Earth Syst. Sci. Data, 17, 517–543, 2025
https://doi.org/10.5194/essd-17-517-2025

Open Access Earth System Science Data

A China dataset of soil properties for land surface modelling (version 2, CSDLv2)

Gaosong Shi[1], Wenye Sun[1], Wei Shangguan[1], Zhongwang Wei[1], Hua Yuan[1], Lu Li[1], Xiaolin Sun[2], Ye Zhang[1], Hongbin Liang[1], Danxi Li[1], Feini Huang[1], Qingliang Li[1,3], and Yongjiu Dai[1]

[1]Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Guangdong Province Key Laboratory for Climate Change and Natural Disaster Studies, School of Atmospheric Sciences, Sun Yat-sen University, Guangzhou 510275, China
[2]School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China
[3]College of Computer Science and Technology, Changchun Normal University, Changchun 130032, China

Correspondence: Wei Shangguan (shgwei@mail.sysu.edu.cn)

Received: 17 July 2024 – Discussion started: 29 August 2024
Revised: 18 November 2024 – Accepted: 12 December 2024 – Published: 7 February 2025

Abstract. Accurate and high-resolution spatial soil information is crucial for efficient and sustainable land use, management, and conservation. Since the establishment of digital soil mapping (DSM) and the GlobalSoilMap working group, significant advances have been made in terms of the availability and quality of spatial soil information globally. However, accurately predicting soil variation over large and complex areas with limited samples remains a challenge, especially for China, which has diverse soil landscapes. To address this challenge, we utilised 11 209 representative multi-source legacy soil profiles (including the Second National Soil Survey of China, the World Soil Information Service, the First National Soil Survey of China, and regional databases) and high-resolution soil-forming environment characterisation. Using advanced ensemble machine learning and a high-performance parallel-computing strategy, we developed comprehensive maps of 23 soil physical and chemical properties at six standard depth layers from 0 to 2 m in China at a 90 m spatial resolution (China dataset of soil properties for land surface modelling version 2, CSDLv2). Data-splitting and independent-sample validation strategies were employed to evaluate the accuracy of the predicted maps' quality. The results showed that the predicted maps were significantly more accurate and detailed compared to traditional soil type linkage methods (i.e. CSDLv1, the first version of the dataset), SoilGrids 2.0, and HWSD 2.0 products, effectively representing the spatial variation of soil properties across China. The prediction accuracy of soil properties at all depth intervals ranged from good to moderate, with median model efficiency coefficients for most soil properties ranging from 0.29 to 0.70 during data-splitting validation and from 0.25 to 0.84 during independent-sample validation. The wide range between the 5 % lower and 95 % upper prediction limits may indicate substantial room for improvement in current predictions. The relative importance of environmental covariates in predictions varied with soil property and depth, indicating the complexity of interactions among multiple factors in the soil formation processes. As the soil profiles used in this study mainly originate from the Second National Soil Survey of China, conducted during the 1970s and 1980s, they could provide new perspectives on soil changes, together with existing maps based on soil profiles from the 2010s. The findings of this study make important contributions to the GlobalSoilMap project and can also be used for regional Earth system modelling and land surface modelling to better represent the role of soil in hydrological and biogeochemical cycles in China. This dataset is freely available at https://www.scidb.cn/s/ZZJzAz (last access: 17 November 2024) or https://doi.org/10.11888/Terre.tpdc.301235 (Shi and Shangguan, 2024).

1 Introduction

Soil plays a pivotal role in the Earth's systems, facilitating the cycling of water, energy, and carbon across varying temporal and spatial scales. Its significance lies in regulating ecosystems by providing vital nutrients to living organisms; storing and cycling water, heat, carbon, and essential nutrients; and serving as a medium for vegetation growth and structural support (Chaney et al., 2019; Crow et al., 2012). Soil data are essential for land surface models (LSMs), which form a part of Earth system models (ESMs) (Dai et al., 2019b; Luo et al., 2016). The diverse range of soil properties and their precise representation are crucial for robust land surface modelling, influencing various environmental, agricultural, and ecological assessments. There is an urgent need for detailed, accurate, and up-to-date soil information to develop solutions for these challenges and to inform decision-making related to natural resource management (Arrouays et al., 2014a; Dai et al., 2019b; Li et al., 2024).

In recent years, the national and global maps of soil properties have gained significant traction in research (Arrouays et al., 2017), with a surge of studies focusing on mapping one or more soil properties at high resolutions, such as 90 m, spanning various countries. These include large-scale endeavours in Australia (Grundy et al., 2015; Viscarra Rossel et al., 2015), France (Chen et al., 2023; Mulder et al., 2016), Chile (Dinamarca et al., 2023; Padarian et al., 2017), Japan (Yamashita et al., 2024), Netherlands (Helfenstein et al., 2024), and the United States (Ramcharan et al., 2018; Thompson et al., 2020). Chaney et al. (2019) developed 30 m probabilistic maps of soil properties across the United States. Denmark has also developed national maps of soil texture at a finer 30 m resolution (Adhikari et al., 2013). Additionally, broader-scale-resolution maps, ranging from 250 to 5000 m, have also been investigated at the national level, exemplified by Brazil (Gomes et al., 2019). These efforts have been expanded to continental scales, including Africa (Hengl et al., 2015, 2021) and Europe (Heuvelink et al., 2016), and ultimately to global levels, as seen in datasets such as the Global Soil Dataset for use in Earth System Models (GSDE, Shangguan et al., 2014), the Harmonized World Soil Database version 2.0 (HWSD 2.0, FAO and IIASA, 2023), and SoilGrids 2.0 (Poggio et al., 2021).

Shangguan et al. (2013) pioneered the development of a comprehensive dataset of soil characteristics specifically designed for land surface modelling over China (i.e. China Soil Dataset for Land Surface Modelling, CSDLv1, the first version dataset of this study). This dataset, based on 8979 legacy soil profiles and the Soil Map of China (1 : 1 000 000), employs the conventional polygon linkage method (Batjes, 1995, 2002; Shangguan et al., 2012) to develop soil physical and chemical properties. It provides a spatial resolution of 30 arcsec (about 1 km at the Equator) and includes over 20 properties at eight vertical soil depths (Shangguan et al., 2013). The dataset has been successfully applied in various fields. Despite its significant contributions to regional land surface modelling and geoscientific research, over time, several issues and shortcomings have been identified. First, while the dataset utilised soil profiles solely from the Second National Soil Survey of China (1979–1985), there is now a broader array of available soil profiles, including those from the World Soil Information Service (WoSIS, Batjes et al., 2020), regional database (Shangguan et al., 2012), and the First National Soil Survey of China (National Soil Survey Office, 1964). The integration of these soil profiles promises to substantially enhance the spatial representation and coverage of the dataset. Second, this dataset relies on the traditional polygon linkage method based on soil transformation rules (Shangguan et al., 2013, 2014), where results depend heavily on the accuracy of soil classification maps and are estimated as the average of a soil class or polygon, leading to discontinuous spatial estimates. The emergence of digital soil mapping (DSM) techniques (McBratney et al., 2003), particularly the success of machine learning in large-scale spatial prediction (Hengl et al., 2017; Poggio et al., 2021; Yan et al., 2020), presents a methodological advancement for this study. Recent studies indicate that advanced machine learning models often outperform simpler ones, with the size of the sample also emerging as a crucial factor influencing model performance (Padarian et al., 2020).

For China, mapping datasets encompassing one or multiple soil properties have already been developed. Liang et al. (2019) and Chen et al. (2019) both developed high-resolution grid maps across China based on about 5000 legacy soil profiles collected from the Second National Soil Survey of China, providing more detailed information for areas with spatial heterogeneity. However, Liang et al. (2019) focused solely on spatial estimates for soil organic carbon in the topsoil (0–20 cm layer), while Chen et al. (2019) concentrated solely on spatial estimates for soil pH in the same layer. Both studies lack estimations for other soil property variables and deeper soil layers. Approximately 4000 legacy soil profiles were utilised by Zhou et al. (2019a) to develop a high-resolution national-scale dataset for total nitrogen in the topsoil (0–20 cm layer) at a 90 m resolution using machine learning methods. Similarly, Song et al. (2020) used over 5000 soil profiles from the 2010s to produce high-resolution maps of soil organic carbon at six standard depths (0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm) across China, achieving explained variances ranging from 0.16 to 0.57. Besides this, Liu et al. (2022a) also employed machine learning methods to develop China's inaugural high-resolution national soil information grid dataset at a 90 m resolution, utilising soil samples from the most recent National Soil Series of China (2009–2019). This dataset has contributed significantly to soil management, agricultural production, hydrological modelling, ecological development, and climate change mitigation. However, the study relied solely on a constrained set of about 4500 soil profiles collected during the recent national soil survey, generating national grid maps for only some fundamental soil properties, including pH (H_2O), organic carbon, cation exchange capacity, total nitrogen, total phosphorus, total potassium, bulk density, gravel content, soil texture, and soil thickness. The limitations stem from the absence of more comprehensive national grid maps for soil properties, including the fractions of total phosphorus and potassium readily available for plant absorption (available phosphorus, AP; available potassium, AK), an index of the potential capacity of the soil to supply nitrogen (alkali-hydrolysable nitrogen, AN), porosity, and others, imposing constraints on applications that necessitate a broader spectrum of soil property information. Additionally, there are abundant legacy soil profiles stored in global or regional databases (e.g. WoSIS, Batjes et al., 2020). These legacy soil profiles serve as a primary data source for digital soil mapping (Lagacherie et al., 2024; Song et al., 2020; Yang et al., 2022). For China, the Second National Soil Survey serves as a significant source of legacy soil profiles, offering valuable insights into soil properties and characteristics (Shangguan et al., 2013). Therefore, these rich legacy soil profile data should be fully utilised as they better reflect historical mapping results, providing a new perspective for studying temporal changes in soil properties (Song et al., 2020). In summary, the existing dataset has several limitations, including its reliance on the traditional polygon linkage method; a limited number of soil profile samples; and the fact that it only contains basic soil property variables, lacking more comprehensive soil properties. Given these limitations, there is a compelling need to develop a new version of the dataset to address these challenges.

This paper aims to develop a new version of the CSDL (CSDLv2), with comprehensive soil physical and chemical properties for China at a 90 m resolution. This work builds on its previous version (CSDLv1, Shangguan et al., 2013), integrating advanced machine learning algorithms, multi-source soil profile samples, and high-resolution environmental covariates related to soil formation. The key advancements of this second-edition dataset, compared to the first edition, are as follows:

1. integration of multi-source soil profile samples, including data from the Second National Soil Survey of China (Shangguan et al., 2013), the World Soil Information Service (Batjes et al., 2020), the First National Soil Survey of China (National Soil Survey Office, 1964), and regional databases (Shangguan et al., 2012), rather than relying solely on data from the Second National Soil Survey, as in CSDLv1, thereby enhancing the spatial representation of soil profiles;

2. application of advanced machine learning methods, replacing the conventional soil polygon linkage method used in CSDLv1;

3. consideration of high-resolution environmental covariates as predictors for the machine learning models, allowing the model to capture more detailed spatial relationships between soil properties and environmental factors;

4. enhancement of the spatial resolution from the original 1 km to 90 m as a result of the improvements in points 1–3, providing more detailed and accurate spatial predictions of soil properties.

Additionally, compared to existing datasets, this second edition offers a major innovation: over 20 comprehensive soil property variables were developed, while most current research focuses on mapping only a few basic soil properties.

2 Materials and methods

The workflow of this study is shown in Fig. 1. Five main processes are involved in this framework:

1. harmonising and preparing soil point data and environmental covariates;

2. incorporating laboratory measurements of multiple soil profiles and overlaying them with covariates to generate a regression matrix for modelling;

3. using cross-validation to obtain optimal modelling parameters;

4. fitting prediction models based on the regression matrix;

5. applying spatial prediction models using high-resolution covariates and evaluating the models using data-splitting and independent-sample validation, as well as uncertainty maps.

2.1 Study area and soil profiles

2.1.1 Study area

China, located in East Asia along the west coast of the Pacific Ocean, extends from 3°51′ to 53°33′ N latitude and from 73°33′ to 135°05′ E longitude, covering an east–west distance of about 5000 km and featuring a continental coastline exceeding 18 000 km. The terrain of the land area of China exhibits a distinctive "ladder" pattern, with higher elevations in the west descending to lower elevations in the east, as shown in Fig. 2b. Mountains, plateaus, and hills comprise about 67 % of the land area, while basins and plains make up the remaining 33 % (Qin et al., 2016). China's topography is highly complex, encompassing an array of landforms such as extensive mountain ranges, vast plateaus, fertile plains, and deep basins. This diverse landscape is further complicated by a range of climatic zones determined by variations in temperature, precipitation, and altitude. These zones include temperate, subtropical, and tropical climates, with the temperate

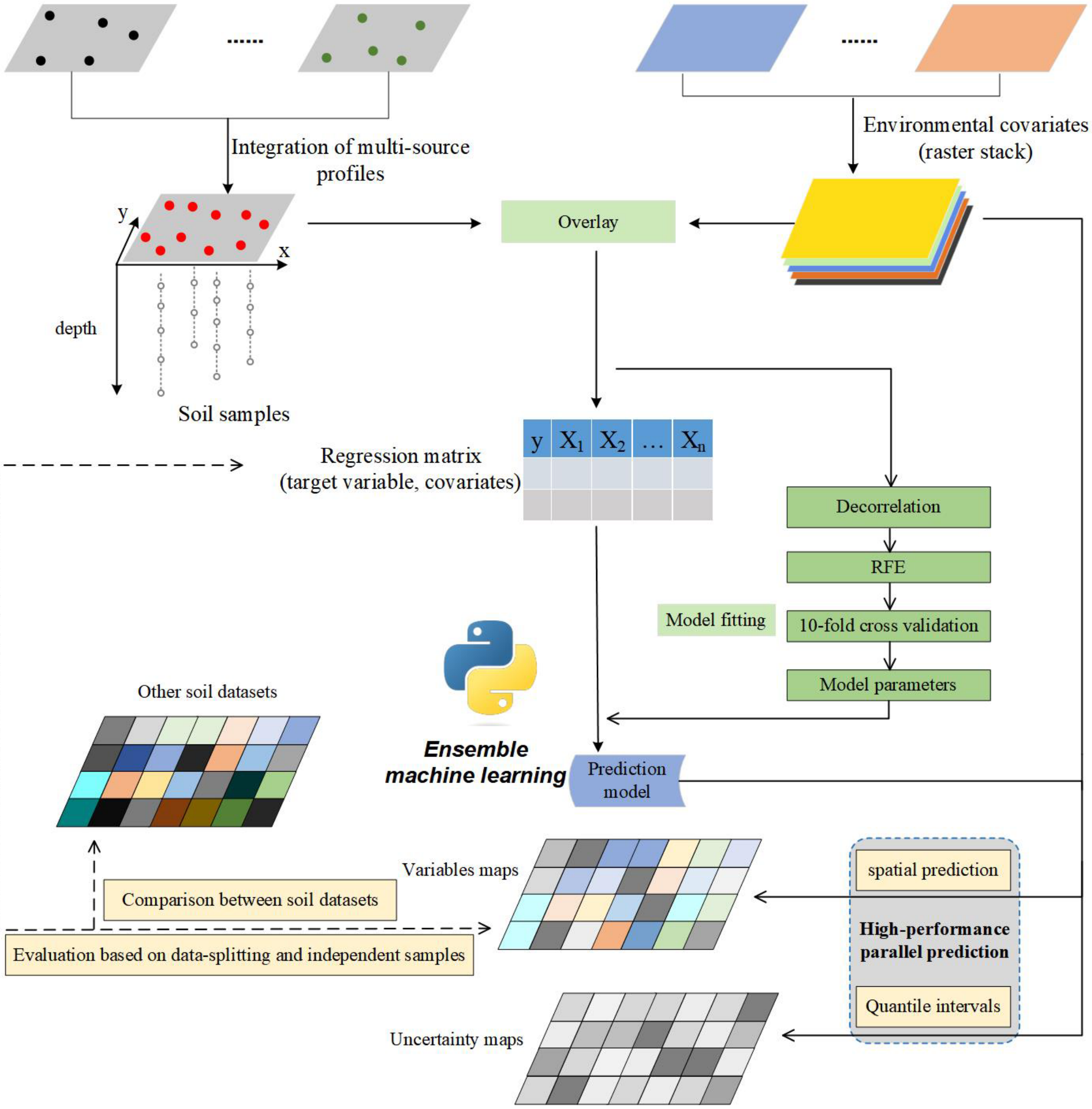

......
......
Integration of multi-source profiles
Environmental covariates (raster stack)
y
x
depth
Soil samples
Overlay
Regression matrix (target variable, covariates)
y
X_1
X_2
…
X_n
Decorrelation
RFE
Model fitting
10-fold cross validation
Model parameters
Ensemble machine learning
Prediction model
Other soil datasets
Variables maps
spatial prediction
High-performance parallel prediction
Quantile intervals
Comparison between soil datasets
Evaluation based on data-splitting and independent samples
Uncertainty maps

Figure 1. The statistical framework for developing national-scale soil property mapping in this study.

zone being the largest (Fan et al., 2016). Given the complexity and diversity of China's geographical and climatic conditions, the study of soil property mapping across this vast nation is of paramount importance.

