Geoderma 473 (2026) 117967

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Systematic review and meta-analysis

Spatiotemporal modelling of soil organic carbon: integrating process-based and machine learning approaches

Yuanyuan Du [a,b,c], Budiman Minasny [d], Wartini Ng [d], Lei Zhang [e], Nicolas P.A. Saby [f], Yue Zhou [g], Bas van Wesemael [h], Yuxin Ma [i], Zheng Wang [a,b], Zhongkui Luo [a,b], Zhou Shi [a,b], Songchao Chen [a,b,c,*]

[a] State Key Laboratory of Soil Pollution Control and Safety, Zhejiang University, Hangzhou 310058, China
[b] Zhejiang Key Laboratory of Agricultural Remote Sensing and Information Technology, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China
[c] ZJU-Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University, Hangzhou 311215, China
[d] Sydney Institute of Agriculture, School of Life and Environmental Sciences, The University of Sydney, New South Wales, 2016, Australia
[e] Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[f] INRAE, Info&Sols, Orléans 45075, France
[g] Laboratoire de Géologie, Ecole Normale Supérieure, CNRS, Institut Pierre-Simon Laplace, Université Paris Sciences et Lettres, Paris 75005, France
[h] Earth and Life Institute, Université catholique de Louvain, Louvain-la-Neuve 1348, Belgium
[i] NSW Department of Climate Change, Energy, the Environment and Water, Parramatta NSW 2150, Australia

ARTICLE INFO

Handling Editor: Dr. H Neely

Keywords:
Spatiotemporal modelling
Knowledge-guided machine learning
Process-based model
Hybrid modelling
Deep learning
Data assimilation

ABSTRACT

Soil organic carbon (SOC) underpins the global carbon cycle and represents a central lever for climate change mitigation and food security. Yet its accurate spatiotemporal quantification remains a challenge, owing to the complex interactions among biological, chemical and physical processes operating across scales. This review critically evaluates the current state of SOC spatiotemporal modelling frameworks and their limitations, and future directions. Process-based models provide mechanistic insight into carbon dynamics but are constrained by parameterisation, structural assumptions and computational demands. In contrast, machine-learning (ML) approaches excel at capturing spatial patterns from large datasets but often struggle to represent temporal kinetics, enforce physical consistency or generalise across scales and environmental contexts. We argue that the next generation of SOC modelling will emerge from the convergence of process-based understanding and data-driven inference through knowledge-guided ML and hybrid modelling strategies. We synthesise four principal integration pathways: (1) *meta*-modelling to accelerate computationally intensive simulations; (2) sequential hybridisation to embed mechanistic trends as dynamic covariates; (3) ensemble frameworks to reduce structural uncertainty; and (4) data assimilation to constrain model trajectories with observations. We further highlight emerging frontiers, including residual and parameter learning, physics-informed neural networks, foundation models for earth observations, and omics-informed frameworks that link microbial functional potential to carbon turnover processes. We conclude that progress toward causally interpretable, uncertainty-aware monitoring frameworks will require tighter integration of mechanistic theory, interpretable artificial intelligence and cross-scale data synthesis.

1. Introduction

Soil organic carbon (SOC) plays a central role in terrestrial ecosystems, governing the global carbon (C) cycle and regulating essential soil functions and services (Evangelista et al., 2023; Lehmann et al., 2020; Lal et al., 2021). Soil represents the largest terrestrial C pool, storing about 2,500 Pg of C in the top two meters, more than three times the amount currently in the atmosphere and four times that in the biotic pool (Friedlingstein et al., 2024). While soils possess a potential pathway for atmospheric CO_2 drawdown, they also harbour the risk of becoming a significant source of CO_2 due to unsustainable anthropogenic practices and positive carbon cycle-climate feedbacks (Amelung et al., 2020).

* Corresponding author at: State Key Laboratory of Soil Pollution Control and Safety, Zhejiang University, Hangzhou 310058, China.
E-mail address: chensongchao@zju.edu.cn (S. Chen).

https://doi.org/10.1016/j.geoderma.2026.117967
Received 5 March 2026; Received in revised form 22 July 2026; Accepted 27 July 2026
Available online 30 July 2026
0016-7061/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Furthermore, SOC is central to soil condition or health assessment, agricultural resilience and carbon-crediting schemes (van Noordwijk et al., 2023).

Given these roles, accurate quantification of SOC stocks and characterization of their spatiotemporal dynamics are essential for assessing ecosystem resilience, agricultural productivity, and food security under global change. SOC is inherently spatially heterogeneous and temporally dynamic, governed by the balance between continuous carbon inputs (e.g., plant litter, rhizodeposition and atmospheric deposition) and loss processes (e.g., microbial respiration, leaching, and erosion). A critical determinant of these stock sizes is the SOC turnover time, commonly expressed as the average time that organic carbon molecules remain in the soil (Sierra et al., 2017). This rate shows significant spatial heterogeneity and is driven by mechanisms that differ substantially across regions. Consequently, many process-based models rely on SOC turnover time to accurately simulate carbon decay and storage potential (Zhang et al., 2025a; Zhou et al., 2024). This balance is strongly modulated by climate, soil properties, land use, and management, resulting in pronounced spatial and temporal variability in SOC stocks and fluxes (Smith, 2008; Lehmann et al., 2020). Crucially, this balance is no longer governed by gradual environmental change alone but is increasingly disrupted by climate extremes.

Recent literature further demonstrates that SOC dynamics are increasingly shaped by climate extremes. Heatwaves, droughts, and extreme precipitation events have been shown to accelerate SOC losses induced by directional climate change, particularly through pulses of microbial activity and destabilization of previously protected C pools (García-Palacios et al., 2021; Sáez-Sandino et al., 2024; Wang et al., 2024). Land-use and land-cover change remains another dominant driver, with conversion of forests and grasslands to croplands causing substantial reductions in topsoil (0–30 cm) SOC globally over the past four decades (Huang et al., 2024). Management can either amplify these losses through intensive tillage, residue removal and simplified rotations, or buffer them through reduced tillage, cover cropping, diversified rotations and organic amendments. These episodic and non-linear responses challenge models calibrated mainly on gradual environmental gradients or legacy observations. High-carbon systems such as peatlands further highlight this challenge: despite covering only a small fraction of the global land area, they store a disproportionate share of global soil carbon and are highly sensitive to drainage, fire and warming (Widyastuti et al., 2025; Yu, 2012).

Despite major advances in understanding SOC processes, substantial limitations persist in how SOC is modelled, mapped, and monitored. Digital soil mapping (DSM) has greatly improved the spatial representation of soil properties from regional to global scales (Arrouays et al., 2014; Chen et al., 2022), yet most applications implicitly assume temporal stationarity and rely on legacy datasets collected across disparate time periods. Consequently, many DSM products are not suited to capturing SOC dynamics and detecting emerging risks. Explicitly spatiotemporal SOC models offer a pathway forward by identifying where and when SOC stocks are most vulnerable. This spatiotemporal capability enables trend detection and supports monitoring designs focused on change detection rather than long-term averages (McBratney et al., 2003; Minasny et al., 2017).

Current space–time SOC modelling approaches remain fragmented. Process-based models (PBMs), geostatistical frameworks, and machine-learning (ML) methods have largely evolved in isolation. PBMs are grounded in mechanistic understanding but face challenges related to spatial scaling, parameter uncertainty, and equifinality (Van de Broek et al., 2025; Wang et al., 2025). Geostatistical approaches provide rigorous spatial inference but require explicit assumptions about space–time trends. Data-driven machine learning (ML) models excel at pattern recognition but often lack explicit process representation, limiting their reliability for temporal extrapolation and scenario analysis. A further challenge is that many SOC modelling studies report predictive accuracy but provide limited assessment of uncertainty propagation, temporal transferability, causal interpretability or suitability for monitoring and reporting applications. Recent advancements have shifted towards hybrid modelling frameworks that integrate the mechanistic rigor of PBMs with the predictive power of ML.

Against this background, there is a need to move beyond treating process-based models, statistical models and machine learning as separate approaches. This review is to evaluate existing approaches for spatiotemporal SOC quantification and synthesise emerging strategies for their integration. We focus on hybrid frameworks that combine process understanding, statistical structure and machine-learning flexibility to develop SOC models that are mechanistically informed, data-efficient, interpretable and scalable. In doing so, this review identifies key methodological challenges, evaluates opportunities for knowledge-guided and hybrid modelling, and proposes research priorities for robust SOC monitoring systems capable of supporting soil security, climate mitigation and sustainable land management.

2. Existing approaches for spatiotemporal SOC modelling

A range of modelling approaches has been developed to quantify SOC content and stock and their variation across space and time. These approaches differ in their underlying assumptions, mathematical structure, and treatment of soil processes, spatial dependence, and temporal change. Broadly, they can be grouped into process-based (mechanistic) models, geostatistical frameworks, and data-driven machine-learning methods (Table 1).

2.1. Process-based (mechanistic) models

PBMs are grounded in the mechanistic understanding of biogeochemical processes. Traditional SOC PBMs, exemplified by RothC (Coleman and Jenkinson, 1996), CENTURY (Parton, 1996), Yasso (Liski et al., 2005), DNDC (Li, 1996), etc., employ a multi-pool, first-order decay framework. SOC, usually expressed as stock (t/ha), is partitioned into discrete conceptual C pools (e.g., active, slow, passive) with specific decomposition rates modulated by environmental factors such as temperature and moisture (Fig. 1a). Their C pools are conceptually defined by intrinsic turnover times rather than physically measurable fractions (e.g., the "slow" pool cannot be isolated and quantified directly).

In response, a new generation of models has emerged that explicitly represents microbial processes and links them to measurable SOC pools (Schimel and Weintraub, 2003), including MEND (Wang et al., 2013), MIMICS (Wieder et al., 2014), CORPSE (Sulman et al., 2014), Millennial (Abramoff et al., 2018), MEMS (Robertson et al., 2019), and SOILcarb (Van de Broek et al., 2025). Many of these models shift the model from first-order decay formulations to Michaelis-Menten kinetics. They explicitly simulate microbial biomass, enzyme production, and carbon use efficiency (CUE), allowing them to capture phenomena such as the priming effect (where fresh inputs accelerate old C loss) and thermal adaptation, as well as dynamics that linear models fail to predict (Kyker-Snowman et al., 2020; Tao et al., 2023). Models such as MEMS v2.0 and Millennial align model pools with physically measurable soil fractions (i.e., POM, MAOM). These measurable pools enable researchers to directly calibrate models against empirically derived soil carbon fractions (Lavallee et al., 2020).

Recent modeling frameworks have also increasingly included representations of deep-soil carbon dynamics. Models such as RothPC-1 (Jenkinson and Coleman, 2008), COMISSION (Ahrens et al., 2015), OC-VGEN (Keyvanshokouhi et al., 2019), and MEMS v2.0 (Zhang et al., 2021) have moved towards multi-layer architectures that partition the soil profile into discrete vertical intervals. Deep-soil carbon dynamics are modelled based on the vertical transport of water-soluble organic matter, bioturbations and vertical distribution of root litter (Ryzhova et al., 2024).

For spatial applications, PBMs typically rely on gridded soil, climate, vegetation and land-management data as model inputs. PBMs are then

Table 1
A summary of spatiotemporal models of SOC.

Category	Method class	Core principle	Representative models / techniques	Key strengths	Key limitations	Best-suited applications
Conventional	Process-based models (PBMs)	Based on biogeochemical and physical processes	Traditional: RothC, CENTURY, Yasso, DNDC; Microbial-explicit: MIMICS, MEND, CORPSE, Millennial, MEMS, SOILcarb	Mechanistic causality; mass balance; climate sensitivity; scenario projection	Structural rigidities; parameter equifinality; pool validation	Long-term SOC projections; climate change scenario analysis
	Geostatistical models	Quantify spatial autocorrelation and covariance structure	OK, STK; INLA-SPDE	Rigorous uncertainty quantification; best linear unbiased predictor	Require regular space–time data; correlation-based; limited mechanistic insight	Regional to continental SOC mapping with sufficient space–time coverage
	Machine learning (ML)	Nonlinear statistical learning from covariates	RF, Cubist, XGBoost; Deep Learning: CNN, LSTM, GNN	High spatial accuracy; flexible; scalable	Poor extrapolation; no kinetics; black-box behaviour	High-resolution SOC maps, space for time substitution
Hybrid	Meta (surrogate) models	ML emulates PBM behaviour	RF-surrogates for APSIM	Orders-of-magnitude speed-up; preserves PBM logic	Inherits PBM biases; lacks explicit physical constraints; complex parameter tuning	Global upscaling; uncertainty analysis
	Sequential (physics-guided ML)	PBM outputs used as ML features	POML, RothC-RF	Improves temporal trends; physics-informed features	Systematic bias propagation from PBM; high computational intensity	SOC trend detection; gap filling
	Ensemble / Stacking	ML learns optimal weighting of multiple models	RothC-MIMICS-RF stack	Reduces structural uncertainty; robust predictions	Feature overshadowing; uncertainty accumulation	Cross-ecosystem modelling
	Data assimilation (DA)	Bayesian fusion of PBMs with observations	EnKF, Bayesian MCMC, PRODA	Strong uncertainty reduction; parameter convergence	Computational cost; equifinality persists; limited by observations	SOC MRV; long-term monitoring
Emerging	Soil-science-informed ML (Soil-ML)	Physical laws embedded in ML training, architecture, or loss	KGML-ag-Carbon	Physically consistent AI; strong generalization	Model design complexity	Data-sparse regions; policy-relevant forecasting
	Residual learning	ML used to predict the residuals of PBM	/	Modular; computationally efficient	Lack of physical knowledge in ML	Spatial refinement of PBM outputs
	Parameter learning	ML estimates uncertain parameters of PBM	dPL	Parameter constraint; Equifinality mitigation	Dependent on parameter-driver relationship	Parameter estimation in PBM
	Physics-informed neural networks (PINNs)	Neural networks constrained by SOC PDEs	Dual-neural network architecture	Works with sparse data; enforces mass balance	Methodologically complex; sensitivity to loss function weighting	Single profile
	Foundation models (FMs)	Self-supervised pre-training on Earth system data	FoundationSoil	Few-shot learning; strong transferability	Training cost; reliance on unlabeled datasets	Global upscaling; uncertainty analysis

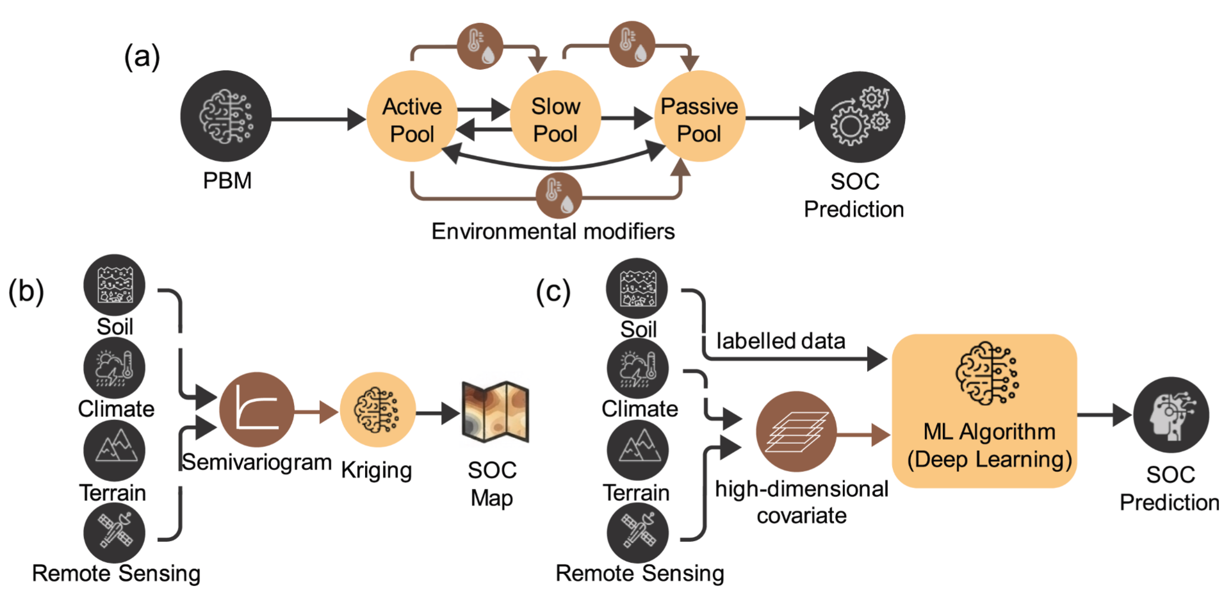

Fig. 1. Conceptual diagram of process-based model (a), geostatistical model (b), and machine learning model (c) for SOC spatiotemporal modelling.

executed independently for each grid cell, usually treating every pixel as a one-dimensional soil column. While this strategy enables large-scale application and straightforward integration with spatial datasets, it generally neglects horizontal fluxes of carbon, water, and sediments, and assumes that model parameters and processes operate identically across heterogeneous landscapes. Despite these spatial simplifications, PBMs remain the most practical and mechanistically grounded tools for projecting SOC sequestration under future climate scenarios.

