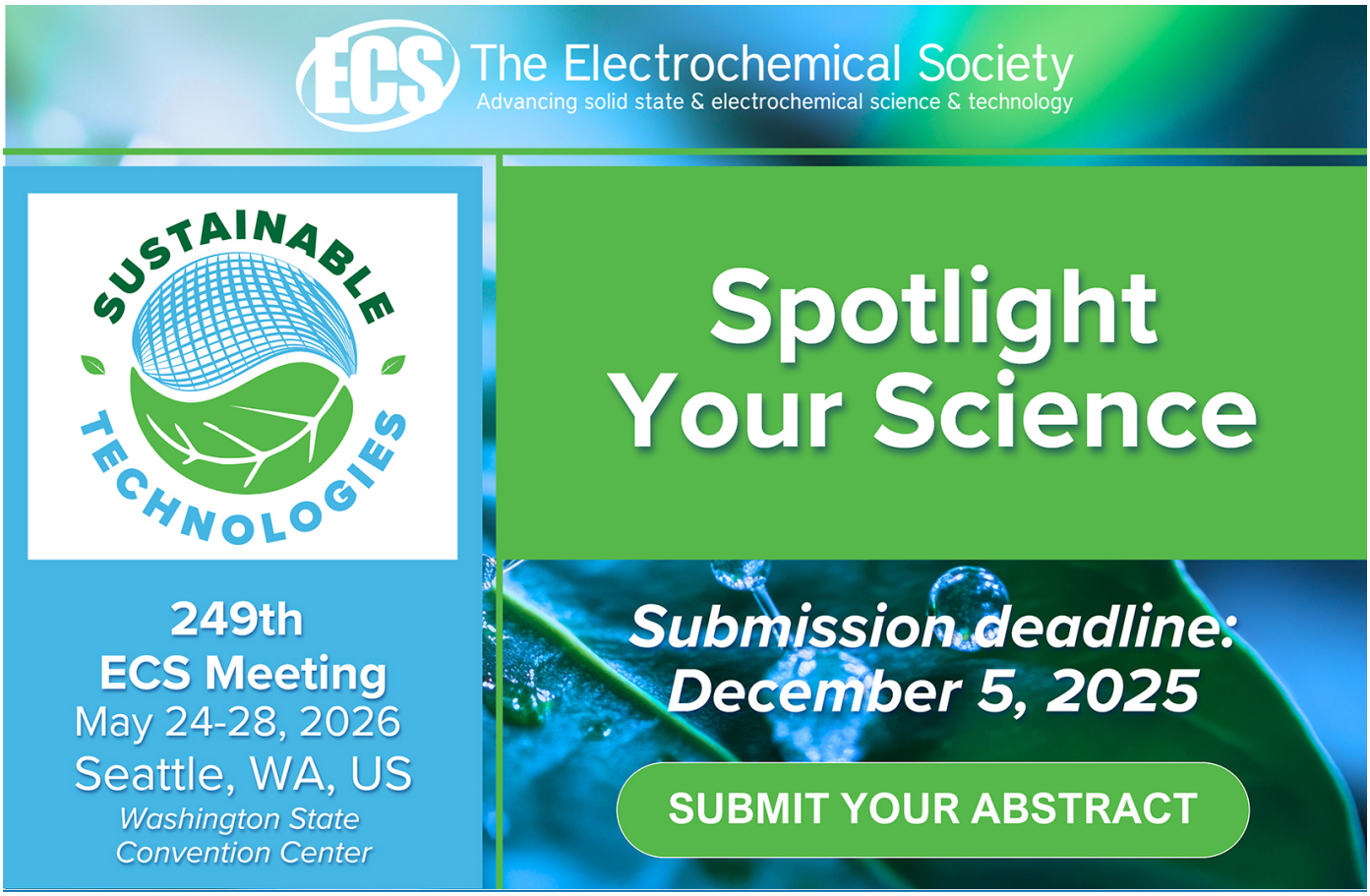
ENVIRONMENTAL RESEARCH
LETTERS

PURPOSE-LED PUBLISHING™ Part of the ENVIRONMENTAL RESEARCH SERIES

PERSPECTIVE • **OPEN ACCESS**

Mining global soil carbon datasets: can modern machine learning uncover the missing pieces of process-based models?