2.1.2 Soil profiles

Typical soil profiles representing the main soil landscapes were collected from four data sources: the Second National Soil Survey of China (SNSSC, National Soil Survey Office, 1996), the World Soil Information Service (WoSIS, Batjes et al., 2020), regional datasets (Shangguan et al., 2012), and the First National Soil Survey of China (FNSSC, National Soil Survey Office, 1964). A total of 11 209 soil profiles were gathered, with the distribution details being as follows: 8979 from the SNSSC, 1540 from the WoSIS database, 614 from regional datasets, and 76 from the FNSSC. Their spatial distribution is illustrated in Fig. 2a, with different colours representing each data source. The soil property variables considered in this study are listed in Table 1. The SNSSC, conducted primarily between 1979 and 1985, provided the majority of soil profiles, although coordinates were approximated due to GPS limitations at the time, impacting mapping accuracy (Lagacherie et al., 2024). Shi et al. (2024) improved the location accuracy of soil profiles in the SNSSC by aligning detailed profile descriptions with environmental covariates. The WoSIS, managed by the International Soil Reference and Information Centre (ISRIC), is a comprehensive global database that consolidates soil profile data from various sources under a common standard (Batjes et al., 2020).

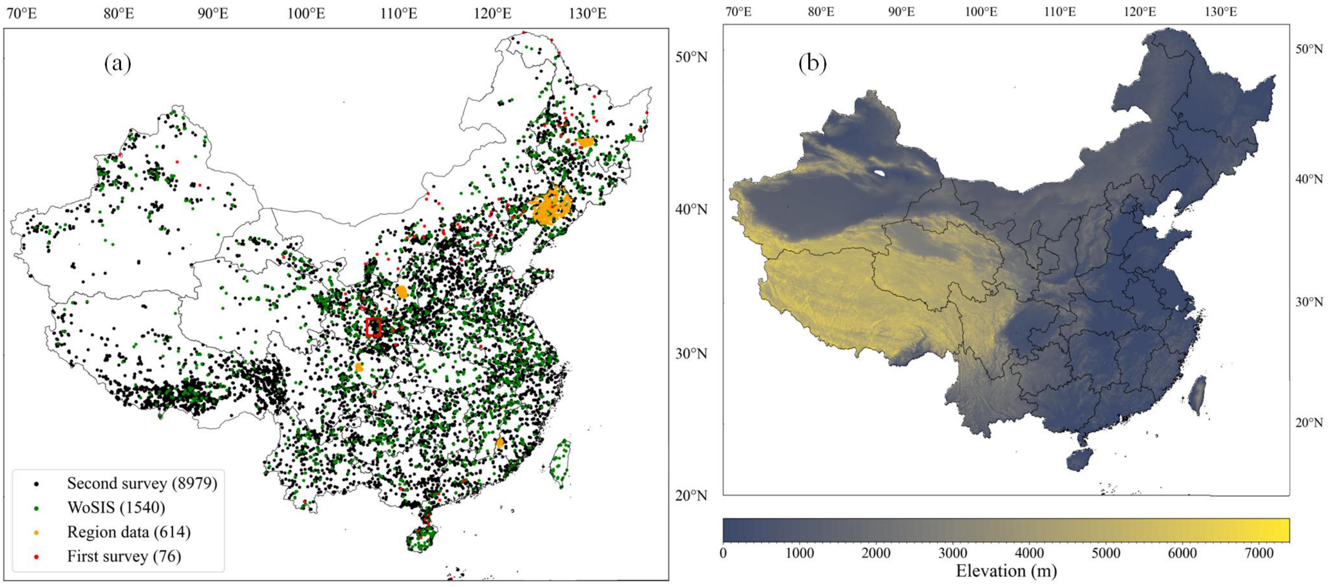


Figure 2. **(a)** Spatial distribution of the 11 209 soil profiles collected from various data sources: black dots indicate the Second National Soil Survey of China (second survey), green dots correspond to the World Soil Information Service (WoSIS), orange dots denote regional data, and red dots represent the First National Soil Survey of China (first survey). The red window indicates the area selected for visualising the spatial patterns of soil properties. **(b)** Geographical map of the land area of China.

These data are standardised and harmonised to facilitate global soil research and to enhance the accuracy of digital soil mapping efforts. It is worth noting that the WoSIS contains soil profiles from the SNSSC. The following approach was employed to determine and eliminate potentially duplicate soil profiles in the WoSIS database that may overlap with those in the SNSSC: soil profiles were considered to be duplicates if they had identical depths of soil horizons or included at least three identical depths, exhibited similar soil property values, and had close geographic coordinates (latitude and longitude). Consequently, 101 duplicate soil profiles were removed from the WoSIS database, leaving 1540 soil profiles for this study. The regional dataset was collected from five areas in 2008 and 2009 (Shangguan et al., 2012). The FNSSC, initiated in 1958, laid the foundation for China's soil science database and agricultural soil classification. The laboratory methods for obtaining and assessing soil profile data from the SNSSC and WoSIS databases are detailed in Shangguan et al. (2013) and Batjes et al. (2020), respectively. All data are exclusively from soil profiles, with no inclusion of boreholes or augerings. The regional database includes only surface data, while the SNSSC and WoSIS datasets contain full soil profiles. Because soil profiles are extracted from soil survey books of the FNSSC or the SNSSC, there may be one or several soil profiles for each soil type. As a result, though there is no sampling design for the major data sources, it may be considered to be soil-type-based stratified sampling for the final soil profile database. For soil properties sensitive to temporal changes, such as soil pH, organic carbon (OC) content, cation exchange capacity (CEC), total nitrogen (TN), total phosphorus (TP), total potassium (TK), alkali-hydrolysable nitrogen (AN), available phosphorus (AP), and available potassium (AK), we used only soil profile data from the SNSSC. In contrast, for properties less sensitive to temporal changes, such as sand, silt, clay, bulk density (BD), gravel, and porosity, we combined data from multiple sources. Since most soil profiles are from the SNSSC, the maps in the CSDLv2 mainly represent the status of soil in the 1980s. The probability density distribution of topsoil (0–5 cm) properties from different data sources is provided in Fig. S1 in the Supplement. To align with international soil mapping standards, a continuous depth function using equal-area splines was applied to horizon data, defining six standard layers (0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm) (Arrouays et al., 2014b, 2015). Detailed descriptions of the equal-area splines can be found in Bishop et al. (1999) and Liu et al. (2022a).

2.2 Environmental covariates

Following the SCORPAN (soil, climate, organisms, topography, parent material, age, and location) concept (McBratney et al., 2003), over 150 environmental covariates associated with soil formation were collected to investigate the spatial distribution of soil properties for this work. A summary of some high-resolution covariates is provided in Table 2, while the complete list can be found in Table S1. These environmental covariates offer information on the factors related to soil properties.

Table 1. List of information of soil profile data.

Soil property	Notation	Units	Description	Maps
Bulk density	BD	$g\,cm^{-3}$	Bulk density of the fine-earth fraction oven dried	Fig. S2
Sand	Sand	%	Gravimetric percentage of sand (2–0.05 mm) in the fine-earth fraction of the soil	Fig. S5
Silt	Silt	%	Gravimetric percentage of silt (0.05–0.02 mm) in the fine-earth fraction of the soil	Fig. S8
Clay	Clay	%	Gravimetric percentage of clay (<0.02 mm) in the fine-earth fraction of the soil	Fig. S11
Rock fragment	Gravel	g per 100 g	Volumetric content of fragments >2 mm in the whole soil	Fig. S14
Porosity	Porosity	$cm^3\,cm^{-3}$	Volume fraction of void space (pores) in a material	Fig. S17
Wet colour	R (wet), G (wet), B (wet)	–	RGB-quantified soil colour for wet soil	Figs. S20, S23, S26
Dry colour	R (dry), G (dry), B (dry)	–	RGB-quantified soil colour for dry soil	Figs. S29, S32, S35
Wet colour	Hue, value, chroma	–	Soil colour of wet soil is represented by the Munsell notation with three dimensions: hue, value, and chroma	Fig. S38
Dry colour	Hue, value, chroma	–	Soil colour of dry soil is represented by the Munsell notation with three dimensions: hue, value, and chroma	Fig. S19
pH value (H_2O)	pH	–	Negative common logarithm of the activity of hydronium ions (H^+) in water	Fig. S40
Soil organic carbon	OC	g per 100 g	Gravimetric content of organic carbon in the fine-earth fraction	Fig. S43
Cation exchange capacity	CEC	me per 100 g	Capacity of the fine-earth fraction to hold exchangeable cations	Fig. S46
Total nitrogen	TN	g per 100 g	Total nitrogen in soil, comprising organic, inorganic, and ammonium nitrogen, among other forms	Fig. S49
Total phosphorus	TP	g per 100 g	Total phosphorus in soil includes all phosphorus compounds, both organic and inorganic, irrespective of their plant availability	Fig. S52
Total potassium	TK	g per 100 g	Total potassium in a soil sample comprises both exchangeable (plant-available) and non-exchangeable forms	Fig. S55
Alkali-hydrolysable nitrogen	AN	$mg\,kg^{-1}$	Total amount of nitrogen released from soil through alkali treatment (i.e. sodium hydroxide or potassium hydroxide)	Fig. S58
Available potassium	AK	$mg\,kg^{-1}$	Portion of potassium in the soil that is readily accessible for plant uptake	Fig. S61
Available phosphorous	AP	$mg\,kg^{-1}$	Fraction of phosphorus in the soil that is soluble in a chemical extract and readily accessible for plant uptake	Fig. S64

Table 2. Summary of the main high-resolution environmental covariates. For the complete list of soil-forming factors, see Table S1.

Factor definitions	Description	Resolution (m)	Source
BDTICM	Depth to bedrock of China	90	http://globalchange.bnu.edu.cn/research/cdtb.jsp (last access: 15 February 2024)
B5/B7	The ratio of band 5 (near-infrared) to band 7 (shortwave infrared 2) surface reflectance	90	https://www.usgs.gov/landsat-missions/landsat-collection-2 (last access: 18 February 2024)
NDVI	Normalised difference vegetation index	90	Calculated from Landsat 8 Collection 2 Level 2 (LC08C02) on the GEE platform
NDWI	Normalised difference water index	90	Calculated from LC08C02 on the GEE platform
surR	Surface reflectance	250	https://modis.gsfc.nasa.gov/data/dataprod/mod09.php (last access: 18 February 2024)
EVI	Enhanced vegetation index	90	Calculated from LC08C02 on the GEE platform
SAI	Snow area index	90	Calculated from LC08C02 on the GEE platform
NPP	Net primary productivity	500	https://lpdaac.usgs.gov/products/mod17a3hgfv061/ (last access: 18 February 2024)
Canopy height	Canopy height	10	https://doi.org/10.3929/ethz-b-000609802 (last access: 17 February 2024)
Land cover	Land cover	30	http://www.sciencemag.org/content/342/6160/850 (last access: 17 February 2024)
Sentinel-2 (B2, B3, B4, B8, B9)	Bands 2, 3, 4, 8, and 9 from Sentinel-2	30	Derived from Sentinel-2 on the GEE platform
QA_PIXEL	Landsat 8 Collection 2 Level 2 pixel quality band	90	Derived from LC08C02 on the GEE platform
QA_RADSAT	Radiometric saturation quality control	90	Derived from LC08C02 on the GEE platform
SR (B4, B5, B6, B7)	Surface reflectance of bands 4, 5, 6, and 7	90	Derived from LC08C02 on the GEE platform
ST_ATRAN	Atmospheric transmittance	90	Derived from LC08C02 on the GEE platform
ST_B10	Band-10 surface temperature	90	Derived from LC08C02 on the GEE platform
ST_EMSD	Emissivity standard deviation	90	Derived from LC08C02 on the GEE platform
ST_TRAD	Thermal radiance	90	Derived from LC08C02 on the GEE platform
ST_URAD	Downwelled radiance	90	Derived from LC08C02 on the GEE platform
DEM	Land surface elevation	90	https://hydro.iis.u-tokyo.ac.jp/~yamadai/MERIT_DEM/ (last access: 17 February 2024)
Slope	Terrain slope	90	Derived from DEM
Land use	Land use type	30	https://www.resdc.cn/DOI/DOL.aspx?DOIID=54 (last access: 17 February 2024)
RTMUSG15	Rock type	250	https://doi.pangaea.de/10.1594/PANGAEA.788537 (last access: 17 February 2024) (Hartmann and Moosdorf, 2012)

Relief covariates were primarily derived from the MERIT digital elevation model (DEM) dataset (https://hydro.iis.u-tokyo.ac.jp/~yamadai/MERIT_DEM/, last access: 17 February 2024), a high-precision global DEM with a resolution of 3 arcsec (∼ 90 m at the Equator), vertically referenced to the EGM96 geoid and horizontally referenced to the World Geodetic System 1984 (Yamazaki et al., 2019). This dataset serves as an improved spaceborne DEM that significantly reduces the major error components found in other DEMs, such as NASA's SRTM3 DEM and the Viewfinder Panoramas DEMs (Li et al., 2023). Based on this study's DEM, other relief covariates such as slope, plan curvature, profile curvature, and terrain wetness index were calculated using SAGA GIS (Conrad et al., 2015).

Organism-related covariates were primarily sourced from six datasets: Landsat 8 Collection 2 Level 2 (LC08C02), MODIS, GLOBELLAND30, the Global Accessibility Map, and GlobCover. L8C2L2 is an advanced satellite data product released by the United States Geological Survey (USGS). Landsat 8, part of the Landsat satellite series, is specifically designed for Earth observation and monitoring. Collection 2 represents an updated version of Landsat data products, incorporating various improvements and enhancements. High-resolution data, such as normalised difference vegetation index (NDVI), normalised difference water index (NDWI), band 5 (near-infrared), and band 7 (shortwave infrared 2), at a 90 m spatial resolution were obtained from this database via the Google Earth Engine (GEE) platform. MODIS data offer an efficient method for monitoring biosphere changes and understanding Earth's climate system, available at a spatial resolution of 1 km. GLOBELLAND30, a significant achievement from China's global and local land cover remote sensing mapping and technology research project, provides comprehensive global land surface coverage at a 30 m resolution. The Global Accessibility Map illustrates urban and rural population gradients at a 1 km resolution from the year 2000 to present. Developed by the European Space Agency, the GlobCover dataset provides a global land cover map at a 1 km resolution.

Climate factors were chiefly obtained from the MODIS, WorldClim, and CHELSA datasets (DAAC, 2018; Karger et al., 2020), primarily offered at a 1 km spatial resolution and covering the years 1970–2000. Soil factors, i.e. soil classifications, were mainly derived from the Harmonized World Soil Database, also available at a 1 km spatial resolution (Nachtergaele et al., 2012). Parent material factors were represented by the depth-to-bedrock maps and a lithological map (Yan et al., 2020).

All environmental covariates were reprojected to a unified coordinate reference system, specifically Goode's homolosine projection applied to the World Geodetic System (WGS) 1984 projection. This projection was chosen as it is the most effective at minimising distortions over land among the equal-area projections available in open-source software (Moreira De Sousa et al., 2019). Additionally, the nearest-interpolation and bilinear-interpolation algorithms were applied to the subtype data (e.g. vegetation type) and continuous variables, respectively, to resample these environmental covariates to a raster cell size of 90 m resolution for spatial modelling and map prediction.

Considering the substantial number of available environmental covariates, those with an absolute Pearson correlation coefficient of less than 0.05 in relation to the target variable were excluded. Subsequently, redundant covariates with a Pearson correlation coefficient greater than 0.8 in relation to any other covariate were removed to eliminate autocorrelation among them. For each pair of environmental covariates with a correlation exceeding this threshold, only the first one in alphabetical order was retained for the modelling phase (Poggio et al., 2021). This process reduced the initial number of environmental covariates to approximately 80 layers.