A persistent challenge for PBMs is equifinality, in which different combinations of parameters or processes can yield similar SOC stock trajectories, sometimes for mechanistically incorrect reasons. To address this limitation, recent frameworks such as SOILcarb (Van de Broek et al., 2025) and PRODA (Tao et al., 2023) employ data assimilation (DA) techniques (see details in section 4.4). For example, the SOILcarb model constrained the residence times of deep soil C by integrating depth-resolved isotope Δ^{14}C constraints. Specifically, it used the bomb-spike gradient in Δ^{14}C to infer carbon turnover rates, thereby providing a robust observation for model calibration. Such approaches illustrate how specific isotopic, fraction-based and depth-resolved observations are required to reduce parameter uncertainty and improve the mechanistic credibility of SOC projections.

2.2. Geostatistical spatiotemporal models

Geostatistics is grounded in the theory of regionalised variables, especially the Gaussian Random Field (GRF), which considers that any variable studied, such as SOC, is the result of a stochastic process that can be linked to a mathematical function (the variogram) quantifying the spatial dependence between observations (Matheron, 1960). Subsequently, ordinary kriging (OK) uses the spatial structure described by a variogram to produce estimates by spatial interpolation at unvisited locations. OK is therefore based solely on sampled SOC data and does not take into account associated covariates (Fig. 1b). Regression kriging (RK) extends this framework by modelling the deterministic trend using environmental covariates, such as SCORPAN factors, and then applying kriging to the regression residuals to account for remaining spatial autocorrelation (Heuvelink and Webster, 2022). Under the assumed model and variogram, OK and RK provide the best linear unbiased predictor, offering not only a prediction but also a rigorous estimate of the variance of the prediction error (uncertainty).

Because SOC is temporally dynamic, purely two-dimensional spatial frameworks are often insufficient. Space-time Kriging (STK) extends the framework by modelling the covariance between points in both space and time. However, exact STK typically requires operations on dense space–time covariance matrices. Direct covariance-matrix computations scale as approximately $\mathscr{O}(n^3)$ for matrix inversion and $\mathscr{O}(n^2)$ for storage, where n is the number of observations. As SOC datasets increase in spatial coverage and temporal resolution, these computational demands can become prohibitive, although iterative solvers, including Krylov subspace methods, can reduce the burden in practice (Sun et al., 2012).

Approximation strategies, including local kriging, covariance tapering and composite likelihood approaches, have been developed to improve scalability, but they often require dataset-specific tuning and may compromise statistical optimality. To overcome these limitations, geostatistics has increasingly adopted stochastic partial differential equation (SPDE) approaches. The SPDE framework represents a continuously indexed Gaussian random field (GRF) as a Gaussian Markov random field (GMRF) with sparse precision matrices. Lindgren et al. (2011) established the explicit link between Matérn GRFs and GMRFs through an SPDE formulation solved using finite element methods. This representation transforms kriging and stochastic simulation into sparse linear algebra problems, reducing computational complexity to approximately $\mathscr{O}(n^{3/2})$ in two dimensions (and near-linear scaling in memory), enabling scalable spatiotemporal modelling for large environmental datasets.

SPDE models can be implemented within a fully Bayesian estimation framework based on the integrated nested Laplace approximation (INLA), yielding an INLA-SPDE model. The INLA method, developed by Rue et al. (2009), provides an alternative to traditional Markov chain Monte Carlo methods for Bayesian estimation and offers fast, accurate, off-the-shelf deterministic approximations of posterior inferences for a large class of models. INLA allows joint estimation of all model components using many data points, ensuring that estimation uncertainty is propagated across all components. A recent example of spatio-temporal modelling of SOC with INLA-SPDE employed separable spatio-temporal covariance models, meaning the covariance matrix can be expressed as the product of the spatial and temporal covariance matrices. Sun et al. (2021) included time as an autoregression model of order 1.

Saby et al. (2020) applied their approach to a French SOC dataset to model national-scale SOC dynamics. The statistical spatio-temporal model used basis functions representation to combine several effects: e.g., regression with covariates $\beta_i^{sc} z_i^{sc}$, a global temporal slope term $\beta_0 + \beta_1 \times t$, and intercept $W_0(s)$ and slope $W_1(s)$ coefficients that are allowed to vary across space ($W_0(s) + W_1(s) \times t$). A final effect corresponds to a nonlinear spatiotemporal structure, represented with a deterministic spline basis $B_k(t)$ for nonlinear modelling of temporal trends, and each of the temporal spline coefficients has spatial structure defined through a Gaussian random field W_k. In full, the model can be written as:

$$Y(s,t) = \beta_0 + \beta_1 \times t + \sum_{i=1}^{S} \beta_i^{sc} z_i^{sc}(s) + W_0(s) + W_1(s) \times t + \sum_{k=1}^{K} B_k(t) W_k(s) + \varepsilon(s,t) \quad (1)$$

where $\varepsilon(s,t)$ denotes a Gaussian white-noise process representing measurement errors, with a mean of 0 and a variance of σ_ε^2.

Fig. 2 presents the predicted distributions of soil organic matter (SOM) contents using the model fitted with INLA-SPDE. It shows that the temporal change of the SOM is important between 1995 and 2016, while the spatial patterns of the SOM distributions are basically similar during the period. The spatial gradient of SOM is mainly driven by the oceanic climate and elevation, with higher values in the western part of the study area. SOM contents decreased significantly from 1995 to 2016; that is, the red colour in the western area of the study region changes to yellow. Primarily, this is attributed to the fact that, in this area of the study, a number of plots underwent grassland conversion for cultivation.

INLA-SPDE currently represents the state of the art in geostatistics, offering a flexible framework for modelling complex anisotropic correlation structures and non-Gaussian data (e.g., zero-inflated soil counts) while maintaining full probabilistic quantification of uncertainty (Rue et al., 2017). However, especially when no covariates are inserted in the model, it is highly sensitive to the SOC training data. When observations across sampling periods are scarce, the uncertainty in spatio-temporal predictions remains high. Mechanistically, INLA-SPDE lacks physical constraints and relies on predefined statistical structures to model temporal trends.

2.3. Machine learning (data-driven) models

In contrast to both mechanistic and geostatistical frameworks, machine-learning approaches prioritise predictive performance by learning empirical relationships between SOC observations and large sets of environmental covariates. The growing availability of spatial, temporal, and remote-sensing datasets has accelerated the adoption of ML in SOC modelling, extending earlier geostatistical approaches (Chen et al., 2022). ML models relate observed SOC content or stock to predictor variables (e.g., climate, terrain, and remote-sensing indices; Fig. 1c). Static covariates primarily represent spatial variation, whereas dynamic covariates, particularly remote-sensing time series, capture temporal variability in environmental conditions and land management. Most applications have used tree-based ML algorithms such as Random Forest (RF), Cubist, and XGBoost (Chen et al., 2024a; Heuvelink et al., 2021; Liu et al., 2021; Wiesmeier et al., 2011).

Recent work also extended SOC modeling toward deep learning (DL). Unlike conventional ML, which treats each pixel as an isolated data point, DL architectures such as convolutional neural networks (CNNs) can ingest spatial context, recognizing that SOC at one location is

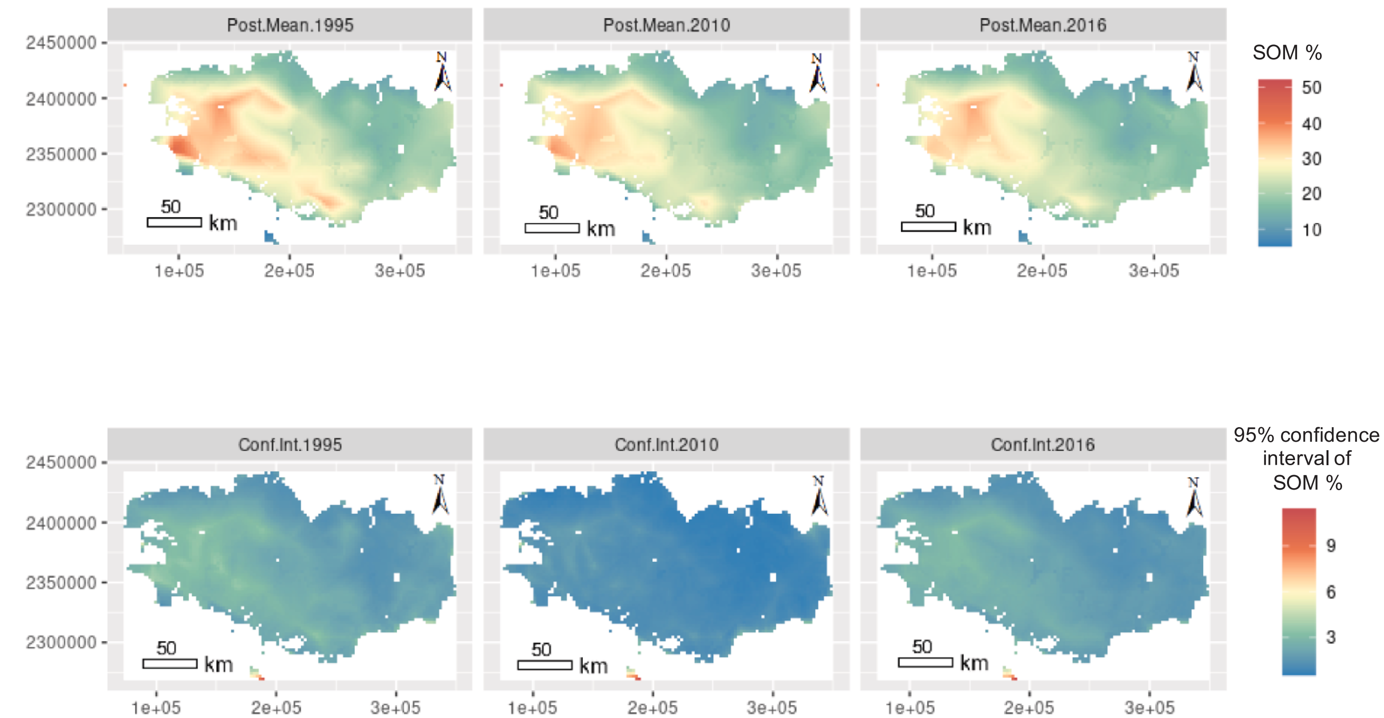

Fig. 2. Posterior mean of the soil organic matter (SOM in %) contents for 1995, 2010 and 2016 and relevant 95% confidence interval over the study area in France (revised from Saby et al., 2020) under the Lambert-93 projection.

influenced by the textures and patterns of its neighbours (Ng et al., 2020; Padarian et al., 2019). In addition, graph neural networks (GNNs) offer another advantage by capturing complex landscape topology and non-Euclidean connectivity that grid-based CNNs often miss. By modelling the landscape as a connected network, GNNs can explicitly represent directional fluxes, such as hydrological flow paths, providing a more physically realistic depiction of soil formation processes. Such architectures are promising for SOC prediction because they can capture multiscale spatial dependencies and landscape connectivity, although their advantages depend on data density, model design, and independent validation (Kipf and Welling, 2017; Shen et al., 2023).

Temporal information has also been incorporated through recurrent architectures such as Long Short-Term Memory (LSTM) networks. Zhang et al. (2022), for example, demonstrated that land-surface phenology features extracted using LSTM improved SOC prediction relative to RF. However, the application of LSTM to SOC modelling remains limited by the scarcity and irregularity of long-term SOC time-series data.

Remote sensing has become central to ML-based SOC mapping to capture temporal variation. Optical satellite platforms such as Landsat, Sentinel, MODIS, and PlanetScope provide covariates that capture vegetation, surface condition, and spectral soil properties. A major development has been the move from single-date imagery to multi-temporal analysis and bare-soil composites (Rogge et al., 2018; Xue et al., 2024). By screening observations across seasons and years, these approaches identify periods when exposed soil dominates the surface reflectance signal, improving the retrieval of soil-related spectral information (Safanelli et al., 2020; Vaudour et al., 2019). Dvorakova et al. (2023), for example, used Sentinel-2 spectral indices to reduce confounding effects from soil moisture and crop residues. Such spectral soilscapes support SOC mapping using visible–near-infrared (VNIR) and shortwave-infrared (SWIR) bands (Demattê et al., 2018).

Nevertheless, bare-soil detection remains challenging. Vegetation can often be masked using NDVI thresholds, but reliable SOC retrieval also requires filtering residues, roughness, and moisture effects (Chabrillat et al., 2019). Dvorakova et al. (2023) showed that locally calibrated thresholds based on the normalized burn ratio 2 (NBR2), derived from SWIR bands, can reduce moisture and residue effects. However, the broad SWIR bands available on current multispectral sensors cannot fully separate these influences. Future hyperspectral satellite missions may improve residue detection through narrow-band indices such as the cellulose absorption index. Sentinel-1 synthetic aperture radar (SAR) can also provide soil-moisture time series, which may serve as dynamic covariates for SOC prediction (Fan et al., 2025).

A further advance is the derivation of spatially explicit management variables from remote sensing (Kc et al., 2021; Zhou et al., 2025). These variables can be used both as covariates in ML models (Yang et al., 2020; Hu et al., 2021; Zhou et al., 2022) and as inputs for PBMs, thereby improving the spatiotemporal representation of management drivers for SOC prediction (Xia et al., 2025; Zhou et al., 2024). For example, Zhou et al. (2025) combined time-series optical and radar data to estimate cover-crop duration and tillage practices using an RF framework (Fig. 3). When these RS-derived management variables were used in RothC, the model yielded a limited bias of 1.03t C ha^{-1} when upscaling SOC simulations across 10,102 agricultural fields covering 50,656 ha in Wallonia, Belgium (Zhou, 2024).

2.4. Synthesis: Complementary strengths and structural limits of existing approaches

The modelling approaches discussed above have complementary strengths and limitations. PBMs and ML approaches have often developed along separate trajectories, yet neither is sufficient on its own to represent the full complexity of SOC dynamics. Recent work in earth system modelling therefore points toward closer integration of mechanistic understanding and data-driven prediction (Reichstein et al., 2019; Irrgang et al., 2021; Minasny et al., 2024).

