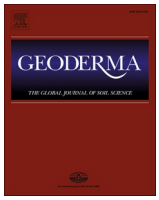
Geoderma 461 (2025) 117466

Contents lists available at ScienceDirect

Geoderma

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



Regional-scale soil carbon predictions can be enhanced by transferring global-scale soil–environment relationships

Lei Zhang [a,b], Lin Yang [a,*], Yuxin Ma [c], A-Xing Zhu [d], Ren Wei [a], Jie Liu [a], Mogens H. Greve [e], Chenghu Zhou [a,f]

[a] School of Geography and Ocean Science, Nanjing University, Nanjing, China
[b] Climate and Ecosystem Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[c] New South Wales Department of Climate Change, Energy, the Environment and Water, Parramatta, NSW, Australia
[d] Department of Geography, University of Wisconsin-Madison, Madison, WI, USA
[e] Department of Agroecology, Aarhus University, Tjele, Denmark
[f] State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Keywords:
Soil organic carbon
Soil–environment relationships
Soil mapping
Pre-trained model
Model transferability
Deep learning
Global and regional scales

ABSTRACT

Accurate modelling and mapping soil organic carbon are crucial for supporting soil health restoration and climate change mitigation at both regional and global scales. However, regional soil predictions often suffer from data scarcity and high prediction uncertainty. Utilizing a pre-trained global-to-regional soil carbon predictive model can be a potential solution to address this challenge. Despite its promise, how to construct and apply the global-scale model to enhance regional-scale soil carbon mapping remains largely unexplored. Here, we propose the Global Soil Carbon Pre-trained Model (GSoilCPM), a deep-learning-based domain adaptative model, to enhance regional-scale soil carbon predictions. Based on large amount of environmental covariate data and 106,167 soil samples across the globe, we verify our hypothesis of the effectiveness of this 'global-to-regional' modelling strategy. The pre-trained model can be then transferred and fine-tuned to bridge the regional- and global-scale soil–environment relationships. We applied and validated this modelling strategy in four regional-scale study areas, three in the Northern Hemisphere and one in the Southern Hemisphere, each with distinct environmental background. Compared to traditional modelling approaches as a baseline, four case studies all demonstrated significant improvement in prediction accuracy across diverse environments and varying data availabilities. The average percentage improvement across all regions is 10.93% (absolute values decreased by 1.20 g kg^{-1} averagely) in MAE and 29.04% (absolute values increased by 0.10 averagely) in CCC. The applicability and future horizons of using GSoilCPM were further discussed. We further reveal that regions with fewer soil samples or lower baseline accuracy benefit more from the pre-trained global model. Our findings highlight the advantages of leveraging the generalized knowledge from global models to enhance specifically localized soil modelling, positioning a potential paradigm shift in digital soil mapping, and far-reaching implications for soil monitoring and land management.

1. Introduction

Soil organic carbon (SOC), a critical component of the Earth's carbon cycle, plays a fundamental role in maintaining soil fertility, regulating greenhouse gas emissions, and supporting overall ecosystem health (Tiessen et al., 1994; Schmidt et al., 2011; Sanderman et al., 2017). Accurate predicting and mapping SOC are essential for understanding carbon dynamics at both regional and global scales, guiding sustainable land management practices, and contributing to climate change mitigation strategies (Sanchez et al., 2009; Chen et al., 2022; Huang et al., 2022; Zhang et al., 2025). Nevertheless, our current understanding of SOC distributions is frequently impeded by high uncertainties stemming from limited sample data and the underdevelopment of modelling techniques (Wadoux et al., 2021).

Traditional soil mapping methods rely on experienced soil surveyors who are familiar with the local area and spend considerable time in field

* Corresponding author.
E-mail addresses: lei.zhang.geo@outlook.com (L. Zhang), yanglin@nju.edu.cn (L. Yang).

https://doi.org/10.1016/j.geoderma.2025.117466
Received 9 April 2025; Received in revised form 19 June 2025; Accepted 29 July 2025
Available online 5 August 2025

work, manually creating maps using the polygon data model (lines, points, and polygon shapes) (Bliss et al., 1995; Minasny and McBratney, 2016; Brevik et al., 2017). Recent advancements in geostatistics (Heuvelink and Webster, 2001; Hengl et al., 2004; Heuvelink et al., 2016; Ma et al., 2017), machine learning algorithms (Heung et al., 2016; Wadoux et al., 2020; Meng et al., 2022, 2024a; Helfenstein et al., 2024; Guo et al., 2025; Zhang et al., 2025), and increasingly accessible land surface datasets (e.g., geographical variables derived from remote sensing observations) (Mulder et al., 2011; Ivushkin et al., 2019; He et al., 2021) have introduced alternative ways to predict and map soil information with less manual labor. Data-driven approaches facilitated us to build soil–environment relationships from soil samples and environmental covariates, and then to apply these fitted relationships to generate predicted soil maps. That is where digital soil mapping (DSM) methods come into play (McBratney et al., 2003; Minasny and McBratney, 2016; Ma et al., 2019; Liu et al., 2022). DSM technique has been widely applied to soil carbon mapping and has yielded satisfactory results at both global and regional scales (Minasny et al., 2013; Huang et al., 2022; Wang et al., 2023; Meng et al., 2024b; Zhang et al., 2022a, 2024, 2025). Despite these advancements, we are still facing a 'data hungry' problem in soil mapping field, especially when the soil data are limited in specific regions (Zhu et al., 2015; Zhang et al., 2021; Cui et al., 2025). The scarcity of soil observations makes it difficult to construct robust soil–environment relationships, which hampers the ability to accurately capture soil carbon patterns at regional scales.

A potential strategy to overcome this bottleneck is to construct a pre-trained domain adaptive model trained on globally available soil and environmental data. This model would operate as a 'knowledge' on soil–environment relationships across the globe, and could then be adapted (e.g., fine-tuned or transferred) for the regional-scale SOC predictions. Given the success of pre-trained and transfer modelling strategy in fields outside of soil mapping (Awais et al., 2023; Moor et al., 2023; Wu et al., 2023; Hong et al., 2024), it is reasonable to explore whether such an approach could be applied for soil carbon predictions and whether it could help alleviate the issue of insufficient soil observations at regional scales and improve mapping accuracy. While a similar 'global-to-local' approach has been used in soil spectroscopic modelling (Shen et al., 2022; Viscarra Rossel et al., 2024), a clear research gap remains in applying this concept and designing appropriate modelling frameworks for SOC mapping.

Another motivation for this study lies in the fact that the demand for accurate regional-scale soil predictions is often difficult to directly fulfill using existing global-scale soil map products. Some efforts have produced global maps of soil carbon at medium spatial resolutions, such as the well-known Harmonized World Soil Database (HWSD) (FAO, 2012) and SoilGrids product (Hengl et al., 2014, 2017; Poggio et al., 2021). However, recent studies have highlighted significant uncertainties in these global soil gridded datasets (Dai et al., 2019; Lilburne et al., 2024). This is largely due to incomplete coverage of global soil profiles, limiting representation across all areas (Batjes et al., 2017). Moreover, global-scale modelling often underestimates or biases the high spatial variability of soil carbon at local scales. This is because spatial predictions tend to smooth distribution tails (Nussbaum et al., 2023) and encounter gaps in predictor space where training sample data is insufficient (Meyer and Pebesma, 2021). In addition, models developed for global applications, such as the random forest (RF) models for producing SoilGrids, cannot be directly transferred to local areas. These factors impede the potential re-utilization of global-scale data and models, which would be a valuable foundational resource for enhancing regional-scale SOC predictions.

To address this challenge, we propose leveraging a Global Soil Carbon Pre-trained Model (GSoilCPM), initially trained on global-scale soil profiles with SOC observations and remote sensing derived environmental covariate datasets, to enhance regional-scale SOC mapping. The model architecture not only adopts soil formation theory (Jenny, 1941; McBratney et al., 2003) to differently process the environmental covariates influencing soil by variable category, but also uses deep convolutional networks to separately extract latent feature from each covariate category with varying window sizes that captures spatial contextual information. The pre-trained model can then be fine-tuned with an additional round of parameter optimization to perform regional SOC mapping tasks.

Here, we aim to determine whether the proposed deep-learning (DL)-based model can achieve competitive performance compared to the widely used ensemble machine learning models at the global scale. Then, we hypothesize that the challenges of data scarcity and complex nonlinear patterns in a target region can be mitigated by utilizing the pre-trained global-level model. To investigate this hypothesis, we applied a transfer learning approach to bridge the gap between global and regional soil–environment relationships. This allows the generalized knowledge captured by the global model to be localized, enhancing regional SOC modelling. Establishing this 'global-to-regional' modelling strategy is grounded in the idea that, while local soil conditions are shaped by site-specific processes, many underlying mechanisms driving soil variations are governed by universal geographical and ecological principles. Therefore, those macro-scale regularities captured in a global model can serve as a knowledge prior and can provide a structured starting point for guiding regional soil predictions.

In this study, we first introduced the GSoilCPM model and trained it using the latest global soil profile database from the World Soil Information Service (WoSIS), integrated with various environmental covariates, including reflectance bands from satellite observations, climate, topography, vegetation and parent materials. We then applied the GSoilCPM model to generate SOC maps in four regions with four distinct landscape backgrounds and varying data availabilities. These results were compared with predictions from SoilGrids, RF models, and DL models without pre-training. Finally, we examined the advantages of using the global-scale model to enhance regional soil carbon modelling, supported by varying sizes of sample data in each case study areas, thereby illustrating the applicability and prospects of this modelling strategy.

2. Materials and methods

2.1. Global-scale datasets

2.1.1. Soil samples

The global-scale soil sample data were sourced from the WoSIS database (Batjes et al., 2020, 2024), one of the largest and most comprehensive repositories of harmonized soil profile data worldwide. For this study, we used the "WoSIS snapshot – September 2019", which comprises 196,498 geo-referenced soil profiles worldwide with over 832,000 soil layers or horizons. We filtered the profiles to include only those with measurements of SOC content (g kg^{-1}), and then checked for the layer observations up to a depth of 0.3 m. The profiles with low accuracy of the geographical coordinates (i.e., missing the information of degree, minute or second data) and those flagged as containing surface litter were excluded. A total number of 106,167 global soil samples were finally obtained for this study. To estimate the mean SOC content in topsoil (0–0.3 m), we employed the equal-area spline algorithm to fit multiple SOC observations at different depth intervals. This method has been proven to be superior to alternative functions (Bishop et al., 1999; Malone et al., 2009). The average of the fitted SOC values from 0 to 0.3 m was calculated as SOC content in topsoil for each sample location. After above data processing, the global soil profile locations, along with their topsoil organic carbon content values are shown in Fig. 1**a**. These profiles provide a broad geographic coverage, ensuring a representative sample set of global soil distributions across different environmental conditions.

2.1.2. Environmental covariates that related with SOC

The environmental covariates contain a series of variables that

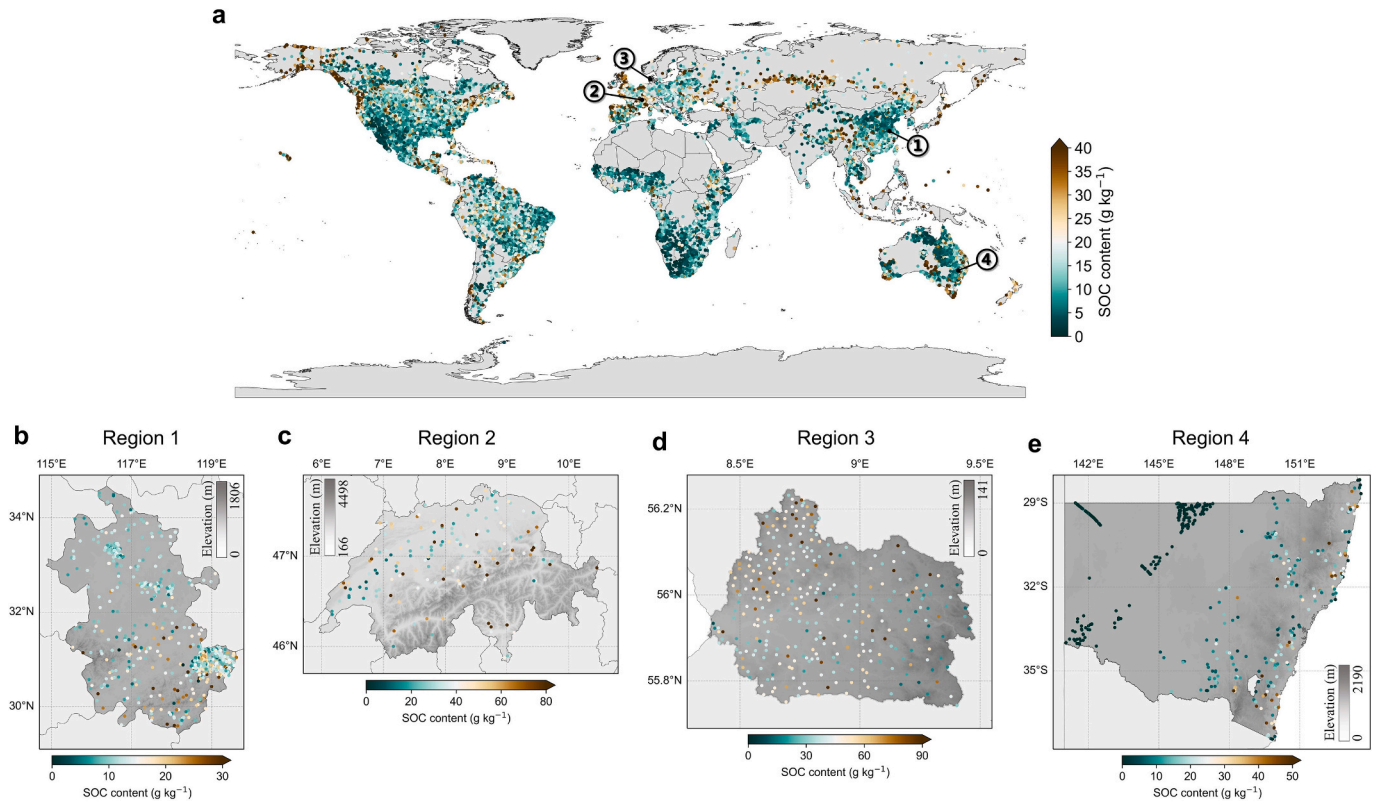


Fig. 1. Spatial distributions of global and regional soil sample data. a, A total number of 106,167 sampling sites (soil profiles) including soil organic carbon (SOC) observations in topsoil were collected from WoSIS database. b-e, Four regional-scale study areas with SOC sample data that are not included in WoSIS database. These study areas are located in Anhui province, China (region 1, $n = 819$) (b), Switzerland (region 2, $n = 150$) (c), Skjern river catchment, Denmark (region 3, $n = 317$) (d), and New South Wales, Australia (region 4, $n = 490$) (e).