In this study, the recursive feature elimination (RFE) method was implemented using the sklearn.feature_selection package in Python, which offers a balanced approach between accuracy and computational efficiency. RFE is a robust technique, widely recognised for its efficacy in selecting optimal covariate sets for regression tree models (Gomes et al., 2019). The RFE process begins by fitting a model that includes all environmental factors, evaluating its performance, and ranking the covariates based on their importance. The least significant factors are systematically eliminated, followed by re-fitting of the model and reassessment of its performance. This iterative procedure continues until the pool is reduced to a set between zero and the total number of environmental covariates. This method relies on out-of-bag (OOB) cross-validation, making it a reliable selection approach for models such as random forests, even though it does not test every possible combination of covariates (Nussbaum et al., 2018). The RFE process is independently conducted on each subset, leveraging the default hyperparameters of the random forest algorithm as provided by the RandomForestRegressor package in Python. The optimal subset of variables is identified when further iterations no longer yield improvements in model performance, defined by the minimisation of the loss function. For this study, the OOB root-mean-square error (RMSE) was used as the loss function. The ultimate set of covariates was identified as the combination that minimised the loss function. The aforementioned analysis was executed for all target variables and depths. For instance, with surface (0–5 cm) soil organic carbon, 35 environmental covariates remained for analysis after the filtering process (Fig. 3), marked with a superscript of "1" in Table S1.

2.3 Digital soil mapping

2.3.1 Spatial prediction and uncertainty

The random forest (RF, Breiman, 2001) and quantile regression forest (QRF) models were employed to evaluate, over

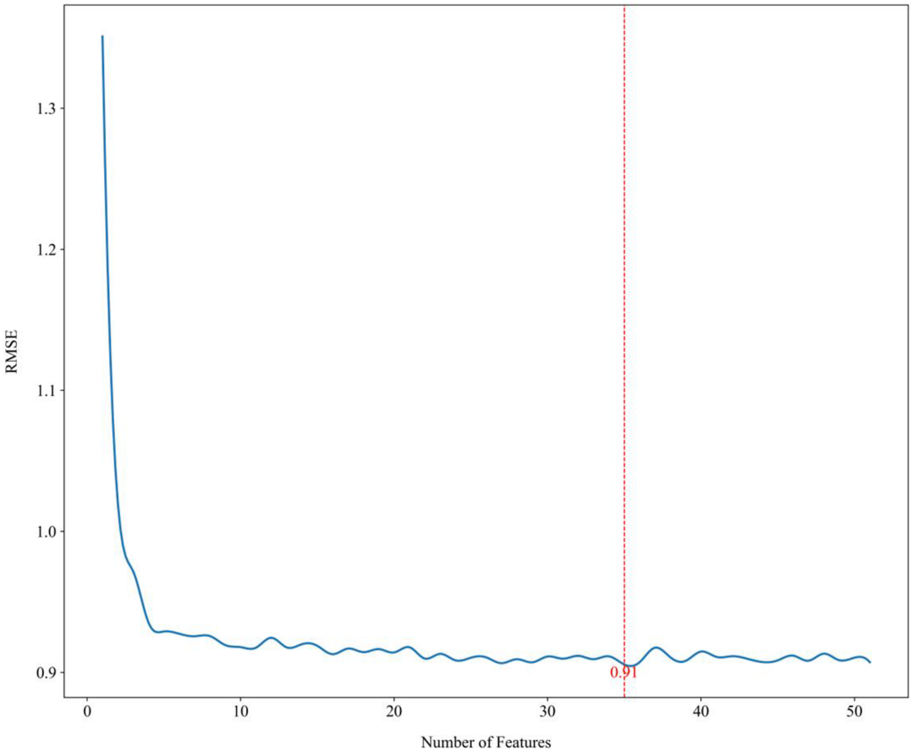


Figure 3. Example of the loss function (RMSE) used in the recursive feature elimination (RFE) step of covariate selection for surface (0–5 cm) soil organic carbon content.

six layers, the statistical relationship between each soil property and environmental covariates. The RF model in this study was used to generate mean predictions, while the QRF model was applied to produce prediction maps at different quantiles, providing a more comprehensive representation of uncertainty. The QRF algorithm, introduced by Meinshausen (2006), is an ensemble machine learning model that utilises tree structures and bootstrapping techniques to create a collection of tree models. Each tree is developed from a learning set generated by repeatedly sampling calibration samples through bootstrapping, with node splits influenced by a randomly selected subset of covariates. The final prediction value for each predetermined quantile is obtained by averaging the predicted values from all trees. Building on the foundation of RF, the QRF algorithm present a novel approach to enhancing regression tree performance (Koenker, 2005). In RF, averaging across multiple tree-based models results in more accurate predictions compared to using a single regression tree. The QRF offers insights into the full conditional distribution of the dependent variable. Consequently, conditional quantiles can be inferred using the QRF algorithm. The conditional distribution of Y given $X=x$ is defined as $F(y|X=x)=P(Y\leq y|X=x)$. To estimate $F(y|X=x)$, a weighted empirical cumulative distribution function is considered:

$$\hat{F}(y|X=x)=\sum\nolimits_{i=1}^{n} w_i(x,\theta)\, Y_{\{Y\leq y\}}. \qquad (1)$$

The tree-based model developed using the QRF algorithm follows the RF methodology. However, unlike RF, where only the mean of the observations within each node is retained, the QRF approach preserves the values of all observations within each node. This comprehensive set of observations in each node is utilised to derive the quantiles, which are subsequently used to construct prediction intervals. These intervals serve as a measure of the prediction uncertainty, providing a more detailed understanding of the conditional distribution of the target variable. Additionally, the uncertainty estimates evaluated by QRF are likely to be more accurate and interpretable than those derived from regression kriging, particularly in areas with sparse samples (Liu et al., 2022a). Furthermore, RF and QRF are capable of handling complex non-linear relationships and multivariate interactions, offering high predictive power (Gyamerah et al., 2020). This distinguishing advantage sets RF and QRF apart from other machine learning algorithms (Liu et al., 2022b).

Separate models were developed independently for each soil layer, ensuring no overlap of observations from the same profile across training and testing datasets. The selection of hyper-parameters, specifically the number of randomly selected variables from all predictors (max_ features) and the minimum node size (min_samples_leaf), plays a crucial role in determining the performance of the RF model. These hyper-parameters significantly influence the model's predictive accuracy. Other parameters, such as the number of trees (n_estimators), were not optimised during the RF's training process. To address potential overfitting concerns, the values of max_features and min_samples_leaf were fine-tuned using a 10-fold cross-validation method. This approach involved randomly dividing the training dataset into 10 folds; 1/10 of these sub-datasets was utilised as the validation sample, while the remaining sub-datasets were applied for training the RF and QRF model. This tuning was conducted using the gridded direct search approach, with max_features being explored within the range of [1, 30] at single intervals and min_ samples_leaf being explored within the range of [5, 30] at intervals of 5. In this study, the aforementioned hyperparameter search was conducted for each of the six soil depth layers for every soil property. These hyperparameters were then used for modelling and spatial prediction of the corresponding soil property variables at their respective depths. To maintain brevity, Table S2 presents the tuned model hyperparameters for each soil property considered at the 0–5 cm depth interval.

The relative importance of covariates in the trained RF and QRF model was assessed to investigate the impact of environmental factors on the spatial variations of soil properties. This importance was determined by evaluating the influence of each covariate on the model's prediction performance. The relative importance of each covariate was quantified using the increase in mean square error (%IncMSE), a metric derived from permuting the values of a covariate to remove its information content. By comparing the model's accuracy before and after permutation, it was possible to determine how crucial each covariate was in predicting soil properties. A higher %IncMSE indicated a greater importance of the covariate, signifying that its presence substantially contributed to the model's predictive accuracy. This relative importance allows for a detailed analysis of how different environmental factors control spatial variations in soil properties, providing valuable insights for digital soil mapping.

Mapping China, which covers approximately 9.6×10^6 km^2, at 90 m resolution requires more than 10^9 pixels for each soil property at each depth, posing a considerable challenge. Due to the extensive geographic coverage and high-resolution requirements in soil mapping for this study, predicting each soil property at a specific depth involves a substantial volume of data, with environmental covariate data reaching up to 470 GB. Faced with such extensive data-processing demands, conventional single-machine resources often prove to be inadequate and challenging to cope with. Therefore, to overcome the memory limitations imposed by high-resolution mapping and to enhance the computational efficiency of spatial prediction, we implemented parallel computing. Initially, we partitioned environmental covariates into distinct $1^\circ \times 1^\circ$ tiles. Using the finalised model, a single core performed spatial predictions within each block. Leveraging multiple-core processing, we simultaneously handled multiple tiles, significantly accelerating spatial predictions. Upon acquiring the outcomes for every tile, we utilised image mosaicking to seamlessly integrate these outputs, ultimately assembling the comprehensive map of various soil properties and depths across China. All the experiments are performed on a Linux server with Intel Core (TM) i9-10980XE, 3.00 GHz × 64 CPU, 512 GB RAM (random access memory), and two NVIDIA RTX A5000 graphics cards. All scripts were written in the open-source Python programming environment with Python version 3.11.4 (https://www.python.org/, last access: 3 December 2024) using PyCharm version 2024.3.28. The RandomForestQuantileRegressor and RandomForestRegressor packages were employed for model construction. The optimisation of the model was performed using the scikit-learn library, while the gdal and matplotlib packages were utilised for data processing and visualisation, respectively.

Using the selected environmental covariates from the aforementioned feature engineering, the constructed models were applied to compute four statistical values – the mean, 0.05 quantile ($q_{0.05}$), median (0.50 quantile, $q_{0.50}$), and 0.95 quantile ($q_{0.95}$) – for every 90 m pixel across all standard depth layers (0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm) as specified by the GlobalSoilMap (Arrouays et al., 2014b) over China, capturing the conditional distribution of soil properties. Although the performance differences between mean predictions using RF and median predictions using QRF are minimal, their ability to capture extreme values (i.e. both high and low values) was considered. In this study, we evaluated the performance of RF and QRF models according to both the overall statistical metrics and their capacity to predict extreme values in order to determine the most suitable model for generating national gridded soil maps of various soil properties at a 90 m resolution. As shown in Table S7, soil properties such as soil pH, silt, clay, TP, red (R) wet-soil colour, blue (B) wet-soil colour, red (R) dry-soil colour, and blue (B) dry-soil colour were modelled using median predictions from QRF as this approach better captured extreme values. Similarly, the study by Helfenstein et al. (2024) also assessed mean predictions by RF and median predictions by QRF, highlighting that, for certain soil properties, median predictions are more appropriate than mean predictions. For most other soil properties in this study – such as sand, BD, OC, gravel, AN, TN, CEC, porosity, TK, AK, AP, green (G) wet-soil colour, and green (G) dry-soil colour – mean predictions from RF were used to generate the 90 m resolution soil maps. The better model was consistent across different depths for the same soil property; thus, Table S7 only presents the performance comparison of mean and median predictions for the surface layer (0–5 cm depth interval), and either the mean or the median is used for the mapping of a soil property for all depths. The calculated median, along with the 0.05 and 0.95 quantiles, was also used to estimate uncertainty. Uncertainty was expressed as the upper and lower limits of a 90 % prediction interval, represented by the empirical distribution's 0.05 and 0.95 quantiles, respectively. Furthermore, to facilitate comparison, the prediction interval relative to the median ($q_{0.50}$) was used as an indicator of uncertainty (Liang et al., 2019; Liu et al., 2022a). A higher ratio for a pixel indicates greater uncertainty in the predicted value for that location (Poggio et al., 2021). When developing the 90 m resolution soil maps in this study, either mean or median predictions were selected for storage efficiency. However, for lower-resolution maps provided at 1 and 10 km, in addition to mean and median predictions, we also included prediction maps for the 0.05 and 0.95 quantiles. These additional maps are helpful for illustrating data uncertainty.

For the sand, silt, and clay contents from the FNSSC and SNSSC, these were measured following the schemes of the International Society of Soil Science (ISSS) and of Katschinski (Katschinski, 1956). Since most land surface models (LSMs) and other applications require soil texture data in the FAO-USDA system, we used several particle size distribution models (Shangguan et al., 2013) to convert the original ISSS and Katschinski particle size distribution data into the FAO-USDA system. A 5 % quality control threshold was applied, excluding soil profile samples where the sum of the three fractions fell outside of the 95 %–105 % range (Shangguan et al., 2013), and they were converted to make sure that their sum was 100 % by using the weighting approach. For the mapping of each particle size fraction (sand, silt, and clay), separate spatial prediction models were developed, and the weighting approach was applied to ensure that the sum of the three fractions equaled 100 %.

2.3.2 Evaluation criteria

To validate the performance of RF and QRF models in generating the CSDLv2, two validation methods were employed to ensure that the CSDLv2 product has low errors at both spatial and vertical depth scales against laboratory measurements values. The first method involved randomly selecting 10 % of the multi-source soil profiles as test samples, while the remaining 90 % were used for training the model (i.e. data splitting). The second method took the WoSIS dataset as an external independent validation dataset, with the rest of the data being used for model training (i.e. independent samples). We choose the WoSIS as the independent validation dataset because it has a spatial distribution close to that of a probability sampling (Brus et al., 2011). Based on the training soil profiles, these two validation approaches were implemented to assess the performance accuracy of predictive mapping for each soil property at various depths. Three statistics, namely, the mean prediction error (ME), root-mean-square prediction error (RMSE), and modelling efficiency coefficient (MEC, Krause et al., 2005), were calculated to evaluate the models' predictive performance. They were calculated as follows:

$$\mathrm{ME} = \frac{1}{N}\sum\nolimits_{i=1}^{N} \varepsilon\left(s_i\right), \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum\nolimits_{i=1}^{N} \varepsilon\left(s_i\right)^2}, \tag{3}$$

$$\mathrm{MEC} = 1 - \frac{\sum_{i=1}^{N}\left(z\left(s_i\right) - \hat{z}\left(s_i\right)\right)^2}{\sum_{i=1}^{N}\left(z\left(s_i\right) - \overline{z}\right)^2}, \tag{4}$$

where z represents the observed soil variable, $\hat{z}$ is the predicted soil variable at location s_i ($i = 1, \ldots, N$; $s_i \in \wp$), and N is the total number of population units in the study area $\wp$. We regard the prediction error as the difference between the observed (z) and predicted ($\hat{z}$) values of a soil property at the ith spatial location, denoted by $\varepsilon\left(s_i\right) = z(s_i) - \hat{z}(s_i)$. To guarantee the accuracy and reliability of our results, we performed 20 repetitions of 10-fold cross-validation and calculated the mean and standard deviation of the measurements.

The soil property maps predicted in this study were compared to three existing soil map datasets. The first dataset is SoilGrids 2.0, accessible at https://soilgrids.org/ (last access: 10 January 2024), which has a 250 m resolution (Poggio et al., 2021). It represents an advancement over previous global soil property maps and is known as SoilGrids250m (Hengl et al., 2017), incorporating up-to-date machine learning methods and benefiting from the expanded availability of standardised soil profile data worldwide, along with environmental covariates (Poggio et al., 2021). The second dataset is the CSDLv1, with a resolution of 1 km (Shangguan et al., 2013), accessible at http://globalchange.bnu.edu.cn (last access: 5 January 2024). Lastly, we considered the Harmonized World Soil Database v2.0 (HWSD 2.0), known for its soil property maps created via a soil type linkage method, available at https://www.fao.org/soils-portal/data-hub/soil-maps-and-databases/harmonized-world-soil-database-v20/en/ (last access: 20 January 2024). The HWSD 2.0 has been synthesised by integrating regional and national soil data globally (FAO and IIASA, 2023). To quantify the enhancement of our predictions over existing soil maps, we calculated the relative improvement (RI) using both RMSE and MEC metrics, employing the following equations:

$$\mathrm{RI}_{\mathrm{RMSE}} = \frac{\mathrm{RMSE}_{\mathrm{existing}} - \mathrm{RMSE}_{\mathrm{CSDLv2}}}{\mathrm{RMSE}_{\mathrm{existing}}}, \tag{5}$$

$$\mathrm{RI}_{\mathrm{MEC}} = \frac{\mathrm{MEC}_{\mathrm{CSDLv2}} - \mathrm{MEC}_{\mathrm{existing}}}{\mathrm{MEC}_{\mathrm{existing}}}, \tag{6}$$

where $\mathrm{RI}_{\mathrm{RMSE}}$ and $\mathrm{RI}_{\mathrm{MEC}}$ denote the relative improvement concerning RMSEand MEC, respectively. $\mathrm{RMSE}_{\mathrm{new}}$ and $\mathrm{MEC}_{\mathrm{new}}$ represent the accuracy statistics for predictions in this study, while $\mathrm{RMSE}_{\mathrm{existing}}$ and $\mathrm{MEC}_{\mathrm{existing}}$ signify the accuracy statistics for the existing soil maps. An RI of > 0 denotes that the CSDLv2 outperforms the existing soil maps.

Considering the unavoidable impact of various error sources on any model for DSM, it is essential to quantify the associated mapping uncertainty (Lilburne et al., 2024; McBratney et al., 2018). To evaluate uncertainty, the prediction interval coverage probability (PICP) was employed based on the randomly held-back soil profile test samples. PICP represents the proportion of observations at each depth encapsulated by the corresponding prediction interval (Li et al., 2023). In this study, the prediction interval was estimated using the aforementioned QRF model. If the uncertainty estimates are reasonably defined, the PICP should yield an estimate of 90 % for a 90 % (or 0.9) prediction interval. A PICP significantly greater than 0.9 suggests that the uncertainty has been underestimated, whereas a PICP significantly less than 0.9 indicates that it has been overestimated (Liu et al., 2020; Poggio et al., 2021).