The case for integration arises from the distinct limitations of each modelling approach. PBMs provide causal structure and physical consistency: they can represent why SOC changes and enforce constraints such as mass balance. However, their performance is often limited by structural assumptions, uncertain process representation, and difficult parameterisation, particularly when kinetic constants are applied across

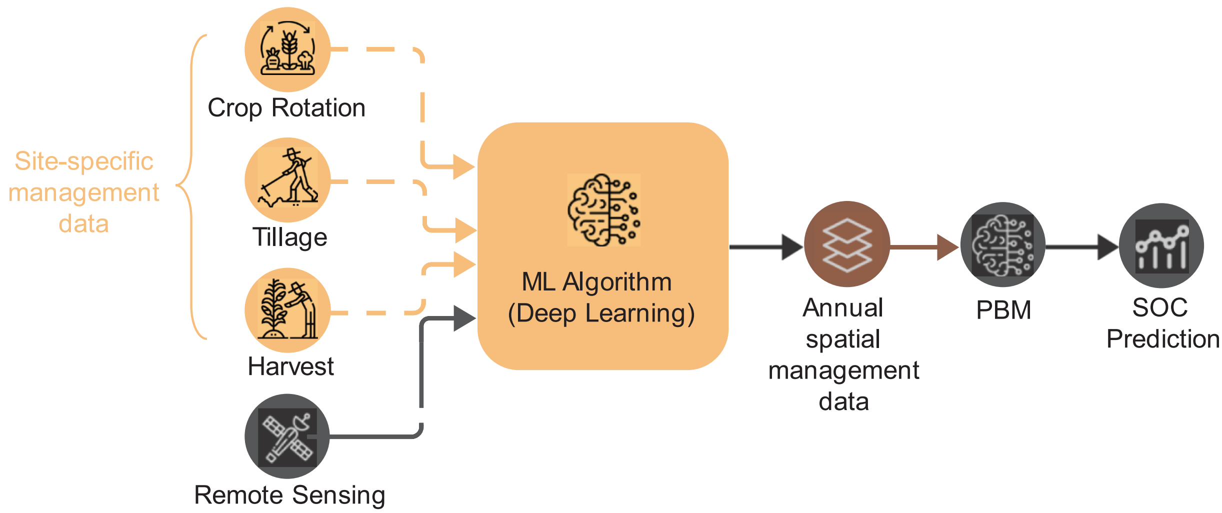

Fig. 3. The framework of generating spatially explicit management data from remote sensing as fundamental inputs for PBMs. Solid arrows denote the RS data used as input covariates, dashed arrows denote the management variables used as training labels.

heterogeneous soils and landscapes (Luo et al., 2022).

Conversely, ML models have demonstrated superior performance in capturing spatial heterogeneity and nonlinear patterns within the training domain (interpolation). Yet, they face a critical generalization barrier. In regions where sample data are limited, ML models are susceptible to overfitting, capturing localized noise rather than broad ecological signals. Furthermore, these algorithms typically predict within the specific range of their training data; they often struggle with extrapolation. Because most ML models infer relationships from statistical correlations rather than biophysical laws, they tend to reproduce average conditions or even generate ecologically implausible predictions when extrapolated to novel climate regimes or data-sparse regions (Liu et al., 2024; Shen et al., 2023).

SOC research is therefore moving toward hybrid modelling strategies. A key motivation is the uneven density of SOC observations, spatial datasets are increasingly available, whereas repeated temporal measurements remain scarce. ML methods can capture spatial patterns, but they often provide a weak representation of temporal SOC dynamics. Mechanistic constraints can help improve extrapolation, fill temporal gaps, and produce SOC predictions that are more robust under changing climate and management conditions (Bodnar et al., 2025; Chen et al., 2023). Section 4 groups these integration approaches into four pathways.

3. Fundamental challenges in spatiotemporal SOC modelling

Despite methodological advances, major challenges remain in the spatiotemporal modelling of SOC content and stock. The accuracy, robustness, and interpretability of SOC assessments are constrained by the limitations in data availability and consistency, difficulties in representing temporal dynamics, and incomplete representation of key soil processes.

3.1. Data limitations and spatiotemporal sparsity

Despite progress in global soil databases, such as WoSIS (Batjes et al., 2024), the scarcity of high-quality, spatiotemporally consistent observations remains a major bottleneck for robust SOC modelling. These limitations arise from three related problems: geographical bias, temporal discontinuity, and methodological inconsistency. Legacy soil databases typically compile observations from multiple surveys, projects, and institutional sources, with sampling concentrated in particular regions or land uses rather than distributed according to a common design (Richer-de-Forges et al., 2025). Because samples were rarely collected with equal or known selection probabilities, their spatial and temporal distributions may introduce bias into model calibration and evaluation.

This sampling imbalance limits model generalisation. Existing global databases are disproportionately concentrated in North America, Europe, and parts of East Asia, whereas large areas of Asia and Africa remain sparsely sampled (Batjes et al., 2024; Nenkam et al., 2024). The problem is not only spatial coverage, but also covariate shift: the environmental conditions represented in the training data often differ from those in target prediction regions. As a result, ML and geostatistical models may be forced to extrapolate into poorly represented feature space. ML models may then learn correlations that do not transfer to new environments, while geostatistical models may suffer from weak spatial or spatiotemporal support in unsampled regions (Wadoux et al., 2020).

Temporal discontinuity is an equally important constraint. Most legacy soil profiles are static observations from single sampling campaigns, often collected decades apart and rarely revisited. This limits direct estimation of SOC change and forces researchers to infer dynamics from spatial contrasts or modelled covariate change. Space-for-time substitution assumes that spatial gradients, for example in climate or land use, can approximate temporal responses (Pickett, 1989). For example, Yigini and Panagos (2016) predicted current SOC across Europe using regression kriging with environmental covariates, then projected SOC to 2050 by modifying climate and land-use inputs. Such projections are inherently extrapolative and often imply movement between discrete equilibrium states. Recent work suggests that space-for-time substitution can miss transient responses, hysteresis, and rapid nonlinear SOC losses triggered by extreme events or abrupt environmental change (Evans et al., 2025; Kreyling, 2025).

Inconsistencies in data quality introduce substantial uncertainty, particularly in SOC stocks, which are calculated as the product of SOC concentration, depth, coarse fraction (CF), and bulk density (BD) (Chen et al., 2024b):

$$SOC_{stock} = SOC^{*}BD^{*}depth^{*}(1 - \mathrm{CF}) \tag{2}$$

where SOC_{stock} is expressed in t ha^{-1}, SOC is SOC concentration (% by mass), BD is bulk density (g cm^{-3}), depth is soil thickness (cm), and CF is the coarse fraction (>2 mm, expressed as a fraction).

Bulk density measurements are missing in more than 50% of global soil profile records, requiring the use of pedotransfer functions (PTFs; Batjes et al., 2024; Chen et al., 2025). Additional inconsistencies arise from differences in how BD and CF are defined and measured. BD may be reported for whole soil or for the fine earth fraction only, while CF may be quantified on either a mass or volumetric basis. When BD is measured for whole soil, CF must be expressed on a mass basis (Hobley et al., 2018). These inconsistencies and estimates could lead to cascading errors, whereby uncertainties in measurements and PTF-derived estimates propagate through SOC stock calculations, often resulting in errors of 20–40% (Hobley et al., 2018; Walter et al., 2016). To contextualize this magnitude, a global synthesis showed that surface SOC storage (0–40 cm) experienced a net increase of only 11.0% over a two-decade period (Liu et al., 2025a).

These issues, including data irregularity, non-stationarity, and uneven observations, represent fundamental data limitations rather than model-specific weaknesses and therefore affect all spatiotemporal SOC modelling approaches. Acknowledging these limitations is necessary to correctly interpret model outputs and assess their reliability. In addition, addressing these data limitations requires methods that explicitly represent and propagate data uncertainty.

3.2. Limits of ML approaches for temporal inference

Although ML has substantially improved spatial SOC mapping, its use for temporal forecasting remains constrained by the mismatch between static or sparsely repeated training data and the dynamic processes that control SOC change. In spatiotemporal SOC modelling, ML models primarily learn associations between observed SOC and high-dimensional environmental covariates. Algorithms such as RF are therefore often better suited to spatial interpolation than temporal extrapolation, especially when trained on static observations distributed across climatic gradients (e.g., Syu et al., 2025; Zhang et al., 2024b). In such cases, the model may learn equilibrium-type relationships, such as higher SOC in cooler environments, without learning the kinetics of SOC change. For example, when applied to future warming scenarios, an ML model may implicitly infer that a 2 °C increase will shift SOC toward values currently observed at sites that are 2 °C warmer, rather than estimating the transient response of SOC to warming.

This limitation exposes ML-based forecasts to two related problems: covariate shift and concept drift (Bayram et al., 2022). Covariate shift occurs when future predictor distributions differ from those represented in the training data, whereas concept drift occurs when the relationship between environmental drivers and SOC changes through time. These issues can lead to substantial projection errors under rapid warming, abrupt management change, or no-analogue future climates (Luo and Weng, 2011; Meyer and Pebesma, 2022).

In spatiotemporal mapping, ML models often use static covariates to represent soil and environmental conditions and remote sensing images to represent dynamic covariates (e.g., Tian et al., 2026; Yang et al., 2021; Yang et al., 2025; Zhang et al., 2023a). A critical, often overlooked challenge is the temporal-scale mismatch between remotely sensed environmental covariates and SOC dynamics (Fig. 4). Satellite sensors (e.g., Landsat, MODIS, Sentinel) capture surface reflectance at high temporal frequencies (days to weeks) and can detect rapid changes in vegetation phenology or land cover. In contrast, SOC is a slow-variable pool, often requiring decades or even centuries to reach a new equilibrium following a disturbance (Smith, 2008). For instance, in a land use change event such as the conversion of grassland to cropland, a satellite detects the conversion almost instantaneously (within days) due to the removal of vegetation. However, the resulting loss of SOC is a gradual process that may persist for 10–20 years (Tang et al., 2019; Jiao et al., 2012), and the rate varies substantially with climate and management. An ML model trained on instantaneous satellite predictors often fails to account for this lag effect (Fig. 5). It may erroneously predict an immediate drop in SOC or, conversely, overestimate SOC by correlating the new land cover with the legacy SOC from the previous ecosystem (memory effects in soil systems; Sanderman et al., 2017). Padarian et al. (2022) address this temporal lag by tracking land use changes and modelling the SOC dynamics based on the time since land use changes. This approach accounts for the lag between rapid vegetation phenological shifts and slow soil carbon turnover, enabling the model to capture the gradual decay or accumulation trajectories that spectra-driven models often miss.

Finally, the black-box nature of complex ML algorithms challenges their reliability for scientific inquiry. Unlike PBMs, which are explicitly constrained by process-based understanding, ML models optimize solely for statistical accuracy. This creates susceptibility to the so-called "Clever Hans" effect (Lapuschkin et al., 2019), in which a model appears to perform well but relies on spurious correlations or proxy signals that are unrelated to the underlying process of interest. For example, an ML model may predict high SOC in a given region not due to soil properties or carbon inputs, but because of systematic artifacts in satellite imagery, such as cloud shadows, sensor geometry, or region-specific acquisition conditions. Similarly, Rentschler and Scholten (2025) demonstrated that even irrelevant covariates can generate apparently strong predictive relationships. Such examples highlight the risk that ML models may exploit coincidental statistical patterns rather than real biogeochemical drivers. In the absence of physical or process-based interpretability, it becomes difficult to determine whether predictions reflect real biogeochemical drivers or coincidental statistical patterns, limiting the credibility of such models for extrapolation, policy

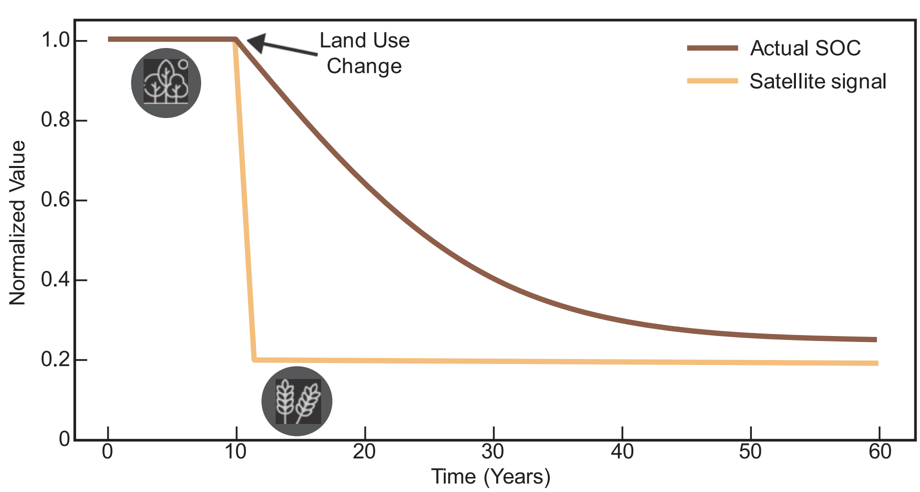

Fig. 4. The schematic of temporal mismatch between remote sensing covariates and soil processes. While satellite sensors detect land cover change (e.g., forest to cropland) instantaneously, SOC responds slowly due to physical protection and turnover kinetics, creating a blind spot for purely data-driven models. The y-axis represents the normalised value for either signal or SOC. The SOC decay curve assumes a turnover time-constant of approximately 36 years, representing the slow SOC pool.

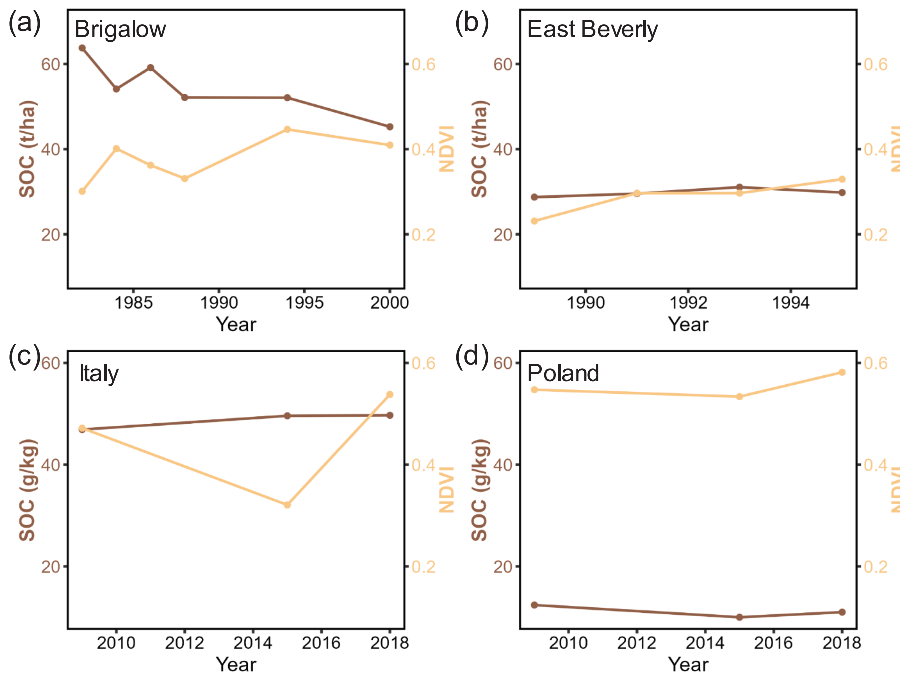

Fig. 5. Temporal dynamics of soil organic carbon (SOC) and Normalized Difference Vegetation Index (NDVI) for two long-term agricultural experimental sites (Brigalow and East Beverly) in Australia (a, b) and two LUCAS Soil sites (Italy and Poland) in Europe (c, d). The Australian sites remained croplands throughout the study period. The LUCAS sites underwent land use changes: the Italian site transitioned from cropland (2009) to grassland (2015) and forest (2018), while the Polish site shifted from forest (2009) to cropland (2015) and shrubland (2018). The units for SOC differ between the Australian (t/ha) and European (g/kg). The Australian data are from Luo et al. (2017), and the LUCAS data are by Orgiazzi et al. (2018) and Fernandez-Ugalde et al. (2022).

support, and land-management decision-making (Wadoux et al., 2020).