To cite this article: Shoji Hashimoto *et al* 2025 *Environ. Res. Lett.* **20** 101003



ENVIRONMENTAL RESEARCH
LETTERS

PERSPECTIVE

Mining global soil carbon datasets: can modern machine learning uncover the missing pieces of process-based models?

Shoji Hashimoto[1,2,*], **Elisa Bruni**[3], **Boris Ťupek**[4], **Naoyuki Yamashita**[5], **Jumpei Toriyama**[1], **Taiki Mori**[6], **Akihiro Imaya**[1], **Bertrand Guenet**[3], **Akihiko Ito**[2] and **Aleksi Lehtonen**[4]

[1] Department of Forest Soils, Forestry and Forest Products Research Institute, Tsukuba, Ibaraki 305-8687, Japan
[2] Graduate School of Agricultural and Life Sciences, The University of Tokyo, Bunkyo-ku, Tokyo 113-8657, Japan
[3] Laboratoire de Geologie ENS, PSL Research University, CNRS, UMR 8538, IPSL, Paris, France
[4] Bioeconomy and Environment Unit, Natural Resources Institute Finland (Luke), Helsinki 00790, Finland
[5] Shikoku Research Center, Forestry and Forest Products Research Institute, Asakuranishi, Kochi 780-8077, Japan
[6] Kyushu Research Center, Forestry and Forest Products Research Institute, Kurokami, Kumamoto 860-0862, Japan
[*] Author to whom any correspondence should be addressed.

E-mail: hashimoto_shoji710@ffpri.go.jp

Keywords: global soil carbon modelling, machine learning, data-mining, global soil carbon map

OPEN ACCESS

RECEIVED
26 June 2025

REVISED
13 August 2025

ACCEPTED FOR PUBLICATION
22 August 2025

PUBLISHED
2 September 2025

Abstract

The future of terrestrial soil carbon stocks plays a crucial role in climate change prediction. Modern machine learning techniques are now widely applied in soil science to predict the spatial distribution of soil properties from observational data. Beyond prediction, the use of machine learning as a data-mining tool offers a promising pathway for improving soil carbon modelling and refining projections of climate–carbon feedbacks. In this paper, we review recent advances in the application of machine learning to global soil carbon modelling as a data-mining tool and highlight its potential to drive an iterative feedback loop that improves the representation of soil carbon dynamics in Earth System Models.

1. The importance of soil data and mismatches between modelled and observed soil organic carbon (SOC) stocks

SOC is the largest organic carbon reservoir in the terrestrial biosphere, with estimates ranging from 500 to 3000 Pg C (Scharlemann *et al* 2014, Friedlingstein *et al* 2025). Although a substantial portion of these stocks are in anaerobic (waterlogged) conditions such as peat soil, most of the stocks in upland soils are shaped as a consequence of long-term processes such as photosynthesis, plant growth, organic matter inputs to soil, decomposition, and land-use change. Even a small loss in SOC could significantly increase atmospheric CO_2 concentrations, making SOC dynamics a critical area of study (Minasny *et al* 2017). Recent years have seen major efforts to develop SOC maps based on tens of thousands of soil profiles (Hengl *et al* 2017, Poggio *et al* 2021), which now serve as essential benchmarks for global carbon modelling (Todd-Brown *et al* 2013, Tian *et al* 2015).

Despite ongoing advancements in soil carbon models and numerous model intercomparison studies, substantial discrepancies persist between data-driven global SOC estimates and outputs from process-based models (Todd-Brown *et al* 2013, Tian *et al* 2015, Ito *et al* 2020, Varney *et al* 2022). These mismatches include differences in the total SOC stock as well as spatial distribution patterns. Such discrepancies reduce the reliability of future climate projections, which rely on accurate representations of soil carbon dynamics.