3 Results

3.1 Statistical analysis

The probability density distributions of topsoil (0–5 cm) properties from different data sources are shown in Fig. S1, with different colours representing different data sources. If a colour representing a data source is absent in some probability density distribution charts, this indicates that the soil property is not available from that data source. As observed in Fig. S1, the probability density distributions of soil properties from multiple sources exhibit a generally similar trend, with minor differences that increase the spatial representativeness of the soil profile samples, rather than representing specific soil types. The abundance of soil profile data allows for a more detailed characterisation of spatial variations in soil properties, particularly in a large and topographically diverse country like China (Liu et al., 2022a). Descriptive statistical analyses of soil properties across six standard depths are presented in Table S3. For most soil property variables at multiple depths, there is an extensive amount of soil profile data. Different soil properties exhibit varying trends with depth, accompanied by a large range and variation (see coefficient of variation). The vertical changes in soil properties vary depending on the specific soil property and soil type. For example, the contents of OC and TN generally decrease with increasing depth in most soil types, exhibiting positive skewness distributions. However, other properties, such as soil pH or BD, show different vertical patterns depending on soil composition and local conditions. Regarding the homogeneity of variance, Levene's test between samples from different depths yielded p values greater than 0.05 for soil property, indicating no statistically significant differences between samples from different depths.

3.2 Predictive performance

After training and optimisation, the effectiveness of the RF and QRF models was evaluated. Using the test set, the model's prediction accuracy across multiple depths was assessed using two validation methods: Tables 3 and S4 present the predictive performance using a data-splitting strategy, where 10 % of aggregated soil profiles were randomly partitioned as the test set. This validation of the CSDLv2 was compared with the validation of the three existing soil map datasets using all soil profiles in this study. Table S5 displays the model's performance when modelling soil profiles from remaining data sources, validated independently using WoSIS data.

Overall, model performance varied depending on the soil properties. The mean ME values were nearly zero, indicating that the predictions were generally unbiased. Soil pH was predicted with the highest accuracy, with MEC performance ranging from 0.75 to 0.68 across depths in the data-splitting validation strategy. That is to say that more than 68 % of the pH variation can be explained, and the predicted values are in good agreement with the laboratory measurements values. This result is consistent with previous studies (Chen et al., 2019; Hu et al., 2024; Lu et al., 2023). The mean MEC values for sand and clay content were slightly higher than those for silt content, indicating that sand and clay are slightly more predictable than silt. As soil depth increased, MEC values showed a decreasing trend, while RMSE values increased, suggesting a vertical decline in the predictability of soil texture. This decline may be attributed to the fact that environmental covariates primarily reflect surface conditions, leading to reduced correlation with deeper soil properties. Additionally, the decrease in sample size at greater depths may also contribute to this trend. Similar observations have been noted in other related studies (Liu et al., 2020; Poggio et al., 2021). The model's predictive performance at the 5–15 cm depth interval was better than at the 0–5 cm depth interval, with higher MEC values and lower RMSE values. The prediction accuracy for OC was relatively high, with approximately 25 % to 60 % of the variation in OC across all depth layers being explained in both the data-splitting and independent validation methods. This performance surpasses the accuracy reported in related literature for OC prediction (Liang et al., 2019; Padarian et al., 2017). The prediction accuracy for soil property contents such as BD, gravel, TN, CEC, TK, and TP is higher at depths of less than 30 cm. These models can explain 30 % to 60 % of the variation in these soil properties, with accuracy comparable to that reported in related studies (Mulder et al., 2016; Ramcharan et al., 2018).

The model's performance varied with soil depth. For most soil property variables, including OC, TN, and BD, predictive accuracy decreased significantly with increasing depth. In contrast, the accuracy for CEC, gravel content, and TK declined only slightly. This decrease in accuracy for deeper layers has been noted in previous studies on soil organic carbon prediction (Mulder et al., 2016; Padarian et al., 2017), primarily because most environmental covariates predominantly characterise surface conditions, leading to weaker correlations with deeper soil layers (Liu et al., 2020). Conversely, the prediction accuracy for soil pH increased slightly with depth. This improvement may be partly due to the increased stability of soil pH in deeper layers over large areas, leading to more consistent relationships with environ-

Table 3. Accuracy evaluation of the selected soil properties with the highest prediction accuracy in the CSDLv2, the CSDLv1, SoilGrids 2.0, and the HWSD 2.0 based on the randomly held-back soil profiles. The "number" column indicates the number of samples used during testing. Refer to Table S4 for the complete accuracy evaluation of the soil properties considered. See Table 1 for the abbreviations and units of the soil properties of interest.

Property	Depth interval	Number	CSDLv2			CSDLv1			SoilGrids 2.0			HWSD 2.0		
			MEC	RMSE	ME	MEC	RMSE	ME	MEC	RMSE	ME	MEC	RMSE	ME
pH	0–5	830	0.69	0.70	0.00	0.48	0.92	−0.03	0.60	0.79	−0.15	0.35	1.03	−0.28
	5–15	830	0.70	0.68	0.00	0.50	0.90	−0.02	0.61	0.77	−0.12	0.36	1.02	−0.13
	15–30	822	0.70	0.68	0.00	0.26	1.21	−0.41	0.60	0.77	−0.16	0.38	1.03	−0.15
	30–60	800	0.68	0.70	−0.00	0.43	0.94	−0.04	0.59	0.78	−0.15	0.38	1.02	−0.17
	60–100	648	0.68	0.70	0.00	0.44	0.94	0.04	0.59	0.78	−0.14	0.39	1.01	−0.18
	100–200	204	0.75	0.60	0.00	0.53	0.84	−0.05	0.63	0.70	−0.09	0.52	0.87	−0.08
Sand	0–5	874	0.67	12.15	0.05	0.19	22.19	−2.24	0.60	13.08	−1.84	0.20	21.84	2.38
	5–15	815	0.71	11.23	0.06	0.18	21.90	−2.28	0.62	11.87	−1.93	0.19	21.43	1.40
	15–30	812	0.71	11.41	0.05	0.15	22.58	−1.67	0.62	11.85	−1.71	0.14	21.89	2.63
	30–60	784	0.69	12.16	0.06	0.13	23.26	−1.31	0.59	12.68	−1.80	0.12	22.57	3.68
	60–100	638	0.68	12.85	0.04	0.11	23.22	−1.30	0.51	13.53	−1.94	0.10	23.45	4.03
	100–200	213	0.64	13.72	0.02	0.10	24.22	−1.42	0.49	14.59	−1.88	0.09	24.11	3.98
Silt	0–5	893	0.61	9.81	0.02	0.11	16.78	2.02	0.55	10.54	−0.58	0.10	17.38	−4.44
	5–15	832	0.65	8.99	−0.00	0.13	16.31	2.29	0.58	9.22	−0.33	0.10	16.90	−5.55
	15–30	830	0.67	8.76	0.00	0.13	16.29	2.12	0.60	9.02	−0.51	0.09	17.30	−6.46
	30–60	802	0.63	9.49	0.00	0.11	16.55	1.76	0.57	9.68	−0.41	0.10	17.53	−6.36
	60–100	656	0.62	10.08	0.00	0.10	17.05	1.49	0.55	10.34	−0.33	0.10	18.07	−6.15
	100–200	221	0.64	10.60	0.01	0.09	17.94	0.70	0.54	11.25	−0.99	0.11	19.14	−5.15
Clay	0–5	914	0.63	6.74	0.01	0.12	11.23	0.21	0.52	7.60	2.49	0.12	11.14	2.06
	5–15	854	0.67	6.50	0.01	0.09	11.28	0.03	0.58	7.18	2.36	0.09	11.89	4.23
	15–30	851	0.68	6.83	0.01	0.10	11.83	0.61	0.60	7.40	2.28	0.09	12.78	3.95
	30–60	523	0.68	7.36	0.02	0.09	12.78	0.14	0.61	7.89	2.22	0.13	13.20	2.70
	60–100	675	0.68	7.79	0.02	0.07	13.43	−0.28	0.61	8.33	2.21	0.12	13.65	1.97
	100–200	230	0.63	7.96	0.03	0.06	13.00	0.86	0.55	8.67	2.74	0.12	13.06	0.91
BD	0–5	153	0.62	0.12	0.00	0.12	0.20	0.01	0.53	0.13	0.01	0.02	0.27	0.15
	5–15	155	0.63	0.11	0.00	0.15	0.19	0.01	0.57	0.12	0.01	0.01	0.29	0.18
	15–30	155	0.60	0.11	−0.00	0.11	0.19	0.01	0.54	0.13	0.01	0.01	0.27	0.12
	30–60	136	0.55	0.12	−0.00	0.10	0.19	−0.01	0.53	0.13	−0.00	0.01	0.24	0.10
	60–100	95	0.57	0.12	−0.00	0.10	0.19	−0.01	0.51	0.13	−0.01	0.02	0.24	0.07
	100–200	33	0.47	0.13	0.00	0.05	0.22	0.02	0.42	0.13	−0.01	0.02	0.24	0.07

mental factors (Liu et al., 2020). This observation aligns with the findings of Padarian et al. (2017). Additionally, independent-sample validation is an effective approach to assess the validity of models and has been utilised in multiple studies (Lamichhane et al., 2019). Table S5 summarises the model's predictive performance based on independent validation and compares it with other data products. These results also demonstrate the reliability of the predictive model.

3.3 Spatial patterns

Figure 4 illustrates the maps of soil physical and chemical properties at the soil surface (0–5 cm) over China at 90 m resolution. The spatial distribution of the complete soil properties (as listed in Table 1) can be found in Figs. S2–24.

As shown in Fig. 4a, the pH values (H_2O) in the topsoil range from 4.3 to 9.8. Soils south of 30° N are predominantly acidic to strongly acidic, while those in the northern and northwestern regions are mostly basic or strongly basic. In some southern hilly and northeastern forested areas, soils appear to be acidic (pH $<$ 7). In certain northern regions, especially in desert areas, soils are alkaline (pH $>$ 7). This distribution aligns with the common understanding that areas with low precipitation tend to have alkaline soils, whereas areas with high precipitation tend to have acidic soils.

As shown in Fig. 4b, for BD, northern regions tend to have higher bulk density due to low organic matter content and frequent agricultural activities. Southern regions generally have lower bulk density owing to higher organic matter content and higher porosity. Northwestern arid regions exhibit high bulk density, while the Qinghai–Tibet Plateau has low bulk density. Southeastern coastal areas show significant variation in surface bulk density, heavily influenced by land use practices.

As shown in Fig. 4c, the spatial predictions of OC content reveal significant regional differences. The highest OC levels are found in the eastern Tibetan Plateau, northeastern China, and northern Xinjiang, where human activities are minimal. In contrast, the lowest OC content is observed in the northwestern desert regions. OC content shows a decreasing trend

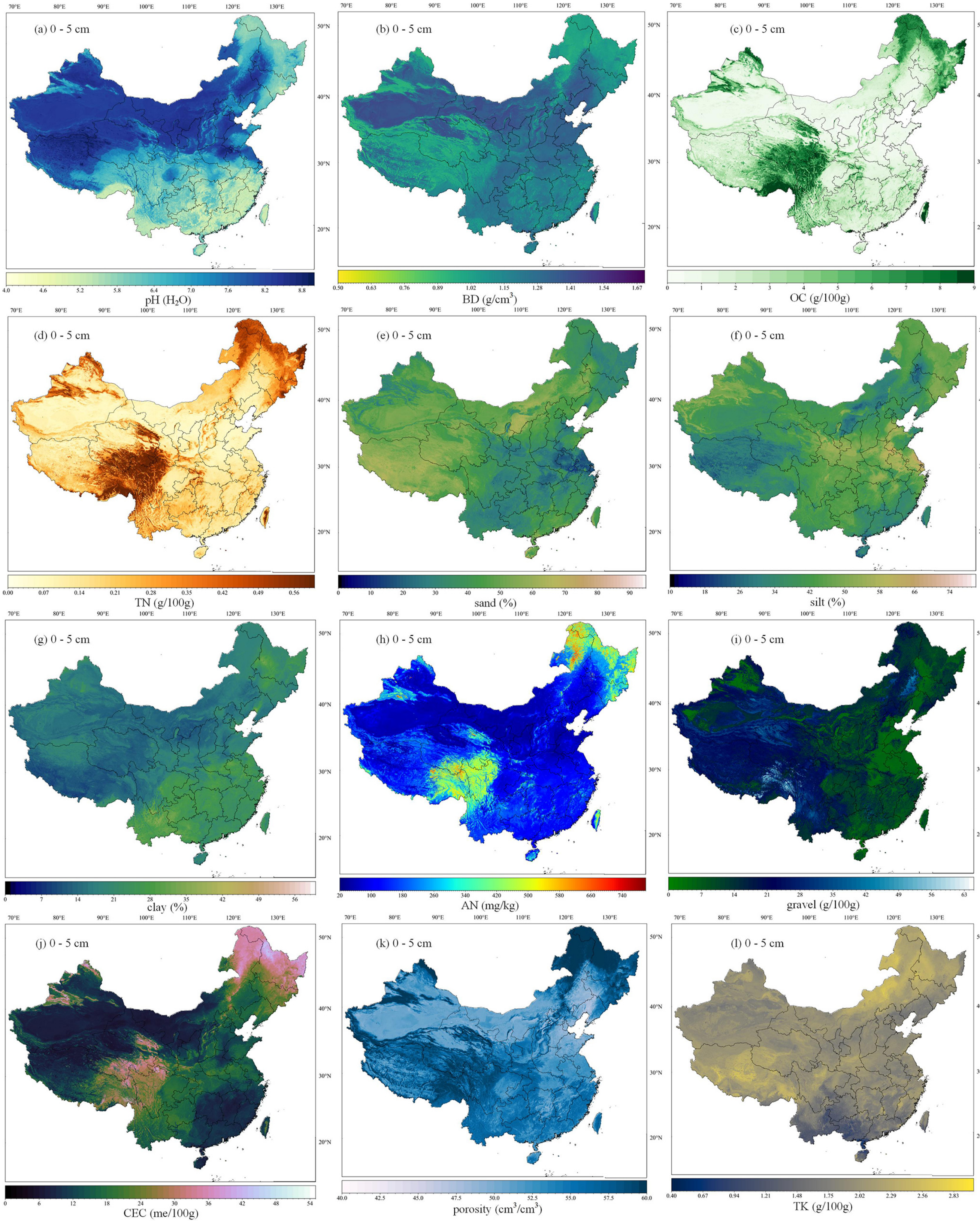
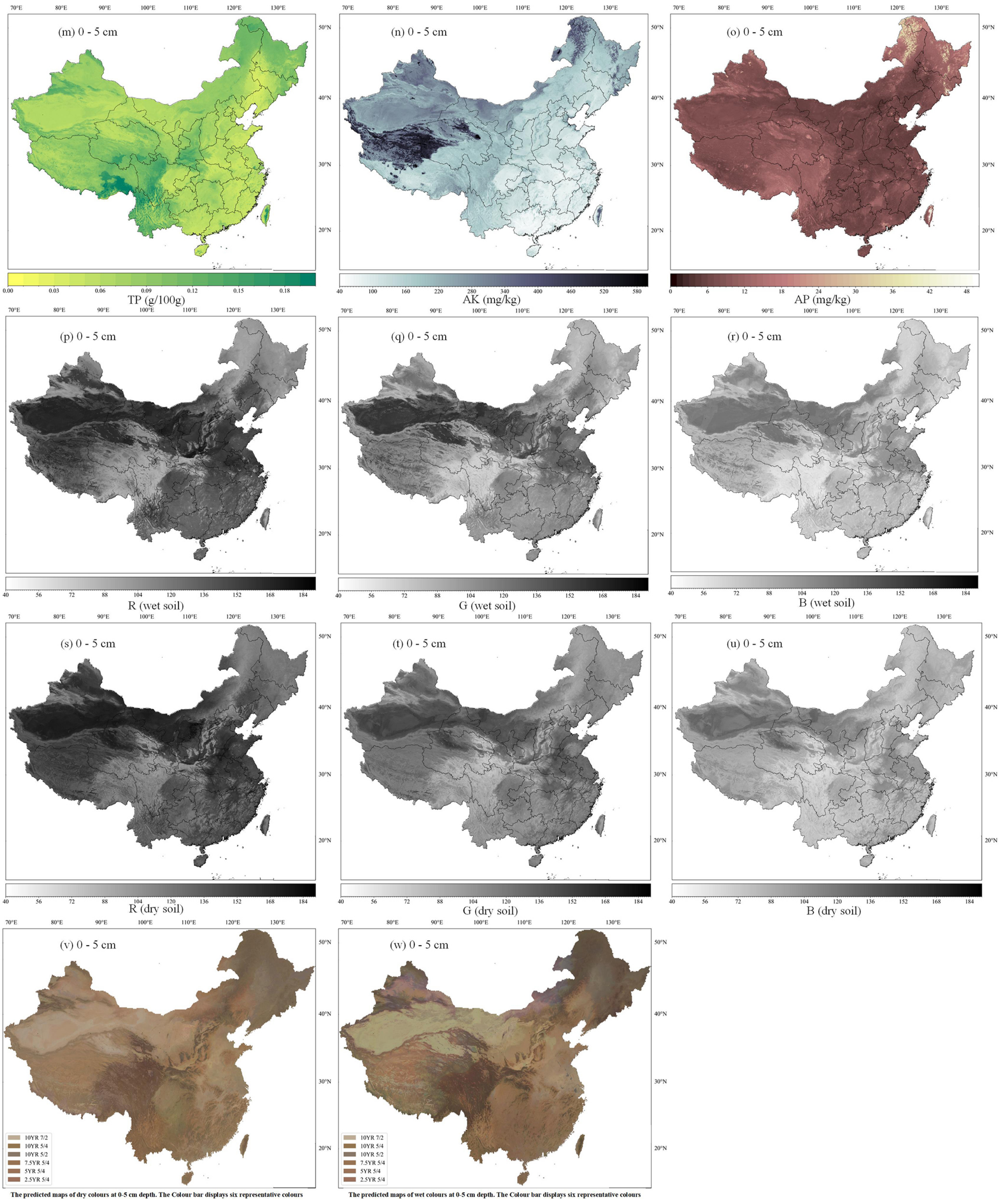

