

RESEARCH ARTICLE

Convergence in simulating global soil organic carbon by structurally different models after data assimilation

Feng Tao^{1,2}  | Benjamin Z. Houlton^{1,3}  | Yuanyuan Huang⁴  | Ying-Ping Wang⁵  |
Stefano Manzoni⁶  | Bernhard Ahrens⁷  | Umakant Mishra^{8,9}  | Lifen Jiang¹⁰  |
Xiaomeng Huang²  | Yiqi Luo¹⁰ 

¹Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York, USA

²Department of Earth System Science, Ministry of Education Key Laboratory for Earth System Modelling, Institute for Global Change Studies, Tsinghua University, Beijing, China

³Department of Global Development, Cornell University, Ithaca, New York, USA

⁴Key Laboratory of Ecosystem Network Observation and Modeling, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China

⁵CSIRO Environment, Clayton South, Victoria, Australia

⁶Department of Physical Geography and Bolin Centre for Climate Research, Stockholm University, Stockholm, Sweden

⁷Max Planck Institute for Biogeochemistry, Jena, Germany

⁸Computational Biology and Biophysics, Sandia National Laboratories, Livermore, California, USA

⁹Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, Emeryville, California, USA

¹⁰Soil and Crop Sciences Section, School of Integrative Plant Science, Cornell University, Ithaca, New York, USA

Correspondence

Feng Tao, Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, NY, USA.
Email: phx.tao@gmail.com

Xiaomeng Huang, Department of Earth System Science, Ministry of Education Key Laboratory for Earth System Modelling, Institute for Global Change Studies, Tsinghua University, Beijing, China.
Email: hxm@tsinghua.edu.cn

Funding information

National Natural Science Foundation of China, Grant/Award Number: 42125503 and 42075137; National Key Research and Development Program of China, Grant/Award Number: 2020YFA0607900, 2020YFA0608000, 2022YFE0195900 and 2021YFC3101600; NYS Connects: Climate Smart Farms & Forestry Project; Sandia National Laboratories, Grant/Award Number: DE-NA-0003525; European Union's Horizon 2020 Research and Innovation Programme, Grant/Award Number: 101001608; Horizon Europe project AI4SoilHealth, Grant/

Abstract

Current biogeochemical models produce carbon–climate feedback projections with large uncertainties, often attributed to their structural differences when simulating soil organic carbon (SOC) dynamics worldwide. However, choices of model parameter values that quantify the strength and represent properties of different soil carbon cycle processes could also contribute to model simulation uncertainties. Here, we demonstrate the critical role of using common observational data in reducing model uncertainty in estimates of global SOC storage. Two structurally different models featuring distinctive carbon pools, decomposition kinetics, and carbon transfer pathways simulate opposite global SOC distributions with their customary parameter values yet converge to similar results after being informed by the same global SOC database using a data assimilation approach. The converged spatial SOC simulations result from similar simulations in key model components such as carbon transfer efficiency, baseline decomposition rate, and environmental effects on carbon fluxes by these two models after data assimilation. Moreover, data assimilation results suggest equally effective simulations of SOC using models following either first-order or Michaelis–Menten kinetics at the global scale. Nevertheless, a wider range of data

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *Global Change Biology* published by John Wiley & Sons Ltd.

Award Number: 101086179; Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship, a Schmidt Futures program; USDA National Institute of Food and Agriculture (NIFA) and NSF National AI Research Institutes Competitive Award, Grant/Award Number: 2023-67021-39829; US Department of Energy, Terrestrial Ecosystem Sciences Grant, Grant/Award Number: DE-SC0023514; US National Science Foundation, Grant/Award Number: DEB 1655499 and DEB 2242034

with high-quality control and assurance are needed to further constrain SOC dynamics simulations and reduce unconstrained parameters. New sets of data, such as microbial genomics-function relationships, may also suggest novel structures to account for in future model development. Overall, our results highlight the importance of observational data in informing model development and constraining model predictions.

KEYWORDS

big data assimilation, deep learning, inter-model uncertainty, model parameterization, model structure, soil organic carbon

1 | INTRODUCTION

Soils store more carbon than the atmosphere and vegetation combined (Ciais et al., 2014; Jackson et al., 2017). A small change in soil carbon storage can significantly impact the atmospheric carbon dioxide concentration and the future trajectory of climate. Substantial research has been conducted to understand the factors underlying the formation of soil organic carbon (SOC) and its persistence. While there is a general agreement that the SOC balance depends on plant carbon input as the source of SOC and organic matter decomposition as the main SOC loss pathway, there are two contrasting paradigms on the regulation of decomposition. The conventional paradigm focuses on chemical recalcitrance and physical protection as the key factors controlling decomposition and, thus, CO₂ emissions back to the atmosphere (Schmidt et al., 2011). A more recent paradigm focuses instead on soil microorganisms and soil carbon stabilization as the key determinants in partitioning carbon inputs between accumulation and loss (Bradford et al., 2016; Cotrufo et al., 2013, 2015; Tao et al., 2023). These two paradigms are the conceptual foundation of two classes of process-based models used to simulate global SOC dynamics (Table 1). Because these model classes have distinctive structures that reflect different underlying theories and assumptions on soil carbon dynamics (Chandel et al., 2023), large differences in the simulated SOC emerge among models, leading to highly uncertain predictions (Wieder et al., 2018). Diverging simulations of SOC storage and its spatial distributions across the globe hinder a better understanding of the soil carbon cycle and its feedback to climate change (Ciais et al., 2014; Luo et al., 2016; Todd-Brown et al., 2013).

In simulating SOC dynamics, state-of-the-art process-based models following the two paradigms differ structurally regarding soil carbon pool classification, SOC decomposition kinetics, and representation of carbon transfer processes (Table 1). Soil organic carbon can be separated into conceptual pools with different turnover rates that reflect heterogeneity in their decomposition rates. For example, models derived from the Century model (Parton et al., 1987) that center their simulations around the “pool turnover” paradigm (Luo, 2022; Schimel, 2023) differentiate substrates according to turnover times, with labile substrates that cycle rapidly (i.e., active SOC) and chemically or physically protected pools that cycle

slowly (i.e., slow and passive SOC). In contrast, recently formulated process-based models that highlight the role of microbial processes define carbon pools as measurable entities that can be validated with field observations (Abramoff et al., 2022)—for example, microbial biomass, dissolved organic carbon, particulate organic carbon, and mineral-associated organic carbon (Table 1).

In representing SOC decomposition, a theory developed back in the 1940s (Jenny, 1941) and consolidated in the 1980s (Parton et al., 1988) portends that organic matter decay in soils follows first-order kinetics: $\frac{dSOC}{dt} \propto -k \times SOC$, where the loss rate of SOC (i.e., k) is independent of its pool size (i.e., SOC). Therefore, with this formulation, the SOC storage changes over time is proportional to its pool size (Forney & Rothman, 2012). With increasing evidence pointing to soil microorganisms as a key factor in soil carbon dynamics, a newer generation of models has explored the possibility of nonlinearity in SOC decomposition (Allison et al., 2010; Georgiou et al., 2017; Schimel & Weintraub, 2003; Wang et al., 2021) (Table 1). Among various nonlinear functions that can be used to describe decomposition, the Michaelis–Menten kinetics (i.e., $\frac{dSOC}{dt} \propto -v \frac{ENZ \times SOC}{K + SOC}$) considers the interplay between the substrate (i.e., SOC) and the extracellular enzymes (i.e., ENZ) that catalyze the decomposition of organic matter. While not new (Briggs & Haldane, 1925), this formulation is now being frequently used in soil carbon cycle models (Schimel & Weintraub, 2003; Wilson & Gerber, 2021). Specifically, parameter v specifies the maximum SOC decomposition rate at its saturated content for a given enzyme content. The inverse of the Michaelis–Menten constant (K) specifies the enzyme's affinity for its substrate in a catalyzed reaction.

Process-based models also differ in allocating the decomposed carbon to other carbon pools or heterotrophic respiration as CO₂ (Table 1). While soil microbes mineralize SOC into CO₂ through their metabolism, transfers of decomposed carbon from one pool to another could result from either an exclusive effect of microbial processes or an integrative effect of biological, chemical, and physical reactions (i.e., including both microbial and non-microbial transfer). Specifically, when a model explicitly defines a microbial biomass carbon pool, carbon received by this pool is partitioned according to the microbial carbon use efficiency (CUE)—that is, the ratio of carbon assimilated in new biomass over the total substrate carbon uptake (Geyer et al., 2016; Manzoni et al., 2018; Tao et al., 2023). Correspondingly, carbon transfers among different soil

compartments that happen without microbial carbon assimilation can be interpreted as results from other biochemical processes (e.g., microbial exudation and mortality) or organo-mineral interactions (Tao et al., 2023). In contrast, for models without explicit representation of microbial biomass and assimilation processes, carbon transfer implicitly integrates the effects of both microbial physiology and other chemical or physical reactions. Depending on the model structure, a range of relations between long-term SOC and microbial traits, such as CUE or carbon inputs to soils, emerge (Georgiou et al., 2017; He et al., 2023; Wutzler & Reichstein, 2008).

In addition to structural differences among varieties of process-based models, parameter values that quantify the strength and represent properties of different processes in the soil carbon cycle also contribute to the uncertainty of model simulations (Luo & Schuur, 2020), especially when they are not well constrained by observations. Most current Earth system models adopt the Century-type model structure using first-order SOC decomposition kinetics.

Notwithstanding their structural similarity, varying parameter values among different models contribute to the divergent estimates of SOC storage both at the site level and across the globe (Luo et al., 2015; Todd-Brown et al., 2013). Moreover, the same model with different choices of parameter values (i.e., parameterization) could also generate varying patterns between SOC and key model components, such as microbial CUE (Tao et al., 2023) and plant carbon input (Tao et al., 2024). However, choices of parameter values and model structure are not fully independent in affecting model simulation: Different model structures can, in some cases, converge to similar results in the long term via parameter adjustments. For example, the Michaelis–Menten kinetics, when the affinity of the enzyme for its substrate is extremely low, such that the Michaelis–Menten constant is much higher than the substrate concentration ($[K]SOC$), the nonlinear decomposition kinetics will converge to linear kinetics with respect to the substrate (Lasaga, 1998; Wilson & Gerber, 2021).

TABLE 1 Major differences in simulating soil carbon cycle among process-based models following two paradigms for SOC loss pathways (see also Figure 1).

	Pool turnover-centered paradigm	Microbe-centered paradigm
Carbon pool classification	Carbon pools are conceptually defined turnover time (i.e., average time a carbon compound stays in the soil). These models usually do not explicitly define microbe-related carbon pools such as microbial biomass, dissolved organic carbon, and enzyme	Carbon pools are defined by their functions in the soil carbon cycle. These models usually explicitly define microbe-related carbon pools such as microbial biomass, dissolved organic carbon, and enzyme by representing specific microbial processes such as assimilation, catabolism, mortality, and enzymatic reactions
Decomposition kinetics	First-order kinetics. Decomposition rate is only dependent on the donor pool size (i.e., the amount of substrate being decomposed)	Microbial explicit kinetics, such as Monod, Michaelis–Menten, reverse Michaelis–Menten, and logistic type kinetics. Decomposition rate is a function of both donor pool size and catalysts
Carbon transfer scheme	Organic carbon is transferred among conceptual pools, and CO ₂ is emitted whenever a transfer happens.	Organic carbon is transferred among functionally explicit pools, and CO ₂ is emitted only when microorganisms assimilate carbon from substrates in metabolism.
Model example used in this contribution	Community Land Model version 5 (CLM5) (Lawrence et al., 2019)	CarbOn cycle and Microbial PARTitioning Soil model (COMPAS) (Tao et al., 2023)

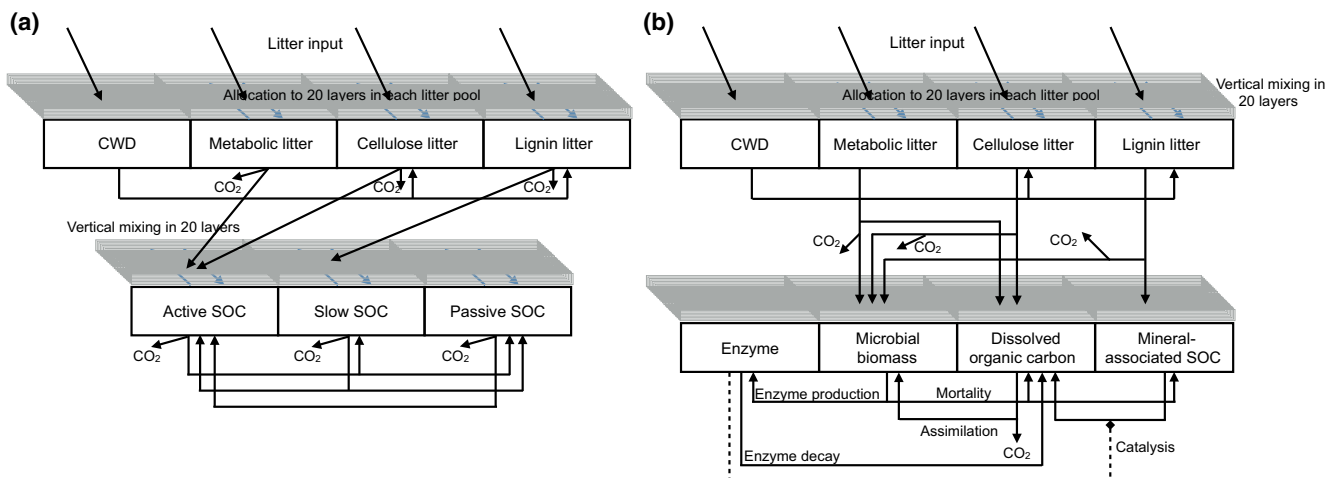


FIGURE 1 Distinctive model structures of CLM5 (a) and COMPAS (b). CWD, coarse wood debris; SOC, soil organic carbon.