3.3. Limits of process-based models for spatial prediction

PBMs provide a mechanistic basis for interpreting causality, but their predictive use is constrained by model complexity, demanding input requirements, and parameter uncertainty. These limitations often render them less accurate than data-driven approaches for spatiotemporal prediction (Abramoff et al., 2022; Hendriks et al., 2021).

A central challenge is the tension between the complexity of SOC dynamics and the simplifications required in model design. PBMs represent SOC change through mathematical descriptions of physico-chemical and biological processes, yet many relevant pathways are difficult to parameterise explicitly. For example, the formation of mineral-associated organic carbon (MAOC) from root exudates involves microbial assimilation, necromass turnover, and mineral-surface adsorption rather than a single transformation step (Zhou et al., 2024). This pathway is only one component of a wider network of interacting SOC transformations (Dungait et al., 2012). As a result, many PBMs still simplify important aspects of soil physics, microbial activity, and organo-mineral interactions (Ding et al., 2025).

To represent spatial heterogeneity, PBMs are typically executed point-by-point or grid-by-grid, treating each location as an independent, one-dimensional soil column. This framework neglects lateral fluxes and spatial connectivity, such as the transport of carbon, water, and sediments along hillslopes, through hydrological flow paths, or via erosion and deposition processes. The importance of these lateral fluxes is illustrated by Ma et al. (2023), who showed that coupling RothPC–1 SOC dynamics model with a cellular automata erosion–deposition modelling substantially altered predicted SOC stocks and reduced systematic overestimation in erosional landscapes. Increasing the spatial resolution or modelling extent further increases computational demands, often requiring additional simplifications that reduce process representation.

The predictive reliability of PBMs is also constrained by uncertainty in parameterisation and initialisation. These models require extensive input data that are frequently sparse, incomplete, or unavailable at the spatial scales of application. Climate data are generally accessible, but detailed management histories, such as cropland tillage and irrigation, forest harvest intensity, and grassland grazing pressure, are rarely available regionally. In forest ecosystems, key inputs such as root turnover rates and litterfall chemistry are also poorly constrained. As a result, simulations often depend on generalized parameters and assumptions, increasing equifinality and reducing confidence in model predictions (Luo et al., 2009).

PBMs are often carefully calibrated but rarely independently validated. Garsia et al. (2023) indicated that about a third of the models they evaluated were never validated under acceptable criteria. An analysis by Le Noë et al. (2023) indicated that about 60% of the models they analysed are not designed for prediction but rather for conceptual understanding of soil processes.

4. Toward integrated and hybrid SOC modelling frameworks

The limitations identified in previous sections indicate that no single modelling approach can adequately represent spatiotemporal SOC dynamics on its own. This has motivated the development of hybrid frameworks that combine mechanistic understanding, statistical structure, and machine-learning flexibility. These approaches differ in how and where integration occurs within the modelling workflow, ranging from surrogate representations of process models to coupling through data assimilation.

4.1. Meta model

A *meta*-model, or surrogate model, trains an ML algorithm to statistically emulate the behaviour of a PBM rather than learning directly from field observations (Razavi et al., 2012). In this framework, the PBM acts as a data generator, producing synthetic input–output pairs (for example, climate, soil and management variables linked to SOC dynamics), which are then used to train the ML model (Fig. 6a). The

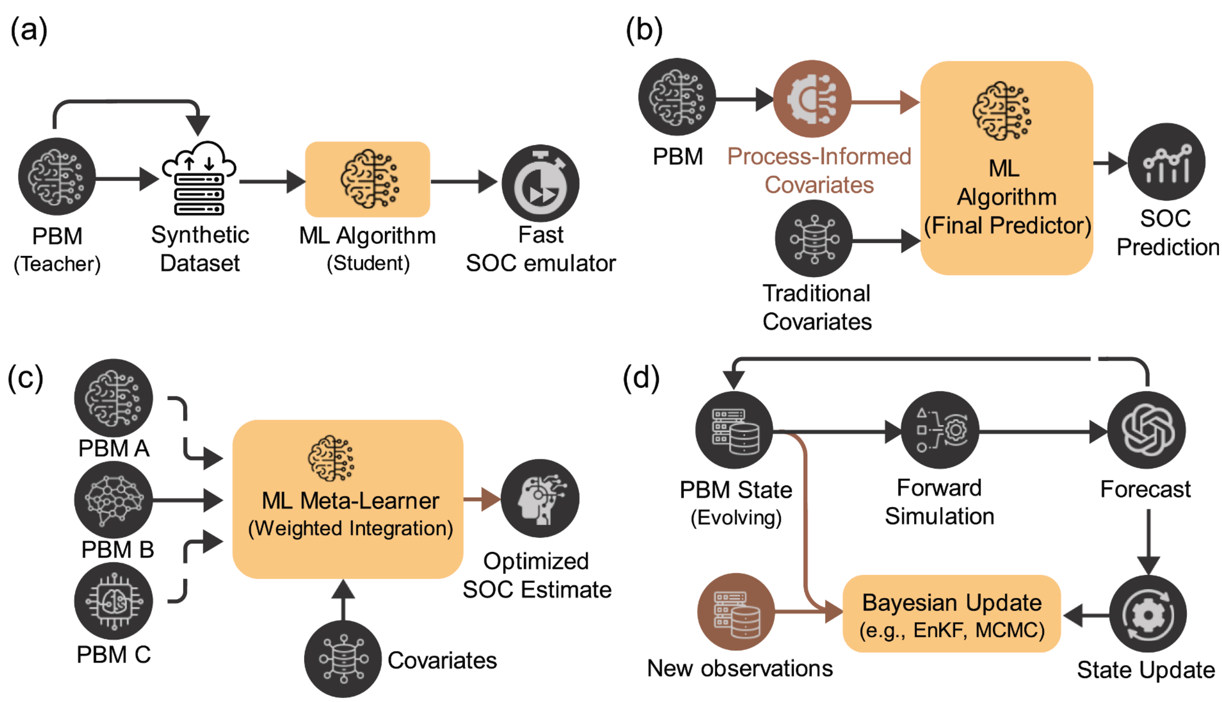

Fig. 6. Four strategies for hybridization, including meta model (a), sequential hybridization (b), ensemble model (c), and data assimilation (d). The abbreviations are defined below: PBM, process-based model; ML, machine learning; SOC, soil organic carbon; EnKF, ensemble Kalman filter; MCMC, Markov chain Monte Carlo. These are the four pathways developed in Section 4.

surrogate effectively approximates the mechanistic simulations at a fraction of the computational cost (Irrgang et al., 2021).

The primary advantage of this strategy is computational efficiency. Complex PBMs such as CENTURY or Earth System Models require solving large systems of differential equations and performing long spin-up runs, often spanning thousands of simulated years, to reach equilibrium. These demands limit high-resolution applications and large ensemble simulations for uncertainty analysis. ML surrogates can reproduce PBM outputs orders of magnitude faster, enabling spatial upscaling and rapid scenario exploration while retaining the structural logic embedded in the original model (Razavi et al., 2021). For example, Sun et al. (2023) demonstrated that ML surrogates can emulate the spin-up phase of land surface models while preserving mass balance, and Xiao et al. (2024) used an RF *meta*-model trained on outputs from a soil-crop model (APSIM; Table 2) to upscale predictions of SOC, crop yield and greenhouse gas emissions across the North China Plain.

Despite these advantages, the *meta*-model entails specific limitations. First, the predictive accuracy of the model is fundamentally constrained by the quality of the PBM-generated training data. If the initial PBM simulations are suboptimal or biased, these uncertainties are directly propagated to the ML. Furthermore, meta models often involve a large number of parameters and lack explicit physical constraints, posing challenges for parameter optimization and model tuning during the development process. As a result, *meta*-models primarily improve computational efficiency but do not fundamentally enhance process representation or error correction, thereby motivating alternative hybridization strategies that integrate physical constraints more directly.

4.2. Sequential hybridization

Sequential hybridization represents a coupling strategy where a PBM acts as a pre-processor for an ML algorithm (Fig. 6b). This approach uses the PBM to generate process-informed covariates to improve the predictive accuracy and physical plausibility. The workflow proceeds in two stages: (1) the PBM is run using standard climatic and soil inputs to generate a theoretical projection of SOC dynamics; (2) this simulated output of SOC, which captures the fundamental biogeochemical trends and constraints, is then fed as a covariate into the ML model, alongside traditional environmental covariates.

This method effectively creates a physics-guided feature space by utilizing PBM outputs that capture fundamental biogeochemical principles (Xie et al., 2022). By including the PBM output as a predictor, the ML model leverages established biogeochemical relationships (e.g., temperature sensitivity) encoded in the process model. This allows the ML to focus on resolving spatial heterogeneities using covariates. An example is Ma et al. (2019), who used particle-size distribution simulated by the State Space Soil Production and Assessment Model (SSSPAM) as an additional covariate in DSM. This approach

Table 2
Evaluation of hybrid models in SOC prediction.

Method class	Model / Method	Predictive accuracy gains over ML/ PBM	Independent validation	Uncertainty	Code accessibility	Source
Meta (surrogate) models	RF-APSIM	/	Held-out site	/	Yes	Xiao et al. (2024)
Sequential (physics-guided ML)	POML	RMSE decrease 19%	Held-out site, Held-out time	/	/	Zhang et al. (2024a)
Ensemble / Stacking	RothC-MIMICS-RF stack	RMSE decrease 9%	Held-out site	/	/	Zhang et al. (2023b)
Data assimilation (DA)	PRODA	RMSE decrease 44%	Held-out site	/	/	Tao and Luo (2022)

demonstrated how process-based landscape models can enrich empirical soil prediction. This is particularly valuable for filling temporal gaps, as the PBM provides a continuous theoretical baseline even for years where no field sampling occurred.

One implementation of this sequential strategy is the process-oriented machine learning (POML, Fig. 7, Table 2) framework proposed by Zhang et al. (2024a). In this approach, RothC model-simulated SOC stock data for unsampled years were used as additional training data for an ML model, while adjusted parameter weights W_p were used to control the contribution of PBM-generated SOC data during model training. The hybrid model improved temporal trend prediction by 19% compared with the standalone ML model.

While this approach addresses the black-box nature of ML, the integration of a PBM introduces its own set of complications. Because the ML model treats the PBM output as a covariate, any structural deficiencies or poor calibrations within the PBM directly introduce systematic biases that the ML may fail to correct. Furthermore, the high computational overhead and complex data requirements of sequential hybridization pose a significant challenge for high-frequency monitoring.

4.3. Ensemble model

An ensemble model (or model stacking) represents a hierarchical integration strategy. Unlike residual learning (which corrects errors), stacking treats the predictions of disparate models as new input features for a higher-level *meta*-learner. In this framework, multiple base models (e.g., mechanistic models with different structural assumptions, such as RothC or MIMICS) generate independent estimates of SOC (Fig. 6c). These estimates are then fed, alongside environmental covariates, into an ML model. The ML algorithm learns the optimal weighting scheme, effectively selecting the best-performing model for each specific environment.

This approach reduces structural uncertainty. No single model structure is perfect for all ecosystems, as SOC turnover processes and parameters differ considerably across different environments. By stacking diverse models, the *meta*-learner can leverage the unique strengths of each model. This results in a final prediction that is statistically more robust and accurate than any single contributing model could achieve alone (also known as super-learning). An application of this strategy is the work by Zhang et al. (2023b), who improved the spatiotemporal dynamics modelling of cropland SOC. They developed a hybrid ensemble by integrating RothC and MIMICS. RothC is designed to simulate SOC turnover in arable topsoil with an emphasis on the impacts of agricultural management, while MIMICS explicitly represents microbial-driven transformations and their response to temperature. Instead of selecting one model, they used the simulated SOC stocks from both models as dynamic covariates for an RF regressor. This RF-RothC-MIMICS (Table 2) stack outperformed both the individual PBM and the standalone ML model.

However, the ensemble model approach has several inherent limitations. First, variable importance analysis in stacked frameworks frequently identifies PBM-simulated SOC as the dominant predictor, which may marginalize the direct influence of underlying environmental covariates (Zhang et al., 2023b). Second, stacking does not rectify the fundamental sources of model error and can therefore lead to the cumulative propagation of hard-to-interpret uncertainties. Finally, incorporating a greater number of models does not guarantee a corresponding improvement in predictive accuracy; rather, it often increases the computational expense and structural complexity of spatiotemporal SOC modelling.

4.4. Data assimilation

Data assimilation (DA) transitions from traditional forward simulation to observation-constrained statistical inference (Fig. 6d). Fundamentally, DA operates within a Bayesian framework that sequentially updates the estimated state of a PBM by integrating new observational data, effectively steering the model trajectory towards reality (Luo et al., 2011).

Established approaches, such as the ensemble Kalman filter (EnKF; Gao et al., 2011) and Bayesian MCMC (Zhao et al., 2023), have been widely employed to assimilate observational data into PBMs. These methods are instrumental in constraining model parameters and reducing predictive uncertainty. More recently, the PRODA approach (Tao and Luo, 2022; Table 2) was proposed to integrate site-level MCMC assimilation results with DL techniques to optimise model parameters for global-scale SOC simulations (Tao et al., 2023). Notably, Tao et al. (2024) demonstrated that when constrained by identical high-quality observational datasets via DA, structurally distinct models exhibit remarkable convergence in their simulations of global SOC stocks and spatial distributions. This finding suggests that effective parameter constraints derived through DA can compensate for structural simplifications or divergences among models. DA serves as a statistical

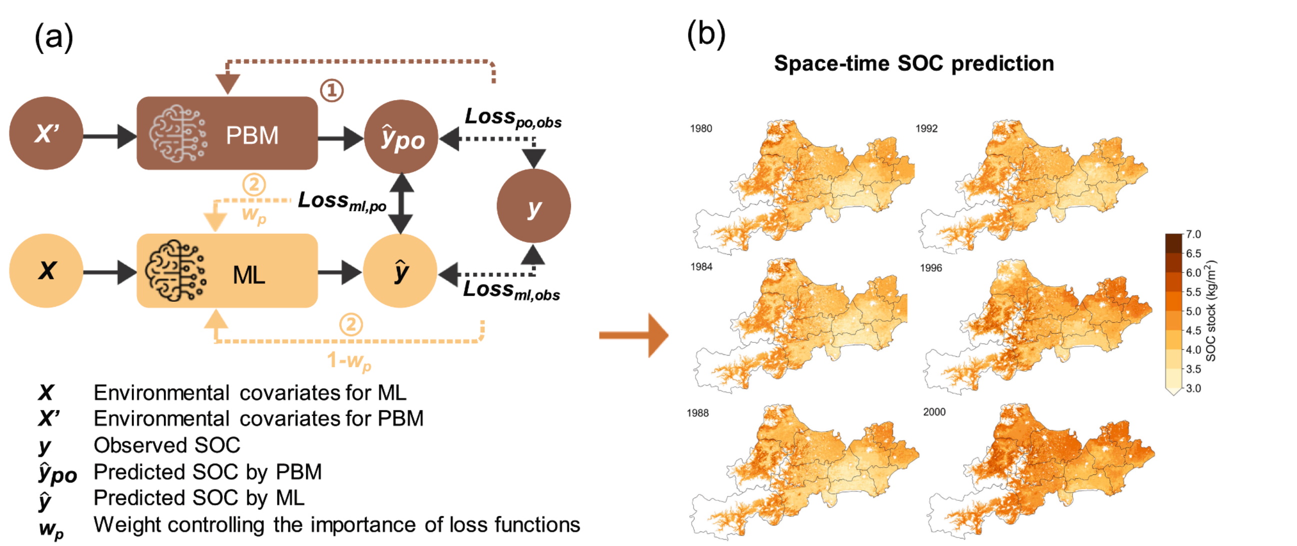

Fig. 7. Process-oriented machine learning (POML) framework that integrates knowledge-guided process-based model and data-driven machine learning model (a). The POML modelling method can produce SOC variations in space and time (b) (adapted from Zhang et al. 2024a). The environmental covariates for ML(x) and PBM (x′) partially overlap.

framework for fusing observations with mechanistic understanding as a way to reduce uncertainties in spatiotemporal SOC predictions.