2. Machine learning as a data-mining tool to identify the causes of mismatch

Although potential mechanisms that should be incorporated into process-based models have been widely discussed, few detailed analyses have attempted to pinpoint the causes of mismatches between modelled and observed SOC data. These discrepancies remain poorly understood, largely because of the absence

of suitable methodologies to disentangle underlying causes across large-scale datasets—those with tens of thousands of grid cells or more—such as high-resolution, data-driven global SOC maps and outputs from sophisticated process-based models.

However, recent studies have highlighted a promising approach: machine learning (Hashimoto *et al* 2017, Hengl *et al* 2017, Reichstein *et al* 2019, Georgiou *et al* 2021). Broadly defined, machine learning encompasses computational methods that enable data-driven prediction or classification by identifying patterns in input data. More narrowly, the term refers to flexible and computer-intensive algorithmic techniques such as random forests, gradient boosting, and neural networks—that are increasingly applied in environmental and ecological research.

These methods are capable of processing larger and more complex datasets than traditional statistical approaches, allowing researchers to uncover nonlinear relationships and hidden patterns that would otherwise remain undetected. That said, machine learning has its limitations. Most notably, it typically requires large datasets to perform effectively, and may be less reliable when applied to small or sparse data. Additionally, algorithms such as random forests are designed to capture general trends, which may limit their usefulness for identifying rare or highly localized cases. Furthermore, many machine learning models function as black boxes: while they provide strong predictive power, they often offer limited insight into how specific input variables contribute to model outputs.

Despite these limitations, machine learning has become the dominant tool in soil mapping research—a major focus in pedometrics—for predicting the spatial distribution of soil properties and generating digital soil maps from observational datasets (*Mapping* in figure 1) (Hengl *et al* 2017, Lamichhane *et al* 2019, Yamashita *et al* 2022). It is also applied to upscale soil respiration measurements from field data (Warner *et al* 2019). These predictive uses of modern machine learning are now well-established and often outperform conventional methods such as linear regression models (Lamichhane *et al* 2019).

Importantly, while machine learning models are often considered 'black boxes,' certain algorithms can quantify the relative importance of covariates that influence the target variable (e.g. SOC stock) and the relationships between the target variable and the influencing factors. This capacity makes them valuable not only for prediction but also as data-mining tools for exploring hidden factors and causal relationships in big data (figure 1). Recent studies have used machine learning in this way to explore the causes of mismatches between observations and predictions of SOC stocks (*Mining map* and Step 1–2 in figure 1). By applying these algorithms to observational SOC maps, global soil profile data, and the outputs of global models, researchers have been able to identify key controlling factors for each dataset and examine how SOC relates to various environmental covariates (Reichstein *et al* 2019, Georgiou *et al* 2021).

3. Identified potential causes of mismatches

To our knowledge, one of the most significant applications of this approach was presented by Georgiou *et al* (2021). In their study, the authors explored the controlling factors of SOC spatial distribution using multivariate linear regressions, gradient boosting machines, and random forests. These methods were applied to raw soil profile data, the Harmonized World Soil Database (HWSD), and the Northern Circumpolar Soil Carbon Database (NCSCD), as well as to outputs from three global models: CASA-CNP, MIMICS, and CORPSE. Their results indicated that models tended to overemphasize the roles of temperature and primary productivity, whereas observational data suggested a greater influence of soil mineralogy. Significant mismatches were identified globally, although some areas showed agreement within specific biomes.

In addition to identifying the important factors, the machine learning analysis quantified the relationships between SOC and factors in observational datasets and model outputs. For example, they found that while all datasets showed a positive relationship between SOC and net primary productivity (NPP), the relationship was saturating and varied in strength across datasets. These findings demonstrate that machine learning can serve as a powerful tool for identifying biomes with, or without, major mismatches between observed and modelled controlling factors, as well as missing processes in models or poorly constrained model parameters and forcing inputs.

