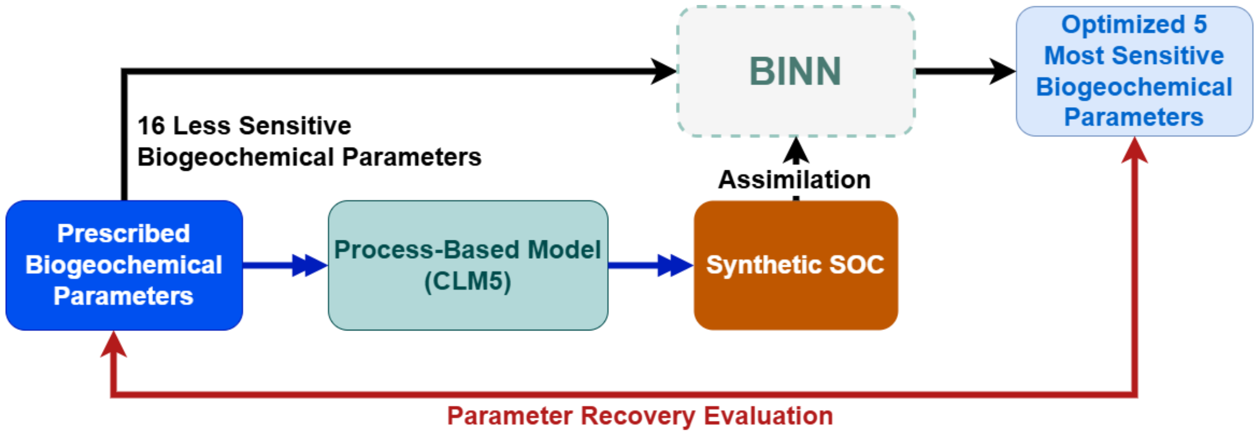
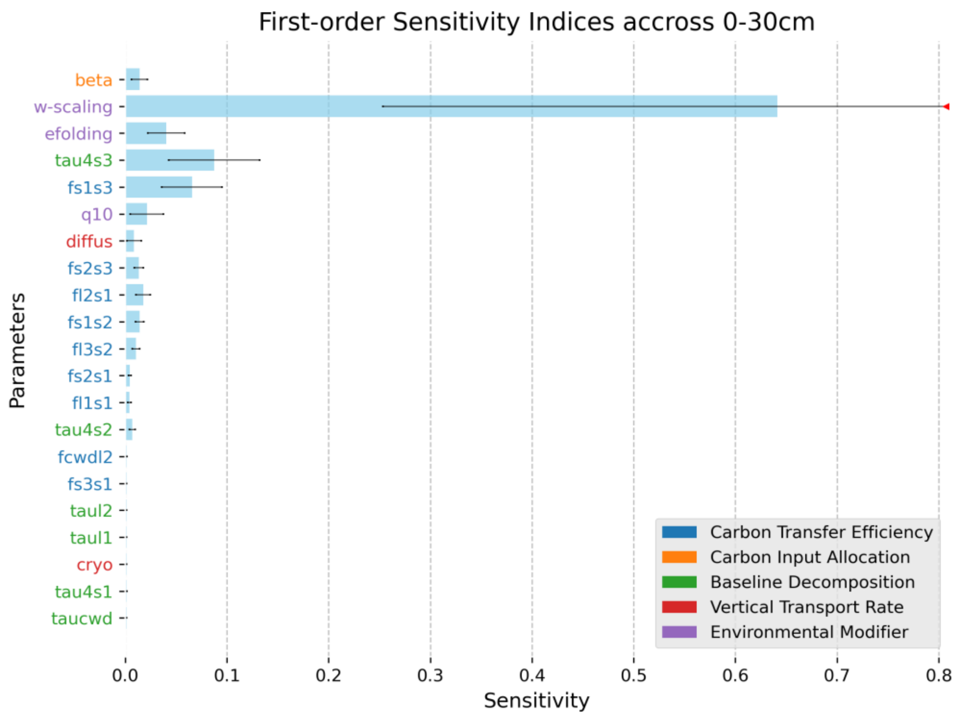


Figure S2: Schematic of the parameter recovery experiment to evaluate BINN's ability to retrieve the model parameters. The parameter recovery experiment involves three main steps. 1). Blue two-headed arrows: Synthesizing a SOC dataset using CLM5 with prescribed parameter values (21 parameters). 2). Single-headed arrows: Using the synthetic SOC dataset to train BINN to predict the 5 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 parameters of processes regulating SOC from observational data.