Despite the potential of DA in reducing simulation uncertainties, it still has several limitations. The effectiveness of DA is strictly limited by the information content of the available observational data. For instance, incorporating C flux data constrains turnover rates but offers limited improvement for pool-size estimation (Xie et al., 2024). Computationally, the high cost of rigorous algorithms such as MCMC necessitates simplifying assumptions, such as steady-state equilibrium, which may compromise the representation of transient dynamics in disturbed agricultural soils (Raoult et al., 2025).

4.5. Validation, uncertainty and reproducibility

Despite the advancements of the hybrid modelling frameworks discussed in Sections 4.1–4.4, their widespread application demands attention to validation, uncertainty and reproducibility. Current evaluations of hybrid models rely on k-fold cross-validation or held-out site splitting within a single study area. While these methods assess local spatial interpolation accuracy, they fail to evaluate model transferability to new regions or temporal horizons. In SOC mapping, the quantification of prediction uncertainty has received little attention. As emphasized by Wadoux et al. (2018), quantifying prediction uncertainty is as critical as the prediction itself. Most hybrid models report only global accuracy metrics or parameter uncertainty, while the uncertainty of the mapping outputs remains unaddressed (Table 2). Furthermore, we call for the release of complete code to improve reproducibility across the scientific community.

5. Recent advancements in hybrid modelling

Building on earlier hybrid strategies that combine models at the workflow level (Section 4), recent advances increasingly integrate soil science knowledge directly into the learning process itself. These approaches move beyond coupling outputs or features and instead embed physical constraints, process structures, or global representations within machine-learning architectures. Here we discuss soil-science informed ML with variations including residual learning, parameter learning, physics-informed ML, and foundation models. These developments aim to improve the extrapolation capacity, interpretability, and robustness of SOC predictions, particularly under conditions of data scarcity and non-stationarity.

5.1. Soil-science informed machine learning

Soil-science Informed ML (SoilML) or knowledge-guided ML (KGML) or also known as theory-guided data science, represents a fundamental shift beyond simple model stacking (Minasny et al., 2024). Unlike *meta*-models, sequential hybridization, or ensemble approaches, SoilML integrates process knowledge directly into the machine-learning training process rather than coupling models externally (Willard et al., 2022). This knowledge guidance can occur through three pathways.

(1) Training initialization: using synthetic data generated by PBMs to pre-train the neural network, thereby transferring physical constraints before the model sees any real-world data.

(2) Architecture design: structuring the neural network layers to mirror the causal flow of the physical system (e.g., ensuring water flows from topsoil to subsoil, not vice versa).

(3) Loss function regularization: modifying the optimization goal to penalize physically impossible predictions.

An application of this approach is the KGML-ag-Carbon framework developed by Liu et al. (2024). They used a PBM (ecosys) to generate millions of synthetic data points representing diverse climate scenarios, and then used this synthetic data to pre-train a gated recurrent unit (GRU) neural network, enabling it to learn biogeochemical processes. When fine-tuned with sparse real-world observations, the KGML-ag-Carbon model outperformed both the original ecosys model (by reducing parameter uncertainty) and standard DL models (by adhering to mass balance). Furthermore, the high-resolution approach captured substantially greater spatial detail in SOC changes, which was reflected in an 86% lower NRMSE than the conventional coarse-resolution baseline. These results illustrate how embedding physical constraints during training can improve both predictive performance and generalization capacity, particularly when observational data are sparse.

5.2. Residual learning

Residual learning represents a pragmatic approach to addressing the imperfections of PBMs. Unlike SoilML, which modifies the internal training processes, residual modelling operates sequentially by using an ML model to predict the residuals of a PBM (Willard et al., 2020). In this framework, the ML model learns the errors in PBM predictions relative to actual observations.

The core concept is to learn the systematic biases of the physical model relative to empirical observations, and subsequently use this learned information to correct the PBM's baseline predictions. For SOC predictions, the final output is formulated as the sum of the mechanistic prediction and the ML-derived correction ϵ_{ML}:

$$SOC_{final}(x,t) = SOC_{PBM}(x,t) + \epsilon_{ML}(x,t) \tag{3}$$

The ML component is trained on the discrepancy between observed SOC and the PBM predictions, typically utilizing a diverse set of static and dynamic environmental covariates to capture localized spatial or temporal heterogeneity that the baseline model failed to resolve.

This approach is highly modular and computationally efficient, combining the long-term temporal extrapolation capabilities of PBMs with the non-linear spatial pattern recognition of ML algorithms. However, the residual modelling approach only learns which components are missing from the model and does not inherently incorporate any informed physical knowledge into the ML algorithm. Consequently, while residual learning can improve SOC prediction accuracy within the bounds of the training data, it remains susceptible to generating physically inconsistent trajectories when extrapolating to novel or extreme environmental conditions.

5.3. Parameter learning

Parameter learning uses ML to estimate uncertain parameters in PBMs, addressing the challenges posed by unobservable and underdetermined parameters. Within this framework, the ML component can be formulated as a mapping function:

$$p = f_{ML}(x; W) \tag{4}$$

where the function transforms environmental covariates x (e.g., soil texture, topography, and climate) into physical parameters p required by the process-based models (e.g. decomposition constants or the amount of inert organic matter in RothC). W denotes the internal learnable weights of the ML model.

An advancement is the differentiable parameter learning (dPL) framework proposed by Tsai et al. (2021), which employs deep learning to construct a global mapping that predicts spatially varying parameters directly from environmental covariates. This method overcomes the inherent inefficiencies and limitations of conventional site-by-site calibration, facilitating a more robust integration of deep learning with physics-based processes. This kind of parameter learning provides an effective mechanism for constraining the parameter space and mitigating the problem of equifinality. The efficacy of this approach is dependent on the availability of physical governing equations and relationships between the parameters and environmental data.

5.4. Physics-informed neural networks

Physics-informed neural networks (PINNs) are a specific subclass of KGML, in which governing equations are embedded explicitly within the learning objective. While standard neural networks are universal function approximators that fit curves to data points, PINNs are trained to solve partial differential equations (PDEs; Raissi et al., 2019). The core innovation lies in the loss function. In a conventional neural network, training minimizes a data-misfit loss,

$$\mathscr{L}_{\text{data}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{C}(x_i,t_i) - C^{\text{obs}}(x_i,t_i)\right\|^2, \tag{5}$$

where $\widehat{C}$ is the predicted SOC concentration and C^{obs} denotes observations. In a PINN, an additional physics-based constraint is imposed by penalizing violations of the governing PDE:

$$\mathscr{L}_{\text{PINN}} = \mathscr{L}_{\text{data}} + \lambda_{\text{phys}}\mathscr{L}_{\text{phys}}, \tag{6}$$

where λ_{phys} is a weighting parameter and

$$\mathscr{L}_{\text{phys}} = \frac{1}{M}\sum_{j=1}^{M}\left\|\mathscr{R}\left(\widehat{C}(x_j,t_j)\right)\right\|^2 \tag{7}$$

represents the residual of the governing SOC equation, evaluated at collocation points. λ_{phys} is usually determined by hyperparameter tuning or adaptive weighting algorithms.

SOC dynamics can be described using PDEs; for example, for a two-pool model:

$$\frac{\partial \mathbf{C}}{\partial t} = \mathbf{I} - \mathbf{K}(\tau)\,\mathbf{C} + \mathbf{T}(\tau)\,\mathbf{C} - \nabla(\mathbf{v}\,\mathbf{C}) + \nabla(\mathbf{D}\,\nabla\mathbf{C}) \tag{8}$$

In this formulation, SOC dynamics are governed by the balance between external inputs $\mathbf{I}$, first-order mineralisation losses $\mathbf{KC}$, internal transfers between fast (e.g. particulate) and slow (e.g. mineral-associated) pools represented by $\mathbf{TC}$, and spatial redistribution through advective transport $\nabla(\mathbf{vC})$ and diffusive mixing $\nabla(\mathbf{D}\nabla\mathbf{C})$. Environmental drivers τ(e.g., temperature, moisture, texture, management) regulate decomposition and transfer rates, while the velocity field $\mathbf{v}$ captures rapid lateral transport processes such as runoff-driven redistribution or erosion.

Additional terms could be added to the loss function, including initial and boundary conditions, and non-negativity C values. By minimizing both data misfit and physical inconsistency, PINNs learn SOC trajectories that satisfy observational constraints while remaining consistent with known biogeochemical laws. The network can infer the correct trajectory of C and decomposition rates from sparse measurements (Sargeant et al., 2024).

A prospective implementation of PINNs for SOC dynamics, as proposed by Minasny et al. (2024), employs a dual neural network architecture to enforce mass balance constraints. In this framework, the problem is split into two coupled learning tasks: (1) Parameter learning: The first neural network estimates the spatially varying kinetic parameters (e.g., decomposition rates **K**, transfer coefficients **T)** based on static environmental covariates (e.g., soil texture, topography), and spatial redistribution parameters from topography and rainfall. (2) State prediction: A second neural network predicts the spatiotemporal SOC stocks using parameters derived from the previous step and environmental controls. These networks can be trained simultaneously using a physics-informed loss function (Eq. (6). This architecture makes PINNs particularly attractive for modelling SOC dynamics in data-poor or physically constrained domains, while also motivating complementary approaches that rely on large-scale representation learning. Nevertheless, PINNs are most effective for one-dimensional systems with well-defined governing equations and may be less suitable for SOC processes that are strongly heterogeneous, threshold-driven, or poorly parameterised.

5.5. Foundation models

A parallel advancement in KGML is the emergence of foundation models (FMs) for the Earth system. In contrast to SoilML, which encodes explicit physical constraints, FMs rely on large-scale representation learning to capture generic spatiotemporal structure across the Earth system. Unlike traditional DL models that are trained for specific tasks, FMs are pre-trained on massive, multi-modal datasets, spanning satellite imagery, climate reanalysis, and soil grids, using self-supervised learning techniques (Barkov et al., 2026; Huang et al., 2026). Bodnar et al. (2025) demonstrated that such models can learn generalizable representations of Earth system dynamics without requiring explicit labels, instead learning directly from the inherent structure of the data by predicting missing pixels or future time steps.

Specifically, the FM workflow operates through three integrated pathways: (1) self-supervised pre-training: learning universal Earth system representations from massive unlabeled datasets (e.g., ERA5, Landsat) by predicting masked pixels or future states. (2) multimodal spatiotemporal encoding: converting heterogeneous data into a unified 3D latent space via Transformer architectures. (3) task-specific adaptation: transferring the pre-trained knowledge to specific SOC prediction tasks through few-shot fine-tuning.

By leveraging the generalized knowledge encoded during pre-training, FMs could achieve high predictive accuracy even in data-sparse regions. A recent application is the FoundationSoil framework developed by Song et al. (2025). They integrated a geospatial foundation model (Prithvi-EO-2.0) with environmental covariates resampled to a 30-m spatial resolution to predict the spatial variability of SOC (0–5 cm and 5–15 cm depth layers). Compared to a standard pixel-based baseline model, FoundationSoil demonstrated better performance, achieving a peak R^2 of 0.72 across the continental United States. Similarly, Zhang et al. (2025b) used a pre-trained global SOC foundation model (GSoilCPM) to enhance localized soil modelling. Their findings highlight that regions with sparse soil sampling or lower baseline accuracy derive the most significant benefits from the generalized knowledge transferred from global models.

While foundation models can improve spatial generalization, they do not inherently enforce physical consistency or mass balance unless combined with process-based or physics-informed constraints. As with other ML models, the reliability and accuracy of FMs depend on the quality of training data, which remains a bottleneck in any modelling approach. Consequently, foundation models can be regarded as representation engines whose outputs can benefit from further integration with process-based constraints or uncertainty-aware frameworks. It's worth noting that foundation models such as Prithvi-EO-2.0 and GSoilCPM are not yet independently benchmarked for SOC outside the studies that produced them.

6. Future directions and research gaps

Hybrid SOC modelling has advanced significantly, yet there are key conceptual, methodological, and data-related gaps. Addressing these gaps is essential for improving the robustness, generalizability, and policy relevance of spatiotemporal SOC assessments. Key priorities include next-generation data acquisition, process representation, model generalization and causality, remote sensing capabilities, and uncertainty quantification.

6.1. Novel approaches for soil data generation

As we discussed in previous sections, data scarcity is the most fundamental bottleneck, and all modelling approaches depend on improved data. The prohibitive cost of traditional wet chemistry continues to limit global SOC and its fractions (e.g., POC and MAOC) modelling, perpetuating data scarcity in critical regions (Tang et al., 2026a). To overcome this, future soil data generation should transition from static surveys to dynamic active learning (AL) frameworks. Unlike conventional sampling, AL algorithms leverage the spatial uncertainty maps generated by models (as discussed in the following section) to autonomously identify high-uncertainty zones. This guides sampling campaigns to locations that maximize information gain per unit effort, effectively allowing the model to request the data it needs most to improve its own performance. Concurrently, the integration of proximal soil sensing (e.g., handheld Vis-NIR and MIR spectrometers) allows for the high-throughput characterization of soil properties directly in the field (Peng et al., 2025; Dai et al., 2025; Tang et al., 2026b). When coupled with cloud-based spectral inference engines, these technologies enable the rapid expansion of global soil databases, thereby filling critical spatial gaps without the latency associated with laboratory analysis (Huang et al., 2026).

6.2. Improving representation of processes

A critical gap in current hybrid models is the oversimplification of the biological formation and transformation of the soil C continuum, which comprises numerous functionally distinct components. While physical protection of SOC (e.g., MAOM) is increasingly represented, the explicit role of the soil microbiome remains largely parameterized as a black box constant. A number of models do not explicitly account for how microbial community composition (e.g., fungal-to-bacterial ratio) or physiology (CUE) shifts under climate stress. This is a significant limitation, as soil microorganisms fundamentally govern organic matter cycling. However, microbial functional traits, which dictate the rates of these processes, can now be inferred from metagenomic sequencing and genomic data (Xue et al., 2025). Early proof-of-concept studies suggest that omics inputs may eventually contribute to hybrid models, although substantial methodological challenges remain. Specifically, scaling omics-to-flux mappings to the landscape scale is hindered by critical barriers, including DNA extraction biases, uncertainties in taxa-to-function mapping, and the inherent difficulty of sample-to-pixel scaling. Recent work by Collart et al. (2025), for example, demonstrated an initial framework in which convolutional neural networks trained on gene abundance data dynamically predicted decomposition rates for a process-based model, thereby directly linking soil biodiversity to SOC dynamics.

6.3. Towards generalizable, interpretable and causal hybrids

While DL models achieve high accuracy locally, they exhibit poor transferability, often failing when applied to regions with different soil parent materials or climates (covariate shift). Current models are often overfitted to data-rich regions (e.g., Europe and the USA), creating a trust gap for global application. Research must prioritize transfer learning and few-shot learning, emerging techniques that allow pre-trained global models to be fine-tuned on sparse local data (e.g., in under-represented areas) without starting from scratch (Huang et al., 2026; Immorlano et al., 2025).