influence soil carbon variations (Table 1). As one of the aims of this study is to build a pre-trained large-scale model that links SOC with environmental conditions, we selected most basic environmental variables as covariates, according to the concepts of soil formation factors and soil–landscape relationships (Jenny, 1941; Hudson, 1992; Zhu et al., 2001; McBratney et al., 2003; Zhu et al., 2018). These covariates include climate, topography, vegetation, and parent material. We also collected remote-sensing-based surface reflectance information with seven spectral bands from the moderate-resolution imaging spectroradiometer (MODIS) data product. Details of these covariates are shown in Table 1 and Table S1, and are also described as follows.

The remote sensing images that include multiple reflectance bands can be used as important predictors for soil mapping. Different soil properties, such as soil texture, moisture content, organic matter and mineral composition, reflect and absorb light differently across various wavelengths (Viscarra Rossel et al., 2006a, 2006b; Mulder et al., 2011). Thus, it is useful to feed the spectral bands into the model for analyzing reflectance patterns and inferring soil characteristics. We collected the reflectance data with seven bands (620–670 nm, 841–876 nm, 459–479 nm, 545–565 nm,1230–1250 nm, 1628–1652 nm, and 2105–2155 nm) from the Moderate Resolution Imaging Spectroradiometer (MODIS) MCD43A4 (version 6.1) data product (Schaaf and Wang, 2021). This dataset provides Nadir Bidirectional Reflectance Distribution Function (BRDF)-Adjusted Reflectance (NBAR) values, with a 500 m spatial resolution.

The temperature and precipitation variables obtained from WorldClim (version 2) (Fick and Hijmans, 2017) were used to represent the climate condition for each sample location. In addition to the annual mean temperature and annual precipitation, the variables of temperature seasonality and precipitation seasonality (calculated as coefficient of variation for annual range of temperature and precipitation) were also adopted for reflecting the seasonal variations of these two climatic factors. The spatial resolution of these climatic factors is 1 km.

Five important topographic variables, including elevation, slope, aspect, vector ruggedness measure, stream power index, were collected from the Multi-Error-Removed Improved Terrain (MERIT) DEM data product (Yamazaki et al., 2017) and Geomorpho90m global dataset (Amatulli et al., 2020). The spatial resolution of these topographic factors is 90 m.

The variables representing vegetation growth were obtained by using vegetation indices derived from the satellite sensors of the MODIS. The MODIS Vegetation Indices product (MOD13A1 v061) (Didan, 2021) was adopted for extracting annual mean Enhanced Vegetation Index (EVI), annual minimum EVI, annual maximum EVI values at each sample location. The spatial resolution of the vegetation index value for each pixel is 500 m.

The bedrock property plays a key role in many processes at the Earth surface, and it is also an important influencing factor for soil predictions. We collected the rock property information from a global lithological map database (GLiM) with a resolution of 0.5 degree (Hartmann and Moosdorf, 2012), and extracted the value of basic lithological class at each soil sample location.

The integration of above datasets allows for the establishment of relationships between SOC and environmental covariates at a global scale, providing a solid foundation for the development and training of our deep learning model. It is noted that the set of covariates collected above for modelling was not designed to be too large for exhaustively covering all surface information, such as more different combinations of features across the spectra. The reason to reduce this complexity is that the goal of our modelling at the global scale is to generate a pre-trained large model reflecting the basic soil–environment relationships. Our emphasis is on constructing a foundational global-to-regional soil

Table 1
Description of environmental covariates associated with soil organic carbon variations used as inputs to train the global-scale pre-trained deep learning model.

Category	Covariate name	Abbreviation	Spatial resolution	Window size (default)	Data source	Reference
Reflectance bands	NBAR band1 (620–670 nm)	*BAND1*	500 *m*	5 × 5	MODIS BRDF/Albedo products (MCD43A4 v061)	(Schaaf and Wang, 2021)
	NBAR band2 (841–876 nm)	*BAND2*				
	NBAR band3 (459–479 nm)	*BAND3*				
	NBAR band4 (545–565 nm)	*BAND4*				
	NBAR band5 (1230–1250 nm)	*BAND5*				
	NBAR band6 (1628–1652 nm)	*BAND6*				
	NBAR band7 (2105–2155 nm)	*BAND7*				
Climate	Annual mean temperature	*BIO01*	1 km	3 × 3	WorldClim (version 2)	(Fick and Hijmans, 2017)
	Temperature Seasonality	*BIO04*				
	Annual Precipitation	*BIO12*				
	Precipitation Seasonality	*BIO15*				
Topography	Elevation	*ELEV*	90 *m*	7 × 7	Geomorpho90m	(Amatulli et al., 2020)
	Slope	*SLP*				
	Aspect	*ASP*				
	Vector ruggedness measure	*VRM*				
	Stream power index	*SPI*				
Vegetation	Annual mean EVI	*EVI_mean*	500 *m*	5 × 5	MODIS Vegetation Indices product (MOD13A1 v061)	(Yamazaki et al., 2017; Didan, 2021)
	Annual minimum EVI	*EVI_min*				
	Annual maximum EVI	*EVI_max*				
Parent material	Bedrock type	*BEDROCK*	0.5°	1 × 1	GLiM	(Hartmann and Moosdorf, 2012)

Note: Moderate Resolution Imaging Spectroradiometer (MODIS); Nadir Bidirectional Reflectance Distribution Function (BRDF)-Adjusted Reflectance (NBAR); Enhanced Vegetation Index (EVI); Global Lithological Map Database (GLiM).

mapping approach that is transferable across regions and built on covariates that are widely recognized as core drivers of soil formation, broadly available and minimally uncertain. In this sense, we intentionally designed the scope of input variable set to include the most fundamental and widely accessible environmental covariates as described above. This way can make it easier for any modelers or users to first collect the same set of basic covariate data to re-utilize the pre-trained model we presented here, and then to adjust or extending the model by using their local-specific datasets.

2.2. Regional-scale study areas and datasets

To evaluate the transferability of the global-scale pre-trained model, four distinct regional-scale study areas were selected in consideration of their varying environmental backgrounds and SOC gradients. Three of these study areas are located in the Northern Hemisphere, and one is in the Southern Hemisphere (Fig. 1**b-e**). These regions were chosen for representing different climates, biomes and land covers, thereby providing a robust test for applying the proposed modelling strategy across different regional scales.

Region 1 represents the coverage of Anhui province in central-eastern China, characterized by a subtropical humid monsoon climate (Yang et al., 2021b). The elevation is roughly from 0 to 1806 m with flat plains in the north, low hills in the middle and rugged mountainous terrain in the south. The average annual temperature in Anhui is around 15 °C, with annual precipitation ranging from 750 to 2000 mm. The land cover types mainly consist of croplands, forests and grass and shrublands. Our research group has collected 819 soil samples with SOC observations in this area (Fig. 1**b**) (Zhang et al., 2022b). Region 2 covers the entire country of Switzerland. The climate in this region is generally temperate but varies greatly across localities, ranging from the near-Mediterranean climate at the southern tip to the glacial conditions on the mountaintops. More than half of this region is dominated by the Swiss Alps with rugged peaks, deep valleys and numerous glaciers. The remaining areas are broadly categorized into the Swiss plateau and the Jura mountains. The two primary terrestrial ecoregions in this region are the western European broadleaf forests and the conifer and mixed forests of the Alps (Dinerstein et al., 2017). There are 150 soil samples collected from the EU project Land Use/Cover Area Frame Survey (LUCAS) (Orgiazzi et al., 2018) in this region (Fig. 1**c**). Region 3 is the Skjern river catchment located in Western Jutland, Denmark. The climate in this area is temperate maritime, with a mean annual temperature of approximately 8 °C and annual precipitation around 990 mm. The land surface elevations are from sea level at the coast to 125 m above sea level in the eastern area (Jensen and Illangasekare, 2011). The land use is mainly agriculture, followed by grasslands and forests. There are 317 soil samples collected in this area (Peng et al., 2015) (Fig. 1**d**). Notably, the soil samples from the above three study areas are not included in the WoSIS database. Region 4 is New South Wales in eastern Australia. The climate varies from warm temperate in the north and east to hot arid in the far west and subalpine in the southeastern highlands. Annual rainfall ranges from less than 200 mm to over 2,000 mm, with average daily maximum temperatures between 12 °C and 30 °C. Elevation goes from sea level along the coast to over 2,000 m inland. The land is used for large-scale agriculture (crops and livestock), forestry (native forests and plantations), urban development in major cities, and conservation areas like national parks (Wang et al., 2022). In this region, 490 soil samples were collected, with half included in the WoSIS database and half not (Fig. 1**e**).

The environmental covariates in these four regions were collected and processed to align with the datasets and variables used in the global dataset. The water bodies and urban areas in these study areas were excluded when applying models to generate the predicted maps of SOC. The descriptive statistics of the soil and environmental covariates in these regions are shown in Table S2-S5.

2.3. Architecture of 'GSoilCPM' deep learning model

The proposed GSoilCPM model is built on a deep learning framework to capture the complex relationships between SOC and environmental covariates. The model uses a deep neural network architecture, specifically a convolutional neural network (CNN), to extract important features from the input data, connect and fuse these features, and finally generate the predicted values. The overall architecture of the model is shown in Fig. 2.

The model inputs are separated by the category of covariate, with each category having a pre-defined 2-D (dimensional) input size. This input consists of a set of pixels in square shape with a center pixel where a soil sample located in, and then includes pixels within a window expanding outward from the center point, with a specific spatial resolution for matching the original imagery dataset. This allows the model to process covariates separately by considering their varying spatial contexts and resolutions. For example, the climatic variables are often not necessary to have a high resolution like topographic variables, as the magnitude of variations in spatial neighborhood information is relatively smaller in climatic variables than that in topographic variables. Therefore, different covariates need be processed with their specific window sizes and resolutions before being used as model inputs.