70°E
80°E
90°E
100°E
110°E
120°E
130°E
50°N
40°N
30°N
20°N
(a) 0 - 5 cm
4.0 4.6 5.2 5.8 6.4 7.0 7.6 8.2 8.8
pH (H_2O)
(b) 0 - 5 cm
0.50 0.63 0.76 0.89 1.02 1.15 1.28 1.41 1.54 1.67
BD (g/cm^3)
(c) 0 - 5 cm
0 1 2 3 4 5 6 7 8 9
OC (g/100g)
(d) 0 - 5 cm
0.00 0.07 0.14 0.21 0.28 0.35 0.42 0.49 0.56
TN (g/100g)
(e) 0 - 5 cm
0 10 20 30 40 50 60 70 80 90
sand (%)
(f) 0 - 5 cm
10 18 26 34 42 50 58 66 74
silt (%)
(g) 0 - 5 cm
0 7 14 21 28 35 42 49 56
clay (%)
(h) 0 - 5 cm
20 100 180 260 340 420 500 580 660 740
AN (mg/kg)
(i) 0 - 5 cm
0 7 14 21 28 35 42 49 56 63
gravel (g/100g)
(j) 0 - 5 cm
0 6 12 18 24 30 36 42 48 54
CEC (me/100g)
(k) 0 - 5 cm
40.0 42.5 45.0 47.5 50.0 52.5 55.0 57.5 60.0
porosity (cm^3/cm^3)
(l) 0 - 5 cm
0.40 0.67 0.94 1.21 1.48 1.75 2.02 2.29 2.56 2.83
TK (g/100g)

Figure 4.



The predicted maps of dry colours at 0-5 cm depth. The Colour bar displays six representative colours

The predicted maps of wet colours at 0-5 cm depth. The Colour bar displays six representative colours

Figure 4. The predicted maps of soil properties considered at the 0–5 cm depth interval for the land area of China. **(a)** pH (H_2O), **(b)** bulk density (BD), **(c)** soil organic carbon content (OC), **(d)** total nitrogen content (TN), **(e, f, g)** soil texture (sand, silt, clay content), **(h)** alkali-hydrolysable nitrogen content (AN), **(i)** rock fragment content (gravel), **(j)** cation exchange capacity content (CEC), **(k)** porosity, **(l)** total potassium content (TK), **(m)** total phosphorus content (TP), **(n)** available potassium content (AK), **(o)** available phosphorous content (AP), **(p, q, r)** wet colour (R, G, B), and **(s, t, u)** dry colour (R, G, B); **(v)** and **(w)** represent the dry and wet colours in the Munsell colour system, respectively. See Figs. S2–S24 in the Supplement for the predicted maps of soil properties at all depth intervals.

from southeast to northwest, corresponding to the influence of the southeastern monsoon. OC content is closely related to climatic conditions and land use practices (Zhang et al., 2023b; Zhou et al., 2019b). The spatial pattern of total nitrogen (TN) is similar to that of OC content. Areas with high precipitation and good vegetation cover tend to have higher OC and TN levels, while areas with low precipitation and poor vegetation cover tend to have lower OC and TN levels. This is because both OC and TN are closely related to organic matter input from vegetation. In regions with high vegetation productivity, organic matter contributes to both carbon and nitrogen accumulation in the soil, resulting in similar spatial patterns for OC and TN.

The mean predicted maps of soil texture (clay, silt, and sand contents) at different depths across China are shown in Fig. 5e–h, respectively. Overall, clay content was predicted to be low in the northern and northwestern regions while being higher in the southern regions. The lowest clay content was found in the deserts of the northwest, and the highest was found on the Yunnan–Guizhou Plateau. Relatively higher clay content was observed in some southern provinces, such as Guangdong and Guangxi. Silt content was predicted to be high on the Loess Plateau and in eastern China, while it was lower in the deserts of the northern and northwestern regions. These findings were consistent with previous studies (Liu et al., 2020). The predicted soil texture patterns fit well with the general characteristics and distribution of known Chinese soils (Gong et al., 2014).

For CEC, the spatial distribution of surface CEC is shown in Fig. 4j. CEC represents the total number of exchangeable cations that soil can absorb, serving as a crucial indicator of soil fertility, nutrient retention capacity, and buffering capacity, thereby influencing plant growth. Lower CEC value indicate that the soil can store fewer nutrients. The CEC levels are closely related to soil type, climatic conditions, and land use practices (Beillouin et al., 2022). Generally, soils with higher clay and organic matter content have higher CEC values. Figure 4j indicates that higher surface soil CEC values are found on the Qinghai–Tibet Plateau and in the peat and forest regions in the northeast (i.e. high-biomass or low-leaching areas). Lower CEC values are observed in the southeastern regions and in the arid and semi-arid areas in the north, with the lowest CEC values being found in desert areas. The relatively low CEC in the southeastern regions is attributed to higher temperatures and rainfall, leading to strong leaching loss of exchangeable substances.

The spatial distributions of TK, TP, and AK are shown in Fig. 4l, m, and n, respectively. Sedimentary rocks in southwestern China are abundant in phosphorus, leading to relatively higher TP levels in soils derived from these rocks. In contrast, southern China's soils typically exhibit lower TP levels due to extensive weathering and leaching. Alpine regions with significant organic matter accumulation are predicted to have relatively high TP content. The concentrations of both TK and AK generally diminish from north to south, despite their distribution patterns being rather different. Low levels of TK are found in tropical regions, whereas high levels are located on the Qinghai–Tibet Plateau and in northeastern China. High values of AK are dispersed throughout the western Tibetan Plateau. The spatial patterns of the variables of interest listed in Table 1 at multiple depths can be found in the Supplement (Figs. S2–S26). These spatial distributions are consistent with those reported in other similar studies (Hu et al., 2024; Liu et al., 2022a, Poggio et al., 2021).

3.4 Prediction uncertainty

Table S6 lists PICP values for different soil properties at multiple depths, calculated based on randomly held-back test samples. For a 90 % (or 0.9) confidence interval, 90 % of the observations are expected to fall within the predicted lower and upper limits. It can be seen that the PICP values for all soil properties at six standard depths are very close to 90 %, indicating that the predicted lower and upper limits estimated by the ensemble machine learning method are appropriate. In other words, the uncertainty estimates are largely reliable. It was observed that different soil properties exhibit distinct spatial patterns of prediction uncertainty, but different depths of the same soil property show similar patterns. The accuracy assessment in Figs. S2–S24 shows the uncertainty maps of soil property predictions. For OC, regions with relatively simple terrain, such as deserts, the North China Plain, and the Northeast Plain, exhibit lower uncertainty. In contrast, the central Qinghai–Tibet Plateau and western Inner Mongolia, where sampling is sparse and OC content is low, show higher uncertainty. The Altai region, with its complex terrain and diverse landscape types, also exhibits relatively high uncertainty. For soil pH, regions with high prediction uncertainty are found in southwestern China, where samples are sparse in complex soil landscapes. As soil depth increases, the uncertainty in predictions for properties like OC and pH generally decreases due to the more stable nature of subsurface layers, reduced influence from external factors, and the fact that deeper soils are less affected by environmental covariates. Additionally, while topsoil is more complex and variable due to its interaction with the environment, subsurface layers tend to have more consistent properties, leading to less uncertainty in predictions at depth (Liu et al., 2022a).

3.5 Relative importance of predictors

The relative importance of environmental covariates for soil property prediction at the 0–5 cm depth interval is shown in Figs. 6 and S26, displaying only the top 15 most important environmental covariates. Overall, organism-type covariates account for a significant proportion among different categories of environmental factors. There are variations in the relative importance of environmental covariates across different soil property variables.

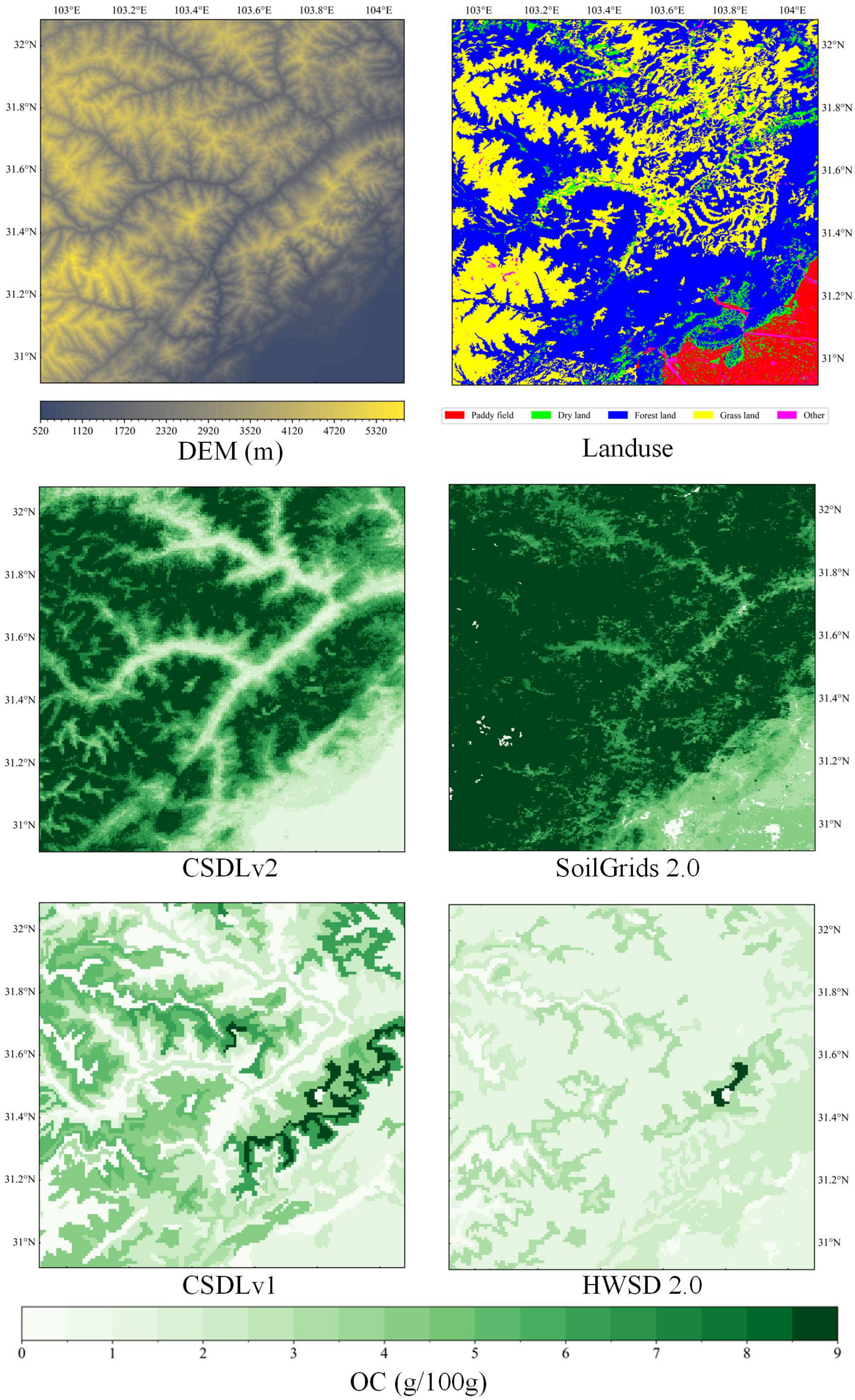


Figure 5. Surface layer (0–5 cm) soil organic carbon (OC) maps derived from our predictions (CSDLv2), SoilGrids 2.0, the CSDLv1, and the HWSD 2.0, respectively, in a selected area (30.92–32.08° N and 102.92–104.08° E) located in Sichuan Province. This selected area corresponds to the red window shown in Fig. 1. DEM and land use refer to the land surface elevation and land use type of the selected area, respectively. The spatial resolutions are 90 m for the CSDLv2, 250 m for SoilGrids 2.0, and 1 km for both the CSDLv1 and the HWSD 2.0.

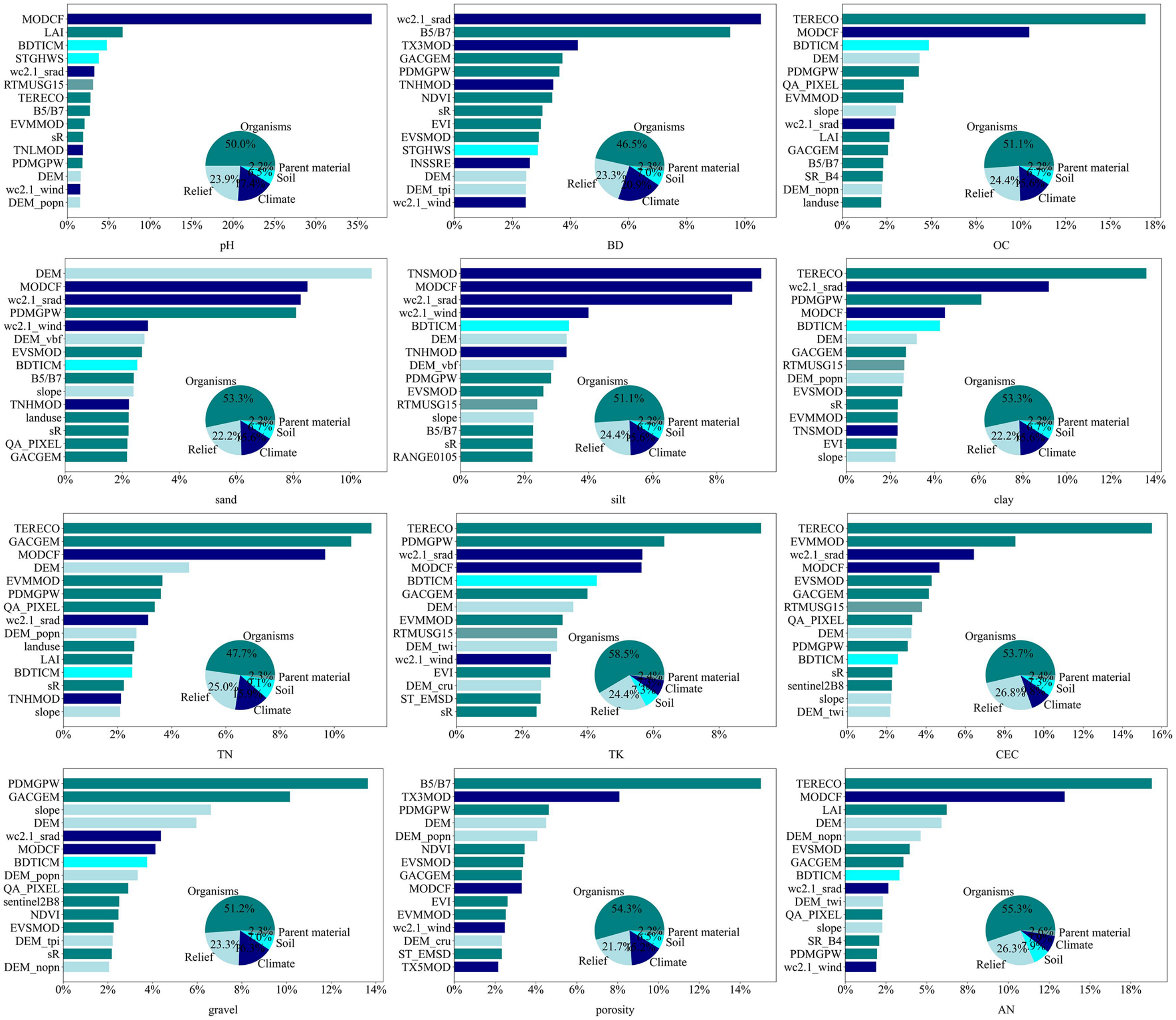


Figure 6. Relative importance of the top 15 predictors for the quantile regression forest model in the spatial predictions of soil pH, bulk density (BD), soil organic carbon (OC), soil texture (sand, silt, clay), total nitrogen (TN), total potassium (TK), cation exchange capacity (CEC), rock fragment (gravel), porosity, and alkali-hydrolysable nitrogen (AN) at the surface layer (0–5 cm). For other surface soil properties of interest, including total phosphorus (TP), available potassium (AK), available phosphorus (AP), dry colour (R, G, B), and wet colour (R, G, B), see Fig. S26. Refer to Table S1 in the Supplement for abbreviations of the environmental covariates.