However, generalizability alone is insufficient; models must also be trustworthy. Here, a critical distinction must be drawn between interpretability and causality. Zhang and Wadoux (2026) distinguish between two fundamental viewpoints for causal interpretation in the DSM framework: a successionist and a generative one. The successionist view interprets causal relationships as stable and repeatable regularities, treating reliable associations as causal factors. Post-hoc frameworks like Shapley Additive exPlanations (SHAP), exemplify this approach by diagnosing model behaviour through the statistical importance of covariates for prediction. As argued by Kakhani et al. (2025), these techniques reveal which variables drive predictions, allowing researchers to detect whether a model relies on spurious correlations rather than pedological drivers. While SHAP explains associations, it does not confirm mechanisms.

By contrast, causal inference requires an explicit causal model, causal sufficiency, and statistical faithfulness. The generative perspective is better aligned with these requirements, as it represents soil-forming factors as operating through mechanistic layers, including microbial processing and nutrient cycling. Advancing toward causal ML, such as structural causal models, is a key research priority. For example, Liu et al. (2025b) used structural equation modelling to identify relevant predictors, which were then fed into a ML model for temporal SOC forecasting. Unlike explainable AI, causal ML enables explicit testing of cause–and–effect relationships, allowing models to support actionable interventions (e.g., whether reducing tillage causes SOC to increase) rather than descriptive correlations. Finally, the integration of causal ML with process-based mechanisms points toward the future development of causal digital twins for soil. However, it's worth noting that landscape-scale causal SOC inference is at an early, methodological development stage.

6.4. The role of next-generation remote sensing

The performance of hybrid SOC models remains limited by the spectral resolution of widely used multispectral satellite sensors. Broadband sensors such as Sentinel-2 provide valuable spatial and temporal coverage, but they cannot reliably separate SOC signals from confounding effects of soil moisture, iron oxides, crop residues, and surface roughness. They also provide limited information on mineralogical controls on SOC stabilisation.

Spaceborne hyperspectral missions, including EnMAP, PRISMA, Copernicus CHIME, and NASA SBG, offer a major opportunity to overcome these constraints. These sensors can better resolve diagnostic features associated with clay mineralogy, organic matter, and crop residues. Because hyperspectral satellites measure continuous reflectance spectra that are more comparable to laboratory and field Vis–NIR instruments, they create new opportunities for proximal–remote sensing fusion. Integrating soil spectral libraries with hyperspectral imagery could allow hybrid models to move beyond total SOC estimation toward spatially explicit mapping of functional SOC pools, including POC and MAOC, at regional to global scales (Hong et al., 2025; Tang et al., 2026a; Zhao et al., 2025). However, improved spectral resolution alone will not resolve all modelling challenges. The temporal mismatch between frequently observed remote-sensing covariates and slowly changing SOC stocks remains a critical limitation.

6.5. Rigorous uncertainty quantification

Finally, as SOC maps are increasingly used for C credits and climate policy (MRV), point estimates of SOC stocks are no longer sufficient. Most hybrid SOC studies report point predictions with global metrics (e.g., overall RMSE or R^2), rather than the prediction uncertainty of the outputs. There is an urgent need to quantify spatial uncertainty. In recent years, the pedometrics community has made significant efforts to quantify and assess uncertainty in SOC modelling at various spatial scales (Szatmári et al., 2021; Szatmári et al., 2024; Wadoux and Heuvelink, 2023). For example, Rau et al. (2025) proposed using a Bayesian deep learning technique, Laplace approximations, to generate uncertainty maps that explicitly highlight areas where model predictions are unreliable. Moving forward, establishing rigorous uncertainty bounds will be essential to translate hybrid model outputs into reliable soil carbon management strategies (Huang et al., 2025). These research directions highlight that progress in hybrid SOC modelling will depend not on a single methodological breakthrough but on coordinated advances in new data acquisition methods, process understanding, learning approaches, and uncertainty quantification.

7. Conclusion

Accurate spatiotemporal quantification of SOC remains a central challenge for climate mitigation, carbon accounting, and sustainable land management. Current modelling approaches have complementary strengths and weaknesses: PBMs provide mechanistic consistency but are limited by parameter uncertainty, scaling constraints, and limited validation, whereas ML approaches efficiently capture spatial heterogeneity but often lack physical constraints, causal interpretability, and robustness under extrapolation.

This review contributes to the field in three key ways. First, it provides a structured synthesis of spatiotemporal SOC modelling approaches across mechanistic, geostatistical, and data-driven approaches, clarifying their different inferential foundations. Second, it organises hybrid strategies—ranging from *meta*-models and sequential coupling to data assimilation, soil-science-informed ML, residual learning, parameter learning, PINNs, and foundation models—and shows how integration can be implemented at the workflow, architectural, or loss-function level. Third, it identifies bottlenecks beyond modelling, including data irregularity, non-stationarity, spatial transport processes, and uncertainty propagation, and links these to emerging solutions in causal ML, uncertainty-aware modelling, hyperspectral sensing, and active learning.

The convergence of mechanistic and data-driven approaches offers a pathway toward models that are both physically constrained and computationally scalable. However, methodological innovation alone is insufficient. Progress will depend equally on improving data quality and coverage, strengthening representation of microbial and transport processes, developing interpretable and causal frameworks, and implementing rigorous spatial uncertainty quantification.

These advances shift SOC modelling from static mapping toward dynamic, uncertainty-aware monitoring systems. Such systems are essential for robust measurement, reporting, and verification, and for ensuring that SOC-based climate mitigation strategies are scientifically defensible, transparent, and scalable under global change.

CRediT authorship contribution statement

Yuanyuan Du: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation. **Budiman Minasny:** Writing – review & editing, Validation, Methodology, Conceptualization. **Wartini Ng:** Writing – review & editing, Validation. **Lei Zhang:** Writing – review & editing, Validation, Formal analysis. **Nicolas P.A. Saby:** Writing – review & editing, Validation, Formal analysis. **Yue Zhou:** Writing – review & editing, Validation, Formal analysis. **Bas van Wesemael:** Writing – review & editing, Validation. **Yuxin Ma:** Writing – review & editing, Validation. **Zheng Wang:** Writing – review & editing, Validation. **Zhongkui Luo:** Writing – review & editing, Validation. **Zhou Shi:** Writing – review & editing, Validation, Supervision, Funding acquisition. **Songchao Chen:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Songchao Chen reports financial support was provided by National Natural Science Foundation of China. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is funded by the National Natural Science Foundation of China (U24A20575, 42201054), the Innovation and Development Special Fund for Hangzhou West Science and Technology Innovation Corridor, and the ZJU-USYD Ignition Grants. Gemini was employed to refine language. BM acknowledges the support of the Soil Science Challenge Grants Program, High-resolution evaluation and monitoring of soil change for Australia.

Data availability

No data was used for the research described in the article.

References

Abramoff, R., Xu, X., Hartman, M., O'Brien, S., Feng, W., Davidson, E., Finzi, A., Moorhead, D., Schimel, J., Torn, M., Mayes, M.A., 2018. The Millennial model: in search of measurable pools and transformations for modeling soil carbon in the new century. Biogeochemistry 137, 51–71.

Abramoff, R.Z., Guenet, B., Zhang, H., Georgiou, K., Xu, X., Viscarra Rossel, R.A., Yuan, W., Ciais, P., 2022. Improved global-scale predictions of soil carbon stocks with Millennial Version 2. Soil Biol. Biochem. 164, 108466.

Ahrens, B., Braakhekke, M.C., Guggenberger, G., Schrumpf, M., Reichstein, M., 2015. Contribution of sorption, DOC transport and microbial interactions to the 14C age of a soil organic carbon profile: Insights from a calibrated process model. Soil Biol. Biochem. 88, 390–402.

Amelung, W., Bossio, D., de Vries, W., Kögel-Knabner, I., Lehmann, J., Amundson, R., Bol, R., Collins, C., Lal, R., Leifeld, J., Minasny, B., 2020. Towards a global-scale soil climate mitigation strategy. Nat. Commun. 11, 5427.

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B.M., Hong, S. Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J., Mendonca-Santos, M. L., Minasny, B., Montanarella, L., Odeh, I.O.A., Sanchez, P.A., Thompson, J.A., Zhang, G.-L., 2014. GlobalSoilMap: Toward a fine-resolution global grid of soil properties. In: Sparks, D.L. (Ed.), Advances in Agronomy. Academic Press, pp. 93–134.

Barkov, V., Schmidinger, J., Gebbers, R., Atzmueller, M., 2026. Modern Neural Networks for Small Tabular Datasets: the New Default for Field-Scale Digital Soil Mapping? Eur. J. Soil Sci. 77 (2), e70299.

Batjes, N.H., Calisto, L., de Sousa, L.M., 2024. Providing quality-assessed and standardised soil data to support global mapping and modelling (WoSIS snapshot 2023). Earth Syst. Sci. Data 16, 4735–4765.

Bayram, F., Ahmed, B.S., Kassler, A., 2022. From concept drift to model degradation: an overview on performance-aware drift detectors. Knowl.-Based Syst. 245, 108632.

Bodnar, C., Bruinsma, W.P., Lucic, A., Stanley, M., Allen, A., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J.A., Dong, H., Gupta, J.K., 2025. A foundation model for the Earth system. Nature 1–8.

Chabrillat, S., Ben-Dor, E., Cierniewski, J., Gomez, C., Schmid, T., van Wesemael, B., 2019. Imaging spectroscopy for soil mapping and monitoring. Surv. Geophys. 40, 361–399.

Chen, M., Qian, Z., Boers, N., Jakeman, A.J., Kettner, A.J., Brandt, M., Kwan, M.P., Batty, M., Li, W., Zhu, R., Luo, W., 2023. Iterative integration of deep learning in hybrid Earth surface system modelling. Nat. Rev. Earth Environ. 4, 568–581.

Chen, Q., Wang, Y., Zhu, X., 2024a. Soil organic carbon estimation using remote sensing data-driven machine learning. PeerJ 12, e17836.

Chen, S., Arrouays, D., Leatitia Mulder, V., Poggio, L., Minasny, B., Roudier, P., Libohova, Z., Lagacherie, P., Shi, Z., Hannam, J., Meersmans, J., Richer-de-Forges, A. C., Walter, C., 2022. Digital mapping of GlobalSoilMap soil properties at a broad scale: a review. Geoderma 409, 115567.

Chen, S., Chen, Z., Zhang, X., Luo, Z., Schillaci, C., Arrouays, D., Richer-de-Forges, A.C., Shi, Z., 2024b. European topsoil bulk density and organic carbon stock database (0-20 cm) using machine-learning-based pedotransfer functions. Earth Syst. Sci. Data 16, 2367–2383.

Chen, Z., Chen, L., Lu, R., Lou, Z., Zhou, F., Jin, Y., Xue, J., Guo, H., Wang, Z., Wang, Y., Liu, F., 2025. A national soil organic carbon density dataset (2010–2024) in China. Sci. Data 12, 1480.

Coleman, K., Jenkinson, D.S., 1996. RothC-26.3 - a Model for the turnover of carbon in soil. In: Powlson, D.S., Smith, P., Smith, J.U. (Eds.), Evaluation of Soil Organic Matter Models. Springer, Berlin Heidelberg, pp. 237–246.

Collart, P., Gall, J., Schnepf, A., Sousa Rocha, A. V., Herold, M., Buckeridge, K., Pagel, H., 2025. Hybrid Soil Microbiome Modeling - Combining process-based models with machine learning to predict microbial dynamics and organic matter turnover in soil systems. EGU General Assembly 2025, Vienna, Austria, EGU25-10523.

Dai, L., Wang, Z., Zhuo, Z., Ma, Y., Shi, Z., Chen, S., 2025. Prediction of soil organic carbon fractions in tropical cropland using a regional visible and near-infrared spectral library and machine learning. Soil till. Res. 245, 106297.

Demattê, J.A.M., Fongaro, C.T., Rizzo, R., Safanelli, J.L., 2018. Geospatial Soil Sensing System (GEOS3): a powerful data mining procedure to retrieve soil spectral reflectance from satellite images. Remote Sens. Environ. 212, 161–175.

Ding, Z., Liu, K., Grunwald, S., Smith, P., Ciais, P., Wang, B., Wadoux, A.M.J.C., Ferreira, C., Karunaratne, S., Shurpali, N., Yin, X., Roberts, D., Madgett, O., Duncan, S., Zhou, M., Liu, Z., Harrison, M.T., 2025. Advancing Soil Organic Carbon Prediction: a Comprehensive Review of Technologies, AI, Process-based and Hybrid Modelling Approaches. Adv. Sci. 12, e04152.

Dungait, J.A.J., Hopkins, D.W., Gregory, A.S., Whitmore, A.P., 2012. Soil organic matter turnover is governed by accessibility not recalcitrance. Glob. Chang. Biol. 18, 1781–1796.

Dvorakova, K., Heiden, U., Pepers, K., Staats, G., van Os, G., van Wesemael, B., 2023. Improving soil organic carbon predictions from a Sentinel–2 soil composite by assessing surface conditions and uncertainties. Geoderma 429, 116128.