While simulations by structurally distinctive models with different parameter values present a range of possibilities under specific theories and assumptions, calibrating model simulations against observational data helps identify the most probable mechanistic explanation that fits reality. Data assimilation is a suite of techniques that compare the model simulation results with different parameter values against observed counterparts and adjust the model parameter values to the set with which the process-based model simulations best-fit observations (Luo et al., 2011). Conventional data assimilation techniques such as the Bayesian inference-based Markov Chain Monte Carlo (MCMC) method have been used at the site level to tune process-based models for better performance in simulating soil carbon cycle (Li et al., 2016; Xu et al., 2006). Recently, the newly developed PROcess-guided deep learning and Data-driven modeling (PRODA) approach (Tao & Luo, 2022) integrates the site-level MCMC-based data assimilation results with deep learning to optimize the model parameter values for global SOC simulations and reveals key mechanisms underlying global SOC storage (Tao et al., 2020, 2023).

To investigate the roles of model structure versus parameters in causing the large inter-model uncertainty, we leverage two models (i.e., CLM5 and COMPAS; Figure 1; see Section 2 for detailed descriptions) that are structurally different in describing carbon pools, decomposition kinetics, and carbon transfer pathways in a data assimilation framework. We hypothesize that being informed by a common observational SOC dataset using the PRODA approach, simulations of global SOC by CLM5 and COMPAS can converge. Despite structural differences among models, we expect that well-calibrated parameters representing key processes in the soil carbon cycle will contribute to converging SOC simulations. Meanwhile, results of PRODA-optimized model simulations can also identify the most probable model structure that best fits observed SOC data across the globe.

2 | MATERIALS AND METHODS

2.1 | Global vertical soil organic carbon profiles

We obtained SOC data in globally distributed soil profiles from the World Soil Information Service (WoSIS) and other data sources. WoSIS compiled soil data, after quality assessment, from soil profiles distributed across 173 countries (Batjes et al., 2020). The 2019 snapshot of the WoSIS dataset consists of 111,380 soil profiles with SOC content information (unit: gCkg^{-1} soil). We estimated the SOC stock (gCm^{-3}) by $\text{SOC Stock} = \text{SOC Content} \times \text{BD}$ (Yigini et al., 2018), where BD is the bulk density of soil (gm^{-3}). Note that SOC stock was multiplied by $1 - \frac{G}{100}$ to account for the volumetric coarse fragment fraction (G, unit: %) at each grid of the global map (data source: SoilGrids, <https://soilgrids.org>). When the measured bulk density was absent in the dataset, we used a pedo-transfer function to estimate it (Grigal et al., 1989; Yigini et al., 2018): $\text{BD} = \alpha + \beta \times \exp(-\gamma \times \text{OM})$, where OM is organic matter, calculated as $\text{SOC} \times 1.724$, with SOC content in percent (%); α , β ,

and γ are fitting parameters. After fitting data of WoSIS (i.e., 78,913 layers from 16,248 profiles that simultaneously recorded bulk density and SOC content) to this equation, we obtained that $\alpha=0.32$, $\beta=1.30$, and $\gamma=0.0089$. The pedo-transfer function explained 55% of the variation in the bulk density. Using the pedo-transfer function does not introduce substantial extra uncertainties in the SOC stock database. At those 16,248 soil sampling sites that recorded bulk density and, thus, SOC stocks, we compared the field measurements with their corresponding values estimated from the pedo-transfer function. The pedo-transferred estimates explained 68% of variation in field-measured SOC stocks. We conducted a *t*-test to quantify whether the difference between field-measured and pedo-transferred SOC stocks (i.e., pedo-transferred estimates minus field measurements) differ from 0. The results suggested that the mean difference is -0.05 kgCm^{-3} , but such a small bias was not significantly different from 0 (p -value = .30, $df = 78,192$, $t = -1.03$).

In addition, we obtained an additional dataset of SOC stock in permafrost regions, which combined the data from a previous study (Mishra et al., 2020) and the Northern Circumpolar Soil Carbon Database (NCSCD) (Hugelius et al., 2013). This dataset contained 2546 soil profiles with SOC stock (gCm^{-3}) information for permafrost regions in North America, northern Eurasia, and Qinghai-Tibet Plateau. Combining this dataset with the WoSIS dataset, in total, we obtained data from 113,926 soil profiles as the raw data. The geographical distributions of all soil profiles are shown in Figure S1.

Not all the soil profiles are used in this study. We pre-processed the 113,926 SOC profiles to ensure the quality of the data before we conducted our analysis. We first excluded SOC profiles with no more than two observation layers or the maximum observation depths of no deeper than 50 cm from this study as such data do not provide enough information on key processes underlying SOC storage. After this screening, we retained 72,377 profiles.

To further examine the suitability of the data for model optimization, we conducted data assimilation for each of the 72,377 SOC vertical profiles with both the Community Land Model version 5 (CLM5) and the CarbOn cycle and Microbial PARTitioning Soil model (COMPAS) using the Markov Chain Monte Carlo (MCMC) method. Model structures of CLM5 and COMPAS are described in Sections 2.2 and 2.3, respectively. The method of data assimilation is briefly described in Section 2.4 below and in detail by Tao et al. (2020).

We used two statistics, that is, Gelman-Rubin (G-R) statistic and Nash-Sutcliffe modeling efficiency (NSE) coefficient, to ensure the quality of model calibration against SOC data along the vertical profiles. We calculated the G-R value (Gelman et al., 2014) for each of the SOC profiles to test the convergence of the site-level data assimilation results after running three independent series of MCMC simulations (see Section 2.6 for details of MCMC). A G-R value approaching 1.0 suggests well-converged data assimilation results. A large G-R value, in contrast, indicates inconsistent data assimilation results from these independent MCMC simulations, and such results may not be trusted. Therefore, we set a threshold of $G-R=1.05$ and excluded SOC profiles with $G-R > 1.05$, with 66,935 profiles remained for CLM5 and 59,476 remained for COMPAS to be included in further analyses. We found that it was more difficult

for the independent MCMC simulations to converge when using COMPAS model than using CLM5 in data assimilation, probably because of the nonlinearity and a lack of vertical transport for the mineral-soil carbon part in COMPAS (see Section 2.3). Thus, the final adopted profiles for COMPAS are fewer than those for CLM5.

We used the NSE coefficient (Janssen & Heuberger, 1995) (NSE) to evaluate the effectiveness of retrieving information from observations by process-based models. NSE is expressed as:

$$\text{NSE} = 1 - \frac{\sum (\text{obs}_i - \text{mod}_i)^2}{\sum (\text{obs}_i - \overline{\text{obs}_i})^2}. \quad (1)$$

At the site-level data assimilation, the summation in Equation 1 extends to all sampling depths at a given site. A value of NSE close to 1 indicates that SOC distributions with depth can be well captured by process-based models so that information contained in the observations can be retrieved to evaluate processes underlying SOC storage. In contrast, a small value of NSE indicates that the model cannot capture the variability in the data, suggesting that such SOC vertical profiles may not offer enough information on the investigated processes underlying SOC storage. While it is possible that the negative NSE values could also result from the fact that process-based models are still not sophisticated enough to capture extreme irregularities in observations, we set the threshold as $\text{NSE} = 0.0$ to include as many profiles as possible in the analysis. Moreover, the soil profiles included in this study are inclusive to diverse vertical shapes in SOC. For example, for the COMPAS model, 66.2% of the 57,267 profiles show monotonically decreasing SOC stocks with soil depths, 4.4% of them record the highest SOC stock at the middle of the soil depths and 29.4% of them show zigzagged SOC stock with increasing soil depths (Tao et al., 2023). Eventually, only 4% (2209 out of 59,476) and 6% (4004 out of 66,935) of the profiles for CLM5 and COMPAS, respectively, were excluded due to negative NSE values. Moreover, we randomly selected a subset of these excluded SOC profiles to visually cross-check their shapes. We found that the thresholds are effective for controlling the suitability of data.

After all the data pre-processing procedures, we obtained data assimilation results from 62,931 soil profiles for CLM5 and 57,267 soil profiles for COMPAS with which we estimated global SOC storage and its components. Our data pre-processing criteria did not cause significant discrimination against profiles belonging to specific soil orders or ecosystems or different vertical shapes (Tao et al., 2023). Meanwhile, the coverages of selected soil profiles across multi-dimensional covariate spaces do not differ much between CLM5 and COMPAS (Figure S2). Thus, the main conclusions drawn from this study are unlikely influenced by our data pre-processing criteria.

2.2 | Model structure of CLM5

CLM5 is the latest version of the land model of the Community Earth System Model version 2 (CESM2) (Lawrence et al., 2018, 2019). The soil carbon part of CLM5 centers its simulations around the pool turnover paradigm (Table 1). Similar structures have been widely used in

most of the state-of-the-art Earth system models. CLM5 uses conceptual soil carbon pools (i.e., active, slow, and passive SOC), and thus, microbial processes are only implicitly represented in the model structure. Meanwhile, CLM5 adopts first-order kinetics in simulating SOC decomposition. SOC dynamics in CLM5 can be expressed in a uniform matrix equation (Huang et al., 2018; Lu et al., 2020; Luo et al., 2022):

$$\frac{d\mathbf{X}(t)}{dt} = \mathbf{B}I(t) + \mathbf{A}\xi(t)\mathbf{K}\mathbf{X}(t) + \mathbf{V}(t)\mathbf{X}(t). \quad (2)$$

This matrix equation has six components (Table S1), including plant carbon inputs ($I(t)$), carbon input allocation to different pools and depths (\mathbf{B}), substrate decomposability (or baseline decomposition rates) (\mathbf{K}), carbon transfer efficiency (\mathbf{A}), environmental modifier ($\xi(t)$), and vertical transport ($\mathbf{V}(t)$).

CLM5 describes seven carbon pools in the soil, including four litter pools (i.e., coarse woody debris (indicated by subscript CWD), metabolic litter (ML), cellulose litter (CL), and lignin litter (LL)) and three soil organic carbon pools (i.e., active (aSOC), slow (sSOC), and passive (pSOC) soil organic carbon pools). Each of the carbon pools is simulated in 20 layers to a maximum depth of 8.4 m. The state of different carbon pools (i.e., carbon stocks) can be expressed as:

$$\mathbf{X}(t) = \begin{pmatrix} \mathbf{x}_{\text{CWD}}(t) \\ \mathbf{x}_{\text{ML}}(t) \\ \mathbf{x}_{\text{CL}}(t) \\ \mathbf{x}_{\text{LL}}(t) \\ \mathbf{x}_{\text{aSOC}}(t) \\ \mathbf{x}_{\text{sSOC}}(t) \\ \mathbf{x}_{\text{pSOC}}(t) \end{pmatrix} \quad (3)$$

where each of the seven block elements (i.e., $\mathbf{x}_i(t)$) of $\mathbf{X}(t)$ has 20 elements to represent the 20 soil layers. In total, CLM5 simulates carbon transfer among 140 pools. Consequently, there are 140 dimensions for vector \mathbf{B} of carbon input allocation, matrix \mathbf{K} of substrate decomposability, matrix \mathbf{A} of carbon transfer from one carbon pool to another, matrix $\xi(t)$ of environmental modifiers, and matrix $\mathbf{V}(t)$ of vertical transport. Plant carbon input ($I(t)$) is a scalar. In this study, parameters (Table S1) that generate the above elements in the matrix equation will be optimized by the PRODA approach.

Specifically, $I(t)$ is allocated to different litter pools in different layers along the soil profile via the allocation vector \mathbf{B} . Organic carbon in pool vector $\mathbf{X}(t)$ is decomposed following first-order kinetics as described by matrix \mathbf{K} :

$$\mathbf{K} = \begin{pmatrix} k_{\text{CWD}} \\ k_{\text{ML}} \\ k_{\text{CL}} \\ k_{\text{LL}} \\ k_{\text{aSOC}} \\ k_{\text{sSOC}} \\ k_{\text{pSOC}} \end{pmatrix}, \quad (4)$$

where k_i is independent from the state of its corresponding substrate $x_i(t)$. Moreover, we used the environmental modifier (i.e., $\xi(t)$) to account for the effects of environmental conditions on the decomposition processes. $\xi(t)$ is calculated from functions of soil temperature (ξ_T), soil water potential (ξ_W), nitrogen and oxygen availability (ξ_{N-O}), and soil depth (ξ_D).