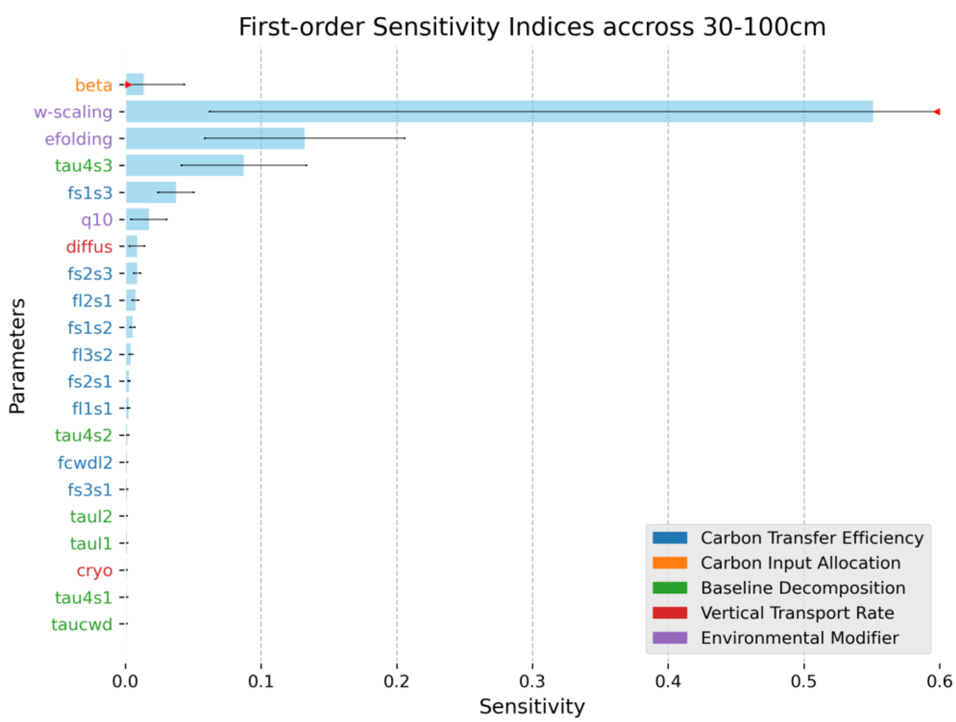
Table S4: Eight environmental forcings for CLM5

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

(a)


First-order Sensitivity Indices accross 0-30cm
beta
w-scaling
efolding
tau4s3
fs1s3
q10
diffus
fs2s3
fl2s1
fs1s2
fl3s2
fs2s1
fl1s1
tau4s2
fcwdl2
fs3s1
taul2
taul1
cryo
tau4s1
taucwd
Parameters
Carbon Transfer Efficiency
Carbon Input Allocation
Baseline Decomposition
Vertical Transport Rate
Environmental Modifier
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Sensitivity

(b)


First-order Sensitivity Indices accross 30-100cm
beta
w-scaling
efolding
tau4s3
fs1s3
q10
diffus
fs2s3
fl2s1
fs1s2
fl3s2
fs2s1
fl1s1
tau4s2
fcwdl2
fs3s1
taul2
taul1
cryo
tau4s1
taucwd
Parameters
Carbon Transfer Efficiency
Carbon Input Allocation
Baseline Decomposition
Vertical Transport Rate
Environmental Modifier
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Sensitivity

(c)