Each category of covariates includes multiple 2-D input variables with the same window size and resolution, allowing them to be combined into a 3-D input for that category. The CNNs were adopted to process these inputs separately and extract their latent features. This includes multiple times of convolutional and pooling operations, which are effective in feature extraction (LeCun et al., 2015). More details on the operations in CNNs refer to Appendix A1. Each category of covariates can be processed to produce a vector of extracted feature. All feature vectors are then concatenated into a single, longer vector. A second round of CNN processes this combined vector to extract the

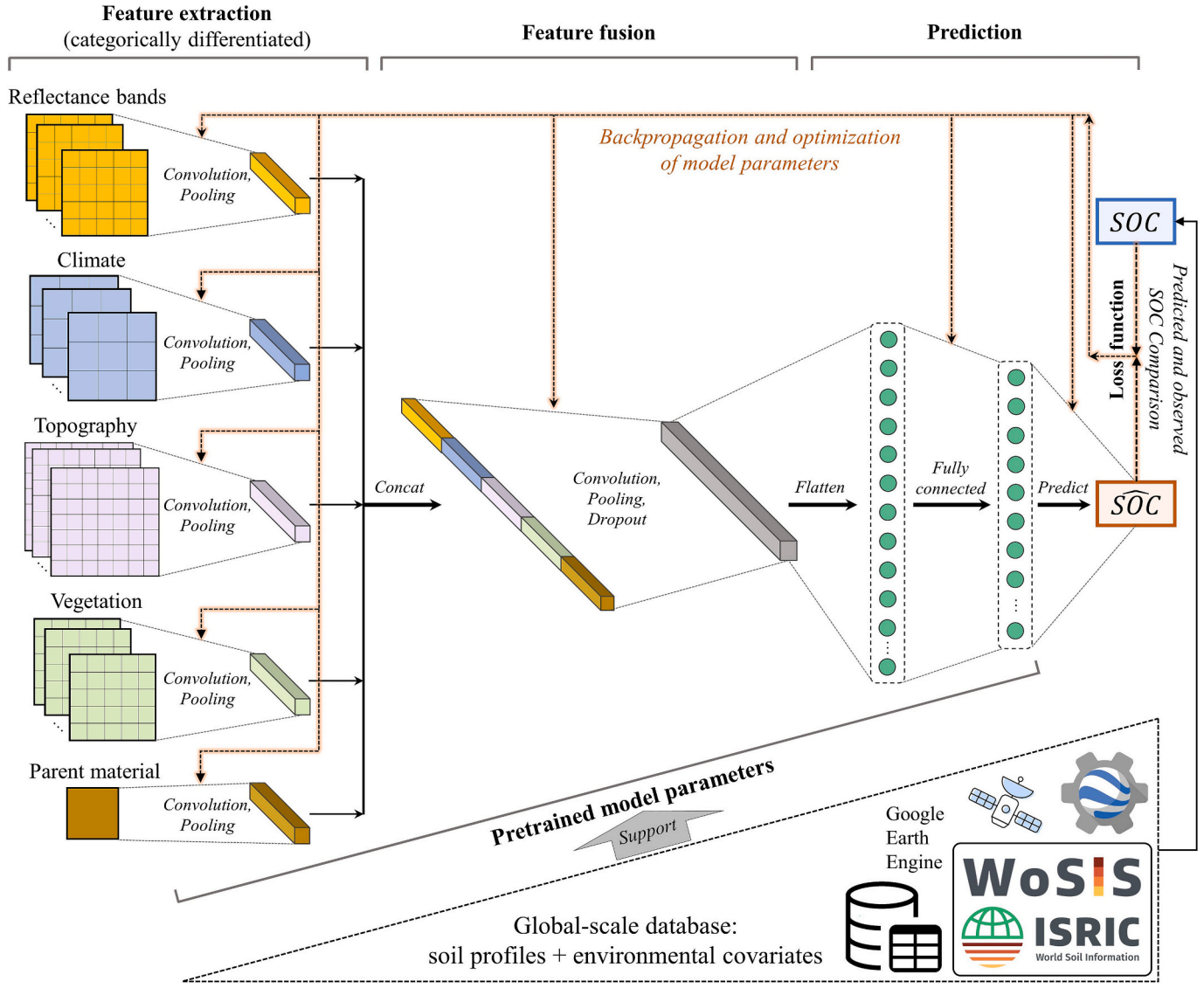


Fig. 2. Architecture of the Global Soil Carbon Pre-trained Model (GSoilCPM). This global-scale model for predicting soil organic carbon (SOC) is designed by using a deep learning framework with a structure that allows for differentiated processing environmental covariates influencing SOC based on soil formation theory. The model uses convolutional networks to extract the latent features from multiple 2-D covariate data for each input category separately, with adjustable window size to account for spatial context and resolution of inputs. These extracted features are concatenated together and processed through convolution and fully connected layers to extract their interactions, and then output the final predicted SOC ($\widehat{SOC}$). The global soil profile observations and remote sensing derived environmental covariates are collected as the global-scale database to support the training procedure of the model. The lighted arrow lines in orange represent that using backpropagation to optimize model parameters by minimizing the loss function generated from comparing the difference between predicted and observed SOC values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

interactions among these features. The dropout layers are included to prevent overfitting. We call this the feature fusion process, which supports the subsequent stage that uses a multilayer neural network (Appendix A2) for continued forward computation. Finally, fully connected layers perform the regression to output the predicted SOC ($\widehat{y}$).

2.4. Global-to-regional training strategy

Different from the usual way of regional soil predictions and soil mapping based on data only in those regions, in this study, the regional soil predictive mapping task is performed by using the regional data with the support of global data. This modelling idea can be called as a 'global-to-regional' model training strategy. The basic concept to implement it is using domain adaptation via transfer learning. It aims to adapt a deep learning model trained on a source domain (i.e., global-scale soil prediction) to perform well on a target domain (i.e., regional-scale soil prediction), when data distribution in the two domains are different but the prediction tasks are similar. Based on the deep learning architecture, domain adaptation can be done by reusing the learned feature representations from a data-rich source domain and fine-tuning the model using relatively limited data from target domain (Ganin et al., 2016; Farahani et al., 2021). The specific methodology and implementation of the above basic concept are described as follows.

Typically, DSM relies on datasets in a certain study area. A data-driven approach is often adopted, employing a machine learning model to establish a soil–environment relationship (f) based on soil samples and environmental covariate datasets only from that region (Fig. 3**a**). This conventional approach can generate a predictive model ($f_{\theta_r} : \boldsymbol{X}_r \mapsto \boldsymbol{y}_r$) for a region as follows:

$$\widehat{\boldsymbol{y}}_r = f_{\theta_r}(\boldsymbol{X}_r) \tag{1}$$

where $\boldsymbol{X}_r$ represents the environmental covariates in the region; $\widehat{\boldsymbol{y}}_r$ represents the predicted SOC values in the region; θ_r represents the model parameters (e.g., the weights that are parameterized in a deep neural network), which can be trained by minimizing the loss function as follows:

$$\widehat{\theta}_r = \underset{\theta_r}{\operatorname{argmin}} \mathscr{L}\left(\boldsymbol{y}_r, f_{\theta_r}(\boldsymbol{X}_r)\right) = \underset{\theta_r}{\operatorname{argmin}} \frac{1}{n_r} \sum_{i=1}^{n_r} \left(y_r^{(i)} - f_{\theta_r}\left(\boldsymbol{x}_r^{(i)}\right)\right)^2 \tag{2}$$

where $\mathscr{L}(\bullet)$ is the loss function measuring the mean squared difference between the predicted values $f_{\theta_r}(\boldsymbol{X}_r)$ and observed SOC values ($\boldsymbol{y}_r$); n_r is the number of samples in the region; $\boldsymbol{x}_r^{(i)}$ and $y_r^{(i)}$ represent the environmental covariates and the observed SOC values for the i-th sample in the regional data, respectively. Here, θ_r is optimized from random starting values, without any prior support from the data out of the region. The backpropagation procedure is adopted to compute the gradient of the loss function with respect to the model parameters to train the model (Appendix A3).

In this study, the global-to-regional training strategy is shown in Fig. 3**b**. We first fit a global-level soil–environmental relationship ($f_{\theta_g} : \boldsymbol{X}_g \mapsto \boldsymbol{y}_g$), which is trained by minimizing the loss function on the global dataset:

$$\widehat{\theta}_g = \underset{\theta_g}{\operatorname{argmin}} \mathscr{L}\left(\boldsymbol{y}_g, f_{\theta_g}(\boldsymbol{X}_g)\right) = \underset{\theta_g}{\operatorname{argmin}} \frac{1}{n_g} \sum_{j=1}^{n_g} \left(y_g^{(j)} - f_{\theta_g}\left(\boldsymbol{x}_g^{(j)}\right)\right)^2 \tag{3}$$

where $\boldsymbol{X}_g$ and $\boldsymbol{y}_g$ represent the environmental covariates and the observed SOC values for all global samples, respectively; n_g is the number of global samples (generally $n_g \gg n_r$); $\boldsymbol{x}_g^{(j)}$ and $y_g^{(j)}$ represent the environmental covariates and the observed SOC values for the j-th sample in the global data, respectively.

Then, the global model can be fine-tuned using the dataset in a target regional area to obtain the regional-level soil–environment relationship ($f_{\theta_{g\to r}}$). Therefore, to transfer the global model to a regional setting, we define an adaptation procedure $\mathscr{T}_{adapt}$ that transforms the pre-trained global model into a regional model by considering limited regional sample data:

$$f_{\theta_g \to \theta_r} = \mathscr{T}_{adapt}\left(f_{\theta_g}; \boldsymbol{X}_r, \boldsymbol{y}_r\right) \tag{4}$$

Alternatively, the adaptation can be expressed as optimizing the regional-specific parameter shift $\Delta\theta$:

$$\theta_r = \theta_g + \Delta\theta \tag{5}$$

$$\widehat{\Delta\theta} = \underset{\theta_g}{\operatorname{argmin}} \mathscr{L}\left(\boldsymbol{y}_r, f_{\theta_g + \Delta\theta}(\boldsymbol{X}_r)\right) \tag{6}$$

This formalism explicitly reflects that the regional model can be built upon the global model and optimized further with regional data, via a training procedure of adaptation. It is noted that this training strategy allows the regional model parameters θ_r to be initialized and further optimized from the pre-trained global parameters. This training strategy is particularly valuable for DSM, where regional datasets are often sparse, but global datasets are more comprehensive. By employing domain adaptation via transfer learning, it enables knowledge reuse of the generalized soil–environment relationships learned globally, thereby reducing local data requirements and enhancing predictive

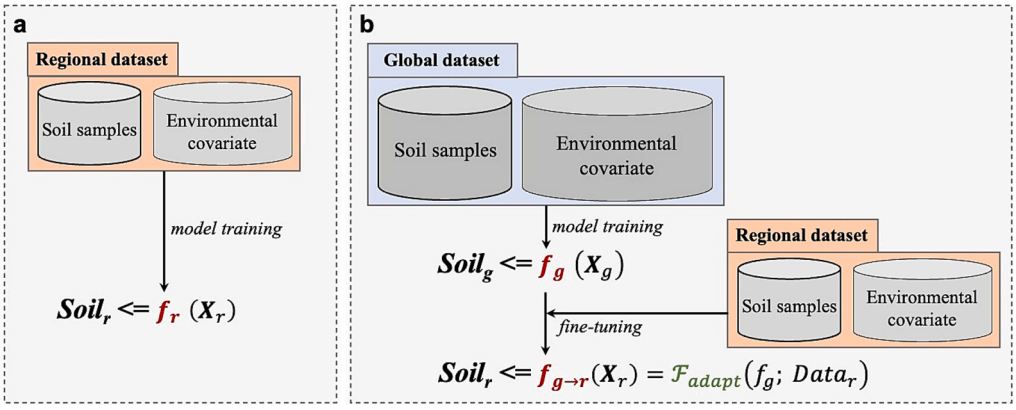


Fig. 3. The difference of basic concepts between the conventional and the proposed new soil mapping methods. (a) The conventional data-driven approach which trains a soil–environmental relationship (f_r) at the regional level only using the soil sample and environmental covariate datasets in a certain region. (b) The proposed new soil prediction/mapping approach first fits a global-level soil–environment relationship (f_g) (it can be fitted by the proposed GSoilCPM that is illustrated in Fig. 2), and then fine-tunes the model by using the dataset in a target regional area to obtain the regional-level soil–environment relationship ($f_{g\to r}$). An adaptation procedure $\mathscr{T}_{adapt}$ is used to transform the pre-trained global model into a regional model by considering limited regional sample data ($Data_r$).

performance.

The previously described global soil profile data with SOC observations and remote sensing derived environmental covariates, serve as the global- and regional scale database to support this training process. The backpropagation is used to optimize model parameters based on the loss function, which is represented as the lighted arrow lines in orange color in Fig. 2.

2.5. Evaluation of SOC predictions

A ten-fold cross-validation was adopted to assess the prediction accuracy by using different modelling methods. This approach helps prevent bias in evaluation caused by the possible overfitting of the model to a single validation set. In each region, all samples were partitioned into ten subsets. Nine subsets were used as the training data to fit the model, and the model performance was validated on the remaining fold. The final assessment of the model was based on the mean accuracy across the ten validation sets. To ensure a representative spatial distribution of the training and validating sample sets, we adopted a global grid with 5-by-5-degree latitude-longitude space, covering the globe, and then split the training and validating sample sets in each grid cell for ensuring uniform spatial coverage. This procedure was repeated ten times, and each time using a different subset for validation. The accuracy metrics of the mean absolute error (MAE) and the concordance correlation coefficient (CCC) were computed on the validation samples to assess the model performance:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - \widehat{y}_i|}{n} \tag{7}$$

$$\mathrm{CCC} = \frac{2r\sigma_y\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\bar{y} - \bar{\widehat{y}})^2} \tag{8}$$

where n is the number of validation samples; y_i and $\widehat{y}_i$ are i-th observed and predicted SOC value respectively; $\bar{y}$ and $\bar{\widehat{y}}$ are the averages of the observed and predicted values; σ_y and $\sigma_{\hat{y}}$ are the corresponding standard deviations; and r is the correlation coefficient value between the predicted and observed values. A smaller value of MAE or a larger value of CCC means a higher prediction accuracy.