Organic carbon from any carbon pool is further partitioned by either microbial or non-microbial processes between a receiver carbon pool and CO_2 released to the atmosphere. All these processes can be summarized in the \mathbf{A} matrix:

$$\mathbf{A} = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ a_{\text{CL,CWD}} & 0 & -1 & 0 & 0 & 0 & 0 \\ a_{\text{LL,CWD}} & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & a_{\text{aSOC,ML}} & a_{\text{aSOC,CL}} & 0 & -1 & a_{\text{aSOC,sSOC}} & a_{\text{aSOC,pSOC}} \\ 0 & 0 & 0 & a_{\text{sSOC,LL}} & a_{\text{sSOC,aSOC}} & -1 & 0 \\ 0 & 0 & 0 & 0 & a_{\text{pSOC,aSOC}} & a_{\text{pSOC,sSOC}} & -1 \end{bmatrix} \quad (5)$$

where all the block elements in the \mathbf{A} matrix (a_{ij}) are diagonal matrices with the dimension of 20. a_{ij} represents the carbon transfer fraction from the donor (j) pool to the recipient (i) pool (see carbon transfer flows in Figure 1). Note that CLM5 does not differentiate carbon transfers mediated by microbial processes from those mediated by non-microbial processes (e.g., organo-mineral interactions). Thus, a_{ij} in Equation 5 are integrative values reflecting carbon transfers contributed by both microbial and non-microbial processes.

The transport matrix \mathbf{V} of CLM5 is a tridiagonal matrix that describes vertical carbon movement between adjacent soil layers within the same carbon pool via bioturbation and cryoturbation. At steady state, the analytical solution of SOC stock by CLM5 was calculated as $\mathbf{X}_{\text{steady state}} = [\mathbf{A}\xi(t)\mathbf{K} + \mathbf{V}(t)]^{-1} [-\mathbf{B}l(t)]$, where the overbars indicate the mean values of related matrices ($\xi(t)$ and $\mathbf{V}(t)$) and scalar ($l(t)$) over the period of forcing data. The matrix representation for process-based soil carbon cycle models has been described in detail by Huang et al. (2018), Lu et al. (2020), and Luo et al. (2022).

2.3 | Structure of COMPAS model

COMPAS explicitly represents the microbial-driven carbon partitions in soil carbon cycle simulations. In addition to applying Michaelis-Menten kinetics in representing organic matter assimilation and decomposition, COMPAS differentiates soil organic carbon into field-measurable components, such as microbial biomass, extracellular enzyme, dissolved organic carbon, and mineral-associated organic carbon. Thus, we choose COMPAS as the representative model based on the microbe-centered paradigm.

Specifically, COMPAS follows the same structure proposed by Allison et al. (2010) for SOC dynamics, which is further embedded within the structure for 20-layered vertical soil profiles. The description of vertical layers was adopted from CLM5. Organic carbon dynamics represented by COMPAS can be expressed by the same

matrix framework as shown in Equation 2 (Table S2). Yet, COMPAS structurally differs from CLM5 in classifying soil carbon pools, expressing substrate decomposition, and explicitly describing microbial partitioning processes in carbon transfer (Table 1 and Figure 1).

Equation 2 describes COMPAS with 160 dimensions to represent eight pools in each of the 20 soil layers. Vector $\mathbf{X}(t)$ has eight block elements to represent four litter carbon pools (indicated by subscripts CWD, ML, CL, and LL) and four soil organic carbon pools (i.e., dissolved organic carbon (DOC), mineral-associated soil organic carbon (mSOC), microbial biomass (MIC), and extracellular enzymes (ENZ)):

$$\mathbf{X}(t) = \begin{bmatrix} \mathbf{x}_{\text{CWD}}(t) \\ \mathbf{x}_{\text{ML}}(t) \\ \mathbf{x}_{\text{CL}}(t) \\ \mathbf{x}_{\text{LL}}(t) \\ \mathbf{x}_{\text{DOC}}(t) \\ \mathbf{x}_{\text{MIC}}(t) \\ \mathbf{x}_{\text{ENZ}}(t) \\ \mathbf{x}_{\text{mSOC}}(t) \end{bmatrix} \quad (6)$$

Each of the eight block elements (i.e., $\mathbf{x}_i(t)$) of $\mathbf{X}(t)$ has 20 elements to represent the 20 soil layers. Similarly, there are 160 dimensions for vector \mathbf{B} , matrix \mathbf{K} , matrix \mathbf{A} , matrix $\xi(t)$, and matrix $\mathbf{V}(t)$. Plant carbon input ($l(t)$) is still a scalar as in CLM5. Parameters (Table S2) that generate the above elements in the matrix equation will be optimized by the PRODA approach.

Different from CLM5, organic carbon pools in vector $\mathbf{X}(t)$ of COMPAS can be transferred to recipient pools either through microbial- or enzyme-mediated kinetics, or without going through microbial metabolism. These transfers are described by the baseline decomposition matrix \mathbf{K} :

$$\mathbf{K} = \text{diag} \left(\begin{array}{c} k_{\text{CWD}} \\ k_{\text{ML}} \\ k_{\text{CL}} \\ k_{\text{LL}} \\ k_{\text{DOC}}(\mathbf{x}_{\text{DOC}}, \mathbf{x}_{\text{MIC}}) \\ k_{\text{MIC}} \\ k_{\text{ENZ}} \\ k_{\text{mSOC}}(\mathbf{x}_{\text{mSOC}}, \mathbf{x}_{\text{ENZ}}) \end{array} \right) \quad (7)$$

While all the litter organic carbon pools and two mineral-soil organic carbon pools (i.e., MIC and ENZ) are decomposed following first-order kinetics with constant baseline decomposition rates, the baseline decomposition rates of DOC and mSOC are functions of carbon pool states. Specifically, the baseline decomposition rate of DOC (i.e., the baseline rate of microbial assimilation of DOC) is: $k_{\text{DOC}}(\mathbf{x}_{\text{DOC}}, \mathbf{x}_{\text{MIC}}) = \frac{v_{\text{max,assim}} \mathbf{x}_{\text{MIC}}}{K_{\text{m,assim}} \xi + \mathbf{x}_{\text{DOC}}}$; the baseline decomposition rate of mSOC is: $k_{\text{mSOC}}(\mathbf{x}_{\text{mSOC}}, \mathbf{x}_{\text{ENZ}}) = \frac{v_{\text{max,decom}} \mathbf{x}_{\text{ENZ}}}{K_{\text{m,decom}} \xi + \mathbf{x}_{\text{mSOC}}}$. Parameters $v_{\text{max,assim}}$ and $v_{\text{max,decom}}$ represent the maximum DOC assimilation and mSOC

decomposition rates, respectively. $K_{m,assim}$ and $K_{m,decom}$ are the Michaelis constants for DOC assimilation and mSOC decomposition, respectively.

The COMPAS model also explicitly differentiates carbon transfers by microbial processes from those in non-microbial processes. The decomposed organic carbon is either partitioned by microorganisms to microbial biomass growth versus respiration (i.e., according to the microbial CUE), or alternatively, transferred to other carbon pools with a fraction that is not mediated by microbial processes (i.e., non-microbial carbon transfer). All these processes are summarized in the **A** matrix:

$$\mathbf{A} = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ a_{CL,CWD} & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ a_{LL,CWD} & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & a_{DOC,ML} & a_{DOC,CL} & 0 & -1 & a_{DOC,MIC} & 1 & a_{DOC,mSOC} \\ 0 & a_{MIC,ML} & a_{MIC,CL} & a_{MIC,LL} & a_{MIC,DOC} & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & a_{ENZ,MIC} & -1 & 0 \\ 0 & 0 & 0 & a_{mSOC,LL} & 0 & a_{mSOC,MIC} & 0 & -1 \end{bmatrix} \quad (8)$$

Because DOC is always assimilated by the microbes with release of CO_2 (Figure 1), the microbial CUE for DOC (η_{DOC}) equals $a_{MIC,DOC}$. In contrast, organic carbon in the metabolic, cellulose, and lignin litter pools is decomposed by microbes following first-order kinetics to generate CO_2 and grow biomass while a fraction of litter organic carbon is broken down without going through microbial metabolism and, thus, directly transferred to DOC or mSOC. In this case, the microbial CUE for the three litter carbon pools can still be expressed as: $\eta_{ML} = \frac{a_{MIC,ML}}{1 - a_{DOC,ML}}$, $\eta_{CL} = \frac{a_{MIC,CL}}{1 - a_{DOC,CL}}$, and $\eta_{LL} = \frac{a_{MIC,LL}}{1 - a_{mSOC,LL}}$, respectively.

COMPAS applies the same approach to simulate carbon input allocation (**B**), environmental modifier (i.e., $\xi(t)$) and transport matrix **V** as that used in CLM5. It should be noted that while COMPAS and CLM5 use the same scheme to simulate **B**, $\xi(t)$, and **V**, parameter values (Tables S1 and S2) that were used to calculate the above elements in the matrix equation were estimated independently by the PRODA approach.

In calculating the steady state of different carbon pools by COMPAS, Equation 2 can be separated into two equations: one for litter carbon cycle and transport, and the other for mineral-soil SOC cycle, because there is no carbon transfer from mineral-soil carbon pools to litter carbon pools (i.e., $a_{litter\ pool, soil\ pool} = 0$ in the **A** matrix). Since **A**, **K**, $\xi(t)$, and **V** are all independent from litter carbon pool states (i.e., **X**), the analytical solution of litter carbon stock at the steady state (SS) can be calculated as $\mathbf{X}_{litter,SS} = [\mathbf{A}_{litter} \overline{\xi(t)}_{litter} \mathbf{K}_{litter} + \mathbf{V}(t)_{litter}]^{-1} [-\mathbf{B}_{litter} \overline{I(t)}_{litter}]$. For the soil organic carbon pools, the related **K** matrix is carbon pool state-dependent (see Equation 7). We assumed that there is no vertical transport for mineral-soil organic carbon pools such that litter is added to different soil layers and transported vertically, and then, it is transferred to soil pools that are immobile in that layer. According to a method reported by Georgiou et al. (2017), the steady-state solutions for soil organic carbon pools are:

$$\mathbf{X}_{soil,SS} = \begin{bmatrix} x_{DOC,SS} \\ x_{MIC,SS} \\ x_{ENZ,SS} \\ x_{mSOC,SS} \end{bmatrix} = \begin{bmatrix} \frac{k_{MIC} \xi K_{m,assim} \xi x_{MIC,SS} - u_{MIC} K_{m,assim} \xi}{(\eta_{DOC} v_{max,assim} - k_{MIC}) \xi x_{MIC,SS} + u_{MIC}} \\ \frac{u_{MIC} + \eta_{DOC} (u_{mSOC} + u_{DOC})}{(1 - \eta_{DOC}) k_{MIC} \xi} \\ \frac{a_{ENZ,MIC} k_{MIC} x_{MIC,SS}}{K_{ENZ}} \\ \frac{(u_{mSOC} + a_{mSOC,MIC} k_{MIC} \xi x_{MIC,SS}) K_{m,decom} \xi}{(v_{max,decom} \xi x_{ENZ,SS} - a_{mSOC,MIC} k_{MIC} \xi x_{MIC,SS} - u_{mSOC})} \end{bmatrix} \quad (9)$$

where u_s is the carbon input from litter pools (L_i) to a mineral-soil carbon pool (S_i ; see Extended Data Figure 3 for corresponding carbon flows for each mineral-soil carbon pool) and is expressed as $\sum_l (a_{s,l} k_l \xi x_{l_i})$. Note that all the elements with bold font indicate vectors of the corresponding variables or parameters for the 20 soil layers. All the multiplications shown in Equation 9 are element-wise operations.

2.4 | Inputs and environmental conditions

For both CLM5 and COMPAS, the carbon input for the litter carbon pools (i.e., net primary productivity, NPP) and environmental forcings (e.g., soil temperature and moisture) are from 20 years of monthly model outputs (Table S3) by CLM5 at the steady state using a preindustrial forcing (i.e., I1850Clm50Bgc, from year 1901 to 1920) at 0.5-degree resolution. We used the 20-year annual mean values of different components in Equation 2 to calculate total soil organic carbon stock at steady state.

2.5 | Customary parameter values for model simulations

We compared the model simulation results of CLM5 and COMPAS by (1) applying customary parameter values and (2) the parameter values optimized by the PRODA approach. For CLM5, we applied the parameter values used in its current version (Lawrence et al., 2019). In the default scheme, most of the selected 21 parameters of CLM5 are constants across the globe, except two carbon transfers that depend on sand content and the parameter controlling plant carbon input allocation that depends on plant functional types (Table S1). For COMPAS, it is a newly constructed model and thus does not have well-tuned parametrization for global simulation. We applied the global mean values of the selected 23 parameters after site-level data assimilation as the customary parameter values for COMPAS to drive the global simulation.

2.6 | PROcess-guided deep learning and DAta-driven modeling (PRODA)

The PRODA approach integrates big data with Bayesian data assimilation and deep learning to optimize soil carbon cycle simulation with process-based models (Tao & Luo, 2022). We used the PRODA approach to optimize both CLM5 and COMPAS at the global scale. Data

assimilation was first applied at each SOC profile to estimate parameter values. Twenty-one parameters for CLM5 and 23 parameters for COMPAS were optimized for each SOC profile so that the process-based model simulations can best fit local observations. Because we conducted data assimilation independently at each observation site, optimized values of the same parameter vary across space. We further used a neural network to generalize those estimated parameter values after the site-level data assimilation to the global scale. The global parameter maps predicted by the neural network were then used in the process-based models to simulate global SOC storage and retrieve the spatial patterns of related model components across the globe.

