Science Bulletin xxx (xxxx) xxx

Contents lists available at ScienceDirect

Science Bulletin

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

Article

Mapping high resolution National Soil Information Grids of China

Feng Liu [a,b], Huayong Wu [a], Yuguo Zhao [a,b], Decheng Li [a], Jin-Ling Yang [a,b], Xiaodong Song [a], Zhou Shi [c], A-Xing Zhu [d,e], Gan-Lin Zhang [a,b,f,*]

[a] State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China
[c] Institute of Agricultural Remote Sensing and Information Technology Application, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China
[d] Key Laboratory of Virtual Geographic Environment of Ministry of Education, Nanjing Normal University, Nanjing 210023, China
[e] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[f] Key Laboratory of Watershed Geographic Science, Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, Nanjing 210008, China

ARTICLE INFO

Article history:
Received 9 February 2021
Received in revised form 9 August 2021
Accepted 19 August 2021
Available online xxxx

Keywords:
Predictive soil mapping
Soil-landscape model
Machine learning
Depth function
Large and complex areas
Soil spatial variation

ABSTRACT

Soil spatial information has traditionally been presented as polygon maps at coarse scales. Solving global and local issues, including food security, water regulation, land degradation, and climate change requires higher quality, more consistent and detailed soil information. Accurate prediction of soil variation over large and complex areas with limited samples remains a challenge, which is especially significant for China due to its vast land area which contains the most diverse soil landscapes in the world. Here, we integrated predictive soil mapping paradigm with adaptive depth function fitting, state-of-the-art ensemble machine learning and high-resolution soil-forming environment characterization in a high-performance parallel computing environment to generate 90-m resolution national gridded maps of nine soil properties (pH, organic carbon, nitrogen, phosphorus, potassium, cation exchange capacity, bulk density, coarse fragments, and thickness) at multiple depths across China. This was based on approximately 5000 representative soil profiles collected in a recent national soil survey and a suite of detailed covariates to characterize soil-forming environments. The predictive accuracy ranged from very good to moderate (Model Efficiency Coefficients from 0.71 to 0.36) at 0–5 cm. The predictive accuracy for most soil properties declined with depth. Compared with previous soil maps, we achieved significantly more detailed and accurate predictions which could well represent soil variations across the territory and are a significant contribution to the GlobalSoilMap.net project. The relative importance of soil-forming factors in the predictions varied by specific soil property and depth, suggesting the complexity and non-stationarity of comprehensive multi-factor interactions in the process of soil development.

1. Introduction

Soil is being highlighted in the global agenda, for example in the United Nations Sustainable Development Goals, with at least nine out of the 17 goals directly related to soil use and management. Soil is fundamental for global and regional issues such as food security, land degradation, water cycling, biodiversity, carbon sequestration, and ecosystem health. Detailed, accurate, and up-to-date soil information is urgently needed to aid in developing solutions for these issues and to support decision making concerning natural resource management [1,2]. However, soil is highly heterogeneous in geographical space. Its spatial variation has traditionally been presented as polygon maps of soil classes (technically, choropleth maps) usually at coarse scales, where mapping units are shown as polygons. Most existing soil information has been generated from historical soil surveys decades ago using the traditional soil survey mapping paradigm [3,4]. It is spatially coarse and out of date.

The GlobalSoilMap.net project plans to make a global digital soil map using digital soil mapping techniques through the contribution of soil scientists around the world. It specifies predictions of soil properties at a 90 m spatial resolution and for depth layers 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm. The targeted soil properties include organic carbon (SOC), pH, cation exchange capacity (CEC), bulk density (BD), coarse fragments (CF), available water capacity, electrical conductivity and soil texture fractions [5]. Efforts have been made on 90 m resolution predictive mapping

* Corresponding author.
E-mail address: glzhang@issas.ac.cn (G.-L. Zhang).