The random forest (RF) model was taken as a reference model for comparing with our proposed modelling approach. As an ensemble model, RF generates random subsets of data to train different base learners and then aggregates their predictions, which allows it to have a good ability in generalization. Considering RF has been tested to be superior to many other conventional machine learning models for soil mapping tasks (Brungard et al., 2015; Heung et al., 2016; Zhang et al., 2021), it is reasonable and more challenging to adopt it as an reference model or a baseline to evaluate the performance of the proposed deep learning model. The RF model was trained only on regional data, considering it exemplifies a conventional "local-only" machine learning-based modelling strategy, which is commonly used in many regional DSM tasks. To illustrate the efficiency and advantages of using pre-trained global-scale model for enhancing regional-scale SOC predictions, the deep learning model with the same architecture and built on regional data but without using the pre-training was also taken for comparing. The data split for training and validation in the cross-validation was identical for both the RF and DL models, ensuring a consistent basis for model evaluation and comparison. We further compared our regional modelling results with the SoilGrids product to highlight how widely used global soil data products suffer from potential biases in regional-scale predictions.

To test how modelling improvement (measured as the percentage improvement of prediction accuracy from GSoilCPM versus the baseline RF model) changes in response to different regional sample sizes and baseline accuracies, we determined different numbers of sample data for training models and thus to detect the potential relationships. We randomly selected different proportions of sample data (ranging from 10% to 100% with an interval of 10%) to the total available training data in each region. For each sample size, we repeated the random sampling process 100 times, allowing us to observe statistical differences in the modelling results across sample sizes. We further aggregated all modelling results to analyze the relationship between model improvements and changes in the baseline accuracy of SOC prediction in each study area.

3. Results and Discussion

3.1. Assessment of global-scale modelling

The global-scale deep learning model (GSoilCPM) exhibited competitive model performance of SOC compared with the widely used RF model at the global scale (Fig. 4), both of which were trained on the same WoSIS global soil sample data and environmental covariates. Specifically, cross-validation results illustrate that the deep learning model achieved an accuracy of 7.15 g kg^{-1} in mean absolute error (MAE) and 0.57 in concordance correlation coefficient (CCC), while the RF model had a MAE of 7.56 g kg^{-1} in and a CCC of 0.57. The RF model is well-known for its strong generalization capability and has been widely used. While previous studies have showed challenges in surpassing RF performance with deep learning models (Wadoux, 2019; Yang et al., 2021a), our results indicate that deep learning model can achieve a comparable or slightly better prediction accuracy in both metrics when large amounts of global-scale data are available. Both models effectively captured the global variability in SOC by modelling the complex relationships between SOC and various environmental covariates.

3.2. Global-scale pre-trained GSoilCPM allows enhancement in regional-scale SOC predictions

SOC prediction accuracies for different models across the four regions were assessed using ten-fold cross-validation (Fig. 5). For comparison, we also evaluated the accuracy of the SoilGrids product, validated by our regional sample data (column 1 in Fig. 5). In general, across all four study areas, models trained on regional sample data produced significantly better predictions than using SoilGrids directly. This is because SoilGrids was derived from a model trained on a global soil sample database (i.e., WoSIS data) that does not include samples from our study areas. While previous studies have shown that SoilGrids can provide good SOC predictions (Hengl et al., 2017; Poggio et al., 2021), our findings suggest that its global model cannot guarantee acceptable accuracy in regions lacking WoSIS sample data.

Comparing the baseline RF model with the DL model without pre-training (i.e., models are trained only based on the data in a region) (columns 2 and 3 in Fig. 5), we found that the proposed DL model architecture (without pre-training) achieved competitive result compared to RF, even with limited regional sample data. More importantly, we found that when we used the DL model with pre-training, which means a pre-trained GSoilCPM was constructed and then it was transferred into these regional study areas, model performance shows a large improvement (column 4 in Fig. 5). These results support our hypothesis that the regional-scale soil carbon predictions can be enhanced by transferring global-scale soil–environment relationships via using a DL-based modelling approach. Specifically, in four case studies from Region 1 to Region 4, the metrics of MAE (unit: g kg^{-1}) decreased by 0.13, 0.14, 0.08 and 0.04, and the metrics of CCC increased by 0.43, 2.23, 1.75 and 0.40 when comparing the RF and fine-tuned pre-trained DL models, respectively. Similar improvements in accuracy metrics were observed in all other study areas. In general, the percentage improvements are 10.21%, 12.92%, 11.25% and 9.32% in MAE and 41.94%, 32.56%, 36.36% and 5.33% in CCC for Regions 1–4, respectively. The average improvement across all regions is 10.93% (decreased by 1.20 g kg^{-1}) in MAE and

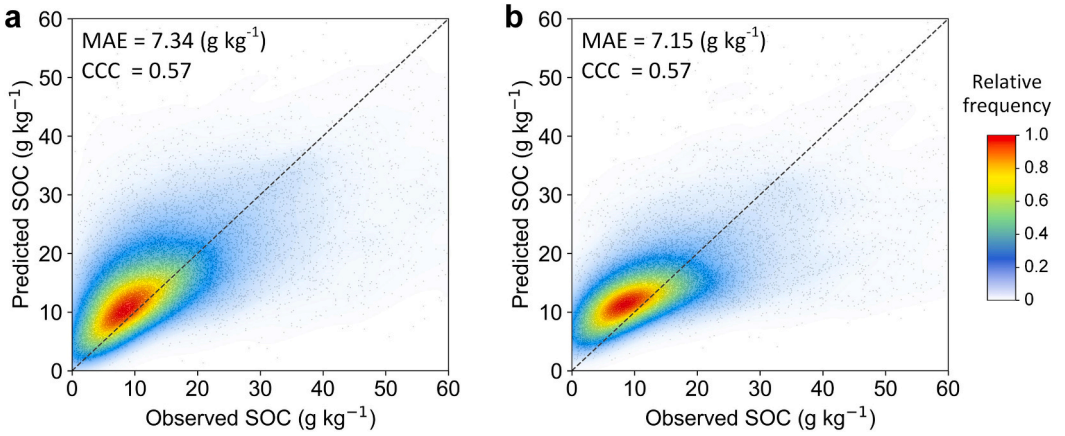


Fig. 4. The validation accuracies of modelling soil organic carbon content (SOC) by using random forest model (a) and the proposed GSoilCPM deep learning model (b) at the global scale. All dots are the observed and predicted SOC values for all validation samples from ten-fold cross-validation.

29.04% (increased by 0.10) in CCC. These results substantiate that, by leveraging the GSoilCPM deep learning model pre-trained based on a vast repository of global SOC data, the model can effectively transfer its nuanced understanding of complex global soil–environment relationships to capture regional soil variations.

When comparing mapping results from SoilGrids, RF models, and pre-trained DL models, similar general pattern of SOC variation across space are observed in each region. However, substantial differences in value ranges and some local details of SOC distributions are observed (Fig. 6). It shows that SoilGrids tend to overestimate SOC in regions with lower SOC (relative to the global mean), and underestimate SOC in regions with higher SOC. For example, across the mountainous areas of Switzerland (Region 2) and arid areas in New South Wales, Australia (Region 4), SoilGrids shows a large bias compared to the prediction maps generated using local sample data (Fig. 6**d-f, j-l**). This bias may result from the RF model's tendency to compress the range of predicted values due to its intrinsic method of averaging multiple predictions from individual trees, smoothing out extremes and reducing variability in the predictions (Nussbaum et al., 2023). This problem is more prominent when comparing the regional and global mapping results, as the regionally extreme low and high SOC values might be compressed to reduce the global-level bias in average.

The mapping results from RF and GSoilCPM (pre-trained DL model) also show notable differences. The SOC maps derived from the pre-trained DL model show more detailed spatial heterogeneity, especially can be found in mountainous and rugged areas across the southern part in Region 1 and 2. This can be partly explained by the fact that the proposed DL model can produce a higher spatial resolution of mapping results. This advantage comes from the GSoilCPM model's ability to incorporate covariate data with different spatial contexts and different original resolutions, whereas traditional models usually require all covariates to be pre-processed to the same resolution. This pre-processing step may lead to the loss of detailed information from local neighborhood pixels (Wadoux, 2019; Yang et al., 2021a; Zhang et al., 2022a), such as high resolution topographic variables. Due to that different factors may influence SOC at different spatial scales (Behrens et al., 2010; Tan et al., 2024), resampling all the covariates into a same fixed resolution may cause mismatches between the affecting scale of environmental factors on SOC. Moreover, the density distribution of SOC values mapped by the pre-trained DL model better matches the sample-level SOC distribution than that from RF model (**Fig. S1**), and the DL model also demonstrated higher validation accuracy. Therefore, the prediction maps of SOC generated by the pre-trained DL models can be considered more reliable. By combining global and regional soil and environmental datasets via transferring a pre-trained DL model, the final prediction maps reconciled the global knowledge of soil–environment relationships extracted from global databases with the region-specific SOC patterns derived from local observations.

3.3. Modelling improvement related with sample size and baseline accuracy

The relationships between model improvement (i.e., the accuracy improvement of the pre-trained DL model over the baseline as represented by the RF model), regional sample size, and baseline accuracy are crucial for understanding how the performance of GSoilCPM changes under different conditions. In our analysis, we observed that applying GSoilCPM in all four regions consistently resulted in better performance on average, compared to the baseline (Fig. 7**a**). More importantly, it shows that the degree of enhancement in SOC prediction particularly became greater as the number of samples in a region decreased (Fig. 7**a**). This trend highlights the advantages of transferability and robustness of the GSoilCPM, particularly when regional sample data are limited. For instance, in Region 2 and 3, models trained with less than half the total sample size reported an over 50% model improvement on average. This trend was less evident in Region 1, where the degree of improvement remained relatively consistent across different sample sizes. This is mainly because Region 1 has the largest number of sample data compared to the other three regions, leading to less contrast in model performance when using samples sizes greater than ten percentage of the total. In addition, it also indicates that a larger number of available regional samples may sometimes continuously contribute to further improvement in DL model fitting.

The observed improvements are also closely linked to baseline accuracy — specifically, the prediction accuracy derived from the RF model. In cases where baseline accuracy was lower, the degree of prediction enhancement when utilizing GSoilCPM tended to be higher (Fig. 7**b**). Conversely, in areas with a higher baseline accuracy, the relative improvements were more modest. This phenomenon was found in all four study areas. Interestingly, the slope of this negative relationship was more pronounced when the mean prediction accuracy of the baseline RF model was low, such as the steep slope shown in Region 2 and 3 when the CCC of RF model result was below 0.3 (Fig. 7**b**). Generally, these results demonstrate that as the number of soil samples decreases, incorporating a pre-trained DL model, such as GSoilCPM proposed in this study, not only enhances SOC prediction accuracy but does so more effectively where the model performance is relatively low when only using regional sample data.

3.4. Modelling from globe to regions: A new paradigm for soil mapping

The modelling strategy employed in this study represents a potential

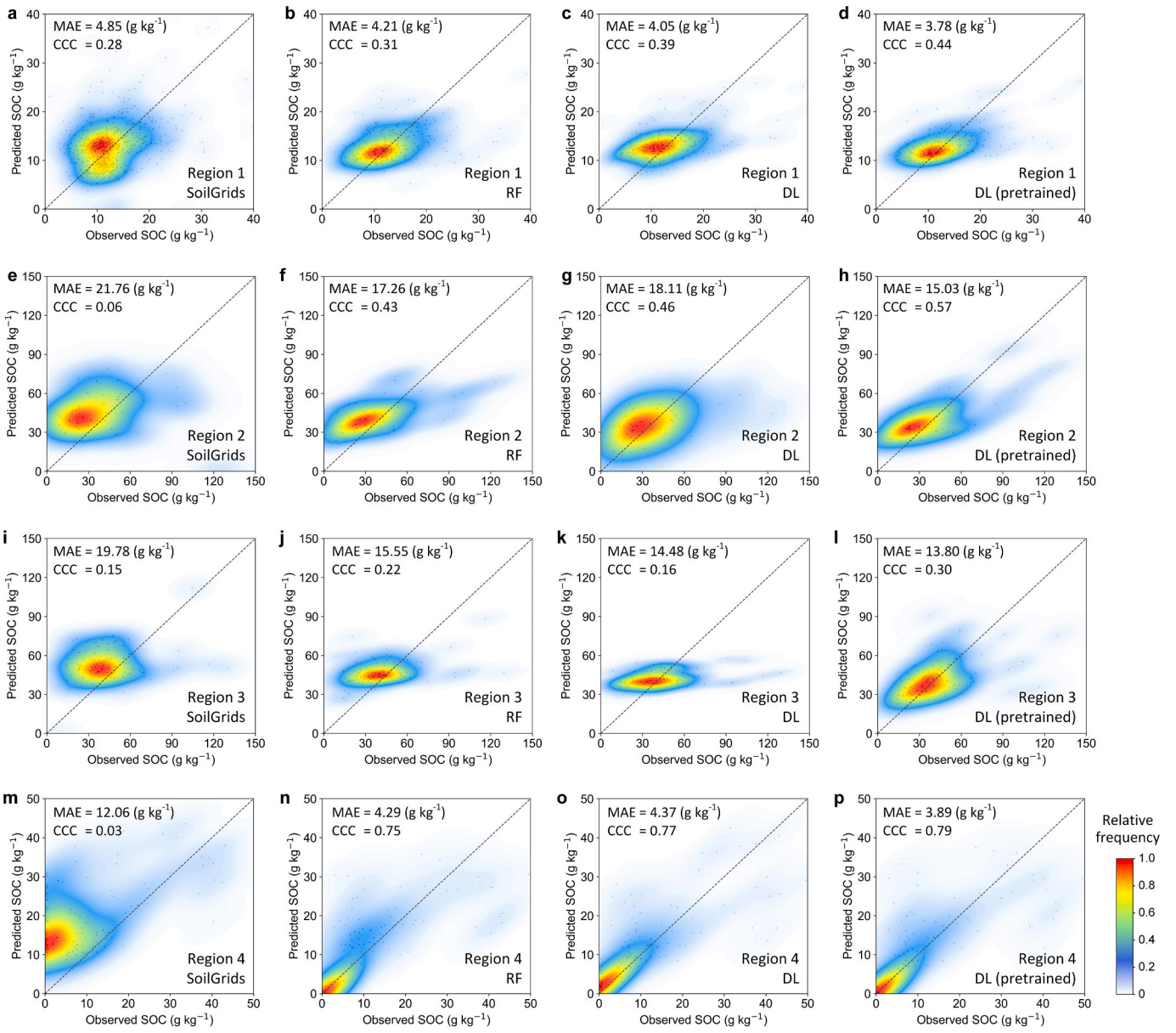


Fig. 5. Comparisons of validation accuracies for predicting soil organic carbon content (SOC) using different modelling strategies across four regional-scale study areas. Each row represents the results for one regional study area. The cross-validation accuracies for SoilGrids (a, e, i, m), random forest (RF) models (b, f, j, n), deep learning (DL) models without pre-training (c, g, k, o), and pre-trained DL model (i.e., GSoilCPM) (d, h, l, p) are shown in the first, second, third and fourth columns, respectively.