We conducted Bayesian data assimilation by using the MCMC method for each of the SOC profiles to estimate the parameter values of the process-based models that best-fit model simulations with SOC observations. Because the soil profile data collected from field measurements of soil organic carbon include all components of organic matter (e.g., microbial biomass carbon), we used the sum of modeled mineral-soil carbon pools classified in CLM5 and COMPAS for each layer to be compared with soil profile data at the corresponding sampling layer.

Specifically, at site-level data assimilation, for each SOC profile, we applied an adaptive Metropolis algorithm (Haario et al., 2001) to generate the posterior distributions of a total of 21 parameters of CLM5 (Table S1) and 23 parameters of COMPAS (Table S2) related to six model components with two phases of simulations (i.e., a test run and a formal run). We first conducted a test run assuming uniform distributions for each of the preselected parameters as the proposal distributions (i.e., prior knowledge). The prior ranges of the uniform distributions for each parameter are shown in Tables S1 and S2. The proposal distributions continuously generated a set of parameter values for the process-based models to simulate SOC storage. We then evaluated whether the proposed parameter values should be accepted or not by comparing their model simulation results with SOC observations. In the formal run, we used the accepted sets of parameter values obtained in the test run as the proposal distributions and assumed that all the target parameters are multivariate Gaussian distributed. We proposed new sets of parameter values and evaluated them to be accepted or not following the same rule in the test run. Unlike the test run, the proposal distributions in the formal run were continuously adjusted according to the newly accepted sets of parameters.

We set 20,000 iterations for the test run and 50,000 iterations for the formal run. Eventually, we controlled the acceptance ratio (i.e., the ratio of accepted sets of parameters out of the total number of iterations) of the formal run between 10% and 50%. We set the burn-in coefficient as 50%, where the first half of the accepted parameter values in the formal run was discarded, and the second half was used to generate the posterior distributions of parameters. We calculated the mean values of the posterior distributions of parameters as the final estimates of parameter values. We ran three independent series of MCMC for each SOC profile and calculated the G-R statistic to test the convergence of data assimilation results. The mean G-R values of the target parameters were further calculated as the holistic performance of MCMC for each SOC profile.

The mathematical foundations of Bayesian data assimilation and technical details of the MCMC method were documented by Tao et al. (2020).

It should be noted that the data assimilation was conducted under the assumption that SOC profiles are at steady state (i.e., $\frac{dX(t)}{dt} = 0$). This assumption makes data assimilation computationally more feasible than that under non-steady state (see the non-steady-state data assimilation in Zhou et al. (2013) and Zhou et al. (2015)). While soil carbon stocks in some ecosystems (e.g., agricultural soils) may not be at the steady state because of the concurrent climate change and human activities, previous research demonstrated that such disequilibrium component of the transient carbon cycle dynamics, especially in SOC pools, is minor in comparison with the amount of SOC storage that was developed over thousands of years (Lu et al., 2018).

We included parameters related to both non-microbial and microbial processes (Tables S1 and S2) in the site-level data assimilation and the following global optimization with the PRODA approach. While we acknowledge that biological processes (and thus their related parameter values) may change in response to external disturbance, in this study, we focus on the long-term spatial patterns of vertically distributed SOC under the steady-state assumption. We used multi-year mean values of a preindustrial forcing (no climate change happened yet) to simulate SOC storage. Therefore, the optimized parameter values should be regarded as long-term averages.

Moreover, we designed a parameter recovery experiment to confirm whether parameters related to microbial processes (e.g., the Michaelis–Menten constants) can be recovered from data assimilation under the steady-state assumption. We randomly chose 200 sites across the world for COMPAS and used prescribed parameter values with different across-site variability to generate a set of synthetic SOC data. The synthetic vertical SOC profile (20 datapoints at the 20 prescribed soil layers in COMPAS) was further used in the MCMC data assimilation to retrieve optimized parameter values. We found a satisfactory agreement between the retrieved parameter values and their prescribed values (e.g., “mm_const_assim” and “mm_const_decom” in Figure S3). For parameters whose prescribed values did not show much across-site variability (e.g., “tau4s1” and “pl1s1” in Figure S3), MCMC method also refrained from assigning them extra variation across sites. The results of the recovery experiment supported the efficacy of using the MCMC method to retrieve optimized parameter values from observations under the steady-state assumption.

We trained a fully connected multilayer neural network to predict the site-level parameter values estimated from data assimilation with a suite of 60 environmental variables (Table S4). We chose variables that represent the climatic, vegetation, edaphic, and geographic conditions at different sites because they are conventionally regarded as the driving factors that regulate the formation and stabilization of SOC (Jackson et al., 2017). Parameters in process-based models quantify the strength of different soil carbon cycle processes and therefore should also have relationships with these environmental variables (Luo & Schuur, 2020). To achieve a better training effectiveness, we first normalized all the environmental variables and parameters to the interval of [0, 1] according to their maximum and minimum values.

We then conducted a set of pre-experiments to determine the best configuration of the neural network. The neural network used in the final training consisted of four hidden layers. The node numbers for each hidden layer were 256, 512, 512, and 256, respectively. We used a rectified linear unit (ReLU) as the activation function and a gradient descent optimization algorithm (adadelat) as the optimizer. The loss function was designed as the multiplication of $L1$ (i.e., ratio loss (RL): $RL = \frac{\sum_{i=1}^N \frac{para_{i,true} - para_{i,pred}}{para_{i,true}}}{N}$) and $L2$ (i.e., mean squared error (MSE): $MSE = \frac{\sum_{i=1}^N (para_{i,true} - para_{i,pred})^2}{N}$) errors, where $para_{i,true}$ is the i th parameter value optimized in the site-level data assimilation, $para_{i,pred}$ is the i th parameter predicted by the neural network, and N is the total number of parameters of the process-based models to be predicted by the neural network ($N = \text{training size} \times 23$ for COMPAS and $\text{training size} \times 21$ for CLM5). While both $L1$ and $L2$ are additive loss functions, we decided to use their multiplicative composite (i.e., $L1 \times L2$) as the loss function because training with either $L1$ or $L2$ loss alone did not yield sufficient prediction accuracy. The batch size for each iteration of optimization was 32. We set a maximum of 6000 epochs to train the neural network and selected the model with the lowest validation loss as the final training result. To avoid overfitting in training the neural network, we set a drop-out ratio of 20% for each of the hidden layers.

2.7 | Global maps of SOC, residence time, and related model components

Global maps of parameters predicted by the best-guess neural network using the gridded environmental variables were applied to the two process-based models to generate global maps of SOC storage and its related components (i.e., 57,267 sets of site-level data assimilation results for COMPAS and 62,931 for CLM5). In addition, we conducted bootstrapping experiments to quantify the simulation uncertainties of CLM5 and COMPAS after being optimized by the PRODA approach. The original SOC database used by CLM5 and COMPAS was sampled with replacement 200 times and was used to train and validate the neural network. Following a common practice in neural network training, for each bootstrapping, 90% of the data were used as training data, and the remaining 10% were used for validation. The predicted parameter values after neural network training were then applied to CLM5 and COMPAS to simulate SOC stock and its related model components. The uncertainty maps of SOC storage and its related components are shown in Figures S4 and S5.

It should be noted that uncertainties shown in the global generalization by the PRODA approach only quantify the variation of trained neural networks in predicting site-level data assimilation results (i.e., the mean value of parameters' posterior distribution). Limited by its optimization algorithm (Tao & Luo, 2022), PRODA is not able to consider propagating the uncertainties in parameters' posterior distribution in the site-level data assimilation to the global scale. Developing the next-generation data assimilation approach that can directly integrate process-based models into deep learning algorithms will be the solution to retrieve process understanding and simultaneously address parameter uncertainties in optimization.

We retrieved the system-level carbon transfer efficiency (CTE), plant carbon inputs, allocation of input carbon to different soil layers, substrate decomposability, environmental modifications, and vertical transport from the optimized parameters of COMPAS and CLM5 (Tables S1 and S2) via the PRODA approach. All the six model components investigated in this study are ensembles of processes that were represented by different parameters in the process-based model. Note that all the system-level components discussed in this study are for the soil system that integrates both litter organic carbon and mineral-soil organic carbon.

Specifically, we calculated the system-level carbon transfer efficiency as the sum of carbon transfer coefficients along each carbon transformation pathway (i.e., a_{ij} in Equations 5 and 8) weighted by the carbon fluxes over all the pathways in the soil system:

$$CTE_{\text{system}} = \sum_{ij} a_{ij} \frac{\sum_z x_{j,z} k_j \xi_z \Delta z}{\sum_j \sum_z x_{j,z} k_j \xi_z \Delta z} \quad (10)$$

where a_{ij} represents the carbon transfer fraction from the donor pool (j) to the recipient pool (i); $x_{j,z}$ is the carbon pool size at depth z (g C m^{-3}); k_j is the depth-independent baseline decomposition rate (yr^{-1}) of the corresponding carbon pool; ξ_z represents the environmental modifier at depth z ; and Δz is the thickness of z th soil layer. Note that CTE along the carbon transfer pathway from donor pool j to recipient pool i (i.e., a_{ij}) is weighted by the flux size from donor pool j (i.e., $\sum_z x_{j,z} k_j \xi_z \Delta z$), which measures the amount of decomposed carbon along the j to i transfer pathway, normalized by the total flux in the soil system (i.e., $\sum_j \sum_z x_{j,z} k_j \xi_z \Delta z$). A higher CTE value indicates a larger amount of carbon remained in the recipient soil pool after organic carbon is decomposed or transformed by biological and/or chemical and physical reactions, which, by definition, also associates with less CO_2 released back to the atmosphere. It should be noted that this weighted average transfer efficiency is defined differently from the system CUE in Tao et al. (2023), which was instead calculated as ratio between the sum of carbon fluxes entering the microbial pool and the sum of carbon fluxes leaving the donor pools.

The baseline decomposition rate (unit: year^{-1}) expresses the rate of organic carbon decomposition at optimal soil temperature and water conditions. We calculated the system-level baseline decomposition rate (K_{system} , unit: year^{-1}) by weighting the baseline decomposition rate of SOC pools by their carbon pool sizes:

$$K_{\text{system}} = \sum_i k_i \frac{x_i}{\sum_i x_i} \quad (11)$$

Similarly, we weighted the vertical transport rate (year^{-1}) and environmental modifiers (unitless) at different soil depths (z) by their corresponding sizes of SOC stock (i.e., x_z , with unit of g C m^{-2}):

$$V_{\text{system}} = \sum_z \left(v_z \frac{x_z}{\sum_z x_z} \right) \quad (12)$$

$$\xi_{\text{system}} = \sum_z \left(\xi_{T,z} \xi_{W,z} \xi_{D,z} \frac{X_z}{\sum_z X_z} \right) \quad (13)$$

Carbon input is distributed vertically according to the distribution of root biomass at different soil depths (Jackson et al., 1996). Therefore, to quantify how effectively the input allocation process distributes litterfall and root exudation to different soil depths, we calculated the fraction of carbon input allocated to soil layers below 5 cm as the system-level index for plant carbon input allocation:

$$B_{\text{system}} = \left[\frac{\sum_z \exp\left[\frac{\ln(1-Y_z)}{D_z}\right]}{n} \right]^5 \quad (14)$$

where Y_z is the cumulative fraction of input carbon at soil depth of D_z ; n is the number of soil layers. A larger system-level input allocation index indicates that more carbon from litterfall and root exudation will be allocated to deeper soils. This index differs between models because the parameters describing the vertical distribution of carbon inputs are optimized independently in the two models, even if we used the simulated total litterfall (equivalent to NPP) in CLM5 as the plant carbon input for both models.

3 | RESULTS

Process-based models with different structures and customary parameter values show diverging results in representing global SOC storage and spatial patterns. With its customary parameter values, CLM5 simulates much more SOC in the boreal regions than in the tropics. In East Siberia and Alaska, SOC storage is more than 50 kg C m⁻² for the first meter, whereas in the Amazon and Congo basins and Indonesia, the average SOC storage is less than 10 kg C

m⁻² (Figure 2a,c). As COMPAS does not have well-tuned parameter values at the global scale, we used the global mean values of the selected parameters after site-level data assimilation as the customary parameter values. With such customary parameterization, COMPAS simulates distinctively different SOC patterns from CLM5 across latitudes. Tropical regions with the highest carbon input are simulated to store the largest amount of SOC. The average SOC storage declines from more than 20 kg C m⁻² in Amazon, Congo, and Indonesia to less than 5 kg C m⁻² in boreal regions (Figure 2b,c). The correlation between the simulated spatial patterns of SOC by CLM5 and COMPAS is -0.026 (logarithmically transformed SOC values, $df=45,213$, $p<.0001$). Despite the contrasting spatial patterns, both models reasonably estimate the total global SOC storage with their customary parameter values. CLM5 and COMPAS simulate 1281 Pg C and 1308 Pg C preserved as SOC for the first-meter soils across the globe, respectively. For comparison, as two commonly used observation-based statistical products, HWSD (FAO/IIASA/ISRIC/ISSCAS/JRC, 2012) and WISE (Batjes, 2016) estimate 1260 Pg C and 1408 Pg C for the global first-meter SOC storage, respectively.