Evangelista, S.J., Field, D.J., McBratney, A.B., Minasny, B., Ng, W., Padarian, J., Dobarco, M.R., Wadoux, A.M.C., 2023. A proposal for the assessment of soil security: soil functions, soil services and threats to soil. Soil Secur. 10, 100086.
Evans, M.E., Adler, P.B., Angert, A.L., Dey, S.M., Girardin, M.P., Heilman, K.A., Klesse, S., Perret, D.L., Sax, D.F., Sheth, S.N., Stemkovski, M., 2025. Reconsidering space-for-time substitution in climate change ecology. Nat. Clim. Chang. 15, 809–812.
Fan, D., Zhao, T., Jiang, X., García-García, A., Schmidt, T., Samaniego, L., Attinger, S., Wu, H., Jiang, Y., Shi, J., Fan, L., Tang, B.-H., Wagner, W., Dorigo, W., Gruber, A., Mattia, F., Balenzano, A., Brocca, L., Jagdhuber, T., Wigneron, J.-P., Montzka, C., Peng, J., 2025. A Sentinel-1 SAR-based global 1-km resolution soil moisture data product: Algorithm and preliminary assessment. Remote Sens. Environ. 318, 114579.
Fernandez-Ugalde, O., Scarpa, S., Orgiazzi, A., Panagos, P., Van Liedekerke, M., Marechal, A., Jones, A., 2022. LUCAS 2018 Soil Module. Publications Office of the European Union, Luxembourg. ISBN.
Friedlingstein, P., O'Sullivan, M., Jones, M.W., Andrew, R.M., Hauck, J., Landschützer, P., Zeng, J., et al., 2024. Global Carbon Budget 2024. Earth Syst. Sci. Data 17, 965–1039.
Gao, C., Wang, H., Weng, E., Lakshmivarahan, S., Zhang, Y., Luo, Y., 2011. Assimilation of multiple data sets with the ensemble Kalman filter to improve forecasts of forest carbon dynamics. Ecol. Appl. 21, 1461–1473.
García-Palacios, P., Crowther, T.W., Dacal, M., Hartley, I.P., Reinsch, S., Rinnan, R., Rousk, J., van den Hoogen, J., Ye, J.-S., Bradford, M.A., 2021. Evidence for large microbial-mediated losses of soil carbon under anthropogenic warming. Nat. Rev. Earth Environ. 2, 507–517.
Garsia, A., Moinet, A., Vazquez, C., Creamer, R.E., Moinet, G.Y., 2023. The challenge of selecting an appropriate soil organic carbon simulation model: a comprehensive global review and validation assessment. Glob. Chang. Biol. 29 (20), 5760–5774.
Hendriks, C.M.J., Stoorvogel, J.J., Álvarez-Martínez, J.M., Claessens, L., Pérez-Silos, I., Barquín, J., 2021. Introducing a mechanistic model in digital soil mapping to predict soil organic matter stocks in the Cantabrian region (Spain). Eur. J. Soil Sci. 72, 704–719.
Heuvelink, G.B.M., Angelini, M.E., Poggio, L., Bai, Z., Batjes, N.H., van den Bosch, R., Bossio, D., Estella, S., Lehmann, J., Olmedo, G.F., Sanderman, J., 2021. Machine learning in space and time for modelling soil organic carbon change. Eur. J. Soil Sci. 72, 1607–1623.
Heuvelink, G.B.M., Webster, R., 2022. Spatial statistics and soil mapping: a blossoming partnership under pressure. Spat. Stat. 50, 100639.
Hobley, E.U., Murphy, B., Simmons, A., 2018. Comment on "Soil organic stocks are systematically overestimated by misuse of the parameters bulk density and rock fragment content" by Poeplau et al. (2017). SOIL 4, 169–171.
Hong, Y., Chen, Y., Chen, S., Wang, Y., Hu, W., Ye, S., Song, X., Liu, F., Zhao, Y., Demattê, J.A., Shi, Z., 2025. Bridging the gap between laboratory VNIR-SWIR spectra and Landsat-8 bare soil composite image for soil organic carbon prediction. Remote Sens. Environ. 328, 114874.
Hu, B., Zhou, Q., He, C., Duan, L., Li, W., Zhang, G., Ji, W., Peng, J., Xie, H., 2021. Spatial variability and potential controls of soil organic matter in the Eastern Dongting Lake Plain in southern China. J. Soil. Sediment. 21, 2791–2804.
Huang, X., Ibrahim, M.M., Luo, Y., Jiang, L., Chen, J., Hou, E., 2024. Land Use Change Alters Soil Organic Carbon: Constrained Global patterns and Predictors. Earth's Future 12, e2023EF004386.
Huang, Y.C., Padarian, J., Minasny, B., McBratney, A.B., 2025. Using Monte Carlo conformal prediction to evaluate the uncertainty of deep-learning soil spectral models. Soil 11 (2), 553–563.
Huang, Y.C., Padarian, J., Ng, W., Minasny, B., McBratney, A.B., 2026. Zero-shot inference with Tabular Prior-data Fitted Network (TabPFN) for soil MIR spectral analysis. Geoderma 471, 117880.
Immorlano, F., Eyring, V., le Monnier de Gouville, T., Accarino, G., Elia, D., Mandt, S., Aloisio, G., Gentine, P., 2025. Transferring climate change physical knowledge. Proc. Natl. Acad. Sci. 122 (15), e2413503122.
Irrgang, C., Boers, N., Sonnewald, M., Barnes, E.A., Kadow, C., Staneva, J., Saynisch-Wagner, J., 2021. Towards neural Earth system modelling by integrating artificial intelligence in Earth system science. Nat. Mach. Intell. 3, 667–674.
Jenkinson, D.S., Coleman, K., 2008. The turnover of organic carbon in subsoils. Part 2. Modelling carbon turnover. Eur. J. Soil Sci. 59 (2), 400–413.
Jiao, Y., Xu, Z., Zhao, J., Yang, W., 2012. Changes in soil carbon stocks and related soil properties along a 50-year grassland-to-cropland conversion chronosequence in an agro-pastoral ecotone of Inner Mongolia. China. J. Arid Land 4, 420–430.
Kakhani, N., Taghizadeh-Mehrjardi, R., Omarzadeh, D., Ryo, M., Heiden, U., Scholten, T., 2025. Towards explainable AI: interpreting soil organic carbon prediction models using a learning-based explanation method. Eur. J. Soil Sci. 76, e70071.
Keyvanshokouhi, S., Cornu, S., Lafolie, F., Balesdent, J., Guenet, B., Moitrier, N., Nougier, C., Finke, P., 2019. Effects of soil process formalisms and forcing factors on simulated organic carbon depth-distributions in soils. Sci. Total Environ. 652, 523–537.
Kc, K., Zhao, K., Romanko, M., Khanal, S., 2021. Assessment of the spatial and temporal patterns of cover crops using remote sensing. Remote Sens. 13, 2689.
Kipf, T.N., Welling, M., 2017. Semi-supervised classification with graph convolutional networks. In: Int. Conf. Learn. Represent. (ICLR).
Kreyling, J., 2025. Space-for-time substitution misleads projections of plant community and stand-structure development after disturbance in a slow-growing environment. J. Ecol. 113, 68–80.
Kyker-Snowman, E., Wieder, W.R., Frey, S.D., Grandy, A.S., 2020. Stoichiometrically coupled carbon and nitrogen cycling in the Microbial-Mineral Carbon Stabilization model version 1.0 (MIMICS-CN v1.0). Geosci. Model Dev. 13, 4413–4434.
Lal, R., Monger, C., Nave, L., Smith, P., 2021. The role of soil in regulation of climate. Philos. Trans. R. Soc. B 376, 20210084.
Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R., 2019. Unmasking Clever Hans predictors and assessing what machines really learn. Nat. Commun. 10, 1096.
Lavallee, J.M., Soong, J.L., Cotrufo, M.F., 2020. Conceptualizing soil organic matter into particulate and mineral-associated forms to address global change in the 21st century. Glob. Chang. Biol. 26, 261–273.
Le Noë, J., Manzoni, S., Abramoff, R., Bölscher, T., Bruni, E., Cardinael, R., Ciais, P., Chenu, C., Clivot, H., Derrien, D., Ferchaud, F., Garnier, P., Goll, D., Lashermes, G., Martin, M., Rasse, D., Rees, F., Sainte-Marie, J., Salmon, E., Schiedung, M., Schimel, J., Wieder, W., Abiven, S., Barré, P., Cécillon, L., Guenet, B., 2023. Soil organic carbon models need independent time-series validation for reliable prediction. Commun. Earth Environ. 4, 158.
Lehmann, J., Bossio, D.A., Kögel-Knabner, I., Rillig, M.C., 2020. The concept and future prospects of soil health. Nat. Rev. Earth Environ. 1, 544–553.
C. Li The DNDC model Evaluation of Soil Organic Matter Models: Using Existing Long-Term Datasets 1996 Springer Berlin Heidelberg 263 267.
Lindgren, F., Rue, H., Lindström, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J. r. Stat. Soc. B 73, 423–498.
Liski, J., Palosuo, T., Peltoniemi, M., Sievänen, R., 2005. Carbon and decomposition model Yasso for forest soils. Ecol. Model. 189, 168–182.
Liu, F., Wu, H., Zhao, Y., Li, D., Yang, J.L., Song, X., Shi, Z., Zhu, A.X., Zhang, G.L., 2021. Mapping high resolution national soil information grids of China. Sci. Bull. 67, 328–340.
Liu, L., Zhou, W., Guan, K., Peng, B., Xu, S., Tang, J., Zhu, Q., Till, J., Jia, X., Jiang, C., Wang, S., Qin, Z., Kong, H., Grant, R., Mezbahuddin, S., Kumar, V., Jin, Z., 2024. Knowledge-guided machine learning can improve carbon cycle quantification in agroecosystems. Nat. Commun. 15, 357.
Liu, M., Zheng, S., Pendall, E., Smith, P., Liu, J., Li, J., Fang, C., Li, B., Nie, M., 2025a. Unprotected carbon dominates decadal soil carbon increase. Nat. Commun. 16, 2008.
Liu, Y., Jiang, C., Feng, A., Xu, H., Wang, Y., Yin, Y., Wang, C., Xie, D., Gao, B., 2025b. A causal prediction method for soil organic carbon storage change estimation, with Shaanxi Province as a case study. Comput. Electron. Agric. 234, 110271.
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. Ecol. Appl. 21, 1429–1442.
Luo, Y., Weng, E., 2011. Dynamic disequilibrium of the terrestrial carbon cycle under global change. Trends Ecol. Evol. 26, 96–104.
Luo, Y., Weng, E., Wu, X., Gao, C., Zhou, X., Zhang, L., 2009. Parameter identifiability, constraint, and equifinality in data assimilation with ecosystem models. Ecol. Appl. 19, 571–574.
Luo, Y., Huang, Y., Sierra, C.A., Xia, J., Ahlström, A., Chen, Y., et al., 2022. Matrix approach to land carbon cycle modeling. J. Adv. Model. Earth Syst. 14, e2022MS003008.
Luo, Z., Feng, W., Luo, Y., Baldock, J., Wang, E., 2017. Soil organic carbon dynamics jointly controlled by climate, carbon inputs, soil properties and soil carbon fractions. Glob. Chang. Biol. 23 (10), 4430–4439.
Ma, Y., Minasny, B., Viaud, V., Walter, C., Malone, B., McBratney, A.B., 2023. Modelling the whole profile soil organic carbon dynamics considering soil redistribution under future climate change and landscape projections over the lower Hunter Valley. Australia. Land 12, 255.
Ma, Y., Minasny, B., Welivitiya, W.D.D.P., Malone, B., Willgoose, G.R., McBratney, A.B., 2019. The feasibility of predicting the spatial pattern of soil particle-size distribution using a pedogenesis model. Geoderma 341, 195–205.
G. Matheron Krigeage d'un panneau rectangulaire par sa périphérie. note géostatistique no 28 Centre De Géostatistique 1960 Fontainebleau, France.
McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.
Meyer, H., Pebesma, E., 2022. Machine learning-based global maps of ecological variables and the challenge of assessing them. Nat. Commun. 13, 2208.
Minasny, B., Bandai, T., Ghezzehei, T.A., Huang, Y.C., Ma, Y., McBratney, A.B., Ng, W., Norouzi, S., Padarian, J., Sharififar, A., Styc, Q., 2024. Soil science-informed machine learning. Geoderma 452, 117094.
Minasny, B., Malone, B.P., McBratney, A.B., Angers, D.A., Arrouays, D., Chambers, A., Chaplot, V., Chen, Z.-S., Cheng, K., Das, B.S., Field, D.J., Gimona, A., Hedley, C.B., Hong, S.Y., Mandal, B., Marchant, B.P., Martin, M., McConkey, B.G., Mulder, V.L., O'Rourke, S., Richer-de-Forges, A.C., Odeh, I., Padarian, J., Paustian, K., Pan, G., Poggio, L., Savin, I., Stolbovoy, V., Stockmann, U., Sulaeman, Y., Tsui, C.-C., Vågen, T.-G., van Wesemael, B., Winowiecki, L., 2017. Soil carbon 4 per mille. Geoderma 292, 59–86.
Nenkam, A.M., Wadoux, A.M.C., Minasny, B., Silatsa, F.B., Yemefack, M., Ugbaje, S.U., Akpa, S., Van Zijl, G., Bouasria, A., Bouslihim, Y., Chabala, L.M., 2024. Applications and challenges of digital soil mapping in Africa. Geoderma 449, 117007.
Ng, W., Minasny, B., Mendes, W.D.S., Demattê, J.A.M., 2020. The influence of training sample size on the accuracy of deep learning models for the prediction of soil properties with near-infrared spectroscopy data. Soil 6 (2), 565–578.
Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., Fernández-Ugalde, O., 2018. LUCAS Soil, the largest expandable soil dataset for Europe: a review. Eur. J. Soil Sci. 69 (1), 140–153.
Padarian, J., Minasny, B., McBratney, A.B., 2019. Using deep learning to predict soil properties from regional spectral data. Geoderma Reg. 16, e00198.
Padarian, J., Stockmann, U., Minasny, B., McBratney, A.B., 2022. Monitoring changes in global soil organic carbon stocks from space. Remote Sens. Environ. 281, 113260.

W.J. Parton The CENTURY model D.S. Powlson P. Smith J.U. Smith Evaluation of Soil Organic Matter Models 1996 Springer Berlin Heidelberg 283 291.

Peng, Y., Ben-Dor, E., Biswas, A., Chabrillat, S., Demattê, J.A., Ge, Y., Gholizadeh, A., Gomez, C., Guerrero, C., Herrick, J., Maynard, J.J., 2025. Spectroscopic solutions for generating new global soil information. The Innovation 6 (5), 100839.

Pickett, S.T.A., 1989. Space-for-Time Substitution as an Alternative to Long-Term Studies. In: Likens, G.E. (Ed.), Long-Term Studies in Ecology: Approaches and Alternatives. Springer, New York, pp. 110–135.

Raissi, M., Perdikaris, P., Karniadakis, G.E., 2019. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J. Comput. Phys. 378, 686–707.

Raoult, N., Douglas, N., MacBean, N., Kolassa, J., Quaife, T., Roberts, A.G., Fisher, R., Fer, I., Bacour, C., Dagon, K., Hawkins, L., 2025. Parameter estimation in land surface models: challenges and opportunities with data assimilation and machine learning. J. Adv. Model. Earth Syst. 17, e2024MS004733.

Rau, K., Eggensperger, K., Schneider, F., Blaschek, M., Hennig, P., Scholten, T., 2025. Quantifying spatial uncertainty to improve soil predictions in data-sparse regions. Soil 11, 833–847.

Razavi, S., Jakeman, A., Saltelli, A., Prieur, C., Iooss, B., Borgonovo, E., Plischke, E., Piano, S.L., Iwanaga, T., Becker, W., Tarantola, S., 2021. The future of sensitivity analysis: an essential discipline for systems modeling and policy support. Environ Model Softw. 137, 104954.

Razavi, S., Tolson, B.A., Burn, D.H., 2012. Review of surrogate modeling in water resources. Water Resour. Res. 48.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat,, 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204.

Rentschler, T., Scholten, T., 2025. A note on spurious correlations and explainable machine learning in digital soil mapping. Eur. J. Soil Sci. 76, e70172.

Richer-de-Forges, A.C., Chen, S., Heuvelink, G.B., van der Westhuizen, S., Orton, T.G., Bourennane, H., Arrouays, D., 2025. Does digital soil mapping prediction performance of soil texture improve when adding uncertain field texture estimates? a study based on clay content. Geoderma 456, 117277.

Robertson, A.D., Paustian, K., Ogle, S., Wallenstein, M.D., Lugato, E., Cotrufo, M.F., 2019. Unifying soil organic matter formation and persistence frameworks: the MEMS model. Biogeosciences 16, 1225–1248.

Rogge, D., Bauer, A., Zeidler, J., Mueller, A., Esch, T., Heiden, U., 2018. Building an exposed soil composite processor (SCMaP) for mapping spatial and temporal characteristics of soils with Landsat imagery (1984–2014). Remote Sens. Environ. 205, 1–17.

Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. r. Stat. Soc. B 71, 319–392.

Rue, H., Riebler, A., Sørbye, S.H., Illian, J.B., Simpson, D.P., Lindgren, F.K., 2017. Bayesian computing with INLA: a review. Annu. Rev. Stat. Appl. 4, 395–421.

Ryzhova, I.M., Romanenkov, V.A., Stepanenko, V.M., 2024. Modern development of soil organic matter dynamics models (review). Mosc. Univ. Soil Sci. Bull. 79, 493–499.

Saby, N.P., Opitz, T., Hu, B., Lemercier, B., Bourennane, H., 2020. Bayesian uncertainty quantification of spatio-temporal trends in soil organic carbon using INLA and SPDE. EGU General Assembly 2020, EGU2020-9154.

Sáez-Sandino, T., Maestre, F.T., Berdugo, M., Gallardo, A., Plaza, C., García-Palacios, P., Guirado, E., Zhou, G., Mueller, C.W., Tedersoo, L., Crowther, T.W., Delgado-Baquerizo, M., 2024. Increasing numbers of global change stressors reduce soil carbon worldwide. Nat. Clim. Chang. 14, 740–745.

Safanelli, J.L., Chabrillat, S., Ben-Dor, E., Demattê, J.A., 2020. Multispectral models from bare soil composites for mapping topsoil properties over Europe. Remote Sens. 12, 1369.

Sanderman, J., Hengl, T., Fiske, G.J., 2017. Soil carbon debt of 12,000 years of human land use. PNAS 114, 9575–9580.