https://doi.org/10.1016/j.scib.2021.10.013
2095-9273/

Please cite this article as: F. Liu, H. Wu, Y. Zhao et al., Mapping high resolution National Soil Information Grids of China, Science Bulletin, https://doi.org/10.1016/j.scib.2021.10.013

studies of one or more soil properties over several countries recently, including Australia (7.62 million km^2), France (0.67 million km^2), Chile (0.76 million km^2), the United States (9.37 million km^2), and China (9.6 million km^2) [6–10]. Denmark (0.04 million km^2) even made 30 m resolution national maps of soil texture [11]. Besides, coarser (250 to 5000 m) resolution maps were explored at national (Brazil with 8.54 million km^2) [12], continental (Africa with 30.2 million km^2, Europe with 10.16 million km^2) [13,14], and global [15] extents.

Despite these efforts, the gap between soil information demand and availability is still very large. For large areas with complex soil landscapes where there is usually a limited number of sparse soil survey points, how to accurately predict soil spatial variation at a high resolution remains a challenging issue. First, almost all the efforts have been based on legacy soil samples without precision geographical positioning. Their reported geographical coordinates may have location errors of much more than 90 m [16]. This makes the legacy samples not appropriate for high resolution predictive soil mapping. Second, environmental covariates are critical for predictive soil mapping. It is difficult to adequately characterize soil-forming environments in complex soil landscapes. Third, it is also difficult for predictive algorithms to adjust flexibly to a variety of soil landscapes and make accurate predictions. Some studies use depth as a covariate in model construction [9,15], but there is some debate about how this strategy performs in comparison to pseudo-three-dimensional mapping [17,18]. Lastly, most models in these efforts cannot straightforwardly estimate prediction uncertainty. A computation-intensive bootstrapping technique [19] has been frequently used, but the resulting uncertainty could be itself highly uncertain when using sparse samples [20]. Although geostatistical models can directly produce prediction variance as a measure of uncertainty, they may not be suitable because of the difficulty in meeting stationarity assumption and calibrating a reliable model for such areas.

China is typical for the above challenges. It has a vast land of over 9.6 million km^2. It covers almost all kinds of thermal conditions including tropical, subtropical, warm temperate, middle temperate, cold temperate zones from south to north and the Qinghai-Tibet Plateau zone, and moisture conditions including humid, semihumid, semiarid, and arid regions from southeast to northwest due to the influence of the southeast monsoon. It includes a variety of geomorphological types and mountainous areas occupy nearly two-thirds of the land. It also has a long history of over 5000 years of agriculture. The spatial overlap of the factors results in the most diverse and complex soil-forming environments in the world. There are 14 soil orders, 138 great groups, and 588 subgroups according to Chinese Soil Taxonomy [21,22] and a similar diversity in other soil taxonomies. But the number of soil survey sites is often very limited due to China's large area and difficult accessibility in many areas.

Therefore, the objective of this study is to develop high resolution national gridded maps of basic soil properties at multiple depths across China, under the constraints of limited number of sparse soil survey points. As part of this process, the environmental controllers of soil spatial variations will also be revealed.

2. Data sources

2.1. Soil data

A systematic soil survey was conducted through the project of National Soil Series Survey and Compilation of Soil Series of China (2009–2019). In this survey, representative soil profiles were selected according to Chinese Soil Taxonomy [22]. They are central concepts of all soil types down to the level of soil series. These representative soil profiles covered various soil-forming environments across China. Soil pits were generally dug to a depth of 1.5–2 m or until a lithic or paralithic contact. The complete soil profile was described and sampled for each survey location by genetic horizon with depth limits determined by soil surveyors, according to standard field soil survey methods [23]. The number of sampling depths per profile ranged from only 1 to 12, with a mean value of 4.3 and a standard deviation value of 1.4. The maximum depth of profiles ranged from 5 to 750 cm, with a mean value of 117 cm and a standard deviation value of 92 cm. The geographical coordinates of survey locations were recorded using a handheld GPS receiver. Fig. 1 shows the locations of the 4844 soil profiles. Samples were taken to laboratory and air-dried at room temperature and then passed through a 2 mm sieve. Soil pH, SOC, total nitrogen (TN), total phosphorus (TP), total potassium (TK) and CEC were measured in laboratory. BD was measured using undisturbed samples taken by a standard ring cut. The volume percentages of CF in soil horizons were visually estimated in the field. The CF includes broken bedrock, saprolite, alluvial stones, coarse sand and secondary concretions (e.g., iron manganese nodules and lime concretion). Soil thickness was observed in the field, which is defined as the upper limit of non-soil materials in which the volume of CF (>2 mm) is greater than 75%. Table S1 (online) lists measurement methods of the soil properties. They are all important basic properties that associated with soil physical, chemical, and biological processes, on which the assessment of soil functions can be based.