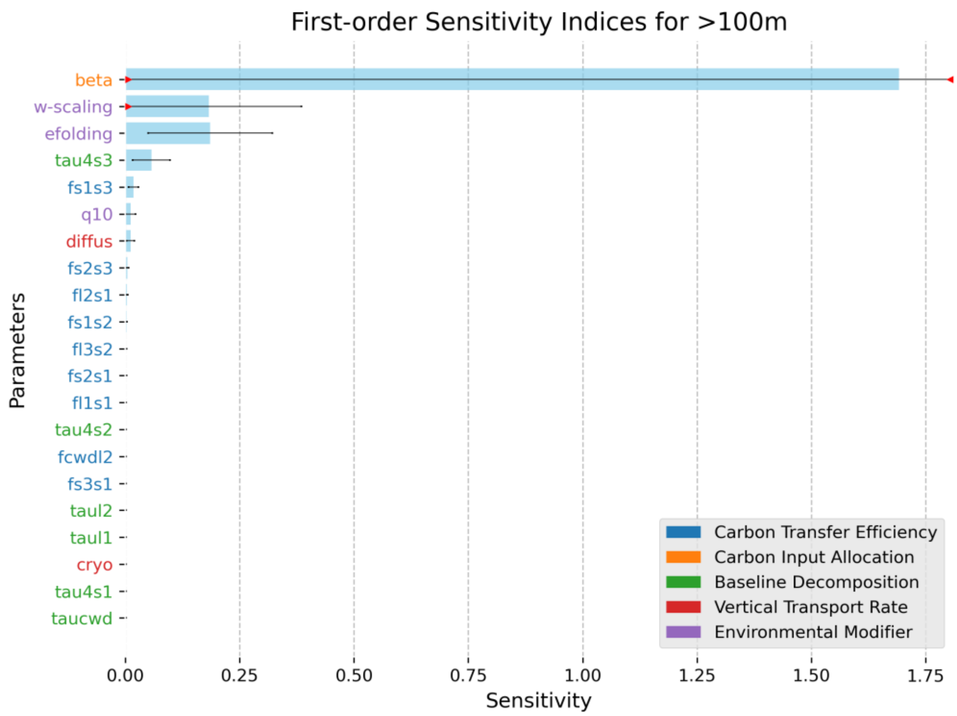


Figure S3: Sensitivity indices for CLM5 biogeochemical parameters across: (a) 0-30cm, (b) 30-100 cm, and (c) >100 cm.