paradigm shift in soil mapping, wherein global-scale pre-training a DL model can serve as a domain adaptation model for more accurate regional-scale soil predictions. By generating the interconnectedness of SOC and environmental covariates on a global scale, we can establish frameworks that facilitate the regional application of soil mapping via the GSoilCPM architecture. Traditionally, regional soil mapping approaches often rely solely on limited regional datasets, which may introduce biases and limit the generalizability of models when regional sample data are insufficient. Conversely, the modelling approach based on GSoilCPM leverages soil–environment relationships obtained from the vast amount of globally available soil and environmental data, providing a helpful foundation for regional fine-tuning.

The proposed 'global-to-regional' modeling strategy derives its effectiveness from the theoretical plausibility of the following basic concepts. First, soil variations are fundamentally shaped by long-term interactions with climatic, topographic, and ecological factors that exhibit consistent patterns across scales. These broad-scale soil–environment relationships, often governed by macro geographical and ecological regularities, can be learned from globally distributed data and serve as transferable knowledge. Second, deep learning models are well-suited for capturing such abstract and hierarchical patterns, enabling them to inductively generalize from diverse environmental contexts. Thus, the integration of transferable global knowledge with region-specific fine-tuning offers a new solution for enhancing regional DSM.

Furthermore, the search process for the optimal hypothesis space (i.e., a set of all possible functions or models that the learning algorithm can produce, and each function in this space represents a potential solution to the problem, such as the problem of SOC prediction) regarding soil–environment relationships can be markedly improved when using GSoilCPM. Fig. 8 conceptually illustrates why GSoilCPM offers an effective modelling strategy for regional soil predictions. In traditional regional modelling, machine learning algorithms search for an optimal predictive function within a hypothesis space using only local data. This

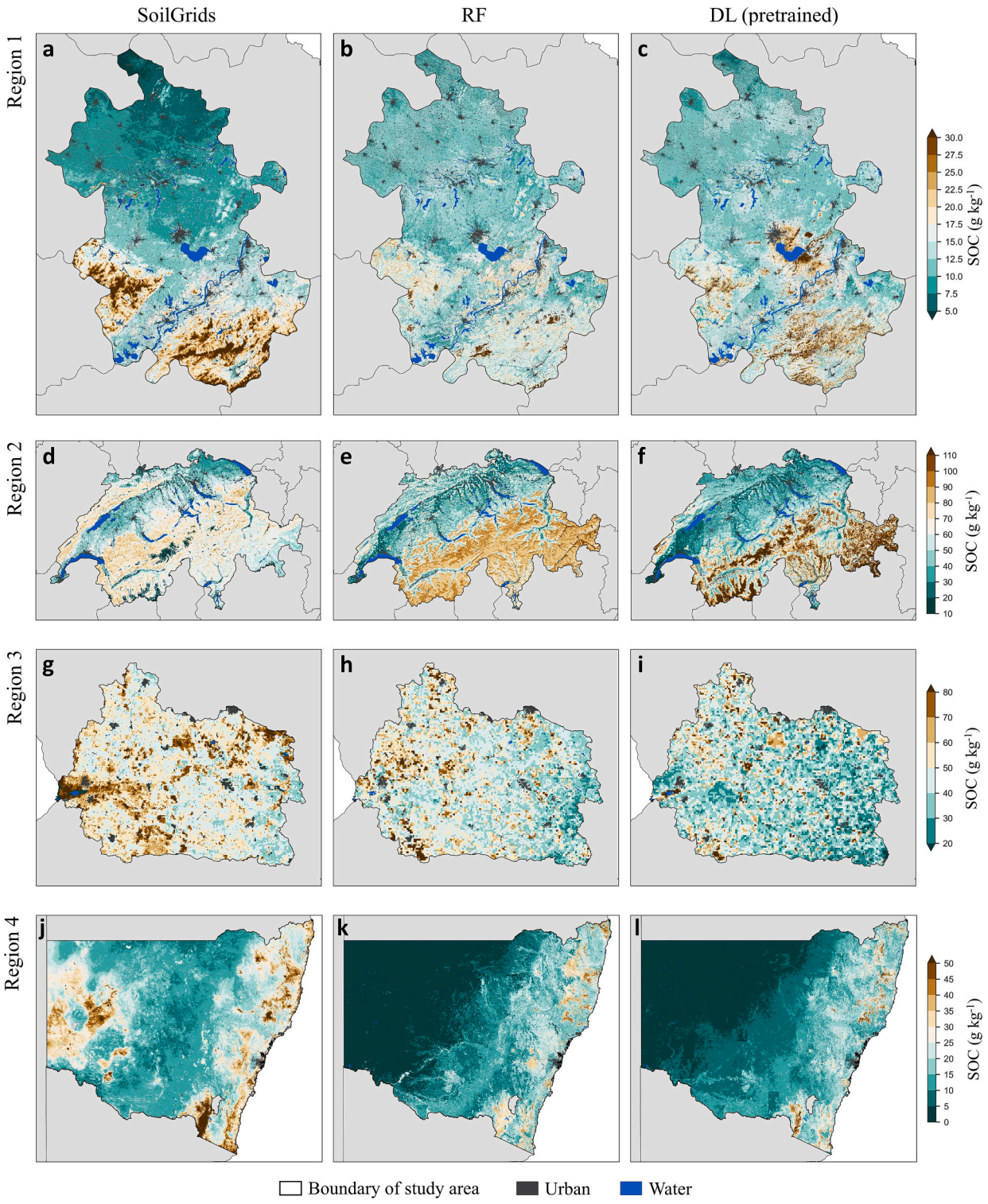


Fig. 6. Mapping results of predicted soil organic carbon content (SOC) derived from SoilGrids (a, d, g, j), random forest (RF) model (b, e, h, k), and the proposed pre-trained deep learning (DL) model (c, f, i, l). Each row shows the comparison of SOC maps in a regional area. The spatial resolution of RF and DL model derived maps is 90 m, compared to the 250 m resolution for SoilGrids product.

process is prone to overfitting or convergence to suboptimal local minima, particularly when observations are sparse. By contrast, the GSoilCPM framework leverages a globally pre-trained model that encodes generalized soil–environment relationships. This global knowledge effectively constrains the hypothesis space, guiding the regional model initialization closer to the optimal solution. As a result, the subsequent fine-tuning process on regional data becomes more stable, sample-efficient, and less susceptible to possible poor convergence. This global-to-regional transfer learning approach provides a plausible way to reduce optimization bias in regional-scale soil mapping.

The advantage in adopting this modelling strategy also stems from the fact that the environmental conditions in a certain region often overlap with those collected from the global database. Using principal component analysis (PCA) to transform the soil and environmental

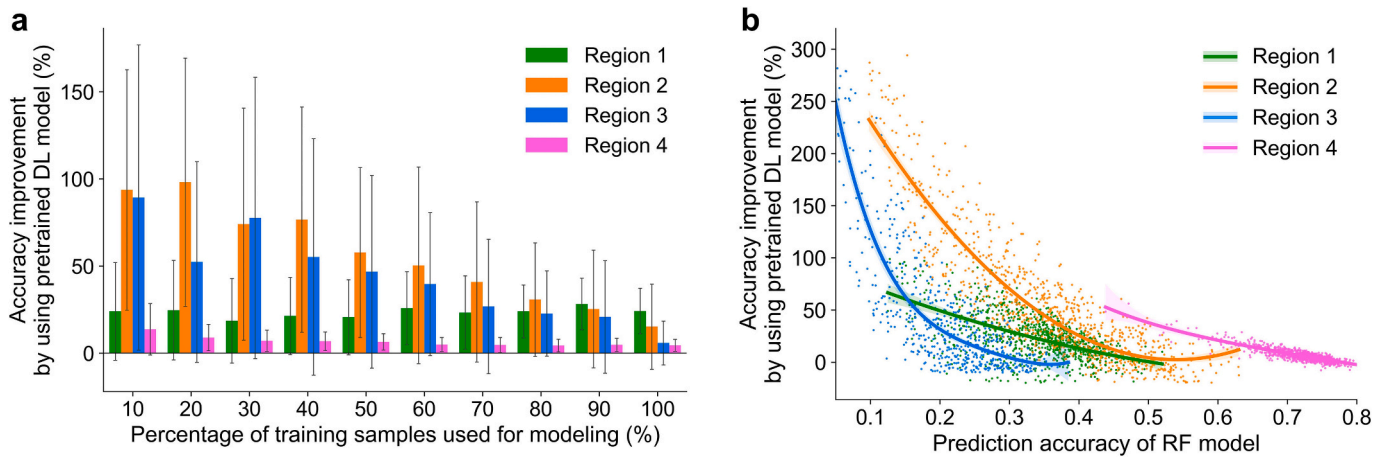


Fig. 7. Accuracy improvement with varying sample sizes (a) and baseline accuracies (b). The accuracy improvement is calculated as the percentage increase by using the pre-trained deep learning model (GSoilCPM) compared to random forest (RF) model. The accuracy (using the metrics of CCC here) of RF model is adopted as the baseline accuracy. Error bars show 90% percentile intervals of accuracy improvement for all modelling runs for each sample size.

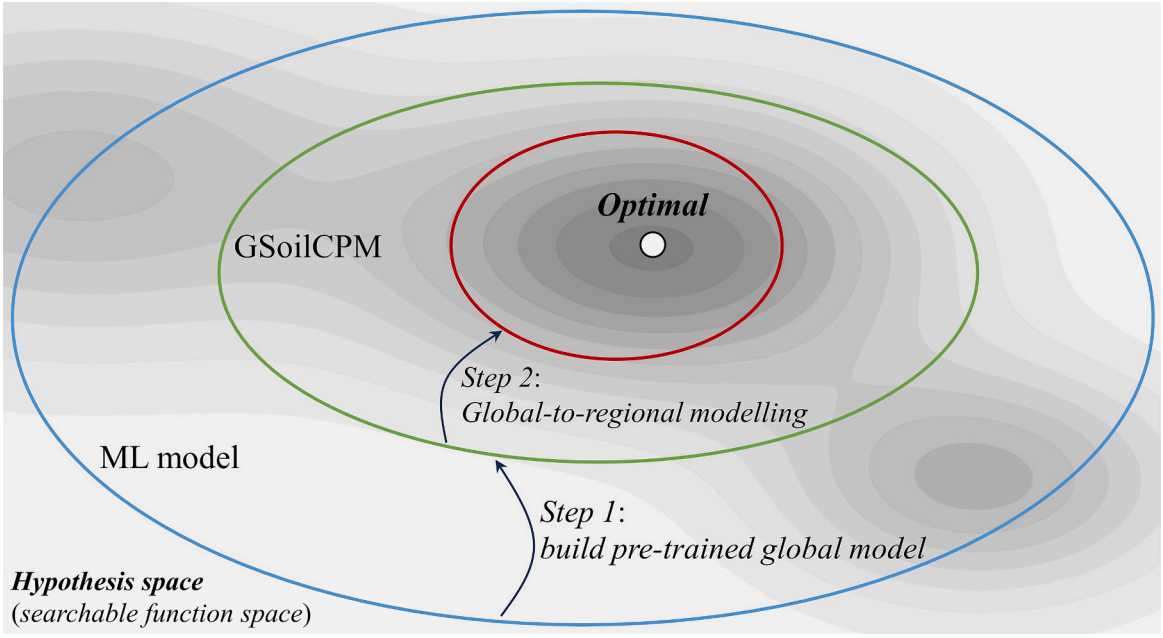


Fig. 8. Why GSoilCPM offers an effective modelling strategy for regional soil predictions. The background color gradient represents the variation of model accuracy in the hypothesis space (a searchable space of the model function). The optimal location indicates the best model function. Building a pre-trained global model to capture the general soil–environment relationships can help narrow the scope of searching regional-scale predictive function. The global-to-regional modelling strategy can reduce the risk of falling into local minima when fitting a machine learning (ML) model based on region data alone.

covariate data into a two-dimensional feature space (Fig. S2) shows that all four regions overlap with the global data, although the extent and location of these overlaps vary. This suggests that some of the knowledge or relationships generated from global data can complement regional-scale SOC modelling, thereby reducing the difficulty in establishing soil–environment relationships given limited samples at regional scales. As an illustration shown in Fig. 9, by initiating the search from a hypothesis space generated by a pre-trained model, optimization algorithms can converge on the best-fitting solution more rapidly and effectively. This pre-defined space incorporates knowledge accumulated from global datasets, providing a robust foundation for exploring model parameters in a local area. It can help the model optimization start from a pre-trained global model, which narrows down the initial hypothesis space and makes it easier to find a good fit for the regional data. In contrast, searches originating from a random starting point in the universal space often yield longer search times and may struggle to reach the optimal result due to the vast number of potential configurations involved. A very large initial hypothesis space might lead to higher possibility of overfitting, where the model may learn the noise in the limited training data rather than the underlying patterns.