The two structurally different models simulate similar SOC storage and spatial patterns after being constrained by the same SOC data using the PRODA approach. At the site level, we found that posterior distributions of selected parameters after data assimilation could differ greatly from their customary values (Figure S6) and from site to site. We further used PRODA to generalize the emerging spatial heterogeneity of optimized parameter values in site-level data assimilation to the global scale and found similar SOC simulations by CLM5 and COMPAS. Based on the best-guess neural network predictions that were trained by all available site-level data assimilation results (see Section 2.7 for details), PRODA-optimized CLM5 explains 57% (median 56%, one-sigma confidence interval 53%–57% in 200-time bootstrapping) of the spatial variations in SOC at measured soil depths across the globe (Figure S7a). The predictive performance of COMPAS after PRODA optimization is similar to that of

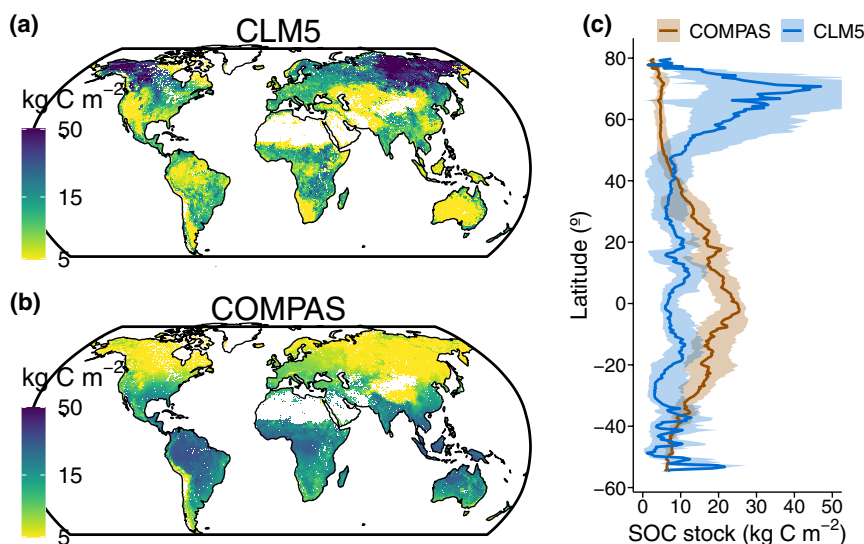


FIGURE 2 Diverging SOC simulation by structurally different models with customary parameter values. (a) SOC estimated by CLM model, (b) SOC estimated by COMPAS, (c) latitudinal variation in estimated SOC by the two models.

CLM5, explaining 55% (median 53%, one-sigma confidence interval 52.5%–54% in 200-time bootstrapping) of the spatial variations in global SOC observations (Figure S7b).

In simulating global SOC patterns, CLM5 continues to simulate higher SOC storage in the boreal regions than in the tropics. In addition to higher SOC in East Siberian and Alaska, PRODA-optimized CLM5 also identifies western Siberian lowlands as areas holding high SOC storage (Figure 3a,c). Meanwhile, after being constrained by observations, the simulated SOC storage in tropical regions increased to an average value of more than 10 kg C m^{-2} (Figure 3b,c). Simulation results by COMPAS after PRODA optimization now follow a pattern similar to that by CLM5. The correlation between simulations by COMPAS and CLM5 is 0.51 (logarithmically transformed SOC values, $df=45,213$, $p<.0001$). Notably, differences still exist in simulating sub-continental patterns by these two models. While both models simulate the highest SOC storage in western Siberian lowlands, Alaska, and Canadian Shield, COMPAS simulates more SOC in the tropics but less SOC in East Siberian than CLM5. The total SOC storage simulated by COMPAS is slightly higher than that by CLM5. Globally, the total SOC storages in the top 1 m of soil estimated by PRODA-optimized CLM5 and COMPAS are 1469 Pg C and 1507 Pg C, respectively.

Simulations of key components related to SOC storage also converge after the two structurally different models are constrained by the same set of SOC data (Figure 4). We assessed the spatial patterns of six components simulated by the two models: carbon transfer efficiency, baseline decomposition, environmental modifier, carbon input allocation, vertical transport rate, and plant carbon input.

The carbon transfer efficiency quantifies the ratio of decomposed carbon being transferred from one carbon pool to another. CLM5 and COMPAS represent the carbon transfer efficiency differently (Figure 1). COMPAS explicitly describes microbial CUE that partitions the metabolized organic carbon into microbial biomass accumulation versus respiration and the non-microbial

carbon transfer related to the transformation of carbon from one carbon pool to another via organo-mineral interactions (Figure 1b). In contrast, CLM5 fuses microbial CUE and non-microbial carbon transfer in its structure, such that the related parameters do not differentiate these two processes but integrate their effects in simulations (Figure 1a). Thus, it is not surprising that the values of the carbon transfer efficiencies are in general different between the two models, with higher values predicted by CLM5 compared with COMPAS (Figure 4c). Yet, despite the difference in structure, CLM5 and COMPAS simulate similar spatial patterns of system-level carbon transfer efficiency (Figure 4c, Pearson correlation coefficient = 0.52, $df=45,228$, $p<.001$) after being constrained by the same observed SOC dataset. Both models show higher carbon transfer efficiency in boreal regions than in the tropics (Figure 4a,b), which indicates that in boreal regions, more carbon is maintained in the soil system after SOC is decomposed or transformed by biological and/or chemical and physical reactions instead of being released back to the atmosphere as CO_2 .

The rate of SOC decomposition is determined by the substrate decomposability (as indicated by the baseline decomposition) and modified by surrounding environmental factors (i.e., soil temperature and moisture). A high baseline decomposition rate indicates the organic substrate is chemically and physically more accessible to soil microorganisms (e.g., simpler chemical compounds or weaker interactions with the soil mineral matrix). In contrast, a lower environmental modifier value indicates that SOC decomposition is more restricted by either low temperature or too much or little soil water. CLM5 and COMPAS assume first-order and Michaelis-Menten kinetics in representing SOC decomposition, respectively. Notwithstanding their different assumptions on the decomposition kinetics, PRODA-optimized CLM5 and COMPAS agree on the highest baseline decomposition rates and the lowest environmental modifier values in boreal regions across the globe (Figure 4d–i). The correlation coefficients between the simulations by the two models are 0.55 ($df=45,228$, $p<.001$) for

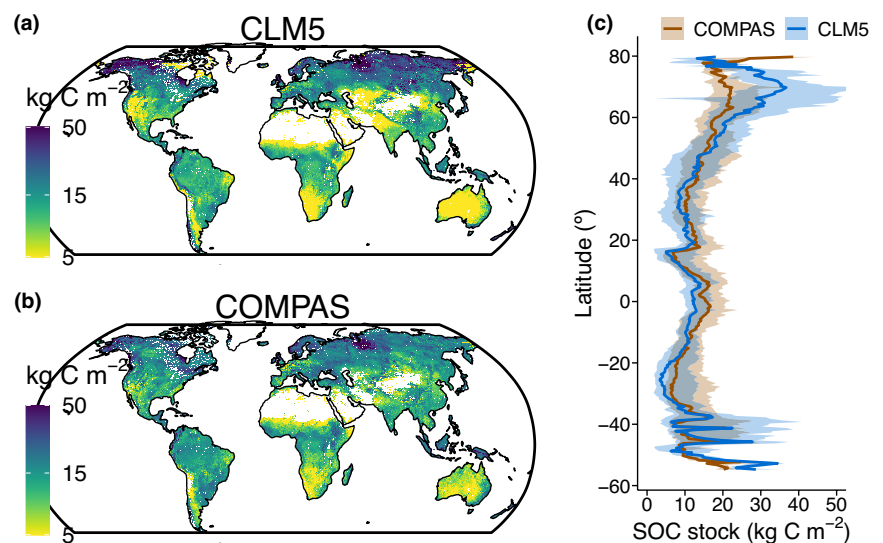


FIGURE 3 Converging SOC simulation by structurally different models after data-model fusion by the PRODA approach. (a) SOC estimated by CLM model, (b) SOC estimated by COMPAS, (c) latitudinal variation in estimated SOC by the two models. Uncertainty maps of SOC storage simulations with CLM5 and COMPAS in a 200-time bootstrapping experiment are shown in Figures S4 and S5.

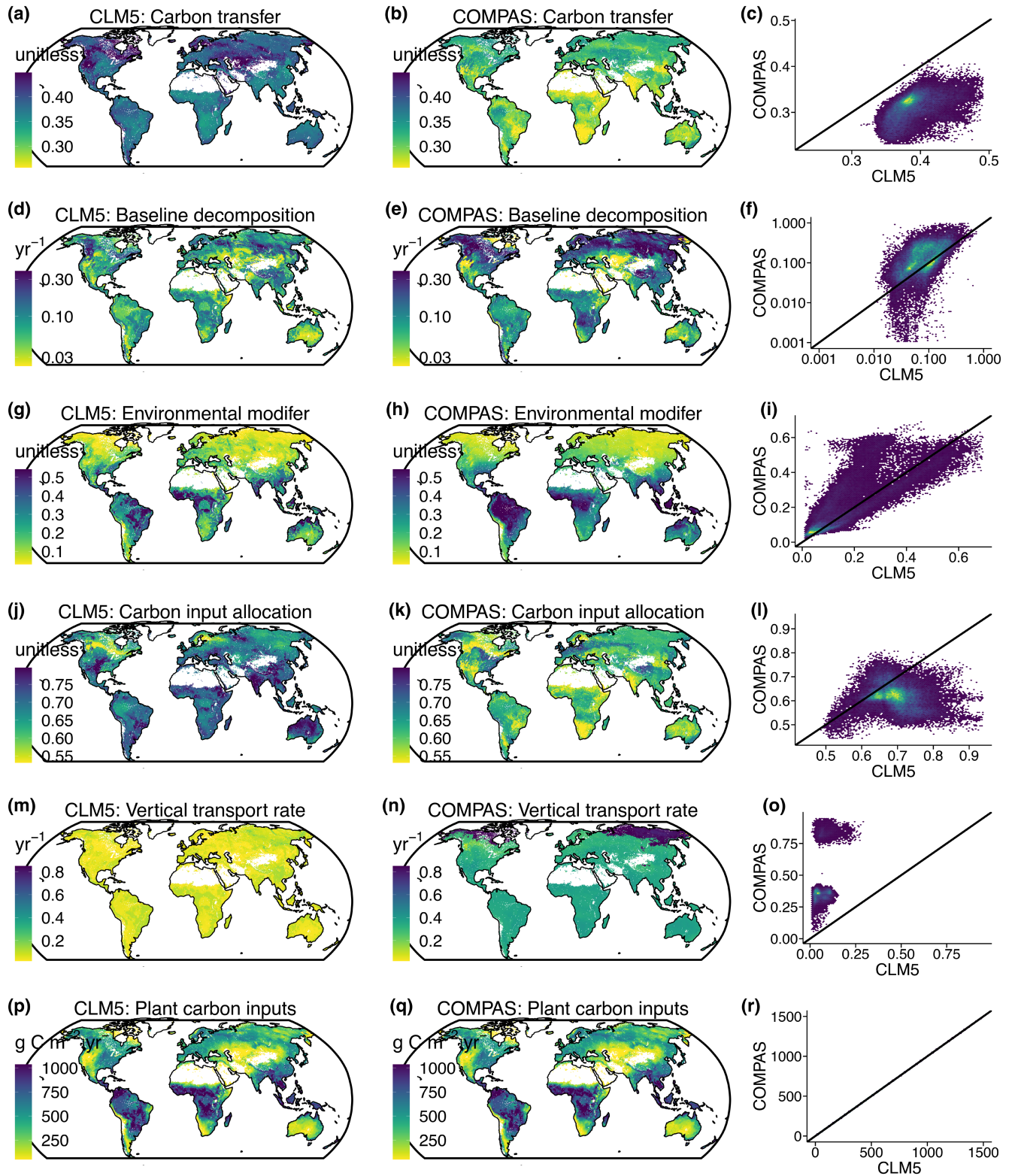


FIGURE 4 Spatial patterns of different model components retrieved by CLM (left column) and COMPAS (central column) models using the PRODA approach. The right column shows comparisons between the model components retrieved from the two models. The model components were: (a–c) carbon transfer efficiency (CTE_{system} , see Equation 10), (d–f) baseline decomposition (K_{system} , see Equation 11), (g–i) environmental modifier (ξ_{system} , see Equation 13), (j–l) carbon input allocation (B_{system} , see Equation 14), (m–o) vertical transport rate (V_{system} , see Equation 12), and (p–r) plant carbon input (same for both models). Uncertainty maps of these components with CLM5 and COMPAS in a 200-time bootstrapping experiment are shown in Figures S4 and S5.

baseline decomposition and 0.80 ($df = 45,228, p < .001$) for the environmental modifier.