A similar approach was taken earlier by Hashimoto *et al* (2017). In that study, the authors applied machine learning—specifically, boosted regression trees—to global SOC maps using datasets such as the HWSD, IGBP-DIS, NCSCD for northern regions, and outputs from 15 Earth System Models (ESMs) participating in the Coupled Model Intercomparison Project Phase 5 (CMIP5). Their analysis suggested that in observational datasets, mean annual temperature, clay content, carbon-to-nitrogen (C:N) ratio, wetland ratio, and land cover were key contributors to SOC distribution. In contrast, for the ESM outputs, mean annual temperature, land cover, and NPP were dominant factors. The greater influence of temperature and NPP in the models reflects the common structure of SOC dynamics in ESMs—where SOC change over time is modelled

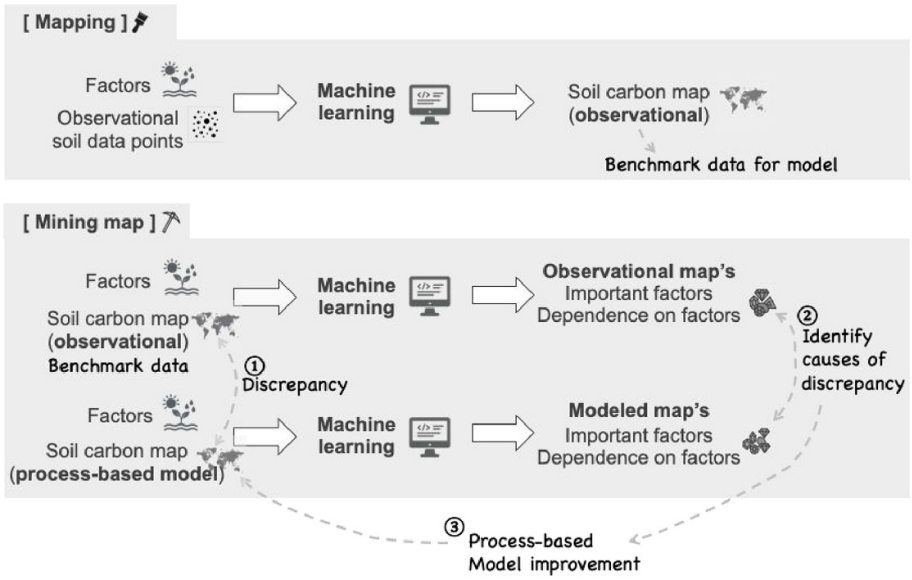


Figure 1. Conceptual diagram illustrating the use of machine learning to investigate the causes of mismatches between observational and modelled soil organic carbon (SOC) distributions. Machine learning is the dominant tool in soil mapping research (*Mapping* in this diagram). Recent studies suggest machine learning as a powerful approach for identifying influential factors and diagnosing mismatches between models and observations (*Mining map* and Step 1–2 in this diagram). Steps 1–3 represent a cycle through which machine learning can inform and improve process-based SOC models.

as a function of carbon inputs and decomposition rates, with temperature serving as the primary modifier. The study also highlighted distinct controlling factors in northern soils, demonstrating the value of this type of analysis.

Similar studies have been conducted using U.S.-scale observational data (Mishra *et al* 2022), global observational data (Luo *et al* 2021), and combined global observational and ESM output data (Nyaupane *et al* 2024). In the 2024 study, machine learning was applied to investigate patterns and controlling factors in both field-based observed SOC data and SOC estimates from CMIP6 Earth System Models. They found that precipitation, temperature, and NPP were the dominant drivers of SOC variability in ESMs.

4. Looking ahead

These pioneering studies revealed a consistent pattern: model outputs were more strongly influenced by climate and NPP, while observational map data showed more complex controlling mechanisms. The consistency of findings across multiple studies—despite differences in machine learning techniques and datasets used—suggests that the causes of these mismatches are likely robust.