Table S2 (online) lists the summary statistics of soil properties. All soil properties had a wide range, resulting from the diverse bioclimatic conditions and soil landscapes across the territory. For example, soil pH values at 0–5 cm depth ranged from 3.1 to 10.4 with a mean value of 6.9 and a standard deviation of 1.5, covering levels of strongly acidic, acidic, weakly acidic, neutral, weakly calcareous, calcareous, and strongly alkaline. The mean values of CF, BD, TK, and pH increased with increasing depth while those of SOC, TN, TP, and CEC decreased with increasing depth. Most soil properties were highly skewed except pH and BD, and thus prior to modeling the sample data were log-transformed for SOC, TN, TP, square-root transformed for CEC, TK, thickness, and cube-root transformed for CF. Fig. S1 (online) shows the histograms of soil properties at 0–5 cm depth before and after the transformations.

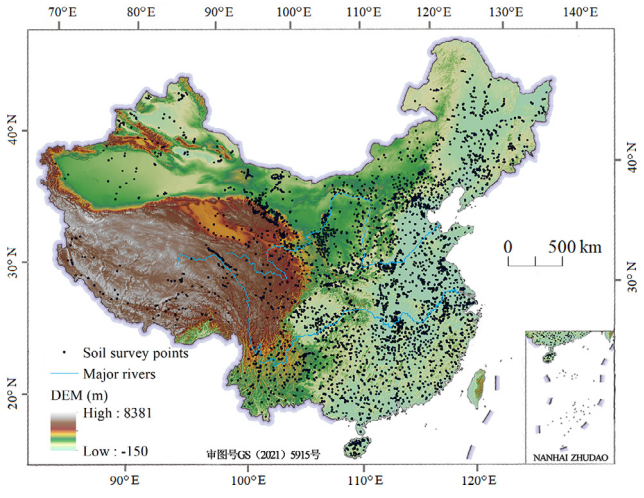


Fig. 1. Locations of soil survey profiles points.

2.2. Environmental covariates

Soil formation is the result of interactions of climate, parent material, topography, vegetation, and human activity over time. We selected the environmental covariates which are associated with these factors and processes. Considering that multi-factor comprehensive interactions exert their influence on soil formation mainly through water and heat, we also included land surface moisture and thermal conditions in the list of environmental factors. We removed redundant environmental variables if their Pearson correlation coefficient values with other variables were greater than 0.82. Table S3 (online) lists the covariates which were used for soil spatial prediction in this study, grouped by the soil-forming factor each represents.

The climatic variables covering the period 1970–2000 at a resolution of 30 arc-second were obtained from WorldClim (https://www.worldclim.org/data/worldclim 21.html). Soil parent material was represented by 30 m resolution Landsat 8 ETM + band 7 (shortwave infrared at 2.08–2.35 μm) and a clay mineral ratio (band 5 / band 7) which are both designed for surficial lithology and minerals detection. They were complemented by 90 m resolution regolith thickness [24] and wind effect. The former is related to the balance between weathering (accumulation) and erosion (removal) and can to a large extent reflect spatial difference of parent materials. The latter, i.e., topographic exposure to wind, can reflect wind contributes to parent materials, especially in deserts, semi-deserts, and extensive loess-affected areas. It was computed in SAGA GIS (http://www.saga-gis.org).