However, not all components we investigated show convergence after data assimilation. Vertical transport quantifies the rate of organic carbon moving from the surface to deeper soil layers. The plant carbon allocation represents the vertical distribution of carbon inputs. While CLM5 and COMPAS adopt identical mathematical functions to describe these two processes (except vertical transport of mineral-soil carbon), no agreement was reached on simulated spatial patterns after the related parameters of the two models were optimized by the PRODA approach (Figure 4j–o). Moreover, it should be noted that the retrieved model components using CLM5 and COMPAS are usually far from 1:1 lines even when they are well correlated. While the two models agree well on the magnitude of the simulated environmental modifier (Figure 4i), the linear CLM5 simulates higher carbon transfer efficiency values (Figure 4c) but lower baseline decomposition rates (Figure 4f) than the nonlinear COMPAS. This pattern may occur because parameters related to carbon transfer efficiency and baseline decomposition compensate each other in CLM5 and COMPAS for a similar SOC storage simulation. Even though we used the same plant carbon input (i.e., the total amount of carbon from plant to litter) from CESM2 outputs in simulating SOC storage by the two models (Figure 4p–r), COMPAS and CLM5 simulated differently how carbon transfers from litter to mineral soils (Figure 1), as quantified by the ratio between the amount of carbon transferred from litter to mineral soils and the total carbon input. COMPAS simulates larger amounts of litter carbon to be transferred to mineral soils than CLM5 (Figure S8), which requires higher baseline decomposition rates in COMAS than CLM5 to reach similar simulated SOC storage, as shown in Figure 4d–f.

The nonlinear decomposition kinetics in COMPAS can be approximated as first-order kinetics with respect to both donor and receiver carbon pools after being constrained by observed SOC data. Compared with the linear first-order kinetics used in CLM5, COMPAS specifies SOC decomposition and DOC assimilation as nonlinear Michaelis–Menten kinetics. Thus, both the catalyst (i.e., microbes for DOC assimilation and enzyme for mSOC decomposition) and the substrate concentration (i.e., DOC for DOC assimilation and mSOC for mSOC decomposition) regulate substrate decomposition. Mathematically, when the Michaelis constants (i.e., $K_{m,decom}$ and $K_{m,assim}$) are much larger (e.g., 100 times larger) than their corresponding substrate concentrations and the catalyst (i.e., DOC in assimilation and MIC in decomposition) concentrations remain stable, the Michaelis–Menten kinetics can be approximated by first-order kinetics with respect to DOC in assimilation and mSOC in decomposition. After data assimilation at each SOC profile using COMPAS, we found that both $K_{m,decom}$ and $K_{m,assim}$ in the Michaelis–Menten equation are more than 100 times that of their substrate concentrations (i.e., SOC and DOC concentrations) for most of the soil profiles (Figure 5). Thus, the nonlinear kinetics for enzyme-based mSOC decomposition and microbe-based DOC assimilation can be approximated by first-order kinetics with respect to mSOC and DOC after COMPAS is constrained by globally distributed SOC vertical profiles.

While losing the nonlinear character of the donor pool effect, these kinetics laws still retain the effect of microbial biomass or enzyme carbon, resulting in multiplicative kinetics.

4 | DISCUSSION

4.1 | Data assimilation enables converged SOC simulations by structurally different models

The divergent simulations by process-based models with different structures and customary parameter values reflect large uncertainties in the current understanding of soil carbon dynamics with different theories and assumptions. In this study, CLM5 and COMPAS structurally differ in classifying soil carbon pools, quantifying SOC decomposition kinetics, and representing carbon transfer processes. The structural differences between these two models contributed to the contrasting SOC spatial patterns across the globe (Figure 2). Uncertainties arise also from poorly constrained parameters. Model parameters quantify the strength or represent the properties of different processes in regulating the soil carbon cycle (Luo & Schuur, 2020). When they are not well constrained, differences in parameter values across models can cause additional large simulation uncertainty. Previous studies have demonstrated that models sharing the same first-order kinetics for SOC decomposition estimated contrasting soil carbon residence time (Wei et al., 2022; Zhou et al., 2018) and age (He et al., 2016; Shi et al., 2020) due to their different choices of parameter values. These differences resulted in large uncertainties in simulating global SOC storage (Todd-Brown et al., 2013). While all these simulations are, to some degree, plausible under given assumptions and theories, we need to identify the most probable

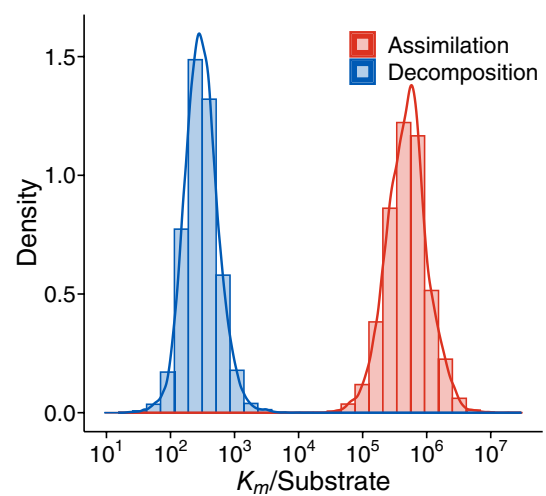


FIGURE 5 Relationship between Michaelis–Menten constants and their corresponding substrate content in COMPAS after being constrained by observational SOC profiles. For decomposition, “Substrate” is mineral-associated organic carbon (mSOC) and $K_m = K_{m,decom}$. For assimilation, “Substrate” is dissolved organic carbon (DOC) and $K_m = K_{m,assim}$

ones to better understand how the soil carbon cycle responds to a changing climate.

Our results show that the vast inter-model uncertainty in simulating global SOC storage is mainly due to the lack of common observational data constraints in major processes. Regardless of their difference in structure, our results show well-converged global SOC simulations by CLM5 and COMPAS after being optimized by the PRODA approach with the same soil carbon observations. The convergence in SOC simulations arises from the fact that the PRODA approach effectively constrains the spatial patterns of parameters of process-based models by the common observational data. Parameters in CLM5 and COMPAS are both conceptually and functionally different from each other due to their structural dissimilarity (e.g., the turnover time values for conceptually different carbon pools and the carbon transfer coefficients in CLM5 and COMPAS; see [Figure 1](#) and Methods for details). However, the spatial distributions of parameters aggregate into six model components defined in the same way, which exhibit some agreement between the models. Carbon transfer efficiency, baseline decomposition rate, and environmental modifiers have been identified as determinants in explaining the spatial patterns of global SOC storage by process-based models (Tao et al., 2023) (see also [Figure S9](#)). In this study, these components show converged spatial patterns despite structurally different models after being informed by observations. In contrast, other model components that are less important for determining global SOC storage (e.g., carbon input allocation and vertical transport) did not converge in the simulations by CLM5 and COMPAS. This difference is probably caused by insufficient information in the data to constrain parameters underlying these specific components (more discussion on this issue in [Section 4.3](#)).

The converged simulations of SOC and its related components demonstrate the fact that although it is impossible to include all the processes in the soil carbon cycle into one process-based model, unresolved processes can be well accounted for in model parameter values at resolved scales after data assimilation (Luo & Schuur, 2020). In this study, COMPAS explicitly describes the microbial CUE that represents the carbon partitioning process in microbial physiology and non-microbial carbon transfer that relates to other biological, chemical, and physical reactions driving organic matter transformations in soils. CLM5, however, does not differentiate these two processes in its structure but represents them through aggregated carbon transfer coefficients (see Methods). After being optimized by the PRODA approach, CLM5 simulates similar spatial patterns of the carbon transfer index with COMPAS ([Figure 4](#)). Similarly, a previous study reported that a process-based model that does not explicitly couple nitrogen-related processes with the soil carbon cycle can still well represent nitrogen limitation after its parameters were constrained by data (Wang et al., 2022).

4.2 | Data assimilation identifies most probable decomposition kinetics at global scale

Representations of organic carbon decomposition in soils has been debated for decades. In this study, we compared two possible SOC

decomposition kinetics at the global scale, namely a linear first-order kinetic model in CLM5 and a nonlinear Michaelis–Menten kinetic model in COMPAS. Our data assimilation results suggest that first-order kinetics may be the simplest and effective mechanism in explaining global SOC storage and its spatial patterns. After PRODA optimization, CLM5 and COMPAS show similar performance in explaining the spatial variability of SOC across the globe. A linear model such as CLM5 that adequately considers the spatial heterogeneity of its parameters can sufficiently capture the variability in space simulating the soil carbon cycle. Indeed, notwithstanding its simplicity, the linear relationship between the decomposition rate and the substrate concentration has been observed from macroscopic litter and soil organic carbon decomposition experiments (Cai et al., 2018; Luo, 2022; Schädel et al., 2014; Xu et al., 2016; Zhang et al., 2008).

Microorganism-centric kinetics (e.g., Michaelis–Menten kinetics) that considers enzymatic depolymerization has been advocated in recent years to account for the nonlinearity in organic carbon decomposition such that the decomposition rate is a function of both the substrate and the enzyme concentrations. Nonlinear kinetics can help capture the spatial variability of soil carbon dynamics (Wieder et al., 2013) and is necessary for understanding lignin decomposition (Liao et al., 2022) and priming effects (Wutzler & Reichstein, 2008). In this study, our data assimilation results show that, at the global scale, nonlinearity in COMPAS does not necessarily lead to more accurate quantification of SOC storage than CLM5. In fact, after being informed by data constraints, the Michaelis constants in COMPAS were much larger than their corresponding substrate concentrations ([Figure 5](#)). In such a case, the Michaelis–Menten kinetics can be mathematically approximated by a linear structure with respect to its corresponding substrate, but also including a first-order effect of the receiver pool, resulting in a multiplicative kinetics.

It should be noted that diversity in model structures is still necessary for a better understanding of the soil carbon cycle at different spatial and temporal scales. Microbial models with nonlinear structures can be useful for studying complex carbon dynamics at small scales that linear models cannot explain (Liao et al., 2022; Manzoni & Porporato, 2007). Meanwhile, microbial responses to environmental fluctuations are highly nonlinear and can be captured only by modeling specific microbial processes (Brangari et al., 2020). Moreover, models simulating SOC storage with different structures can perform differently across subregions, suggesting that some structures are more suitable for certain pedoclimatic conditions. For example, we have detected different patterns of SOC storage simulated by CLM5 and COMPAS in boreal (e.g., East Siberia) and tropical regions (e.g., Amazon and Congo Basins), even though the common observational SOC data constrained both models. The Michaelis–Menten kinetics investigated in this study is only one possibility from an array of theories. How other nonlinear kinetics, such as reverse Michaelis–Menten kinetics (Tang & Riley, 2019), perform in simulating SOC at different scales in comparison with linear models requires more studies in the future.

4.3 | More and high-quality data required to diminish prediction uncertainty

Uncertainty still exists in predicting SOC storage by structurally different models after PRODA optimization (Figure S6). The PRODA approach used in this study reveals the spatial heterogeneity of model parameters after site-level data assimilation. Thus, at the global scale, PRODA optimizes about 1.41 million parameter values (21 selected parameters for each of the 66,935 vertical SOC profiles) for CLM5 and 1.37 million parameter values (23 selected parameters for each of the 59,476 vertical SOC profiles) for COMPAS across observational sites. The posterior distributions of different parameters showed substantial uncertainties after data assimilation at the site level. In an example of data assimilation at one site (Figure S6), while a few parameters can be well constrained by vertical SOC profile data, resulting in narrower posterior distributions than the priors, more than half of the selected parameters had weak identifiability to the observations such that their posterior distribution showed flat shapes within the prior ranges.

The identifiability of different parameters is associated with the convergence of their corresponding model components by structurally different models and further affects the final global SOC simulations (Luo et al., 2009). For parameters well constrained by vertical SOC profiles in data assimilation, their corresponding model components (e.g., carbon transfer efficiency, baseline decomposition, and environmental modifiers) also showed similar spatial patterns between CLM5 and COMPAS despite differences in model structures. The revealed spatial patterns of these model components further presented high explanatory power to predict model-simulated SOC spatial patterns across the globe (Tao et al., 2023) (Figure S9). In contrast, for parameters that are less identifiable after data assimilation, different choices of optimized parameter value could lead to similar simulation of SOC storage, causing the so-called equifinality problem. Even simple models constrained by detailed data face this problem (Marschmann et al., 2019). Thus, the spatial pattern of their corresponding components, such as vertical transport and carbon input allocation, did not agree well between CLM5 and COMPAS after data assimilation in different models. Their spatial variability was also less responsible for the predictive accuracy of global SOC simulations (Figure S9). In the future, improved performance of process-based models in simulating the global patterns of SOC storage relies on a better understanding of those key components (e.g., carbon transfer, baseline decomposition, and environmental modifier) and their underlying mechanisms (e.g., microbial carbon use efficiency and organo-mineral interactions).

The equifinality problem (or weak identifiability of parameters) imposes challenges to using the optimized models to predict future SOC changes under climate change. In this study, we found that the spatial patterns of vertical transport and carbon input allocation may be less consequential to simulating steady-state SOC storage at the global scale. However, both these processes can influence the physical disconnection of SOC from decomposers,

so they could regulate the transient dynamics of SOC in response to climate change, warranting further investigations. Moreover, despite reasonable correlations between results retrieved from the two structurally different models, carbon transfer efficiency and baseline decomposition simulated by CLM5 and COMPAS are numerically different (i.e., not on the 1:1 line in Figure 4). Whether structurally different models after PRODA optimization can also predict converged SOC changes at different temporal scales is still an open question.