It is also crucial to recognize the challenges associated with directly relying on global soil map products such as SoilGrids for regional-scale studies. Our results indicate that the predicted global soil map products may exhibit considerable biases in regions where local sample data were not included in the model training process, despite it has been validated to have strong global-level model performance. Promisingly, GSoilCPM is designed to overcome these biases by effectively integrating local data alongside global observations.

3.5. *Limitations and future horizons*

The current implementation of GSoilCPM primarily focuses on

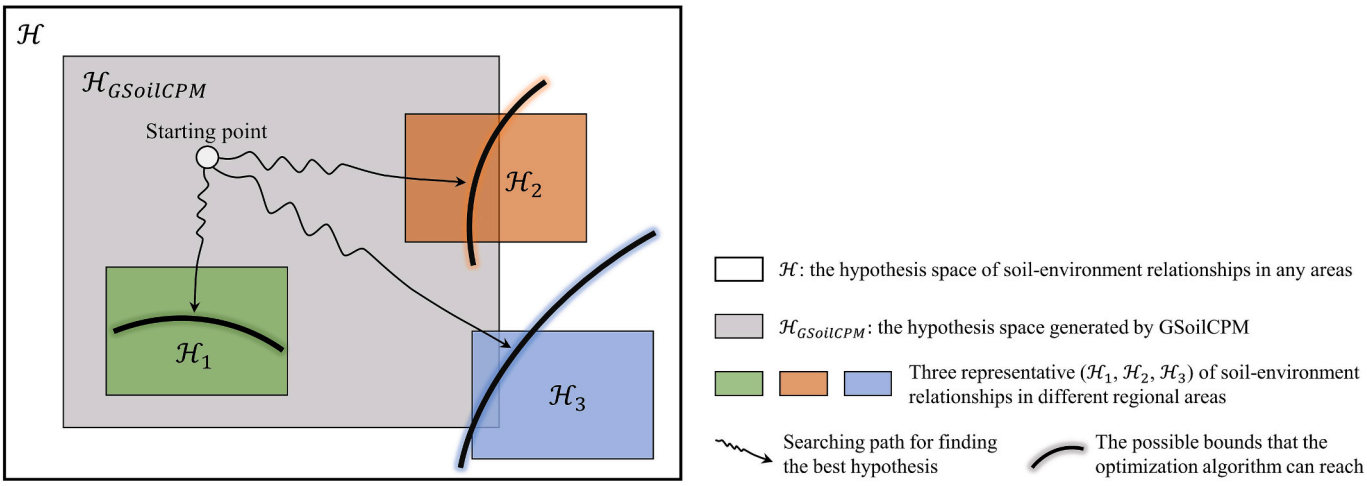


Fig. 9. An illustration explaining the model optimization when taking advantages of GSoilCPM to search the optimal hypothesis ($\mathscr{H}$) of regional soil–environment relationships.

predicting a single soil property, specifically SOC in the topsoil, which is widely concerned for soil health and carbon sequestration. While this model has not been applied to predict other soil properties, our benchmark test suggests a large potential of adopting this modelling concept for enhancing regional predictions of all soil properties. This prompts further investigation to ascertain whether similar modelling techniques can generalize effectively to other soil mapping tasks. Additionally, how to modify the model to take account for the variations of soil properties across different soil depths is an open question. Developing a whole-profile pre-trained model certainly needs further explored in future work. Furthermore, the next version of the GSoilCPM framework could be expanded by adding a modular sub-network designed to handle region-specific inputs (e.g., high-resolution agricultural management data), which can be integrated during the regional fine-tuning stage when reliable local data are available.

While the statistical average of model accuracy show improvement across all four study areas, the results reveal a high degree of variability. The large variance in the degree of model improvement suggest that enhancements cannot be consistently attained across all modelling instances. Some regions or data availability conditions could exhibit less pronounced gains. This variability necessitates caution when transferring models and suggests that further studies are needed to identify situations where the pre-trained model may underperform.

There are promising horizons for future research and development related to the GSoilCPM model. One key direction involves expanding the model to predict multiple soil properties simultaneously. Implementing a multi-task learning framework could allow us to generalize its findings across various soil attributes and depths. Moreover, ongoing investigations into optimizing the model performance under different regional conditions should continue. This includes determining the optimal sizes and resolutions of model inputs, and expanding the model structure to allow the inclusion of specific local dataset, such as the extra inputs reflecting human activities for a region. Maximizing the model applicability from these perspectives will be a promising development direction to improve GSoilCPM's predictability across diverse contexts.

4. Conclusion

This study presents GSoilCPM, an innovative deep learning framework designed to advance the prediction and mapping accuracy of soil organic carbon at regional scales. By integrating global soil datasets with environmental covariates derived from satellite-based observations, our proposed modelling strategy effectively addresses a critical knowledge gap on how to use the global-scale soil–environment relationships learned from global databases to enhance the regional-scale soil predictions with limited regional soil data. The results in four study areas prove that GSoilCPM, when fine-tuned for regional contexts, significantly outperforms models trained exclusively on regional data. The relationships between model improvement, regional sample size, and baseline accuracy that without using pre-trained model, revealed in our study, are crucial for understanding how the performance of GSoilCPM changes under different conditions. Our analyses also indicate that, regions with fewer soil samples and/or lower baseline accuracy can benefit more from using the pre-trained global-scale model, underscoring the importance of leveraging global-scale observations to overcome local data limitations.

In summary, the insights gained from this study are valuable for better informing the next generation of soil modelling and mapping methodology. The advent of the modelling approach using GSoilCPM signifies a transformative advancement in soil mapping, particularly in enhancing regional soil predictions by bridging global- and regional-scale soil–environment relationships. The breakthroughs offered by the 'global-to-regional' transfer deep learning model extend beyond mere predictive enhancements, it represents a potential new paradigm in the methodology of soil mapping itself. By embracing a model that synergizes local insights with global knowledge, scientists and land use managers can achieve a data- and model-reusable approach to understand soil variations more accurate. Such integration not only optimizes existing information but also drastically improves the modelling efficiency in regional soil predictions, positioning GSoilCPM as a new pivotal tool in advancing the field of soil monitoring and land management.

CRediT authorship contribution statement

Lei Zhang: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Lin Yang:** Writing – review & editing, Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization. **Yuxin Ma:** Writing – review & editing, Validation, Resources, Investigation, Data curation. **A-Xing Zhu:** Writing – review & editing, Resources. **Ren Wei:** Validation, Software, Data curation. **Jie Liu:** Validation, Software. **Mogens H. Greve:** Writing – review & editing, Resources, Investigation, Data curation. **Chenghu Zhou:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors 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 was supported by the National Key Research and Development Program Plan (Grant No. 2022YFC3800802), the National Natural Science Foundation of China (Grant No. 42471468), the Leading Funds for the First-Class Universities (020914912203 and 020914902302). Jie Liu was supported by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No. KYCX25_0209). Lei Zhang acknowledges the support from the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, under Award Number DE-AC02-05CH11231.

Appendix A. . More details of GSoilCPM

Appendix A. 1. Convolutional neural networks in GSoilCPM

The convolutional neural networks (CNNs) are adopted in the proposed GSoilCPM framework for extracting latent features from spatial information of environmental covariates. CNNs are designed to handle grid-like data such as raster images. The convolutional layer is the most important building block of a CNN. The convolution operation applies a filter (kernel) over the input. For example, if there is a 2-D data used as the input, a 2-D kernel can be adopted to detect features of the input data (Fig. A1**a**). The kernel is a set of weights, which allows to detect features in the input data. Each unit in the output is a weighted sum of a local patch of input values, defined as:

$$z_{i,j} = (x^{*}k)_{i,j} = \sum_{u=1}^{h}\sum_{v=1}^{w} x_{i',j'} \bullet w_{u,v} \text{with} \begin{cases} i' = u \bullet s_h + h - 1 \\ j' = v \bullet s_w + w - 1 \end{cases} \tag{1}$$

where $z_{i,j}$ is the output of the unit located in row i, column j in feature map of a convolutional layer; h and w are the height and width of the kernel (receptive field); s_h and s_w are the vertical and horizontal strides; $x_{i',j'}$ is the unit located at row i', column j' in the input; $w_{u,v}$ is the weight between any unit in feature map and its input located at row u, column v (relative to the unit's receptive field). The local weighted sum is then passed through a non-linear transfer function. The outputted units in a convolutional layer are organized and are referred to as the feature map. All units in a feature map share the same kernel. This shared-weights strategy was used because we consider the local statistics of input images and their signals are invariant to location (Goodfellow et al., 2016; LeCun et al., 2015). Most importantly it means once the CNN has learned to recognize a pattern in one location, it can recognize it in any other location. In practice, multiple kernels are used in each convolutional layer to extract different types of spatial features, making it capable of detecting multiple features simultaneously in the input data.

After the convolution operation, the max pooling operation is a useful further stage. The role of the pooling layer is to merge semantically similar features into one. This operation can reduce the computational load, the memory usage, and the number of parameters (thereby limiting the risk of overfitting). Max pooling slides a window over the feature map and aggregate values into the maximum value from each window (Fig. A1**b**).

In a typical CNN architecture, two or three stages of convolution, non-linearity and pooling are stacked, then followed by more convolutional layers, non-linearity and pooling. In our study, CNNs allow GSoilCPM to learn spatial features from different environmental covariates that influence soil, such as terrain structure, vegetation patchiness, and climatic gradients, directly from raster inputs with different spatial resolutions.

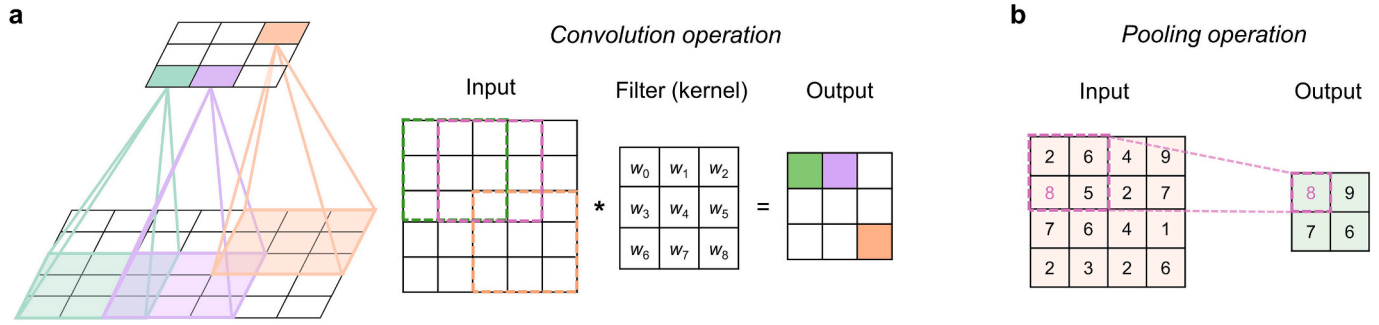


Fig. A1. The convolution layer (a) and the pooling layer (b) as the core building blocks in a convolutional neural network (CNN).

Appendix A. 2. Multi-layer neural network and forward computation

As all feature vectors extracted by CNNs are then concatenated into a single and longer vector, this allows the multi-layer neural network to process this vector for extracting the interactions among these features. A perceptron is a fundamental building block in multi-layer neural network (Fig. A2**a**). It is essentially a single-layer neural network that can compute a weighted sum of its inputs, and then uses an activation function to produce an output:

$$h = \phi\left(w^{T} \bullet x + b\right) \tag{2}$$

where x is the input vector; w provides the weights of a linear transformation and b is the bias; $\phi(\bullet)$ represents an activation function which is often set to be a nonlinear function such as the rectified linear activation function (ReLU); h is the output.

A multi-layer neural network is composed of an input layer, one or more hidden layers, and a final output layer (Fig. A2**b**). All units except a bias in

a layer are fully connected to the previous layer. This multi-layer structure enables learning of abstract, hierarchical representations, essential for generating the complex and nonlinear relationships among multiple environmental covariates influencing soil variations.