Topographic variables were computed based on a 90 m digital elevation model (DEM) of the Shuttle Radar Topographic Mission (http://srtm.csi.cgiar.org/srtmdata/) using the SAGA GIS. Vegetation and land use conditions were represented by 30 m resolution Landsat 5 TM band 3, band 4 and the mean and standard deviation of normalized difference vegetation index (NDVI) during the period 2000–2017. The mean represents an average vegetation status while the standard deviation represents seasonality and is related to land cover and cropping rotation patterns.

Land surface moisture conditions were represented by 30 m resolution shortwave infrared band 5, band 7 and normalized difference water index (NDWI) [25] over 2000–2017. Thermal conditions were represented by 1 km resolution seasonal mean land surface temperatures (LST) computed from 8-day composite MODIS LST data over 2002–2017 (http://modis.gsfc.nasa.gov). All the covariates with national coverage were resampled to a raster cell size of 90 m by bilinear interpolation.

3. Methods

The theory of soil-environment relations [26,27] contends that the same environmental conditions develop the same soil. By detailed characterization of soil-forming environment, it is expected that fine differences in soil properties can be identified, even with a limited number of soil observations in large and complex area of interest. Based on this idea, we designed a methodological framework for mapping high resolution National Soil Information Grids of China (Fig. 2). It considers a soil property (S) and its prediction uncertainty (U) at a geographical location to be a function (f) of its environmental factors (E) at that location (Eq. (1)):

$$S\&U <= f(E). \quad (1)$$

The E represents environmental covariates obtained by geographical information system and remote sensing techniques. The function f represents model structure and parameters. The structure is determined once an algorithm is specified. The parameters are determined through model calibration and optimization based on soil samples. The symbol <= represents the process of implementing soil predictions, i.e., applying the function f over geographical space in a high-performance parallel computing environment. The S & U are outputs including the predicted soil property map and its associated uncertainty map.

3.1. Generating soil samples at a set of standard depths

Equal-area quadratic splines are commonly used to fit a continuous depth function based on the properties measured by genetic horizons [28]. The splines are mainly applicable to a vertically gradual soil variation. However, soil property variation along a profile often includes abrupt changes in reality. To reduce fitting errors, we developed an automatic procedure for adaptive fitting. It identifies abrupt changes using a threshold of the ratio of higher value to lower value of two neighbouring horizons. This threshold was set to 1.225 in this study. For the situation with an abrupt change, a thin layer with 1 cm thickness is added between the neighbouring horizons before the spline fitting. This forces the fitted curve to respect profile morphology. The Spline Tool (https://www.asris.csiro.au/methods.html) does not provide such a procedure although it briefly illustrates the solution of adding thin layers in its "readme" file. Fig. 3 shows examples of the fitting for four SOC profiles. Original equal-area quadratic splines work well for a gradually changing profile (see blue line in Fig. 3a) but not for profiles with abrupt changes (see blue lines in Fig. 3b–d). The adptive fitting procedure gives better results (see red lines in Fig. 3b–d) in this situation. We derived from the fitted curve the mean values of each soil property for six depth layers 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm. They were taken as the standardized sample data for the following soil predictions.

3.2. Predicting soil properties and estimating local uncertainties

Based on the soil samples, quantile regression forest [29], an ensemble tree-based machine learning model, was constructed to model the relationships between each soil property and the environmental covariates at each depth layer. The algorithm was chosen for three reasons. First, few predictive soil mapping studies have tested this algorithm. Second, it directly estimates prediction uncertainty, and the uncertainty estimation may be more accurate and interpretable than that made by regression kriging, especially for areas with sparse samples [30]. Third, it can deal with complex non-linear relations and multivariate interactions and has high predictive power [31].