For soil pH, in the optimal ensemble machine learning model, the climate factor (MODCF) was identified as the most important variable, with an importance exceeding 30 %, which is significantly higher than that of other covariates. The leaf area index (LAI) ranks second in relative importance. Previous studies have also indicated that LAI is a key factor in predicting soil pH (Sun et al., 2023). Other environmental covariates had relatively smaller contributions. In terms of covariates types, organism factors accounted for 50 % of the contribution to soil pH prediction, followed by relief factors (23.9 %) and climate factors (17.4 %).

For OC content, terrestrial ecosystems (TERECOs) and climate factors (MODCFs) are the most important covariates, followed by depth to bedrock and elevation (DEM). Shallow bedrock typically results in thinner soil layers, which can limit soil development and the accumulation of OC. However, soils developed on shallow bedrock do not always have low OC as the rate of OC accumulation can be significantly influenced by the type of vegetation present. In contrast, deeper bedrock allows for thicker soil layers, providing more space and time for OC accumulation. DEMs can indirectly reflect differences in land use and vegetation types, which can also affect the distribution of OC content. This indicates that the prediction of soil organic carbon is influenced by multiple factors. Many studies have shown that organisms factors (e.g. land use) are the most important predictor (Gomes et al., 2019).

For sand prediction, elevation and mean annual cloud frequency (MODCF) rank as the top two most important covariates in the ensemble machine learning model. Altitude primarily affects soil through gravitational and erosional processes, which transport fine particles and leave behind coarse particles (Li et al., 2023). This is evident in the relatively higher sand content in most mountainous areas compared to in adjacent lowland regions. Thermal processes drive physical weathering, while wind, water, and terrain govern erosion processes, predominantly shaping the distribution patterns of sand in China.

For silt prediction, climate-related factors (e.g. MODCF, and wc2.1_srad) are the most important covariates. Apart from climate, terrain factors (e.g. DEM, DEM_vbf, and slope) also play crucial roles in silt prediction. Terrain features largely determine gravitational and hydraulic conditions, thereby influencing the erosion, redistribution, and sorting processes of soil particles. This observation is consistent with previous studies (Hengl et al., 2017), indicating that climate data can enhance the predictive performance of soil texture models.

For clay prediction, organism-type covariates (e.g. TERECO, Table S1) rank as the most important environmental covariate, followed by the climatic variable wc2.1_srad. Terrain-related variables (e.g. DEM, DEM_popn, and slope) rank second in importance overall, exerting their influence by controlling local moisture and thermal conditions, as well as redistributing terrain material (Liu et al., 2020). Other studies have similarly shown that vegetation indices, rock type, bioclimatic zones, and agricultural indices can help characterise changes in soil clay content (Ge et al., 2019; Hengl et al., 2017).

For CEC prediction, the most important covariate is terrestrial ecosystems (i.e. TERECO). Plant roots can alter the chemical environment of the soil by secreting organic acids and other substances, which influence the dissolution and reprecipitation processes of soil minerals. These changes can affect the soil's CEC. Shiri et al. (2017) investigated the relationships of soil carbon content, clay content, and particle size with CEC. They found that higher organic carbon and clay content significantly enhance CEC due to their high specific surface areas and cation retention capacities. This is consistent with our findings, where areas with higher organic content, influenced by plant root activity, showed higher CEC values. The relative importance of the top 15 environmental covariates for other soil properties across all depths is visualised in Figs. S31–S52.

4 Discussion

4.1 Comparison with previous products

Tables 3, S4, and S5 present the accuracy assessments of our predictions (i.e. the CSDLv2), the CSDLv1 (Shangguan et al., 2013), SoilGrids 2.0 (Poggio et al., 2021), and the HWSD 2.0 (FAO and IIASA, 2023) at six standard depth intervals using data-splitting validation and independent-sample validation methods. Table 3 lists the validation accuracy of selected soil properties with the highest prediction accuracy using the data-splitting validation method, while Table S4 provides the complete accuracy assessments for all soil properties of interest. Table S5 identifies the variables for which the WoSIS database can serve as an independent sample. Overall, our predictions, whether using data-splitting validation or independent-sample validation, achieved relatively higher MEC values and lower RMSE values across multiple depths for most target variables, demonstrating much greater accuracy than existing soil property maps (FAO and IIASA, 2023; Poggio et al., 2021; Shangguan et al., 2013; Song et al., 2020; Zhou et al., 2019b). Specifically, using data-splitting validation as an example, our predictions for pH demonstrated an absolute improvement in the mean MEC for all layers, increasing from 0.60 to 0.70, while the RMSE decreased from 0.77 to 0.68 compared to SoilGrids 2.0. In comparison to the CSDLv1, our prediction performance for pH improved from 0.44 to 0.70, with the RMSE being reduced from 0.96 to 0.68. Compared to the HWSD 2.0, the prediction performance showed the greatest improvement in MEC and the most significant reduction in RMSE. The MEC values indicated that SoilGrids 2.0 significantly overestimated TN content, whereas the CSDLv1 and the HWSD 2.0 underestimated it. Additionally, in the independent validation (Table S5), across predictions of various soil properties at different depths, this study demonstrates overall predictive performance that is comparable to or better than SoilGrids 2.0, even though SoilGrids 2.0 used all the soil profiles of the WoSIS in its production. Moreover, it shows superior performance compared to the CSDLv1 and the HWSD 2.0.

Such a national-scale publication of soil maps hides most of the details. Nevertheless, because the soil properties are predicted at a 90 m resolution, portions of the maps can be enlarged to reveal increasingly detailed information up to the limit of that resolution. Using the example of surface (0–5 cm) OC content, Fig. 5 shows a visual comparison within a window of western Sichuan Province (30.92–32.08° N and 102.92–104.08° E). This window corresponds to the red window in Fig. 2a. The comparison is between the dataset developed in this study (CSDLv2) and the widely used SoilGrids 2.0, the CSDLv1, and the HWSD 2.0. The OC map produced in this study clearly reveals spatial variability with local morphology and provides more detailed information than the other three maps. Moreover, the CSDLv2 and SoilGrids 2.0 datasets, both products of advanced digital soil mapping techniques, exhibit notably higher OC content compared to the other two datasets generated through the linkage method across the majority of this region. This finding aligns well with our understanding of the area's environmental conditions: the cold climate at high elevations (Fig. 5, DEM), coupled with extensive forest and grassland cover (Fig. 5, land use), creates an ideal setting for the accumulation of OC in the soil. Figures S67–S71 show the spatial details of other soil properties, including TN, gravel, porosity AN, and AP. Therefore, the fine-soil property map, with a spatial resolution of 90 m, can better present the spatial variability of soil properties in related research, which can aid in precision agriculture and soil management.

To characterise the spatial pattern differences between the CSDLv2 and the CSDLv1, Fig. 7a, c, and e illustrate the spatial difference maps of OC, sand, and clay predictions in the CSDLv2 subtracted from those in the CSDLv1 as an example. For OC, the differences are mainly observed on the Tibetan Plateau, the Yunnan–Guizhou Plateau, and the Northeast Plain, where OC content is higher in the CSDLv2 than in the CSDLv1. For sand, the CSDLv2 shows relatively lower sand content in desert and semi-desert areas (e.g. Taklamakan Desert), while relatively higher sand content is observed in southern coastal regions. For clay, an opposite trend to sand is observed. The possible cause of these differences may be attributed to the linkage method used in developing the CSDLv1, which averaged all soil profiles for a given soil type or soil polygon, neglecting local spatial variation in soil properties. Additionally, as shown in Fig. 5, the two datasets derived by means of DSM technology (i.e. the CSDLv2 and SoilGrids 2.0) had similar spatial patterns and higher values than the other two, indicating an underestimation of OC content by the linkage method in this region. The scatterplots in Fig. 7b, d, and f show the comparison between the CSDLv2, the CSDLv1, and the observed data. From the bivariate kernel density estimates and correlation coefficients, it is evident that the CSDLv2 has a stronger correlation with the observed data. It can also be seen that the scatter points for the CSDLv1, based on the linkage method, are more dispersed, whereas the scatter points for the CSDLv2, based on DSM technology, are more concentrated. Compared to the CSDLv2, the CSDLv1 showed a significant underestimation of OC and a significant overestimation and underestimation of sand and clay, respectively. This may be due to the better fitting ability of DSM technology with the available data. However, the use of the ensemble learning algorithm, which averages predictions from multiple trees, tends to smooth out extreme values during spatial extrapolation, potentially reducing variability in certain regions. On the whole, the CSDLv2 provides a more accurate estimation of soil properties than the CSDLv1; thus, it may have significant influences on land surface modelling due to the large differences in spatial distribution. Further studies are needed to demonstrate the impact of the new soil dataset compared to the old version and compared to global soil datasets by running a land surface model (Li et al., 2020).

Based on the experimental results and analysis, compared to the CSDLv1, the main advantages of the CSDLv2 include the following aspects. First, the CSDLv2's spatial resolution is 90 m, aligning with the resolution of the most important input layers used for the predictions, and this is an improvement over the CSDLv1's 1 km resolution. This addresses the long-standing issue of lacking detailed and accurate soil information and enhances the modelling of energy, water, and momentum processes in the land surface model. Second, high-resolution environmental covariates related to soil formation were used with advanced machine learning algorithms, replacing traditional soil transformation rules. In recent years, digital soil mapping technology has made significant progress, particularly with the success of machine learning in large-scale spatial predictions (Poggio et al., 2021). Numerous studies have shown that advanced machine learning models typically have better predictive performance than simpler models (Yan et al., 2020). Third, an RGB soil colour system (i.e. red, green, and blue) has been added, resolving the inconvenience of only having the Munsell colour system in the first-edition dataset. This addition enhances the visual representation of soil colours and allows for better integration with digital platforms, remote sensing applications, and computer displays (Al-Naji et al., 2021). Finally, global validation was conducted using data splitting and independent samples, and prediction uncertainty was quantitatively provided using QRF rather than merely offering quality control information. Compared to other related data products, the CSDLv2 encompasses more than 20 comprehensive soil physical and chemical properties, whereas most existing studies focus on mapping one or several fundamental soil properties, lacking comprehensive soil property dataset products (Liang et al., 2019; Chen et al., 2019; Zhou et al., 2019a; Liu et al., 2022a; Liu et al., 2020). For instance, AN serves as an indicator of soil fertility, reflecting the potential release of organic nitrogen and ammonium nitrogen in the soil. AK reflects the potassium available for plant uptake, which is crucial for plant growth and development. The extensive soil information has significant applications across various fields. Additionally, another advantage of the CSDLv2 over both the CSDLv1 and other related data products is that a large number of soil profile samples from different data sources were collected, enhancing the spatial representativeness of the soil profiles. Sample size is a critical factor affecting model performance (Padarian et al., 2020).

4.2 Potential applications of the CSDLv2

The national-scale high-resolution soil property maps developed in this study have significant potential for applications in land surface modelling and Earth system modelling. These models simulate interactions between the land surface, atmosphere, and biosphere, making accurate representation of

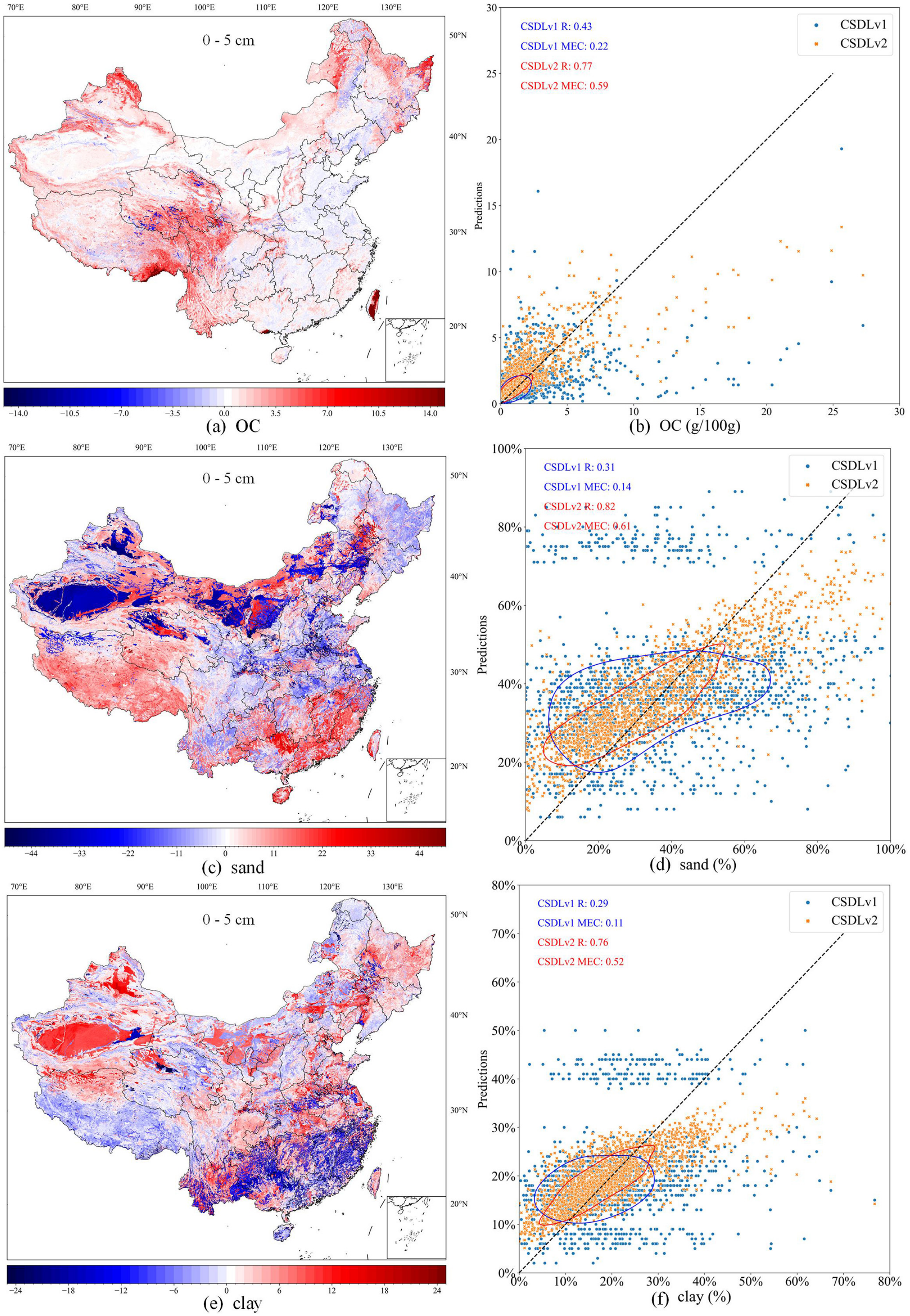


Figure 7. Differences in predicted maps of soil organic carbon **(a)**, sand **(c)**, and clay **(e)** between the CSDLv2 and CSDLv1 at the 0–5 cm depth interval and the corresponding scatterplots **(b, d, f)** indicating how well the predictions of the CSDLv2 and CSDLv1 match the observations. The red and blue circles are bivariate kernel density estimates. Publisher's remark: please note that the above figure contains disputed territories.

soil properties essential for improving model performance and predictions. For instance, soil pH is crucial for nutrient solubility, while CEC indicates fertility and nutrient retention capacity in land surface modelling. In biogeochemical process modelling with land surface modelling, OC, TN, and TP are key parameters and prognostic variables. These soil nutrients can be calculated by running models for thousands of years until an equilibrium state is reached, a process known as model "spin-up" (i.e. a warm-up period) (Dai et al., 2019b; Shangguan et al., 2013). However, the nonlinear feedbacks in biogeochemical cycles make such spin-up time-consuming and less reliable for initialising soil nutrients. Therefore, this dataset can also serve as an important benchmark for initial or calibration variables.