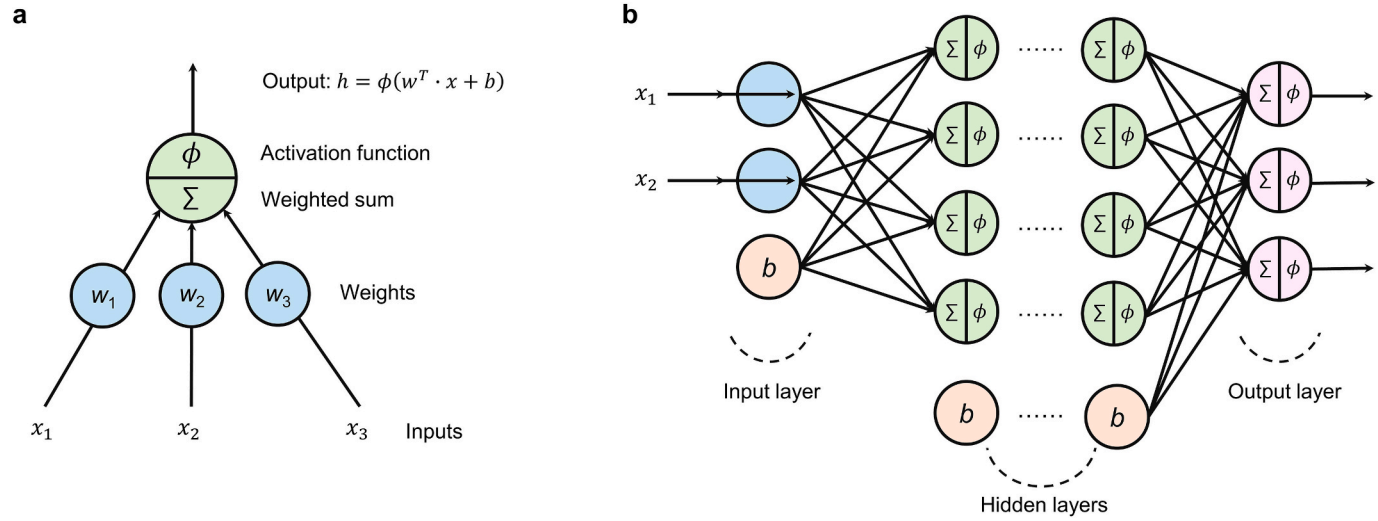


Fig. A2. The perceptron (a) as a fundamental building block in a multi-layer neural network (b).

Appendix A. 3. Backpropagation and gradient-based optimization

Training the deep neural network involves minimizing the error between predicted and observed SOC values using backpropagation and gradient descent. The prediction error can be calculated using the mean squared error which is considered as the loss function $\mathscr{L}$ (described in Section 2.4). The backpropagation procedure to compute the gradient of $\mathscr{L}$ with respect to the weights in GSoilCPM is fundamentally an application of the chain rule for derivatives. The key insight is that the gradient of the loss with respect to the model parameters in a layer can be computed by working backwards from the gradient with respect to the output of that layer. According to Taylor's theorem, for a differentiable function such as $\mathscr{L}(\theta)$, a small perturbation s leads to:

$$\mathscr{L}(\theta+s) \approx \mathscr{L}(\theta) + s \bullet \nabla\mathscr{L}(\theta) \tag{3}$$

where $\nabla\mathscr{L}$ is the gradient of $\mathscr{L}$. This motivates the use of gradients in minimizing $\mathscr{L}$. In neural networks, the chain rule of derivatives allows backpropagation equation can be applied repeatedly to propagate gradients through all layers. In gradient descent, the first order gradient is used, and we set:

$$s = -\eta \bullet \nabla\mathscr{L}(\theta) \tag{4}$$

where η ($\eta > 0$) is called the "step size" or "learning rate", which is usually set to be small to ensure $\mathscr{L}(\theta+s) \leq \mathscr{L}(\theta)$. The direction of the negative gradient moves the function's output toward the local minimum. Therefore, the model parameters θ can be updated by:

$$\theta \leftarrow \theta - \eta \bullet \frac{\partial\mathscr{L}}{\partial\theta} \tag{5}$$

There are some faster and more stable convergence than regular gradient descent optimizer. In this study, we adopted Adam optimizer (Kingma and Ba, 2017), which adjusts the learning rate adaptively during training.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geoderma.2025.117466.

Data availability

All data used in the analysis are publicly accessible. The WoSIS soil profile data were obtained from a published snapshot (September 2019) archived at https://doi.org/10.17027/isric-wdcsoils.20190901. The SoilGrids (version 2.0) database can be accessed at https://soilgrids.org. The climate data are available in the WorldClim version 2.0 database (https://worldclim.org). The global lithological map database is available at https://doi.org/10.1594/PANGAEA.788537. The global digital elevation model (DEM) data are available at https://doi.org/10.5066/F7DF6PQS. The global dataset comprising of multiple topographic features (Geomorpho90m) is obtained from https://doi.org/10.5069/G91R6NPX. The MODIS BRDF/Albedo and vegetation indices products are available at https://ladsweb.modaps.eosdis.nasa.gov.

The codes for processing data have been deposited into a repository at https://github.com/leizhang-geo/GSoilCPM.git. The pre-trained model (GSoilCPM) is also stored in this repository.

References

Amatulli, G., McInerney, D., Sethi, T., Strobl, P., Domisch, S., 2020. Geomorpho90m, empirical evaluation and accuracy assessment of global high-resolution geomorphometric layers. Sci. Data 7, 162. https://doi.org/10.1038/s41597-020-0479-6.

Awais, M., Naseer, M., Khan, S., Anwer, R.M., Cholakkal, H., Shah, M., Yang, M.-H., Khan, F.S., 2023. Foundational Models Defining a New Era in Vision: A Survey and Outlook. https://doi.org/10.48550/arXiv.2307.13721.

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. https://doi.org/10.5194/essd-16-4735-2024.

Batjes, N.H., Ribeiro, E., Van Oostrum, A., 2020. Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019). Earth Syst. Sci. Data 12, 299–320. https://doi.org/10.5194/essd-12-299-2020.

Batjes, N.H., Ribeiro, E., van Oostrum, A., Leenaars, J., Hengl, T., Mendes de Jesus, J., 2017. WoSIS: providing standardised soil profile data for the world. Earth Syst. Sci. Data 9, 1–14. https://doi.org/10.5194/essd-9-1-2017.

Behrens, T., Zhu, A.-X., Schmidt, K., Scholten, T., 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma 155, 175–185. https://doi.org/10.1016/j.geoderma.2009.07.010.

Bishop, T.F.A., McBratney, A.B., Laslett, G.M., 1999. Modelling soil attribute depth functions with equal-area quadratic smoothing splines. Geoderma 91, 27–45. https://doi.org/10.1016/S0016-7061(99)00003-8.

Bliss, N.B., Waltman, S.W., Petersen, G.W., 1995. Preparing a soil carbon inventory for the United States using geographic information systems, in: Soils and Global Change. CRC Press, Boca Raton, FL, pp. 275–295.

Brevik, E.C., Pereira, P., Muñoz-Rojas, M., Miller, B.A., Cerdà, A., Parras-Alcántara, L., Lozano-García, B., 2017. Chapter 1 - Historical Perspectives on Soil Mapping and Process Modeling for Sustainable Land Use Management, in: Pereira, P., Brevik, E.C., Muñoz-Rojas, M., Miller, B.A. (Eds.), Soil Mapping and Process Modeling for Sustainable Land Use Management. Elsevier, pp. 3–28. https://doi.org/10.1016/B978-0-12-805200-6.00001-3.

Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Edwards, T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239, 68–83. https://doi.org/10.1016/j.geoderma.2014.09.019.

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. https://doi.org/10.1016/j.geoderma.2021.115567.

Cui, W., Yang, L., Zhang, L., Yang, C., Zhu, C., Zhou, C., 2025. A novel approach of generating pseudo revisited soil sample data based on environmental similarity for space-time soil organic carbon modelling. Int. J. Appl. Earth Obs. Geoinf. 139, 104542. https://doi.org/10.1016/j.jag.2025.104542.

Dai, Y., Shangguan, W., Wei, N., Xin, Q., Yuan, H., Zhang, S., Liu, S., Lu, X., Wang, D., Yan, F., 2019. A review of the global soil property maps for Earth system models. Soil 5, 137–158. https://doi.org/10.5194/soil-5-137-2019.

Didan, K., 2021. MODIS/Terra Vegetation Indices 16-Day L3 Global 500m SIN Grid V061. https://doi.org/10.5067/MODIS/MOD13A1.061.

Dinerstein, E., Olson, D., Joshi, A., Vynne, C., Burgess, N.D., Wikramanayake, E., Hahn, N., Palminteri, S., Hedao, P., Noss, R., Hansen, M., Locke, H., Ellis, E.C., Jones, B., Barber, C.V., Hayes, R., Kormos, C., Martin, V., Crist, E., Sechrest, W., Price, L., Baillie, J.E.M., Weeden, D., Suckling, K., Davis, C., Sizer, N., Moore, R., Thau, D., Birch, T., Potapov, P., Turubanova, S., Tyukavina, A., de Souza, N., Pintea, L., Brito, J.C., Llewellyn, O.A., Miller, A.G., Patzelt, A., Ghazanfar, S.A., Timberlake, J., Klöser, H., Shennan-Farpón, Y., Kindt, R., Lillesø, J.-P.-B., van Breugel, P., Graudal, L., Voge, M., Al-Shammari, K.F., Saleem, M., 2017. An Ecoregion-based Approach to Protecting half the Terrestrial Realm. Bioscience 67, 534–545. https://doi.org/10.1093/biosci/bix014.

FAO, 2012. Harmonized World Soil Database v1.2 [WWW Document]. URL http://webarchive.iiasa.ac.at/Research/LUC/External-World-soil-database/HTML/ (accessed 6.17.22).

Farahani, A., Voghoei, S., Rasheed, K., Arabnia, H.R., 2021. A Brief Review of Domain Adaptation. In: Stahlbock, R., Weiss, G.M., Abou-Nasr, M., Yang, C.-.-Y., Arabnia, H. R., Deligiannidis, L. (Eds.), Advances in Data Science and Information Engineering. Springer International Publishing, Cham, pp. 877–894. https://doi.org/10.1007/978-3-030-71704-9_65.

Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37, 4302–4315. https://doi.org/10.1002/joc.5086.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V., 2016. Domain-Adversarial Training of Neural Networks. J. Mach. Learn. Res. 17, 1–35.

Goodfellow, I., Courville, A., Bengio, Y., 2016. Deep learning. The MIT Press, Cambridge, Massachusetts, Adaptive computation and machine learning.

Guo, M., Yang, L., Zhang, L., Shen, F., Meadows, M.E., Zhou, C., 2025. Hydrology, vegetation, and soil properties as key drivers of soil organic carbon in coastal wetlands: a high-resolution study. Environ. Sci. Ecotechnol. 23, 100482. https://doi.org/10.1016/j.ese.2024.100482.

Hartmann, J., Moosdorf, N., 2012. The new global lithological map database GLiM: a representation of rock properties at the Earth surface. Geochem. Geophys. Geosyst. 13. https://doi.org/10.1029/2012GC004370.

He, X., Yang, L., Li, A., Zhang, L., Shen, F., Cai, Y., Zhou, C., 2021. Soil organic carbon prediction using phenological parameters and remote sensing variables generated from Sentinel-2 images. Catena 205, 105442. https://doi.org/10.1016/j.catena.2021.105442.

Helfenstein, A., Mulder, V.L., Heuvelink, G.B.M., Hack-ten Broeke, M.J.D., 2024. Three-dimensional space and time mapping reveals soil organic matter decreases across anthropogenic landscapes in the Netherlands. Commun. Earth Environ. 5, 1–16. https://doi.org/10.1038/s43247-024-01293-y.

Hengl, T., Heuvelink, G.B.M., Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120, 75–93. https://doi.org/10.1016/j.geoderma.2003.08.018.

Hengl, T., de Jesus, J.M., Heuvelink, G.B.M., Gonzalez, M.R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS One 12, e0169748. https://doi.org/10.1371/journal.pone.0169748.

Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.R., 2014. SoilGrids1km — Global Soil Information based on Automated Mapping. PLoS One 9, e105992. https://doi.org/10.1371/journal.pone.0105992.

Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265, 62–77. https://doi.org/10.1016/j.geoderma.2015.11.014.

Heuvelink, G.B.M., Kros, J., Reinds, G.J., De Vries, W., 2016. Geostatistical prediction and simulation of European soil property maps. Geoderma Reg. 7, 201–215. https://doi.org/10.1016/j.geodrs.2016.04.002.

Heuvelink, G.B.M., Webster, R., 2001. Modelling soil variation: past, present, and future. Geoderma, Developments and Trends in Soil Science 100, 269–301. https://doi.org/10.1016/S0016-7061(01)00025-8.

Hong, D., Zhang, B., Li, X., Li, Y., Li, C., Yao, J., Yokoya, N., Li, H., Ghamisi, P., Jia, X., Plaza, A., Gamba, P., Benediktsson, J.A., Chanussot, J., 2024. SpectralGPT: Spectral Remote Sensing Foundation Model. IEEE Trans. Pattern Anal. Mach. Intell. 46, 5227–5244. https://doi.org/10.1109/TPAMI.2024.3362475.