The examination of mismatch causes has only recently begun, and further studies are needed. Emerging methods such as Shapley values (Wadoux *et al* 2023), ensemble machine learning (Mishra *et al* 2020), and physics-informed machine learning (Minasny *et al* 2024), may offer deeper insights. Furthermore, incorporating additional covariates—such as fire (Pellegrini *et al* 2022), drought (Canarini *et al* 2017), microbial activity (García-Palacios *et al* 2021), and mineral interactions (Georgiou *et al* 2022)—could shift the perceived importance of controlling factors. Because SOC is shaped by processes operating on centennial to millennial timescales, covariates representing longer-term processes would be also effective. Since machine learning identifies the most important variables from the available covariates, expanding the list of covariates could yield new interpretations. It is also important to recognize that the observational, data-driven maps themselves carry substantial uncertainty, including significant inter-product variability (Fan *et al* 2020, Hashimoto *et al* 2023), and that these datasets are continually updated (*Mapping* in figure 1).

In summary, despite certain limitations, recent advances in machine learning are transforming our ability to predict SOC dynamics (e.g. stocks and respiration) and serve as powerful data-mining tools. Although data-intensive and not fully mechanistic, these methods excel at identifying key drivers and diagnosing mismatches between models and observations, providing a quantitative basis for targeted improvements in global soil carbon models. In parallel with further data-mining using emerging methods and additional covariates, integrating such insights into model development is an urgent priority (Step 3 in figure 1). The cycle of identifying the causes of mismatches and improving models is iterative (Steps 1–3 in figure 1), rather than a one-time process. This machine-learning-assisted feedback loop can accelerate the accurate representation of soil carbon dynamics in Earth System Modelling.

Data availability statement

No new data were created or analysed in this study.

Acknowledgment

This study is supported by JSPS KAKENHI Grant Numbers JP19H03008, JP21H03580, JP23K23665, JP24K01818, and JP24K01817, by the Environment Research and Technology Development Fund (JPMEERF24S12208 and JPMEERF25S12423) of the Environmental Restoration and Conservation Agency provided by Ministry of the Environment of Japan, and also by the Grant Holistic management practices, modelling and monitoring for European forest soils—HoliSoils (EU Horizon 2020 Grant Agreement No 101000289) and by the research and innovation action 'Soil Health and Agriculture Resilience through an Integrated Geographical information systems of Mediterranean Drylands' (SHARInG-MeD) funded by the 'Partnership for Research & Innovation in the Mediterranean Area' (PRIMA Foundation) under the Grant Agreement No. 2211. The authors acknowledge the use of the ChatGPT language model, developed by OpenAI for providing language assistance in preparing the manuscript, and for generating the icons used in the figure.

ORCID iDs

Shoji Hashimoto 0000-0003-3022-7495
Elisa Bruni 0000-0001-8074-0516
Naoyuki Yamashita 0000-0003-3398-6825
Jumpei Toriyama 0000-0001-8061-6398
Akihiro Imaya 0009-0003-5727-2154
Akihiko Ito 0000-0001-5265-0791
Aleksi Lehtonen 0000-0003-1388-0388

References

Canarini A, Kiær L P and Dijkstra F A 2017 Soil carbon loss regulated by drought intensity and available substrate: a meta-analysis *Soil Biol. Biochem.* **112** 90–99

Fan N, Koirala S, Reichstein M, Thurner M, Avitabile V, Santoro M, Ahrens B, Weber U and Carvalhais N 2020 Apparent ecosystem carbon turnover time: uncertainties and robust features *Earth Syst. Sci. Data* **12** 2517–36

Friedlingstein P *et al* 2025 Global carbon budget 2024 *Earth Syst. Sci. Data* **17** 965–1039

García-Palacios P, Crowther T W, Dacal M, Hartley I P, Reinsch S, Rinnan R, Rousk J, van den Hoogen J, Ye J-S and Bradford M A 2021 Evidence for large microbial-mediated losses of soil carbon under anthropogenic warming *Nat. Rev. Earth Environ.* **2** 507–17