Higher oversight of data quality control and broader inclusion of other types of observational data related to soil carbon cycle at different spatial-temporal scales are the keys to resolving the equifinality problem and better predictions of SOC dynamics. Our results demonstrated that applying the PRODA approach with observational constraints can effectively realize converged simulations of SOC storage by structurally different models, even if they could generate contrasting simulation results before PRODA optimization. While providing comprehensive and quality-controlled soil data worldwide, the dataset used in this study still has substantial measurement uncertainty in SOC content data (Batjes et al., 2020). The absence of SOC content information at deeper soils and irregularities of vertical SOC profiles resulting from measurement errors could cause difficulties in data assimilation convergence and parameter optimization to simulate SOC storage accurately (see descriptions in Section 2.1). Thus, higher oversight of quality control and quality assurance is critical to improving prediction and understanding of SOC storage across scales.

Moreover, beyond SOC content data, an array of measurements could be used in the PRODA approach to further improve model predictive ability and inform model development. Measured carbon pools with clear physical meanings, such as particulate and mineral-associated organic carbon, can help constrain their conceptual counterparts in models (Abramoff et al., 2022; Guo et al., 2022). Meanwhile, time series flux data for the decomposition of different soil carbon pools and isotopes could help better understand decomposition kinetics and varying nutrient limitation mechanisms (Manzoni et al., 2021). In addition to carbon pool and flux data, microbial trait data can inform some model parameters or offer avenues for testing emerging properties such as CUE. For example, data related to microbial carbon use efficiency could constrain carbon transfer-related parameters, but only if measurements represent in situ conditions (e.g., using the ^{18}O incorporation method instead of adding labile ^{13}C sources) (Geyer et al., 2019). Moreover, including observations related to vegetation and hydrology dynamics in data assimilation may be more effective in understanding the spatial patterns of carbon input allocation and vertical transport.

5 | CONCLUSION

This study highlights the importance of high-quality field-measured data in informing model development and constraining simulations.

While diverse model structures stemming from different assumptions and theories, as well as the choices of parameter values, generate diverse possibilities in simulating SOC storage, data assimilation identifies the most probable ones that best explain the observations. The PRODA approach used in this study optimizes the parameters of a model based on first-order kinetics (i.e., CLM5) and one based on Michaelis–Menten kinetics (i.e., COMPAS). The two optimized models lead to convergence in simulating spatial patterns of both SOC storage and its related key components (i.e., the main contributing mechanisms), such as carbon transfer and baseline decomposition. Moreover, our PRODA approach reveals that the first-order kinetics has an equally effective explanation of SOC storage as the Michaelis–Menten kinetics at the global scale. In the future, it is still critical to explore various processes of the soil carbon cycle at different scales by developing structurally different models to be tested with new field-measured datasets. The development of tools such as PRODA will be critical in reconciling field observations and theoretical reasoning in modeling. New findings and patterns revealed by the PRODA approach will further stimulate new data acquisition and improvement of models.

AUTHOR CONTRIBUTIONS

Feng Tao: Conceptualization; data curation; formal analysis; investigation; methodology; software; validation; visualization; writing – original draft; writing – review and editing. **Benjamin Z. Houlton:** Writing – review and editing. **Yuanyuan Huang:** Writing – review and editing. **Ying-Ping Wang:** Writing – review and editing. **Stefano Manzoni:** Writing – review and editing. **Bernhard Ahrens:** Writing – review and editing. **Umakant Mishra:** Writing – review and editing. **Lifen Jiang:** Writing – review and editing. **Xiaomeng Huang:** Funding acquisition; resources; supervision; writing – review and editing. **Yiqi Luo:** Conceptualization; supervision; writing – original draft; writing – review and editing.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (42125503, 42075137) and the National Key Research and Development Program of China (2020YFA0607900, 2020YFA0608000, 2022YFE0195900, and 2021YFC3101600). We thank the support by the National Key Scientific and Technological Infrastructure project “Earth System Science Numerical Simulator Facility” (EarthLab). F.T. is supported by the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship, a Schmidt Futures program. S.M. has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 Research and Innovation Programme (grant agreement no.: 101001608). B.A. was funded by the Horizon Europe project AI4SoilHealth (grant no.: 101086179). Contributions of U.M. were supported through a US Department of Energy grant to the Sandia National Laboratories, which is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the US Department of Energy’s National Nuclear Security Administration under contract

DE-NA-0003525. Y.L. is supported by US National Science Foundation (NSF) Grants (DEB 1655499 and DEB 2242034), US Department of Energy, Terrestrial Ecosystem Sciences Grant DE-SC0023514, the subcontract CW39470 from the Oak Ridge National Laboratory to Cornell University, and the project NYS Connects: Climate Smart Farms & Forestry funded by US Department of Agriculture (USDA), New York State Department of Environmental Conservation, and New York State Department of Agriculture and Markets. The research of Y.L. is also part of AI-CLIMATE: “AI Institute for Climate-Land Interactions, Mitigation, Adaptation, Tradeoffs and Economy” supported by USDA National Institute of Food and Agriculture (NIFA) and NSF National AI Research Institutes Competitive Award no. 2023-67021-39829.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no competing interests.










DATA AVAILABILITY STATEMENT

All the data that support the findings of this study is available via <https://doi.org/10.5281/zenodo.10957105>. The raw data of SOC profiles from WoSIS database can be accessed at <https://www.isric.org/explore/wosis>.

CODE AVAILABILITY

All the code used in the analyses presented in this paper is available via https://github.com/phxtao/SOC_Convergence.

ORCID

Feng Tao  <https://orcid.org/0000-0001-6105-860X>
 Benjamin Z. Houlton  <https://orcid.org/0000-0002-1414-0261>
 Yuanyuan Huang  <https://orcid.org/0000-0003-4202-8071>
 Ying-Ping Wang  <https://orcid.org/0000-0002-4614-6203>
 Stefano Manzoni  <https://orcid.org/0000-0002-5960-5712>
 Bernhard Ahrens  <https://orcid.org/0000-0001-7226-6682>
 Umakant Mishra  <https://orcid.org/0000-0001-5123-2803>
 Lifen Jiang  <https://orcid.org/0000-0002-1546-8189>
 Xiaomeng Huang  <https://orcid.org/0000-0002-4158-1089>
 Yiqi Luo  <https://orcid.org/0000-0002-4556-0218>

REFERENCES

- Abramoff, R. Z., Guenet, B., Zhang, H., Georgiou, K., Xu, X., Rossel, R. A. V., Yuan, W., & Ciais, P. (2022). Improved global-scale predictions of soil carbon stocks with millennial version 2. *Soil Biology and Biochemistry*, 164, 108466.
- Allison, S. D., Wallenstein, M. D., & Bradford, M. A. (2010). Soil-carbon response to warming dependent on microbial physiology. *Nature Geoscience*, 3(5), 336–340.
- Batjes, N. (2016). Harmonized soil property values for broad-scale modelling (WISE30sec) with estimates of global soil carbon stocks. *Geoderma*, 269, 61–68.
- Batjes, N. H., Ribeiro, E., & Van Oostrum, A. (2020). Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019). *Earth System Science Data*, 12(1), 299–320.
- Bradford, M. A., Wieder, W. R., Bonan, G. B., Fierer, N., Raymond, P. A., & Crowther, T. W. (2016). Managing uncertainty in soil carbon feedbacks to climate change. *Nature Climate Change*, 6(8), 751–758.

- Brangarí, A. C., Manzoni, S., & Rousk, J. (2020). A soil microbial model to analyze decoupled microbial growth and respiration during soil drying and rewetting. *Soil Biology and Biochemistry*, *148*, 107871. <https://doi.org/10.1016/j.soilbio.2020.107871>
- Briggs, G. E., & Haldane, J. B. S. (1925). A note on the kinetics of enzyme action. *Biochemical Journal*, *19*(2), 338–339.
- Cai, A., Liang, G., Zhang, X., Zhang, W., Li, L., Rui, Y., Xu, M., & Luo, Y. (2018). Long-term straw decomposition in agro-ecosystems described by a unified three-exponentiation equation with thermal time. *Science of the Total Environment*, *636*, 699–708.
- Chandel, A., Jiang, L., & Luo, Y. (2023). Microbial models for simulating soil carbon dynamics: A review. *Journal of Geophysical Research Biogeosciences*, *128*, 1–27.
- Ciais, P., Sabine, C., Bala, G., Bopp, L., Brovkin, V., Canadell, J., Chhabra, A., DeFries, R., Galloway, J., Heimann, M., Jones, C., Quéré, C. L., Myneni, R., Piao, S., Thornton, P., Metz, N., & Wania, R. (2014). Carbon and other biogeochemical cycles. In T.F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex & P.M. Midgley (Eds.), *Climate change 2013: The physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change* (pp. 465–570). Cambridge University Press.
- Cotrufo, M. F., Soong, J. L., Horton, A. J., Campbell, E. E., Haddix, M. L., Wall, D. H., & Parton, W. J. (2015). Formation of soil organic matter via biochemical and physical pathways of litter mass loss. *Nature Geoscience*, *8*(10), 776–779.
- Cotrufo, M. F., Wallenstein, M. D., Boot, C. M., Deneff, K., & Paul, E. (2013). The microbial efficiency-matrix stabilization (MEMS) framework integrates plant litter decomposition with soil organic matter stabilization: Do labile plant inputs form stable soil organic matter? *Global Change Biology*, *19*(4), 988–995.
- FAO/IIASA/ISRIC/ISSCAS/JRC. (2012). *Harmonized world soil database (version 1.2)*. FAO.
- Forney, D. C., & Rothman, D. H. (2012). Common structure in the heterogeneity of plant-matter decay. *Journal of the Royal Society Interface*, *9*(74), 2255–2267.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). CRC press Boca.
- Georgiou, K., Abramoff, R. Z., Harte, J., Riley, W. J., & Torn, M. S. (2017). Microbial community-level regulation explains soil carbon responses to long-term litter manipulations. *Nature Communications*, *8*(1), 1–10.
- Geyer, K. M., Dijkstra, P., Sinsabaugh, R., & Frey, S. D. (2019). Clarifying the interpretation of carbon use efficiency in soil through methods comparison. *Soil Biology and Biochemistry*, *128*, 79–88. <https://doi.org/10.1016/j.soilbio.2018.09.036>
- Geyer, K. M., Kyker-Snowman, E., Grandy, A. S., & Frey, S. D. (2016). Microbial carbon use efficiency: Accounting for population, community, and ecosystem-scale controls over the fate of metabolized organic matter. *Biogeochemistry*, *127*(2–3), 173–188.
- Grigal, D., Brovold, S., Nord, W., & Ohmann, L. (1989). Bulk density of surface soils and peat in the north central United States. *Canadian Journal of Soil Science*, *69*(4), 895–900.
- Guo, X., Viscarra Rossel, R. A., Wang, G., Xiao, L., Wang, M., Zhang, S., & Luo, Z. (2022). Particulate and mineral-associated organic carbon turnover revealed by modelling their long-term dynamics. *Soil Biology and Biochemistry*, *173*, 108780. <https://doi.org/10.1016/j.soilbio.2022.108780>
- Haario, H., Saksman, E., & Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, *7*(2), 223–242.
- He, X., Abramoff, R., Abs, E., Georgiou, K., Zhang, H., & Goll, D. S. (2023). Contribution of carbon inputs to soil carbon accumulation cannot be neglected. *bioRxiv*, 2023.2007.2017.549330.
- He, Y., Trumbore, S. E., Torn, M. S., Harden, J. W., Vaughn, L. J., Allison, S. D., & Randerson, J. T. (2016). Radiocarbon constraints imply reduced carbon uptake by soils during the 21st century. *Science*, *353*(6306), 1419–1424.
- Huang, Y., Lu, X., Shi, Z., Lawrence, D., Koven, C. D., Xia, J., Du, Z., Kluzek, E., & Luo, Y. (2018). Matrix approach to land carbon cycle modeling: A case study with the community land model. *Global Change Biology*, *24*(3), 1394–1404.
- Hugelius, G., Tarnocai, C., Broll, G., Canadell, J., Kuhry, P., & Swanson, D. (2013). The northern circumpolar soil carbon database: Spatially distributed datasets of soil coverage and soil carbon storage in the northern permafrost regions. *Earth System Science Data*, *5*(1), 3–13.
- Jackson, R., Canadell, J., Ehleringer, J. R., Mooney, H., Sala, O., & Schulze, E. (1996). A global analysis of root distributions for terrestrial biomes. *Oecologia*, *108*(3), 389–411.
- Jackson, R. B., Lajtha, K., Crow, S. E., Hugelius, G., Kramer, M. G., & Piñeiro, G. (2017). The ecology of soil carbon: Pools, vulnerabilities, and biotic and abiotic controls. *Annual Review of Ecology, Evolution, and Systematics*, *48*, 419–445.
- Janssen, P., & Heuberger, P. (1995). Calibration of process-oriented models. *Ecological Modelling*, *83*(1–2), 55–66.
- Jenny, H. (1941). *Factors of soil formation*. McGraw-Hill Book Company, Inc.
- Lasaga, A. C. (1998). *Kinetic theory in the earth sciences*. Princeton University Press.
- Lawrence, D., Fisher, R., Koven, C., Oleson, K., Swenson, S., Vertenstein, M., Andre, B., Bonan, G., Ghimire, B., van Kampenhout, L., Kennedy, D., Kluzek, E., Knox, R., Lawrence, P., Li, F., Li, H., Lombardozzi, D., Lu, Y., Perket, J., ... Zeng, X. (2018). Technical description of version 5.0 of the Community Land Model (CLM).
- Lawrence, D. M., Fisher, R. A., Koven, C. D., Oleson, K. W., Swenson, S. C., Bonan, G., Collier, N., Ghimire, B., van Kampenhout, L., & Kennedy, D. (2019). The community land model version 5: Description of new features, benchmarking, and impact of forcing uncertainty. *Journal of Advances in Modeling Earth Systems*, *11*(12), 4245–4287.
- Li, Q., Xia, J., Shi, Z., Huang, K., Du, Z., Lin, G., & Luo, Y. (2016). Variation of parameters in a flux-based ecosystem model across 12 sites of terrestrial ecosystems in the conterminous USA. *Ecological Modelling*, *336*, 57–69.
- Liao, C., Huang, W., Wells, J., Zhao, R., Allen, K., Hou, E., Huang, X., Qiu, H., Tao, F., Jiang, L., Aguilos, M., Lin, L., Huang, X., & Luo, Y. (2022). Microbe-iron interactions control lignin decomposition in soil. *Soil Biology and Biochemistry*, *173*, 108803. <https://doi.org/10.1016/j.soilbio.2022.108803>
- Lu, X., Du, Z., Huang, Y., Lawrence, D., Kluzek, E., Collier, N., Lombardozzi, D., Sobhani, N., Schuur, E. A., & Luo, Y. (2020). Full implementation of matrix approach to biogeochemistry module of CLM5. *Journal of Advances in Modeling Earth Systems*, *12*(11), e2020MS002105.
- Lu, X., Wang, Y.-P., Luo, Y., & Jiang, L. (2018). Ecosystem carbon transit versus turnover times in response to climate warming and rising atmospheric CO₂ concentration. *Biogeosciences*, *15*(21), 6559–6572.
- Luo, Y. (2022). Theoretical foundation of the land carbon cycle and matrix approach. In Y. Luo & B. Smith (Eds.), *Land carbon cycle modeling: Matrix approach, data assimilation, & ecological forecasting*. CPC Press, Taylor & Francis Group.
- Luo, Y., Ahlström, A., Allison, S. D., Batjes, N. H., Brovkin, V., Carvalhais, N., Chappell, A., Ciais, P., Davidson, E. A., & Finzi, A. (2016). Toward more realistic projections of soil carbon dynamics by earth system models. *Global Biogeochemical Cycles*, *30*(1), 40–56.
- Luo, Y., Huang, Y., Sierra, C. A., Xia, J., Ahlström, A., Chen, Y., Hararuk, O., Hou, E., Jiang, L., Liao, C., Lu, X., Shi, Z., Smith, B., Tao, F., & Wang, Y.-P. (2022). Matrix approach to land carbon cycle modeling. *Journal of Advances in Modeling Earth Systems*, *14*, e2022MS003008. <https://doi.org/10.1029/2022MS003008>
- Luo, Y., Keenan, T. F., & Smith, M. (2015). Predictability of the terrestrial carbon cycle. *Global Change Biology*, *21*(5), 1737–1751.