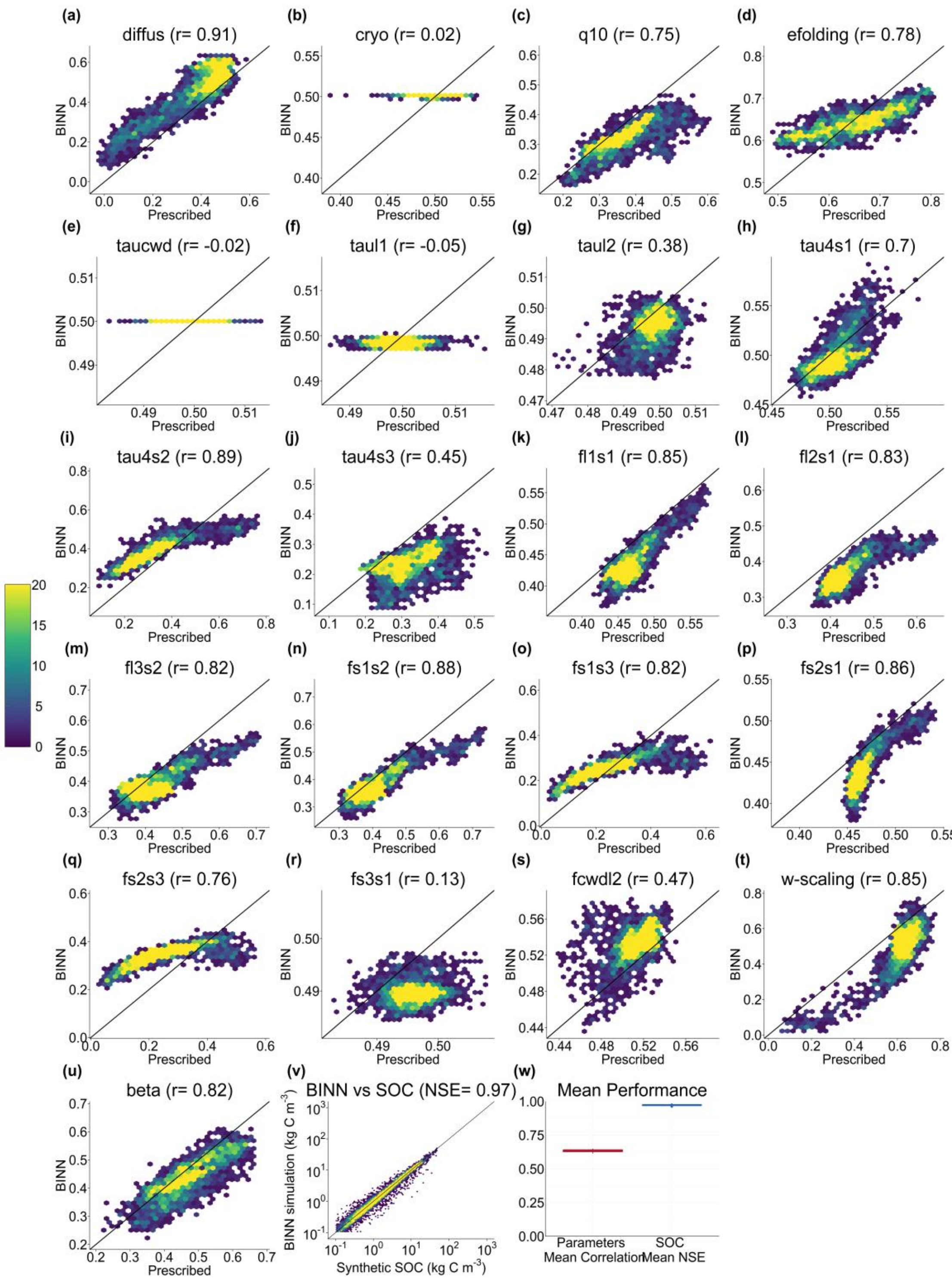


Figure S4: BINN parameter recovery for all CLM5 parameters. (a–u) Scatter plots of BINN-predicted versus prescribed values for each parameter. (v) Density scatter of BINN simulated SOC versus synthetic SOC. (w) Summary across a 10-fold cross-validation: mean correlation for all parameters (predicted vs. prescribed) and NSE for SOC simulations.

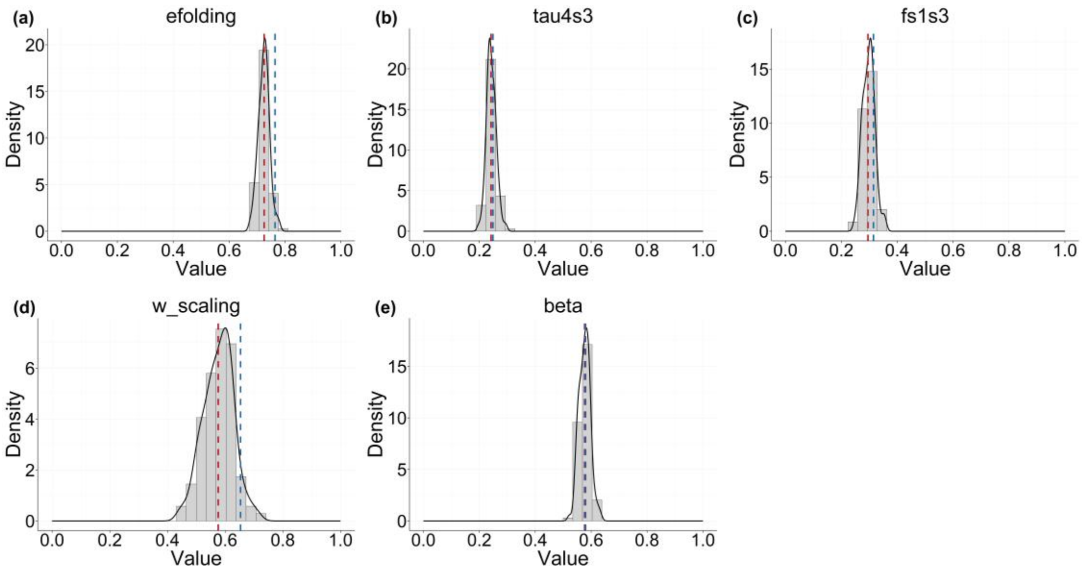


Figure S5: Uncertainty in posterior distributions of BINN predicted parameters in relation to Monte Carlo dropout in the parameter recovery experiment at one site. For each parameter, the marginal posterior is shown, with the vertical blue and red dashed lines indicating the prescribed ("true") and BINN's point estimate, respectively.