Currently, many soil properties are not yet utilised in land surface model simulations, with only soil texture, OC, gravel, and BD being primarily used. However, more soil properties can theoretically be employed as initial variables in Earth system modelling. Each soil property plays an important role in both Earth system modelling and land surface modelling, and although some properties are not yet used, they hold significant potential for future applications. For example, soil albedo is significantly correlated with the Munsell soil colour value (hue, value, chroma). In some Earth system models, parameters derived from pedotransfer functions are used directly as inputs rather than being calculated within the models.

Moreover, the CSDLv2 offers extensive possibilities for research and applications across various fields, including climate change research and carbon cycling (Chen et al., 2023), as well as support for the spatial delineation of management zones in precision agriculture (Piikki et al., 2017). Regarding soil pH, for agricultural departments and farmers, fine mapping of soil pH holds significant value in local and field land use planning and management as different crops exhibit optimal growth in soils with varying pH ranges (Hu et al., 2024). For instance, rice thrives best in soils with pH levels between 6.0 and 7.5, whereas peanuts prefer soils with pH levels between 5.6 and 6.0. Thus, precise soil pH maps provide essential information for agricultural zoning and management. Furthermore, due to the widespread applicability of soil information, the CSDLv2 also holds potential applications in numerous other fields.

4.3 Limitations and outlook

Some advances have been made in this study, but several limitations still need to be addressed in future efforts. First, remote sensing imagery has been used globally for soil property mapping (Guo et al., 2022; Xia and Zhang, 2022). With the advancement of remote sensing technology, more and more high-spatial-resolution free data have become available. For example, Xia and Zhang (2022) found that using high-spatial-resolution GF-2 imagery improved soil property prediction accuracy compared to medium-resolution imagery (e.g. Landsat 8 and Sentinel-2 imagery). Therefore, future digital soil mapping work can focus more on integrating high-resolution remote sensing products, which can enable models to capture the complex statistical relationships between soil properties and environmental covariates at fine scales (Mulder et al., 2016).

Secondly, soil is a three-dimensional volume with property variability in all three dimensions. In this study, the vertical dimension of soil variability was modelled using spline interpolation. It is noteworthy that smoothing spline interpolation standardises soil layer data, which are not error-free, but due to the lack of a "true" depth function for each soil profile (vertically dense samples), the standardisation error cannot be quantitatively estimated (Liu et al., 2022a). Recent publications have considered observation depth as a covariate (Hengl et al., 2017; Nauman and Duniway, 2019), creating a "3D" model, but some studies indicate that this approach may be overly simplistic or may lead to consistency issues in the predicted depth sequences (Ma et al., 2021). This might be true for local datasets, where short-range spatial variability and vertical variability have similar magnitudes (Poggio et al., 2021). Further research is needed to assess the impact of using depth as a covariate on national datasets and models. Additionally, alternatives such as 3D models or geostatistical models utilising 3D spatial autocorrelation are worth exploring (Helfenstein, 2024).

Thirdly, in this study, approximately 150 covariates related to soil properties, topography, climate, biomes, lithology, land use, and existing soil maps were collected. By removing inter-variable correlations and using recursive feature elimination, approximately 40 optimal variables were selected to map soil properties across the country. However, the original environmental variables with a resolution of 90 m did not play a significant role in variable selection or importance ranking. Several reasons may explain this. First, many studies have confirmed that soil properties (e.g. soil pH) are highly correlated with lithology (e.g. soil group and parent material) and climatic factors, especially at large scales (Hu et al., 2024; Lu et al., 2023). Topography downscaling methods can be used to prepare high-resolution climate covariates (Chen et al., 2024). However, fine and reliable maps of these factors are typically unavailable, especially at large spatial scales. Therefore, when introducing these factors to map soil properties, coarse-resolution raster data (e.g. 1 km) often have to be used (Liu et al., 2022a; Lu et al., 2023). Secondly, in this study, some covariates (e.g. elevation and slope) with an original resolution of 90 m are highly correlated with soil properties (e.g. soil pH). However, these factors are also highly correlated with other factors such as mean annual temperature and mean annual precipitation (Guo et al., 2022). These factors were removed by the recursive feature elimination algorithm when selecting the optimal variables because they were highly correlated with the already retained existing variables. This also led to the relatively lower importance of these factors in contributing to the models for soil properties (e.g. soil pH). Therefore, the final maps of soil properties with a 90 m resolution in this study will be useful for practical decision-making. In future work, introducing fine-resolution environmental covariates is expected to improve mapping accuracy.

Last but not least, although this study utilised multi-source soil profile data from different time periods to develop comprehensive static maps of soil properties, the CSDLv2 maps mainly represent the status of soil in the 1980s as most soil profiles come from the SNSSC. For soil properties that change over time, other multi-source soil profile data have not been fully utilised. Together with maps based on data from other periods, such as the 2010s, as in Liu et al. (2022a), the CSDLv2 could provide new perspectives for studying temporal changes in soil properties. However, more efforts are needed to model the temporal change in soil properties with more time slices, especially for those soil properties which may change in the short term. Considering this aspect, the undergoing Third National Soil Survey of China and other legacy soil profiles should be exploited to map time series of soil properties using spatiotemporal modelling technology. As the CSDLv2 is developed on the national scale, the maps are suitable for broad-scale applications, such as national-scale and large-regional-scale (e.g. provincial-level) analyses. Although generated at a high resolution (90 m), these maps may not provide sufficient accuracy for farm- or field-scale applications, where locally calibrated models and detailed surveys are recommended. Users should consider the provided accuracy metrics and uncertainty maps to assess suitability for specific applications (Helfenstein et al., 2024).

5 Code and data availability

All the resources for the ensemble machine learning model, including training and testing code, are publicly available at https://doi.org/10.5281/zenodo.14783774 (Shi and Shangguan, 2025). The soil maps in this study for six depth layers (0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm) at a 90 m spatial resolution across China are openly accessible: https://www.scidb.cn/s/ZZJzAz (last access: 17 November 2024) or https://doi.org/10.11888/Terre.tpdc.301235 (Shi and Shangguan, 2024). Users can efficiently download the datasets provided in the first link of the above statement by using the file transfer protocol (FTP) account information provided at the above links and common FTP client tools such as Filezilla (https://filezilla-project.org/, last access: 15 June 2024) or FlashFXP (https://www.flashfxp.com/, last access: 15 June 2024).

To meet the spatial resolution requirements of different applications, the CSDLv2 not only provides soil properties at a 90 m resolution but also offers 1 and 10 km resolution data, with maps of the mean, median, 0.05 and 0.95 quantiles. These 1 and 10 km resolution data were derived from spatial predictions made by the constructed model using environmental covariates at the corresponding resolutions. The dataset is provided in raster format, available in both network Common Data Form 4 (NetCDF4) and GeoTIFF (GTiff) formats.

6 Conclusions

The second version of the high-resolution national soil information grid for China was developed in this study, utilising a vast number of multi-source legacy soil profile samples and advanced machine learning techniques, as a replacement for the first version of the dataset. This version includes over 20 soil physical and chemical properties, with prediction maps for each soil property covering six standard depths (0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm). By combining ensemble machine learning with currently available high-resolution environmental covariates, the spatial variations of soil properties across China and at different depths can be effectively predicted. Overall, all the soil property maps performed well, accurately representing the spatial variations of soil properties. Under both data-splitting and independent-sample schemes, the CSDLv2 generally outperformed other gridded soil datasets, including the CSDLv1, SoilGrids 2.0, and the HWSD 2.0. The CSDLv2 provided more spatial details and better represented the spatial-variation characteristics of soil properties in China compared to other soil products. Furthermore, as this dataset is primarily based on legacy soil profiles from the Second National Soil Survey of China and describes the state of soil properties in the 1980s, it serves as a valuable complement to maps based on soil profiles from the 2010s, providing new perspectives for studying temporal changes in soil properties. These prediction maps also contribute to China's input to the GlobalSoilMap project and can be used for various hydrological and ecological analyses and for regional Earth system modelling, especially for applications requiring high-resolution soil property maps. Future work can improve soil property mapping by employing advanced deep learning methods and incorporating more observations, particularly in regions with sparse samples, like western China. Additionally, integrating high-resolution remote sensing data, developing more accurate 3D models, and accounting for temporal changes in soil properties will further enhance the mapping accuracy and usefulness of the CSDLv2.

Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/essd-17-517-2025-supplement.

Author contributions. WeiG conceived the research and secured funding for the research. GS and WeiG performed the analyses. GS conducted the research and wrote the initial draft of the paper. WeiG

and GS reviewed and edited the paper before submission. All the other authors joined the discussion of the research.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgements. The authors are grateful to all the data contributors who made it possible to complete this research.

Financial support. This work was supported by the National Natural Science Foundation of China (grant nos. 42375144, 4227515 and 42205149), and the Guangdong Major Project of Basic and Applied Basic Research (grant no. 2021B0301030007).

Review statement. This paper was edited by Hao Shi and reviewed by two anonymous referees.

References

Adhikari, K., Kheir, R. B., Greve, M. B., Bøcher, P. K., Malone, B. P., Minasny, B., McBratney, A. B., and Greve, M. H.: High-Resolution 3-D Mapping of Soil Texture in Denmark, Soil Sci. Soc. Am. J., 77, 860–876, https://doi.org/10.2136/sssaj2012.0275, 2013.

Al-Naji, A., Fakhri, A. B., Gharghan, S. K., and Chahl, J.: Soil color analysis based on a RGB camera and an artificial neural network towards smart irrigation: A pilot study, Heliyon, 7, e06078, https://doi.org/10.1016/j.heliyon.2021.e06078, 2021.

Arrouays, D., Grundy, M. G., Hartemink, A. E., Hempel, J. W., Heuvelink, G. B. M., Hong, S. Y., Lagacherie, P., Lelyk, G., McBratney, A. B., McKenzie, N. J., Mendonca-Santos, M. d.L., Minasny, B., Montanarella, L., Odeh, I. O. A., Sanchez, P. A., Thompson, J. A., and Zhang, G.-L.: GlobalSoilMap: Toward a Fine-Resolution Global Grid of Soil Properties, in: Advances in Agronomy, vol. 125, Elsevier, 93–134, https://doi.org/10.1016/B978-0-12-800137-0.00003-0, 2014a.

Arrouays, D., McKenzie, N., Hempel, J., Richer de Forges, A., and McBratney, A. B. (Eds.): GlobalSoilMap: Basis of the global spatial soil information system, 1st edn., CRC Press, https://doi.org/10.1201/b16500, 2014b.

Arrouays, D., McKenzie, N. J., and Hartemink, A. E.: The GlobalSoilMap project specifications D. Notes, in: Proceedings of the 1st GlobalSoilMap Conference, Orléans, France, 7–9 October 2013, https://www.iuss.org/wp-content/uploads/2024/02/iuss_bulletin_123.pdf (last access: 25 July 2024), 2015.

Arrouays, D., Savin, I., Leenaars, J., and McBratney, A. B.: GlobalSoilMap-Digital Soil Mapping from Country to Globe, in: Proceedings of the Global Soil Map 2017 Conference, Moscow, Russia, 4–6 July 2017, https://doi.org/10.1201/9781351239707, 2017.

Batjes, N. H.: A global data set of soil pH properties, Tech. Pap., 27, Int. Soil Ref. and Int. Soil Ref. And Inf. Cent (ISRIC), Wageningen, Netherlands, https://www.isric.org/sites/default/files/ISRIC_TechPap27.pdf (last access: 25 July 2024), 1995.

Batjes, N. H.: Soil parameter estimates for the soil types of the world for use in global and regional modelling (Version 2.1), ISRIC Rep. 2002/02c, Int. Food Policy Res. Inst. (IFPRI) and Int. Soil Ref. Inf. Cent. (ISRIC), Wageningen, Netherlands, https://www.isric.org/sites/default/files/isric_report_2002_02c.pdf (last access: 25 July 2024), 2002.

Batjes, N. H., Ribeiro, E., and van Oostrum, A.: 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, 2020.

Beillouin, D., Demenois, J., Cardinael, R., Berre, D., Corbeels, M., Fallot, A., Boyer, A., and Feder, F.: A global database of land management, land-use change and climate change effects on soil organic carbon, Sci. Data, 9, 228, https://doi.org/10.1038/s41597-022-01318-1, 2022.

Bishop, T. F. A., McBratney, A. B., and Laslett, G. M.: 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, 1999.

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Brus, D. J., Kempen, B., and Heuvelink, G. B. M.: Sampling for validation of digital soil maps, Eur. J. Soil Sci., 62, 394–407, https://doi.org/10.1111/j.1365-2389.2011.01364.x, 2011.

Chen, S., Li, L., Wei, Z., Wei, N., Zhang, Y., Zhang, S., Yuan, H., Shangguan, W., Zhang, S., Li, Q., and Dai, Y.: Exploring Topography Downscaling Methods for Hyper-Resolution Land Surface Modeling, J. Geophys. Res.-Atmos., 129, e2024JD041338, https://doi.org/10.1029/2024JD041338, 2024.

Chaney, N. W., Minasny, B., Herman, J. D., Nauman, T. W., Brungard, C. W., Morgan, C. L. S., McBratney, A. B., Wood, E. F., and Yimam, Y.: POLARIS Soil Properties: 30-m Probabilistic Maps of Soil Properties Over the Contiguous United States, Water Resour. Res., 55, 2916–2938, https://doi.org/10.1029/2018WR022797, 2019.

Chen, S., Liang, Z., Webster, R., Zhang, G., Zhou, Y., Teng, H., Hu, B., Arrouays, D., and Shi, Z.: A high-resolution map of soil pH in China made by hybrid modelling of sparse soil data and environmental covariates and its implications for pollution, Sci. Total Environ., 655, 273–283, https://doi.org/10.1016/j.scitotenv.2018.11.230, 2019.

Chen, Z., Shuai, Q., Shi, Z., Arrouays, D., Richer-de-Forges, A. C., and Chen, S.: National-scale mapping of soil organic carbon stock in France: New insights and lessons learned by direct and indirect approaches, Soil & Environmental Health, 1, 100049, https://doi.org/10.1016/j.seh.2023.100049, 2023.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., and Böhner, J.: System for Automated Geoscientific Analyses (SAGA) v. 2.1.4, Geosci. Model

Dev., 8, 1991–2007, https://doi.org/10.5194/gmd-8-1991-2015, 2015.

Crow, W. T., Berg, A. A., Cosh, M. H., Loew, A., Mohanty, B. P., Panciera, R., De Rosnay, P., Ryu, D., and Walker, J. P.: Upscaling sparse ground-based soil moisture observations for the validation of coarse-resolution satellite soil moisture products, Rev. Geophys., 50, 2011RG000372, https://doi.org/10.1029/2011RG000372, 2012.

DAAC, O.: MODIS and VIIRS Land Products Global Subsetting and Visualization Tool. In, https://modis.gsfc.nasa.gov (last access: 3 April 2024), 2018.

Dai, Y., Shangguan, W., Wei, N., Xin, Q., Yuan, H., Zhang, S., Liu, S., Lu, X., Wang, D., and Yan, F.: 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, 2019b.

Dinamarca, D. I., Galleguillos, M., Seguel, O., and Faúndez Urbina, C.: CLSoilMaps: A national soil gridded database of physical and hydraulic soil properties for Chile, Sci. Data, 10, 630, https://doi.org/10.1038/s41597-023-02536-x, 2023.

Fan, J., Wu, L., Zhang, F., Xiang, Y., and Zheng, J.: Climate change effects on reference crop evapotranspiration across different climatic zones of China during 1956–2015, J. Hydrol., 542, 923–937, https://doi.org/10.1016/j.jhydrol.2016.09.060, 2016.

FAO and IIASA: Harmonized World Soil Database version 2.0, FAO, International Institute for Applied Systems Analysis (IIASA) [data set], https://doi.org/10.4060/cc3823en, 2023.

Ge, N., Wei, X., Wang, X., Liu, X., Shao, M., Jia, X., Li, X., and Zhang, Q.: Soil texture determines the distribution of aggregate-associated carbon, nitrogen and phosphorous under two contrasting land use types in the Loess Plateau, CATENA, 172, 148–157, https://doi.org/10.1016/j.catena.2018.08.021, 2019.

Gomes, L. C., Faria, R. M., de Souza, E., Veloso, G. V., Schaefer, C. E. G., and Fernandes Filho, E. I.: Modelling and mapping soil organic carbon stocks in Brazil, Geoderma, 340, 337–350, 2019.

Gong, C., Ma, L., Cheng, H., Liu, Y., Xu, D., Li, B., Liu, F., Ren, Y., Liu, Z., Zhao, C., Yang, K., Nie, H., and Lang, C.: Characterization of the particle size fraction associated heavy metals in tropical arable soils from Hainan Island, China, J. Geochem. Explor., 139, 109–114, https://doi.org/10.1016/j.gexplo.2013.01.002, 2014.