Georgiou K *et al* 2022 Global stocks and capacity of mineral-associated soil organic carbon *Nat. Commun.* **13** 3797

Georgiou K *et al* 2021 Divergent controls of soil organic carbon between observations and process-based models *Biogeochemistry* **156** 5–17

Hashimoto S, Ito A and Nishina K 2023 Divergent data-driven estimates of global soil respiration *Commun. Earth Environ.* **4** 460

Hashimoto S, Nanko K, Ťupek B and Lehtonen A 2017 Data-mining analysis of the global distribution of soil carbon in observational databases and Earth system models *Geosci. Model Dev.* **10** 1321–37

Hengl T *et al* 2017 SoilGrids250m: global gridded soil information based on machine learning *PLoS One* **12** e0169748

Ito A *et al* 2020 Soil carbon sequestration simulated in CMIP6-LUMIP models: implications for climatic mitigation *Environ. Res. Lett.* **15** 124061

Lamichhane S, Kumar L and Wilson B 2019 Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: a review *Geoderma* **352** 395–413

Luo Z, Viscarra-Rossel R A and Qian T 2021 Similar importance of edaphic and climatic factors for controlling soil organic carbon stocks of the world *Biogeosciences* **18** 2063–73

Minasny B *et al* 2024 Soil science-informed machine learning *Geoderma* **452** 117094

Minasny B *et al* 2017 Soil carbon 4 per mille *Geoderma* **292** 59–86

Mishra U, Gautam S, Riley W J and Hoffman F M 2020 Ensemble machine learning approach improves predicted spatial variation of surface soil organic carbon stocks in data-limited northern circumpolar region *Front. Big Data* **3** 528441

Mishra U, Yeo K, Adhikari K, Riley W J, Hoffman F M, Hudson C and Gautam S 2022 Empirical relationships between environmental factors and soil organic carbon produce comparable prediction accuracy to machine learning *Soil Sci. Soc. Am. J.* **86** 1611–24

Nyaupane K, Mishra U, Tao F, Yeo K, Riley W J, Hoffman F M and Gautam S 2024 Observational benchmarks inform representation of soil organic carbon dynamics in land surface models *Biogeosciences* **21** 5173–83

Pellegrini A F A, Harden J, Georgiou K, Hemes K S, Malhotra A, Nolan C J and Jackson R B 2022 Fire effects on the persistence of soil organic matter and long-term carbon storage *Nat. Geosci.* **15** 5–13

Poggio L, De Sousa L M, Batjes N H, Heuvelink G B M, Kempen B, Ribeiro E and Rossiter D 2021 SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty *SOIL* **7** 217–40

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

Scharlemann J P, Tanner E V, Hiederer R and Kapos V 2014 Global soil carbon: understanding and managing the largest terrestrial carbon pool *Carbon Manage.* **5** 81–91

Tian H *et al* 2015 Global patterns and controls of soil organic carbon dynamics as simulated by multiple terrestrial biosphere models: current status and future directions *Glob. Biogeochem. Cycles* **29** 775–92

Todd-Brown K E O, Randerson J T, Post W M, Hoffman F M, Tarnocai C, Schuur E A G and Allison S D 2013 Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations *Biogeosciences* **10** 1717–36

Varney R M, Chadburn S E, Burke E J and Cox P M 2022 Evaluation of soil carbon simulation in CMIP6 Earth system models *Biogeosciences* **19** 4671–704

Wadoux A M J-C, Saby N P A and Martin M P 2023 Shapley values reveal the drivers of soil organic carbon stock prediction *Soil* **9** 21–38

Warner D L, Bond-Lamberty B, Jian J, Stell E and Vargas R 2019 Spatial predictions and associated uncertainty of annual soil respiration at the global scale *Glob. Biogeochem. Cycle* **33** 1733–45

Yamashita N *et al* 2022 National-scale 3D mapping of soil organic carbon in a Japanese forest considering microtopography and tephra deposition *Geoderma* **406** 115534