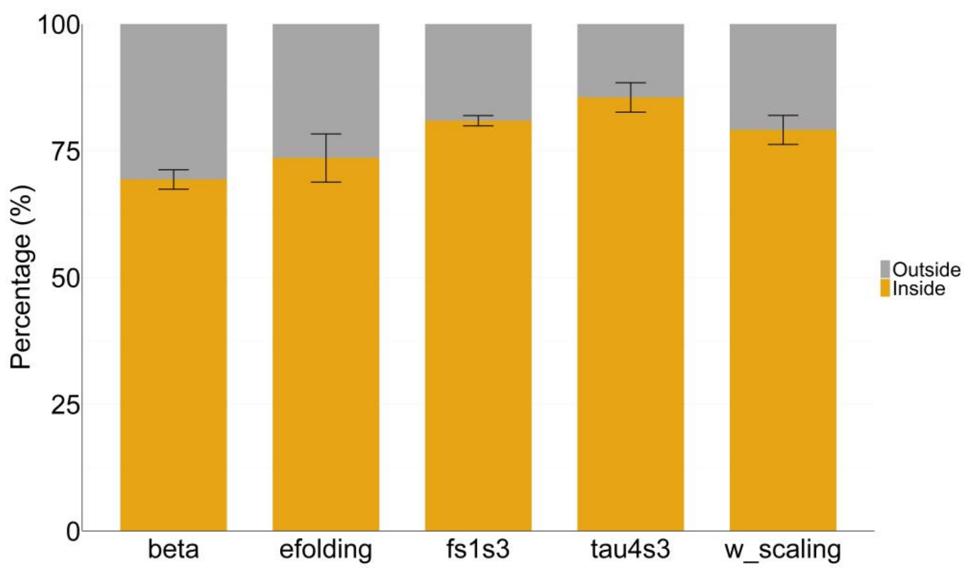


Figure S6: Coverage of parameters' posterior ranges of BINN on the prescribed parameters. The coverage was quantified by the percentage of prescribed ("true") parameter values that fall into the Monte Carlo–dropout posterior range across test sites. Error bars show the mean ± SD across the 10 cross-validation folds of the synthetic parameter-recovery experiment.

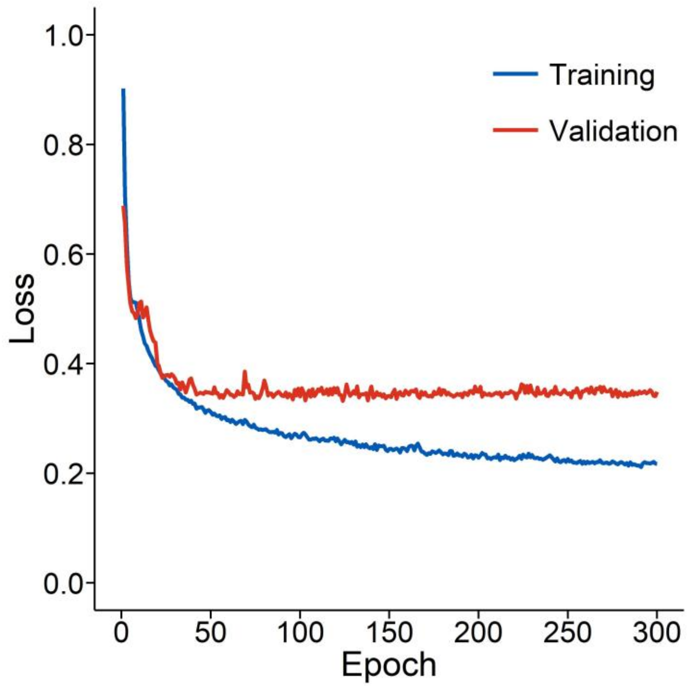


Figure S7: Training and validation inefficiency history for one cross-validation fold with median NSE value.