Grundy, M. J., Rossel, R. A. V., Searle, R. D., Wilson, P. L., Chen, C., and Gregory, L. J.: Soil and Landscape Grid of Australia, Soil Res., 53, 835, https://doi.org/10.1071/SR15191, 2015.

Guo, J., Wang, K., and Jin, S.: Mapping of Soil pH Based on SVM-RFE Feature Selection Algorithm, Agronomy, 12, 2742, https://doi.org/10.3390/agronomy12112742, 2022.

Gyamerah, S. A., Ngare, P., and Ikpe, D.: Probabilistic forecasting of crop yields via quantile random forest and Epanechnikov Kernel function, Agr. Forest Meteorol., 280, 107808, https://doi.org/10.1016/j.agrformet.2019.107808, 2020.

Hartmann, J. and Moosdorf, N.: Global Lithological Map Database v1.0 (gridded to 0.5° spatial resolution), PANGAEA [data set], https://doi.org/10.1594/PANGAEA.788537, 2012.

Helfenstein, A., Mulder, V. L., Heuvelink, G. B. M., and Hack-ten Broeke, M. J. D.: Three-dimensional space and time mapping reveals soil organic matter decreases across anthropogenic landscapes in the Netherlands, Commun. Earth Environ., 5, 130, https://doi.org/10.1038/s43247-024-01293-y, 2024a.

Helfenstein, A., Mulder, V. L., Hack-ten Broeke, M. J. D., Van Doorn, M., Teuling, K., Walvoort, D. J. J., and Heuvelink, G. B. M.: BIS-4D: mapping soil properties and their uncertainties at 25 m resolution in the Netherlands, Earth Syst. Sci. Data, 16, 2941–2970, https://doi.org/10.5194/essd-16-2941-2024, 2024b.

Hengl, T., Heuvelink, G. B. M., Kempen, B., Leenaars, J. G. B., Walsh, M. G., Shepherd, K. D., Sila, A., MacMillan, R. A., Mendes De Jesus, J., Tamene, L., and Tondoh, J. E.: Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions, PLoS ONE, 10, e0125814, https://doi.org/10.1371/journal.pone.0125814, 2015.

Hengl, T., Mendes De Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., 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., and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning, PLoS ONE, 12, e0169748, https://doi.org/10.1371/journal.pone.0169748, 2017.

Hengl, T., Miller, M. A. E., Križan, J., Shepherd, K. D., Sila, A., Kilibarda, M., Antonijević, O., Glušica, L., Dobermann, A., Haefele, S. M., McGrath, S. P., Acquah, G. E., Collinson, J., Parente, L., Sheykhmousa, M., Saito, K., Johnson, J.-M., Chamberlin, J., Silatsa, F. B. T., Yemefack, M., Wendt, J., MacMillan, R. A., Wheeler, I., and Crouch, J.: African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning, Sci. Rep., 11, 6130, https://doi.org/10.1038/s41598-021-85639-y, 2021.

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

Hu, B., Xie, M., Shi, Z., Li, H., Chen, S., Wang, Z., Zhou, Y., Ni, H., Geng, Y., Zhu, Q., and Zhang, X.: Fine-resolution mapping of cropland topsoil pH of Southern China and its environmental application, Geoderma, 442, 116798, https://doi.org/10.1016/j.geoderma.2024.116798, 2024.

Karger, D. N., Schmatz, D. R., Dettling, G., and Zimmermann, N. E.: High-resolution monthly precipitation and temperature time series from 2006 to 2100, Sci. Data, 7, 248, https://doi.org/10.1038/s41597-020-00587-y, 2020.

Katschinski, N. A.: Die mechanische Bodenanalyse und die Klassifikation der Böden nach ihrer mechanischen Zusammensetzung, Pari, B, 321–327, 1956.

Koenker, R.: Quantile Regression, Cambridge University Press, https://doi.org/10.1017/CBO9780511754098, 2005.

Krause, P., Boyle, D. P., and Bäse, F.: Comparison of different efficiency criteria for hydrological model assessment, Adv. Geosci., 5, 89–97, https://doi.org/10.5194/adgeo-5-89-2005, 2005.

Lagacherie, P., Arregui, M., and Fages, D.: Evaluating the quality of soil legacy data used as input of digital soil mapping models, Eur. J. Soil Sci., 75, e13463, https://doi.org/10.1111/ejss.13463, 2024.

Lamichhane, S., Kumar, L., and Wilson, B.: Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review, Geoderma, 352, 395–413, https://doi.org/10.1016/j.geoderma.2019.05.031, 2019.

Li, Q., Zhang, C., Shangguan, W., Li, L., and Dai, Y.: A novel local-global dependency deep learning model for soil mapping, Geoderma, 438, 116649, https://doi.org/10.1016/j.geoderma.2023.116649, 2023.

Li, T., Cui, L., Kuhnert, M., McLaren, T. I., Pandey, R., Liu, H., Wang, W., Xu, Z., Xia, A., Dalal, R. C., and Dang, Y. P.: A comprehensive review of soil organic carbon estimates: Integrating remote sensing and machine learning technologies, J. Soil. Sediment., 24, 3556–3571, https://doi.org/10.1007/s11368-024-03913-8, 2024.

Li, W., Wei, N., Huang L., and Shangguan W.: Impact of Soil Datasets on the Global Simulation of Land Surface Processes, Climatic and Environmental Research, 25, 555–574, https://doi.org/10.3878/j.issn.1006-9585.2020.20025, 2020 (in Chinese).

Liang, Z., Chen, S., Yang, Y., Zhao, R., Shi, Z., and Viscarra Rossel, R. A.: National digital soil map of organic matter in topsoil and its associated uncertainty in 1980's China, Geoderma, 335, 47–56, https://doi.org/10.1016/j.geoderma.2018.08.011, 2019.

Lilburne, L., Helfenstein, A., Heuvelink, G. B. M., and Eger, A.: 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, 2024.

Liu, F., Zhang, G.-L., Song, X., Li, D., Zhao, Y., Yang, J., Wu, H., and Yang, F.: High-resolution and three-dimensional mapping of soil texture of China, Geoderma, 361, 114061, https://doi.org/10.1016/j.geoderma.2019.114061, 2020.

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

Liu, F., Yang, F., Zhao, Y., Zhang, G., and Li, D.: Predicting soil depth in a large and complex area using machine learning and environmental correlations, J. Integr. Agr., 21, 2422–2434, https://doi.org/10.1016/S2095-3119(21)63692-4, 2022b.

Lu, Q., Tian, S., and Wei, L.: Digital mapping of soil pH and carbonates at the European scale using environmental variables and machine learning, Sci. Total Environ., 856, 159171, https://doi.org/10.1016/j.scitotenv.2022.159171, 2023.

Luo, Y., Ahlström, A., Allison, S. D., Batjes, N. H., Brovkin, V., Carvalhais, N., Chappell, A., Ciais, P., Davidson, E. A., Finzi, A., Georgiou, K., Guenet, B., Hararuk, O., Harden, J. W., He, Y., Hopkins, F., Jiang, L., Koven, C., Jackson, R. B., Jones, C. D., Lara, M. J., Liang, J., McGuire, A. D., Parton, W., Peng, C., Randerson, J. T., Salazar, A., Sierra, C. A., Smith, M. J., Tian, H., Todd-Brown, K. E. O., Torn, M., Van Groenigen, K. J., Wang, Y. P., West, T. O., Wei, Y., Wieder, W. R., Xia, J., Xu, X., Xu, X., and Zhou, T.: Toward more realistic projections of soil carbon dynamics by Earth system models, Global Biogeochem. Cy., 30, 40–56, https://doi.org/10.1002/2015GB005239, 2016.

Ma, Y., Minasny, B., McBratney, A., Poggio, L., and Fajardo, M.: Predicting soil properties in 3D: Should depth be a covariate?, Geoderma, 383, 114794, https://doi.org/10.1016/j.geoderma.2020.114794, 2021.

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

McBratney, A. B., Minasny, B., and Stockmann, U. (Eds.): Pedometrics, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-319-63439-5, 2018.

Meinshausen, N.: Quantile Regression Forests, J. Mach. Learn. Res., 7, 983–999, 2006.

Moreira De Sousa, L., Poggio, L., and Kempen, B.: Comparison of FOSS4G Supported Equal-Area Projections Using Discrete Distortion Indicatrices, ISPRS Int. J. Geo-Inf., 8, 351, https://doi.org/10.3390/ijgi8080351, 2019.

Mulder, V. L., Lacoste, M., Richer-de-Forges, A. C., and Arrouays, D.: GlobalSoilMap France: High-resolution spatial modelling the soils of France up to two meter depth, Sci. Total Environ., 573, 1352–1369, https://doi.org/10.1016/j.scitotenv.2016.07.066, 2016.

Nachtergaele, F. O., van Velthuizen, H., Verelst, L., Batjes, N. H., Dijkshoorn, J. A., van Engelen, V. W. P., Fischer, G., Jones, A., Montanarella, L., Petri, M., Prieler, S., Teixeira, E., Wilberg, D., and Shi, X.: Harmonized World Soil Database (version 1.0), ISMC [data set], https://soil-modeling.org/resources-links/data-portal/harmonized-world-soil-database, 2012.

National Soil Survey Office: Agricultural Soils in China, China Agricultural Press, Beijing, 1964.

National Soil Survey Office: Chinese Soil Genus Records, vol. 6, China Agriculture Press, Beijing, 1996 (in Chinese).

Nauman, T. W. and Duniway, M. C.: Relative prediction intervals reveal larger uncertainty in 3D approaches to predictive digital soil mapping of soil properties with legacy data, Geoderma, 347, 170–184, https://doi.org/10.1016/j.geoderma.2019.03.037, 2019.

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M. E., and Papritz, A.: Evaluation of digital soil mapping approaches with large sets of environmental covariates, SOIL, 4, 1–22, https://doi.org/10.5194/soil-4-1-2018, 2018.

Padarian, J., Minasny, B., and McBratney, A. B.: Chile and the Chilean soil grid: A contribution to GlobalSoilMap, Geoderma Regional, 9, 17–28, https://doi.org/10.1016/j.geodrs.2016.12.001, 2017.

Padarian, J., Minasny, B., and McBratney, A. B.: Machine learning and soil sciences: a review aided by machine learning tools, SOIL, 6, 35–52, https://doi.org/10.5194/soil-6-35-2020, 2020.

Piikki, K., Söderström, M., and Stadig, H.: Local adaptation of a national digital soil map for use in precision agriculture, Advances in Animal Biosciences, 8, 430–432, https://doi.org/10.1017/S2040470017000966, 2017.

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., and Rossiter, D.: 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, 2021.

Qin, D., Ding, Y., and Mu, M. (Eds.): Climate and environmental change in China: 1951–2012, Springer, Berlin; Heidelberg, 152 pp., https://doi.org/10.1007/978-3-662-48482-1, 2016.

Ramcharan, A., Hengl, T., Nauman, T., Brungard, C., Waltman, S., Wills, S., and Thompson, J.: Soil Property and Class Maps of the Conterminous United States at 100-Meter Spatial Resolution, Soil Sci. Soc. Am. J., 82, 186–201, https://doi.org/10.2136/sssaj2017.04.0122, 2018.

Shangguan, W., Dai, Y., Liu, B., Ye, A., and Yuan, H.: A soil particle-size distribution dataset for regional land and climate modelling in China, Geoderma, 171–172, 85–91, https://doi.org/10.1016/j.geoderma.2011.01.013, 2012.

Shangguan, W., Dai, Y., Liu, B., Zhu, A., Duan, Q., Wu, L., Ji, D., Ye, A., Yuan, H., Zhang, Q., Chen, D., Chen, M., Chu, J., Dou, Y., Guo, J., Li, H., Li, J., Liang, L., Liang, X., Liu, H., Liu, S., Miao, C., and Zhang, Y.: A China data set of soil properties for land surface modeling, J. Adv. Model. Earth Syst., 5, 212–224, https://doi.org/10.1002/jame.20026, 2013.

Shangguan, W., Dai, Y., Duan, Q., Liu, B., and Yuan, H.: A global soil data set for earth system modeling, J. Adv. Model. Earth Sy., 6, 249–263, https://doi.org/10.1002/2013MS000293, 2014.

Shi, G. and Shangguan, W.: A China dataset of soil properties for land surface modeling (version 2), National Tibetan Plateau/Third Pole Environment Data Center [data set], https://doi.org/10.11888/Terre.tpdc.301235, 2024.

Shi, G. and Shangguan, W.: shgsong/CSDLv2: A China dataset of soil properties for land surface modeling (version 2, CSDLv2), Zenodo [code], https://doi.org/10.5281/zenodo.14783774, 2025.

Shi, G., Shangguan, W., Zhang, Y., Li, Q., Wang, C., and Li, L.: Reducing location error of legacy soil profiles leads to improvement in digital soil mapping, Geoderma, 447, 116912, https://doi.org/10.1016/j.geoderma.2024.116912, 2024.

Shiri, J., Keshavarzi, A., Kisi, O., Iturraran-Viveros, U., Bagherzadeh, A., Mousavi, R., and Karimi, S.: Modeling soil cation exchange capacity using soil parameters: Assessing the heuristic models, Comput. Electron. Agr., 135, 242–251, https://doi.org/10.1016/j.compag.2017.02.016, 2017.

Song, X.-D., Wu, H.-Y., Ju, B., Liu, F., Yang, F., Li, D.-C., Zhao, Y.-G., Yang, J.-L., and Zhang, G.-L.: Pedoclimatic zone-based three-dimensional soil organic carbon mapping in China, Geoderma, 363, 114145, https://doi.org/10.1016/j.geoderma.2019.114145, 2020.

Sun, Y., Ma, J., Zhao, W., Qu, Y., Gou, Z., Chen, H., Tian, Y., and Wu, F.: Digital mapping of soil organic carbon density in China using an ensemble model, Environ. Res., 231, 116131, https://doi.org/10.1016/j.envres.2023.116131, 2023.

Thompson, J. A., Kienast-Brown, S., D'Avello, T., Philippe, J., and Brungard, C.: Soils2026 and digital soil mapping – A foundation for the future of soils information in the United States, Geoderma Regional, 22, e00294, https://doi.org/10.1016/j.geodrs.2020.e00294, 2020.

Viscarra Rossel, R. A., Chen, C., Grundy, M. J., Searle, R., Clifford, D., and Campbell, P. H.: The Australian three-dimensional soil grid: Australia's contribution to the GlobalSoilMap project, Soil Res., 53, 845, https://doi.org/10.1071/SR14366, 2015.

Xia, C. and Zhang, Y.: Comparison of the use of Landsat 8, Sentinel-2, and Gaofen-2 images for mapping soil pH in Dehui, northeastern China, Ecol. Inform., 70, 101705, https://doi.org/10.1016/j.ecoinf.2022.101705, 2022.

Yamashita, N., Ohnuki, Y., Iwahashi, J., and Imaya, A.: National-scale mapping of soil-thickness probability in hilly and mountainous areas of Japan using legacy and modern soil survey, Geoderma, 446, 116896, https://doi.org/10.1016/j.geoderma.2024.116896, 2024.

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., and Pavelsky, T. M.: MERIT Hydro: A High-Resolution Global Hydrography Map Based on Latest Topography Dataset, Water Resour. Res., 55, 5053–5073, https://doi.org/10.1029/2019WR024873, 2019.

Yan, F., Shangguan, W., Zhang, J., and Hu, B.: Depth-to-bedrock map of China at a spatial resolution of 100 meters, Sci. Data, 7, 2, https://doi.org/10.1038/s41597-019-0345-6, 2020.

Yang, J., Guan, X., Luo, M., and Wang, T.: Cross-system legacy data applied to digital soil mapping: A case study of Second National Soil Survey data in China, Geoderma Regional, 28, e00489, https://doi.org/10.1016/j.geodrs.2022.e00489, 2022.

Zhang, Z., Ding, J., Zhu, C., Wang, J., Ge, X., Li, X., Han, L., Chen, X., and Wang, J.: Historical and future variation of soil organic carbon in China, Geoderma, 436, 116557, https://doi.org/10.1016/j.geoderma.2023.116557, 2023b.

Zhou, Y., Xue, J., Chen, S., Zhou, Y., Liang, Z., Wang, N., and Shi, Z.: Fine-Resolution Mapping of Soil Total Nitrogen across China Based on Weighted Model Averaging, Remote Sensing, 12, 85, https://doi.org/10.3390/rs12010085, 2019a.

Zhou, Y., Hartemink, A. E., Shi, Z., Liang, Z., and Lu, Y.: Land use and climate change effects on soil organic carbon in North and Northeast China, Sci. Total Environ., 647, 1230–1238, https://doi.org/10.1016/j.scitotenv.2018.08.016, 2019b.