Sargeant, J., Teng, S.W., Murshed, M., Paul, M., Brennan, D., 2024. Estimating Soil Organic Carbon from Multispectral Images using Physics-Informed Neural Networks. In: Proc. Asian Conf. Comput. Vis, 2632–2649.

Schimel, J.P., Weintraub, M.N., 2003. The implications of exoenzyme activity on microbial carbon and nitrogen limitation in soil: a theoretical model. Soil Biol. Biochem. 35, 549–563.

Shen, C., Appling, A.P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., Baity-Jesi, M., Fenicia, F., Kifer, D., Li, L., Liu, X., 2023. Differentiable modelling to unify machine learning and physical models for geosciences. Nat. Rev. Earth Environ. 4, 552–567.

Sierra, C.A., Müller, M., Metzler, H., Manzoni, S., Trumbore, S.E., 2017. The muddle of ages, turnover, transit, and residence times in the carbon cycle. Glob. Chang. Biol. 23, 1763–1773.

Smith, P., 2008. Land use change and soil organic carbon dynamics. Nutr. Cycl. Agroecosyst. 81, 169–178.

Song, J., Zhao, C., Liao, H.-Y., Shao, W., 2025. FoundationSoil: Enhancing soil organic carbon mapping using a multi-temporal geospatial foundation model. In: Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems. Association for Computing Machinery, 1067–1070.

Sulman, B.N., Phillips, R.P., Oishi, A.C., Shevliakova, E., Pacala, S.W., 2014. Microbe-driven turnover offsets mineral-mediated storage of soil carbon under elevated CO2. Nat. Clim. Chang. 4, 1099–1102.

Sun, X.L., Minasny, B., Wang, H.L., Zhao, Y.G., Zhang, G.L., Wu, Y.J., 2021. Spatiotemporal modelling of soil organic matter changes in Jiangsu, China between 1980 and 2006 using INLA-SPDE. Geoderma 384, 114808.

Y. Sun B. Li M.G. Genton Geostatistics for large datasets E. Porcu J.M. Montero M. Schlather Advances and Challenges in Space-Time Modelling of Natural Events 2012 Springer Berlin Heidelberg, Berlin, Heidelberg 55 77.

Sun, Y., Goll, D.S., Huang, Y., Ciais, P., Wang, Y.P., Bastrikov, V., Wang, Y., 2023. Machine learning for accelerating process-based computation of land biogeochemical cycles. Glob. Chang. Biol. 29 (11), 3221–3234.

Syu, C.H., Yen, C.C., Maisyarah, S., Yang, B.J., Tzou, Y.M., Jien, S.H., 2025. Soil organic carbon projections and climate adaptation strategies across Pacific Rim agro-ecosystems. Soil 11, 1109–1130.

Szatmári, G., Pásztor, L., Heuvelink, G.B.M., 2021. Estimating soil organic carbon stock change at multiple scales using machine learning and multivariate geostatistics. Geoderma 403, 115356.

Szatmári, G., Pásztor, L., Takács, K., Mészáros, J., Benő, A., Laborczi, A., 2024. Space-time modelling of soil organic carbon stock change at multiple scales: Case study from Hungary. Geoderma 451, 117067.

Tang, S., Guo, J., Li, S., Li, J., Xie, S., Zhai, X., Wang, C., Zhang, Y., Wang, K., 2019. Synthesis of soil carbon losses in response to conversion of grassland to agriculture land. Soil till. Res. 185, 29–35.

Tang, Y., Tran, T.K.A., Minasny, B., Bakhshandeh, S., Du, M., Francos, N., Huang, Y.C., Jang, H.J., Ng, W., Xue, P., McBratney, A., 2026a. SOC stabilisation shifts from carbon accumulation in temperate soils to mineral association in subtropical soils. Soil Biol. Biochem. 215, 110101.

Tang, Y., Tran, T.K.A., Minasny, B., Ng, W., Francos, N., Huang, Y.C., Jang, H.J., Xue, P., Du, M., Bakhshandeh, S., McBratney, A., 2026b. Machine learning prediction of soil carbon fractions using bulk-soil and fraction-specific MIR spectra: is physical fractionation necessary? Soil till. Res. 264, 107307.

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. Glob. Chang. Biol. 30, e17297.

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., Viatkin, K., Vargas, R., Yigini, Y., Omuto, C., Malik, A.A., Peralta, G., Cuevas-Corona, R., Di Paolo, L.E., Luotto, I., Liao, C., Liang, Y.-S., Saynes, V.S., Huang, X., Luo, Y., 2023. Microbial carbon use efficiency promotes global soil carbon storage. Nature 618, 981–985.

Tao, F., Luo, Y., 2022. PROcess-Guided Deep Learning and DAta-Driven Modelling (PRODA). In: Machine Learning, Optimization, and Data Science. LOD 2021. Lect. Notes Comput. Sci. 13164, 319-328.

Tian, S., Wei, L., Lu, Q., Wei, Z., Zhong, Y., Zhou, Z., 2026. Spatiotemporal dynamics of soil organic carbon at 30 m-resolution in Europe: responses to land cover stability and conversion. Ecol. Ind. 182, 114495.

Tsai, W.P., Feng, D., Pan, M., et al., 2021. From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. Nat. Commun. 12, 5988.

Van de Broek, M., Govers, G., Schrumpf, M., Six, J., 2025. A microbially driven and depth-explicit soil organic carbon model constrained by carbon isotopes to reduce parameter equifinality. Biogeosciences 22, 1427–1446.

Van Noordwijk, M., Aynekulu, E., Hijbeek, R., Milne, E., Minasny, B., Dwi Saputra, D., 2023. Soils as carbon stores and sinks: expectations, patterns, processes, and prospects of transitions. Annu. Rev. Env. Resour. 48 (1), 177–205.

Vaudour, E., Gomez, C., Fouad, Y., Lagacherie, P., 2019. Sentinel-2 image capacities to predict common topsoil properties of temperate and Mediterranean agroecosystems. Remote Sens. Environ. 223, 21–33.

Wadoux, A.M.J.C., Brus, D.J., Heuvelink, G.B.M., 2018. Accounting for non-stationary variance in geostatistical mapping of soil properties. Geoderma 324, 138–147.

Wadoux, A.M.C., Minasny, B., McBratney, A.B., 2020. Machine learning for digital soil mapping: applications, challenges and suggested solutions. Earth Sci. Rev. 210, 103359.

Wadoux, A.M.J.C., Heuvelink, G.B.M., 2023. Uncertainty of spatial averages and totals of natural resource maps. Methods Ecol. Evol. 14, 1320–1332.

Walter, K., Don, A., Tiemeyer, B., Freibauer, A., 2016. Determining soil bulk density for carbon stock calculations: a systematic method comparison. Soil Sci. Soc. Am. J. 80, 579–591.

Wang, G., Post, W.M., Mayes, M.A., 2013. Development of microbial-enzyme-mediated decomposition model parameters through steady-state and dynamic analyses. Ecol. Appl. 23, 255–272.

Wang, L., Abramowitz, G., Wang, Y.-P., Pitman, A., Ciais, P., Goll, D., 2025. Using explainable AI to diagnose the representation of environmental drivers in process-based soil organic carbon models. Biogeosciences 22, 7845–7863.

Wang, M., Zhang, S., Guo, X., Xiao, L., Yang, Y., Luo, Y., Mishra, U., Luo, Z., 2024. Responses of soil organic carbon to climate extremes under warming across global biomes. Nat. Clim. Chang. 14, 98–105.

Widyastuti, M.T., Minasny, B., Padarian, J., Maggi, F., Aitkenhead, M., Beucher, A., Connolly, J., Fiantis, D., Kidd, D., Ma, Y., Macfarlane, F., 2025. Digital mapping of peat thickness and carbon stock of global peatlands. Catena 258, 109243.

Wieder, W.R., Grandy, A.S., Kallenbach, C.M., Bonan, G.B., 2014. Integrating microbial physiology and physio-chemical principles in soils with the MIcrobial-MIneral Carbon Stabilization (MIMICS) model. Biogeosciences 11, 3899–3917.

Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I., 2011. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. Plant and Soil 340, 7–24.

Willard, J.D., Jia, X., Xu, S., Steinbach, M.S., Kumar, V., 2020. Integrating physics-based modeling with machine learning: A survey. arXiv preprint arXiv:2003.04919.

Willard, J., Jia, X., Xu, S., Steinbach, M., Kumar, V., 2022. Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems. ACM Comput. Surv. 55, 66.

Xia, Y., Sanderman, J., Watts, J.D., Machmuller, M.B., Mullen, A.L., Rivard, C., Endsley, A., Hernandez, H., Kimball, J., Ewing, S.A., Litvak, M., Duman, T., Krishnan, P., Meyers, T., Brunsell, N.A., Mohanty, B., Liu, H., Gao, Z., Chen, J., Abraha, M., Scott, R.L., Flerchinger, G.N., Clark, P.E., Stoy, P.C., Khan, A.M., Brookshire, E.N.J., Zhang, Q., Cook, D.R., Thienelt, T., Mitra, B., Mauritz-Tozer, M., Tweedie, C.E., Torn, M.S., Billesbach, D., 2025. Coupling remote sensing with a process model for the simulation of rangeland carbon dynamics. J. Adv. Model. Earth Syst. 17, e2024MS004342.

Xiao, L., Wang, G., Wang, E., Liu, S., Chang, J., Zhang, P., Zhou, H., Wei, Y., Zhang, H., Zhu, Y., Shi, Z., 2024. Spatiotemporal co-optimization of agricultural management practices towards climate-smart crop production. Nat. Food 5, 59–71.

Xie, E., Zhang, X., Lu, F., Peng, Y., Chen, J., Zhao, Y., 2022. Integration of a process-based model into the digital soil mapping improves the space-time soil organic carbon modelling in intensively human-impacted area. Geoderma 409, 115599.

Xie, E., Chen, J., Peng, Y., Yan, G., Zhao, Y., 2024. Integration of machine learning and process-based model outputs via ensemble Kalman filter enhanced space-time modelling of soil organic carbon in a highly human impacted area. Soil Use Manag. 40, e13127.

Xue, J., Zhang, X., Huang, Y., Chen, S., Dai, L., Chen, X., Yu, Q., Ye, S., Shi, Z., 2024. A two-dimensional bare soil separation framework using multi-temporal Sentinel-2 images across China. Int. J. Appl. Earth Obs. Geoinf. 134, 104181.

Xue, P., Minasny, B., Román Dobarco, M., Wadoux, A.M.C., Padarian Campusano, J., Bissett, A., de Caritat, P., McBratney, A., 2025. The biogeography of soil bacteria in Australia exhibits greater resistance to climate change than fungi. Glob. Chang. Biol. 31 (6), e70268.

Yang, C., Shen, F., Li, X., Cui, W., Zhang, L., Yang, L., Zhou, C., 2025. Spatio-temporal mapping reveals changes in soil organic carbon stocks across the contiguous United States since 1955. Commun. Earth Environ. 6, 615.

Yang, L., He, X., Shen, F., Zhou, C., Zhu, A.X., Gao, B., Chen, Z., Li, M., 2020. Improving prediction of soil organic carbon content in croplands using phenological parameters extracted from NDVI time series data. Soil till. Res. 196, 104465.

Yang, L., Cai, Y., Zhang, L., Guo, M., Li, A., Zhou, C., 2021. A deep learning method to predict soil organic carbon content at a regional scale using satellite-based phenology variables. Int. J. Appl. Earth Obs. Geoinf. 102, 102428.

Yigini, Y., Panagos, P., 2016. Assessment of soil organic carbon stocks under future climate and land cover changes in Europe. Sci. Total Environ. 557, 838–850.

Yu, Z.C., 2012. Northern peatland carbon stocks and dynamics: a review. Biogeosciences 9, 4071–4085.

Zhang, L., Cai, Y., Huang, H., Li, A., Yang, L., Zhou, C., 2022. A CNN-LSTM model for soil organic carbon content prediction with long time series of MODIS-based phenological variables. Remote Sens. 14, 4441.

Zhang, L., Heuvelink, G.B.M., Mulder, V.L., Chen, S., Deng, X., Yang, L., 2024a. Using process-oriented model output to enhance machine learning-based soil organic carbon prediction in space and time. Sci. Total Environ. 922, 170778.

Zhang, L., Yang, L., Crowther, T.W., Zohner, C.M., Doetterl, S., Heuvelink, G.B.M., Wadoux, A.-M.-J.-C., Zhu, A.-X., Pu, Y., Shen, F., Ma, H., Zou, Y., Zhou, C., 2025a. Mapping global distributions, environmental controls, and uncertainties of apparent topsoil and subsoil organic carbon turnover times. Earth Syst. Sci. Data 17, 2605–2623.

Zhang, L., Yang, L., Ma, Y., Zhu, A.X., Wei, R., Liu, J., Greve, M.H., Zhou, C., 2025b. Regional-scale soil carbon predictions can be enhanced by transferring global-scale soil–environment relationships. Geoderma 461, 117466.

Zhang, L., Wadoux, A.M.J.C., 2026. Can digital soil mapping be causal? Eur. J. Soil Sci. 77, e70284.

Zhang, M., Wang, X., Ding, X., Yang, H., Guo, Q., Zeng, L., Cui, Y., Sun, X., 2023a. Monitoring regional soil organic matter content using a spatiotemporal model with time-series synthetic Landsat images. Geoderma Reg. 34, e00702.

Zhang, T., Huang, L.M., Yang, R.M., 2024b. Evaluation of digital soil mapping projection in soil organic carbon change modeling. Ecol. Inform. 79, 102394.

Zhang, X., Xie, E., Chen, J., Peng, Y., Yan, G., Zhao, Y., 2023b. Modelling the spatiotemporal dynamics of cropland soil organic carbon by integrating process-based models differing in structures with machine learning. J. Soil. Sediment. 23, 2816–2831.

Zhang, Y., Lavallee, J.M., Robertson, A.D., Even, R., Ogle, S.M., Paustian, K., Cotrufo, M. F., 2021. Simulating measurable ecosystem carbon and nitrogen dynamics with the mechanistically defined MEMS 2.0 model. Biogeosciences 18, 3147–3171.

Zhao, R., Zhang, W., Duan, Z., Chen, S., Shi, Z., 2023. An improved estimate of soil carbon pool and carbon fluxes in the Qinghai-Tibetan grasslands using data assimilation with an ecosystem biogeochemical model. Geoderma 430, 116283.

Zhao, X., Xiong, Z., Karlshöfer, P., Tziolas, N., Wiesmeier, M., Heiden, U., Zhu, X.X., 2025. Soil organic carbon estimation using spaceborne hyperspectral composites on a large scale. Int. J. Appl. Earth Obs. Geoinf. 140, 104504.

Zhou, Y., Chartin, C., Van Oost, K., van Wesemael, B., 2022. High-resolution soil organic carbon mapping at the field scale in southern Belgium (Wallonia). Geoderma 422, 115929.

Y. Zhou Modelling the space-time dynamics of soil organic carbon 2024 Université catholique de Louvain Available PhD thesis at: http://hdl.handle.net/2078.1/294795.

Zhou, Y., Ferdinand, M.S., van Wesemael, J., Dvorakova, K., Baret, P.V., Van Oost, K., van Wesemael, B., 2025. A framework for mapping conservation agricultural fields using optical and radar time series imagery. Remote Sens. Environ. 328, 114858.

Zhou, Z., Ren, C., Wang, C., Delgado-Baquerizo, M., Luo, Y., Luo, Z., Du, Z., Zhu, B., Yang, Y., Jiao, S., Zhao, F., Cai, A., Yang, G., Wei, G., 2024. Global turnover of soil mineral-associated and particulate organic carbon. Nat. Commun. 15, 5329.