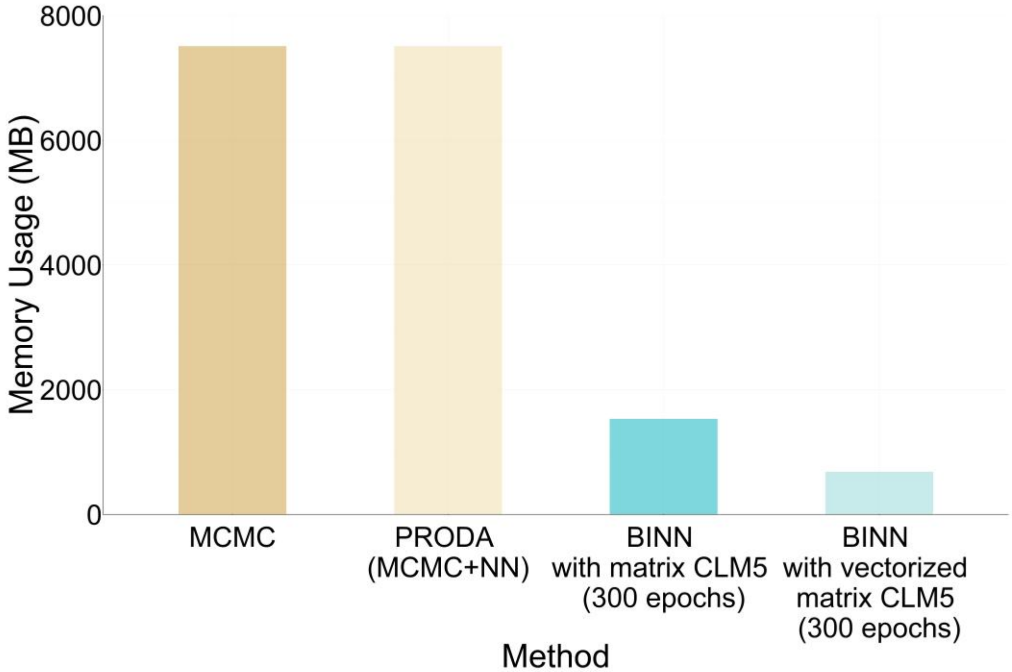


Figure S8: Comparative analysis of the peak memory (peak resident set size, RSS) required for integrating 2,000 soil profiles into the process-based model (CLM5).

Table S5: Comparison among different data assimilation methods

Criteria	BINN	PRODA	MCMC	Kalman Filter	Genetic Algorithm
Optimization Speed	Fast	Slow	Slow	Fast	Slow
Optimization Target	Parameters	Parameters	Parameters	States	Parameters
Recognition of Spatial/Temporal Heterogeneity	Yes	Yes	No	No	No
Uncertainty Assessment	No	No	Yes	Yes	No
Multisource Data	Yes	Yes	Yes	Yes	Yes
Key Refs	This study	Tao et al. 2020 Frontiers(Tao et al., 2020)	Hararuk et al. 2014 JGR(Hararuk et al., 2014)	Williams et al. 2005 GCB(Williams et al., 2005)	Barrett et al. 2002 GBC(Barrett, 2002)

Reference

Barrett, D. J.: Steady state turnover time of carbon in the Australian terrestrial biosphere, Global Biogeochemical Cycles, 16, 55-1-55–21, https://doi.org/10.1029/2002GB001860, 2002.

Hararuk, O., Xia, J., and Luo, Y.: Evaluation and improvement of a global land model against soil carbon data using a Bayesian Markov chain Monte Carlo method, Journal of Geophysical Research: Biogeosciences, 119, 403–417, https://doi.org/10.1002/2013JG002535, 2014.

Tao, F., Zhou, Z., Huang, Y., Li, Q., Lu, X., Ma, S., Huang, X., Liang, Y., Hugelius, G., Jiang, L., Doughty, R., Ren, Z., and Luo, Y.: Deep Learning Optimizes Data-Driven Representation of Soil Organic Carbon in Earth System Model Over the Conterminous United States, Frontiers in Big Data, 3, 2020.

Tao, F., Houlton, B. Z., Huang, Y., Wang, Y.-P., Manzoni, S., Ahrens, B., Mishra, U., Jiang, L., Huang, X., and Luo, Y.: Convergence in simulating global soil organic carbon by structurally different models after data assimilation, Global Change Biology, 30, e17297, https://doi.org/10.1111/gcb.17297, 2024.

Williams, M., Schwarz, P. A., Law, B. E., Irvine, J., and Kurpius, M. R.: An improved analysis of forest carbon dynamics using data assimilation, Global Change Biology, 11, 89–105, https://doi.org/10.1111/j.1365-2486.2004.00891.x, 2005.