- Luo, Y., Ogle, K., Tucker, C., Fei, S., Gao, C., LaDeau, S., Clark, J. S., & Schimel, D. S. (2011). Ecological forecasting and data assimilation in a data-rich era. *Ecological Applications*, 21(5), 1429–1442.
- Luo, Y., & Schuur, E. A. (2020). Model parameterization to represent processes at unresolved scales and changing properties of evolving systems. *Global Change Biology*, 26(3), 1109–1117.
- Luo, Y., Weng, E., Wu, X., Gao, C., Zhou, X., & Zhang, L. (2009). Parameter identifiability, constraint, and equifinality in data assimilation with ecosystem models. *Ecological Applications*, 19(3), 571–574.
- Manzoni, S., Čapek, P., Porada, P., Thurner, M., Winterdahl, M., Beer, C., Brüchert, V., Frouz, J., Herrmann, A. M., & Lindahl, B. D. (2018). Reviews and syntheses: Carbon use efficiency from organisms to ecosystems—definitions, theories, and empirical evidence. *Biogeosciences*, 15(19), 5929–5949.
- Manzoni, S., Chakrawal, A., Spohn, M., & Lindahl, B. D. (2021). Modeling microbial adaptations to nutrient limitation during litter decomposition. *Frontiers in Forests and Global Change*, 4, 686945.
- Manzoni, S., & Porporato, A. (2007). A theoretical analysis of nonlinearities and feedbacks in soil carbon and nitrogen cycles. *Soil Biology and Biochemistry*, 39(7), 1542–1556. <https://doi.org/10.1016/j.soilbio.2007.01.006>
- Marschmann, G. L., Pagel, H., Kügler, P., & Streck, T. (2019). Equifinality, sloppiness, and emergent structures of mechanistic soil biogeochemical models. *Environmental Modelling & Software*, 122, 104518. <https://doi.org/10.1016/j.envsoft.2019.104518>
- Mishra, U., Gautam, S., Riley, W., & Hoffman, F. M. (2020). Ensemble machine learning approach improves predicted spatial variation of surface soil organic carbon stocks in data-limited northern circumpolar region. *Frontiers in Big Data*, 3, 40.
- Parton, W., Schimel, D. S., Cole, C., & Ojima, D. (1987). Analysis of factors controlling soil organic matter levels in great plains grasslands. *Soil Science Society of America Journal*, 51(5), 1173–1179.
- Parton, W. J., Stewart, J. W., & Cole, C. V. (1988). Dynamics of C, N, P and S in grassland soils: A model. *Biogeochemistry*, 5(1), 109–131.
- Schädel, C., Schuur, E. A., Bracho, R., Elberling, B., Knoblauch, C., Lee, H., Luo, Y., Shaver, G. R., & Turetsky, M. R. (2014). Circumpolar assessment of permafrost C quality and its vulnerability over time using long-term incubation data. *Global Change Biology*, 20(2), 641–652.
- Schimel, J. (2023). Modeling ecosystem-scale carbon dynamics in soil: The microbial dimension. *Soil Biology and Biochemistry*, 178, 108948. <https://doi.org/10.1016/j.soilbio.2023.108948>
- Schimel, J. P., & Weintraub, M. N. (2003). The implications of exoenzyme activity on microbial carbon and nitrogen limitation in soil: A theoretical model. *Soil Biology and Biochemistry*, 35(4), 549–563.
- Schmidt, M. W., Torn, M. S., Abiven, S., Dittmar, T., Guggenberger, G., Janssens, I. A., Kleber, M., Kögel-Knabner, I., Lehmann, J., & Manning, D. A. (2011). Persistence of soil organic matter as an ecosystem property. *Nature*, 478(7367), 49–56.
- Shi, Z., Allison, S. D., He, Y., Levine, P. A., Hoyt, A. M., Beem-Miller, J., Zhu, Q., Wieder, W. R., Trumbore, S., & Randerson, J. T. (2020). The age distribution of global soil carbon inferred from radiocarbon measurements. *Nature Geoscience*, 13(8), 555–559.
- Tang, J., & Riley, W. J. (2019). Competitor and substrate sizes and diffusion together define enzymatic depolymerization and microbial substrate uptake rates. *Soil Biology and Biochemistry*, 139, 107624.
- Tao, F., Houlton, B. Z., Frey, S. D., Lehmann, J., Manzoni, S., Huang, Y., Jiang, L., Mishra, U., Hungate, B. A., Schmidt, M. W. I., Reichstein, M., Carvalhais, N., Ciais, P., Wang, Y.-P., Ahrens, B., Hugelius, G., Hocking, T. D., Lu, X., Shi, Z., ... Luo, Y. (2024). Reply to: Model uncertainty obscures major driver of soil carbon. *Nature*, 627(8002), E4–E6. <https://doi.org/10.1038/s41586-023-07000-9>
- Tao, F., Huang, Y., Hungate, B. A., Manzoni, S., Frey, S. D., Schmidt, M. W. I., Reichstein, M., Carvalhais, N., Ciais, P., Jiang, L., Lehmann, J., Wang, Y.-P., Houlton, B. Z., Ahrens, B., Mishra, U., Hugelius, G., Hocking, T. D., Lu, X., Shi, Z., ... Luo, Y. (2023). Microbial carbon use efficiency promotes global soil carbon storage. *Nature*, 618(7967), 981–985. <https://doi.org/10.1038/s41586-023-06042-3>
- Tao, F., & Luo, Y. (2022). PROcess-guided deep learning and DATA-driven modelling (PRODA). In Y. Luo & B. Smith (Eds.), *Land carbon cycle modeling: Matrix approach, data assimilation, and ecological forecasting*. Taylor and Francis.
- Tao, F., Zhou, Z., Huang, Y., Li, Q., Lu, X., Ma, S., Huang, X., Liang, Y., Hugelius, G., Jiang, L., Doughty, R., Ren, Z., & Luo, Y. (2020). Deep learning optimizes data-driven representation of soil organic carbon in earth system model over the conterminous United States. *Frontiers in Big Data*, 3(17), 1–15. <https://doi.org/10.3389/fdata.2020.00017>
- Todd-Brown, K., Randerson, J., Post, W., Hoffman, F., Tarnocai, C., Schuur, E., & Allison, S. (2013). Causes of variation in soil carbon simulations from CMIP5 earth system models and comparison with observations. *Biogeosciences*, 10(3), 1717–1736.
- Wang, S., Luo, Y., & Niu, S. (2022). Reparameterization required after model structure changes from carbon only to carbon-nitrogen coupling. *Journal of Advances in Modeling Earth Systems*, 14(4), e2021MS002798.
- Wang, Y. P., Zhang, H., Ciais, P., Goll, D., Huang, Y., Wood, J. D., Ollinger, S. V., Tang, X., & Prescher, A. K. (2021). Microbial activity and root carbon inputs are more important than soil carbon diffusion in simulating soil carbon profiles. *Journal of Geophysical Research Biogeosciences*, 126(4), e2020JG006205.
- Wei, N., Xia, J., Zhou, J., Jiang, L., Cui, E., Ping, J., & Luo, Y. (2022). Evolution of uncertainty in terrestrial carbon storage in earth system models from CMIP5 to CMIP6. *Journal of Climate*, 35(17), 5483–5499.
- Wieder, W. R., Bonan, G. B., & Allison, S. D. (2013). Global soil carbon projections are improved by modelling microbial processes. *Nature Climate Change*, 3(10), 909–912.
- Wieder, W. R., Hartman, M. D., Sulman, B. N., Wang, Y.-P., Koven, C. D., & Bonan, G. B. (2018). Carbon cycle confidence and uncertainty: Exploring variation among soil biogeochemical models. *Global Change Biology*, 24(4), 1563–1579. <https://doi.org/10.1111/gcb.13979>
- Wilson, C. H., & Gerber, S. (2021). Theoretical insights from upscaling Michaelis–Menten microbial dynamics in biogeochemical models: A dimensionless approach. *Biogeosciences*, 18(20), 5669–5679.
- Wutzler, T., & Reichstein, M. (2008). Colimitation of decomposition by substrate and decomposers—a comparison of model formulations. *Biogeosciences*, 5(3), 749–759.
- Xu, T., White, L., Hui, D., & Luo, Y. (2006). Probabilistic inversion of a terrestrial ecosystem model: Analysis of uncertainty in parameter estimation and model prediction. *Global Biogeochemical Cycles*, 20(2), 1–15.
- Xu, X., Shi, Z., Li, D., Rey, A., Ruan, H., Craine, J. M., Liang, J., Zhou, J., & Luo, Y. (2016). Soil properties control decomposition of soil organic carbon: Results from data-assimilation analysis. *Geoderma*, 262, 235–242.
- Yigini, Y., Olmedo, G., Reiter, S., Baritz, R., Viatkin, K., & Vargas, R. (2018). Soil organic carbon mapping: Cookbook.
- Zhang, D., Hui, D., Luo, Y., & Zhou, G. (2008). Rates of litter decomposition in terrestrial ecosystems: Global patterns and controlling factors. *Journal of Plant Ecology*, 1(2), 85–93.
- Zhou, S., Liang, J., Lu, X., Li, Q., Jiang, L., Zhang, Y., Schwalm, C. R., Fisher, J. B., Tjiputra, J., & Sitch, S. (2018). Sources of uncertainty in modeled land carbon storage within and across three MIPs: Diagnosis with three new techniques. *Journal of Climate*, 31(7), 2833–2851.
- Zhou, T., Shi, P., Jia, G., Dai, Y., Zhao, X., Shangguan, W., Du, L., Wu, H., & Luo, Y. (2015). Age-dependent forest carbon sink: Estimation via inverse modeling. *Journal of Geophysical Research: Biogeosciences*, 120(12), 2473–2492.

Zhou, T., Shi, P., Jia, G., & Luo, Y. (2013). Nonsteady state carbon sequestration in forest ecosystems of China estimated by data assimilation. *Journal of Geophysical Research: Biogeosciences*, 118(4), 1369–1384.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Tao, F., Houlton, B. Z., Huang, Y., Wang, Y.-P., Manzoni, S., Ahrens, B., Mishra, U., Jiang, L., Huang, X., & Luo, Y. (2024). 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>