Huang, H., Yang, L., Zhang, L., Pu, Y., Yang, C., Wu, Q., Cai, Y., Shen, F., Zhou, C., 2022. A review on digital mapping of soil carbon in cropland: progress, challenge, and prospect. Environ. Res. Lett. 17, 123004. https://doi.org/10.1088/1748-9326/aca41e.

Hudson, B.D., 1992. The Soil Survey as Paradigm-based Science. Soil Sci. Soc. Am. J. 56, 836–841. https://doi.org/10.2136/sssaj1992.03615995005600030027x.

Ivushkin, K., Bartholomeus, H., Bregt, A.K., Pulatov, A., Kempen, B., De Sousa, L., 2019. Global mapping of soil salinity change. Remote Sens. Environ. 231, 111260. https://doi.org/10.1016/j.rse.2019.111260.

Jenny, H., 1941. Factors of Soil Formation: a System of Quantitative Pedology. Agron. J. 33, 857–858. https://doi.org/10.2134/agronj1941.00021962003300090016x.

Jensen, K.H., Illangasekare, T.H., 2011. HOBE: a Hydrological Observatory. Vadose Zone J. 10, 1–7. https://doi.org/10.2136/vzj2011.0006.

Kingma, D.P., Ba, J., 2017. Adam: A Method for Stochastic Optimization. https://doi.org/10.48550/arXiv.1412.6980.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539.

Lilburne, L., Helfenstein, A., Heuvelink, G.B.M., Eger, A., 2024. Interpreting and evaluating digital soil mapping prediction uncertainty: a case study using texture from SoilGrids. Geoderma 450, 117052. https://doi.org/10.1016/j.geoderma.2024.117052.

Liu, F., Wu, H., Zhao, Y., Li, D., Yang, J.-L., Song, X., Shi, Z., Zhu, A.-X., Zhang, G.-L., 2022. Mapping high resolution National Soil Information Grids of China. Science Bulletin 67, 328–340. https://doi.org/10.1016/j.scib.2021.10.013.

Ma, Y., Minasny, B., Malone, B.P., Mcbratney, A.B., 2019. Pedology and digital soil mapping (DSM). Eur. J. Soil Sci. 70, 216–235. https://doi.org/10.1111/ejss.12790.

Ma, Y., Minasny, B., Wu, C., 2017. Mapping key soil properties to support agricultural production in Eastern China. Geoderma Reg. 10, 144–153. https://doi.org/10.1016/j.geodrs.2017.06.002.

Malone, B.P., McBratney, A.B., Minasny, B., Laslett, G.M., 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma 154, 138–152. https://doi.org/10.1016/j.geoderma.2009.10.007.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4.

Meng, X., Bao, Y., Luo, C., Zhang, X., Liu, H., 2024a. SOC content of global Mollisols at a 30 m spatial resolution from 1984 to 2021 generated by the novel ML-CNN prediction model. Remote Sens. Environ. 300, 113911. https://doi.org/10.1016/j.rse.2023.113911.

Meng, X., Bao, Y., Luo, C., Zhang, X., Liu, H., 2024b. A new methodology for establishing an SOC content prediction model that is spatiotemporally transferable at multidecadal and intercontinental scales. ISPRS J. Photogramm. Remote Sens. 218, 531–550. https://doi.org/10.1016/j.isprsjprs.2024.09.038.

Meng, X., Bao, Y., Wang, Y., Zhang, X., Liu, H., 2022. An advanced soil organic carbon content prediction model via fused temporal-spatial-spectral (TSS) information based on machine learning and deep learning algorithms. Remote Sens. Environ. 280, 113166. https://doi.org/10.1016/j.rse.2022.113166.

Meyer, H., Pebesma, E., 2021. Predicting into unknown space? estimating the area of applicability of spatial prediction models. Methods Ecol. Evol. 12, 1620–1633. https://doi.org/10.1111/2041-210X.13650.

Minasny, B., McBratney, A.B., Malone, B.P., Wheeler, I., 2013. Digital Mapping of Soil Carbon, in: Advances in Agronomy. Elsevier, pp. 1–47. https://doi.org/10.1016/B978-0-12-405942-9.00001-3.

Minasny, B., McBratney, Alex.B., 2016. Digital soil mapping: A brief history and some lessons. Geoderma, Soil mapping, classification, and modelling: history and future directions 264, 301–311. https://doi.org/10.1016/j.geoderma.2015.07.017.
Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P., 2023. Foundation models for generalist medical artificial intelligence. Nature 616, 259–265. https://doi.org/10.1038/s41586-023-05881-4.
Mulder, V.L., de Bruin, S., Schaepman, M.E., Mayr, T.R., 2011. The use of remote sensing in soil and terrain mapping — a review. Geoderma 162, 1–19. https://doi.org/10.1016/j.geoderma.2010.12.018.
Nussbaum, M., Vogel, S., Oechslin, S., Tanner, S., Burgos, S., 2023. Smoothed predicted distributions in digital soil mapping – a comprehensive comparative study to predict soil texture for irrigation (No. EGU23-5543). Presented at the EGU23, Copernicus Meetings. https://doi.org/10.5194/egusphere-egu23-5543.
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, 140–153. https://doi.org/10.1111/ejss.12499.
Peng, Y., Xiong, X., Adhikari, K., Knadel, M., Grunwald, S., Greve, M.H., 2015. Modeling Soil Organic Carbon at Regional Scale by Combining Multi-Spectral Images with Laboratory Spectra. PLoS One 10, e0142295. https://doi.org/10.1371/journal.pone.0142295.
Poggio, L., de Sousa, L.M., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil 7, 217–240. https://doi.org/10.5194/soil-7-217-2021.
Sanchez, P.A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonça-Santos, M. de L., Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vågen, T.-G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., Zhang, G.-L., 2009. Digital Soil Map of the World. Science 325, 680–681. https://doi.org/10.1126/science.1175084.
Sanderman, J., Hengl, T., Fiske, G.J., 2017. Soil carbon debt of 12,000 years of human land use. Proc. Natl. Acad. Sci. 114, 9575–9580. https://doi.org/10.1073/pnas.1706103114.
Schaaf, C., Wang, Z., 2021. MODIS/Terra+Aqua BRDF/Albedo Nadir BRDF Adjusted Ref Daily L3 Global - 500m V061. https://doi.org/10.5067/MODIS/MCD43A4.061.
Schmidt, M.W.I., Torn, M.S., Abiven, S., Dittmar, T., Guggenberger, G., Janssens, I.A., Kleber, M., Kögel-Knabner, I., Lehmann, J., Manning, D.A.C., Nannipieri, P., Rasse, D.P., Weiner, S., Trumbore, S.E., 2011. Persistence of soil organic matter as an ecosystem property. Nature 478, 49–56. https://doi.org/10.1038/nature10386.
Shen, Z., Ramirez-Lopez, L., Behrens, T., Cui, L., Zhang, M., Walden, L., Wetterlind, J., Shi, Z., Sudduth, K.A., Baumann, P., Song, Y., Catambay, K., Viscarra Rossel, R.A., 2022. Deep transfer learning of global spectra for local soil carbon monitoring. ISPRS J. Photogramm. Remote Sens. 188, 190–200. https://doi.org/10.1016/j.isprsjprs.2022.04.009.
Tan, T., Genova, G., Heuvelink, G.B.M., Lehmann, J., Poggio, L., Woolf, D., You, F., 2024. Importance of Terrain and climate for predicting Soil Organic Carbon is Highly Variable across local to Continental Scales. Environ. Sci. Technol. 58, 11492–11503. https://doi.org/10.1021/acs.est.4c01172.
Tiessen, H., Cuevas, E., Chacon, P., 1994. The role of soil organic matter in sustaining soil fertility. Nature 371, 783–785. https://doi.org/10.1038/371783a0.
Viscarra Rossel, R.A., Minasny, B., Roudier, P., McBratney, A.B., 2006a. Colour space models for soil science. Geoderma 133, 320–337. https://doi.org/10.1016/j.geoderma.2005.07.017.
Viscarra Rossel, R.A., Shen, Z., Ramirez Lopez, L., Behrens, T., Shi, Z., Wetterlind, J., Sudduth, K.A., Stenberg, B., Guerrero, C., Gholizadeh, A., Ben-Dor, E., St Luce, M., Orellano, C., 2024. An imperative for soil spectroscopic modelling is to think global but fit local with transfer learning. Earth Sci. Rev. 254, 104797. https://doi.org/10.1016/j.earscirev.2024.104797.
Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006b. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma 131, 59–75. https://doi.org/10.1016/j.geoderma.2005.03.007.
Wadoux, A.M.J.C., 2019. Using deep learning for multivariate mapping of soil with quantified uncertainty. Geoderma 351, 59–70. https://doi.org/10.1016/j.geoderma.2019.05.012.
Wadoux, A.-M.-J.-C., Heuvelink, G.B.M., Lark, R.M., Lagacherie, P., Bouma, J., Mulder, V.L., Libohova, Z., Yang, L., McBratney, A.B., 2021. Ten challenges for the future of pedometrics. Geoderma 401, 115155. https://doi.org/10.1016/j.geoderma.2021.115155.
Wadoux, A.-M.-J.-C., Minasny, B., Mcbratney, A.B., 2020. Machine learning for digital soil mapping: applications, challenges and suggested solutions. Earth Sci. Rev. 210, 103359. https://doi.org/10.1016/j.earscirev.2020.103359.
Wang, B., Gray, J.M., Waters, C.M., Rajin Anwar, M., Orgill, S.E., Cowie, A.L., Feng, P., Li Liu, D., 2022. Modelling and mapping soil organic carbon stocks under future climate change in south-eastern Australia. Geoderma 405, 115442. https://doi.org/10.1016/j.geoderma.2021.115442.
Wang, X., Li, S., Wang, L., Zheng, M., Wang, Z., Song, K., 2023. Effects of cropland reclamation on soil organic carbon in China's black soil region over the past 35 years. Glob. Chang. Biol. 29, 5460–5477. https://doi.org/10.1111/gcb.16833.
Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., Tang, Y., 2023. A Brief Overview of ChatGPT: the history, Status Quo and potential Future Development. IEEE/CAA J. Autom. Sin. 10, 1122–1136. https://doi.org/10.1109/JAS.2023.123618.
Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J.C., Sampson, C.C., Kanae, S., Bates, P.D., 2017. A high-accuracy map of global terrain elevations. Geophys. Res. Lett. 44, 5844–5853. https://doi.org/10.1002/2017GL072874.
Yang, L., Cai, Y., Zhang, L., Guo, M., Li, A., Zhou, C., 2021a. 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. https://doi.org/10.1016/j.jag.2021.102428.
Yang, L., Li, X., Yang, Q., Zhang, L., Zhang, S., Wu, S., Zhou, C., 2021b. Extracting knowledge from legacy maps to delineate eco-geographical regions. Int. J. Geogr. Inf. Sci. 35, 250–272. https://doi.org/10.1080/13658816.2020.1806284.
Zhang, L., Cai, Y., Huang, H., Li, A., Yang, L., Zhou, C., 2022a. A CNN-LSTM model for soil organic carbon content prediction with long time series of MODIS-based phenological variables. Remote Sens. (Basel) 14, 4441. https://doi.org/10.3390/rs14184441.
Zhang, L., Heuvelink, G.B.M., Mulder, V.L., Chen, S., Deng, X., Yang, L., 2024. Using process-oriented model output to enhance machine learning-based soil organic carbon prediction in space and time. Sci. Total Environ. 922, 170778. https://doi.org/10.1016/j.scitotenv.2024.170778.
Zhang, L., Yang, L., Cai, Y., Huang, H., Shi, J., Zhou, C., 2022b. A multiple soil properties oriented representative sampling strategy for digital soil mapping. Geoderma 406, 115531. https://doi.org/10.1016/j.geoderma.2021.115531.
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., 2025. Mapping global distributions, environmental controls, and uncertainties of apparent topsoil and subsoil organic carbon turnover times. Earth Syst. Sci. Data 17, 2605–2623. https://doi.org/10.5194/essd-17-2605-2025.
Zhang, L., Yang, L., Ma, T., Shen, F., Cai, Y., Zhou, C., 2021. A self-training semi-supervised machine learning method for predictive mapping of soil classes with limited sample data. Geoderma 384, 114809. https://doi.org/10.1016/j.geoderma.2020.114809.
Zhu, A., Lu, G., Liu, J., Qin, C., Zhou, C., 2018. Spatial prediction based on Third Law of Geography. Ann. GIS 24, 225–240. https://doi.org/10.1080/19475683.2018.1534890.
Zhu, A.-X., Hudson, B., Burt, J., Lubich, K., Simonson, D., 2001. Soil Mapping using GIS, Expert Knowledge, and Fuzzy Logic. Soil Sci. Soc. Am. J. 65, 1463–1472. https://doi.org/10.2136/sssaj2001.6551463x.
Zhu, A.X., Liu, J., Du, F., Zhang, S.J., Qin, C.Z., Burt, J., Behrens, T., Scholten, T., 2015. Predictive soil mapping with limited sample data. Eur. J. Soil Sci. 66, 535–547. https://doi.org/10.1111/ejss.12244.