This algorithm grows an ensemble of trees as in standard random forest algorithm [32]. For regression, the prediction of a single tree $T(\theta)$ in the forest for a new data point $X = x$, i.e., the estimate of the conditional mean of response variable Y given covariate $X = x$, is obtained by averaging over original observed values Y_i in leaf $l(x, \theta)$ (Eq. (2)):

$$\hat{Y}_{\text{tree}}(x) = \sum_{i=1}^{n} w_i(x, \theta)\ Y_i. \quad (2)$$

The weight vector $w_i(x, \theta)$ is given by a positive constant if observation X_i is part of leaf $l(x, \theta)$ and 0 if it is not, with the sum of weights equal to one. The θ is random parameter vector that determines how a tree is grown (e.g., which covariates are considered for splitting at each node).

The prediction of the forest is approximated by the averaged prediction of k single trees and is then given by a weighted sum over all original observations (Eq. (3)). The weight vector $w_i(x)$ is the average of $w_i(\theta)$ over the k single trees.

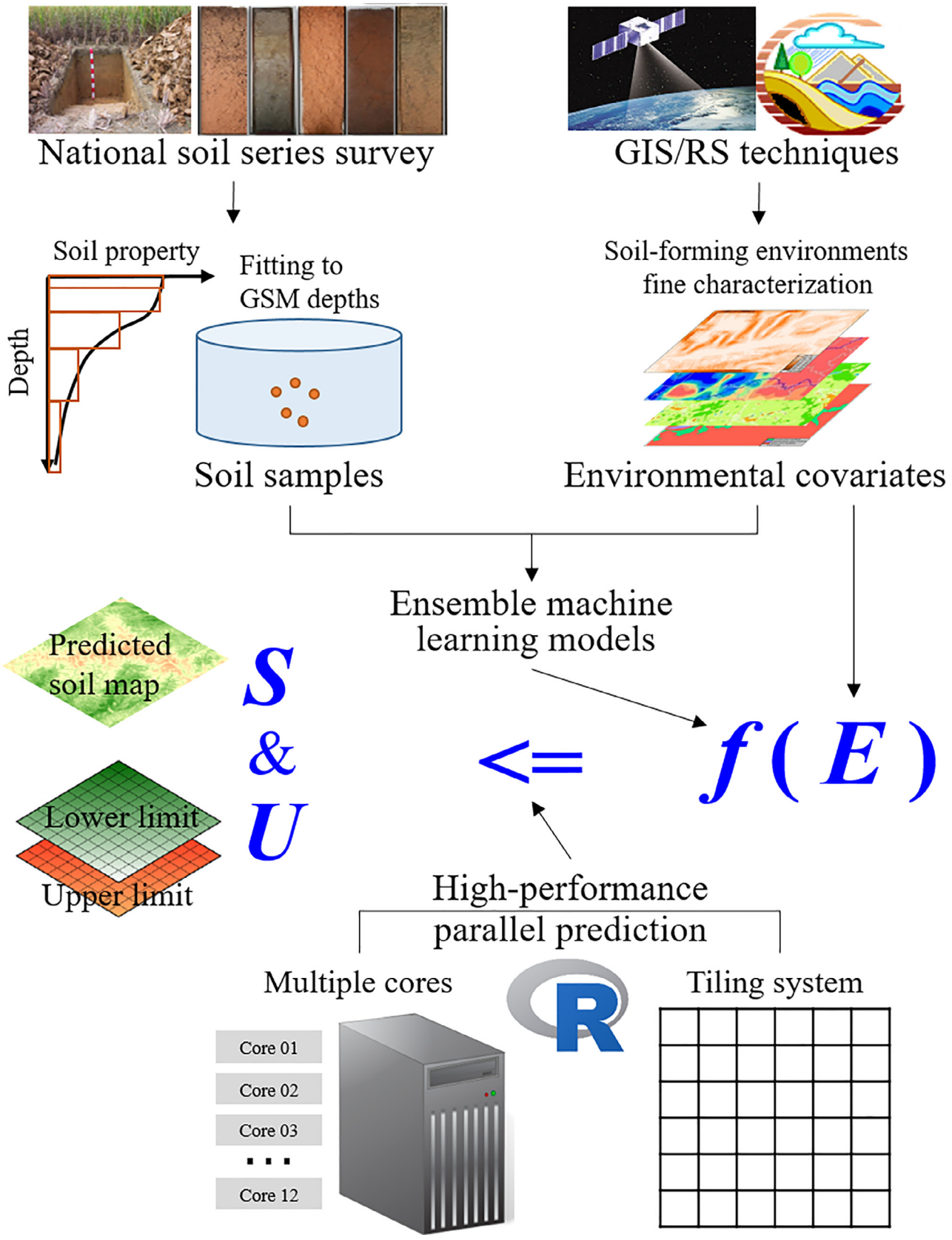


Fig. 2. Methodological framework for mapping high resolution National Soil Information Grids of China.

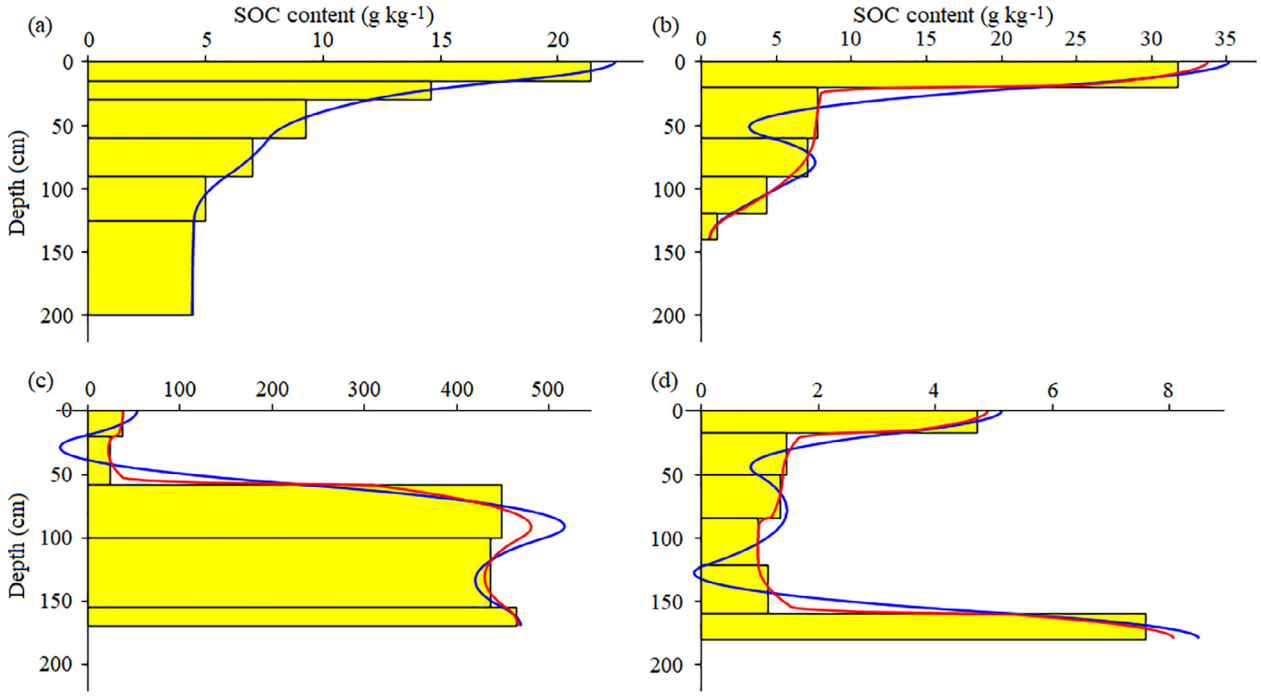


Fig. 3. Examples of splines fittings for four soil organic carbon (SOC) profiles. Blue lines represent the fittings using original equal-area quadratic splines, and red lines represent the fittings using the adaptive fitting procedure.

$$\hat{Y}_{\text{forest}}(x) = \sum_{i=1}^{n} w_i(x)\ Y_i. \tag{3}$$

It has been shown that the weighted observations used for estimating the conditional mean also provide a good approximation of the full conditional distribution. The conditional distribution function of Y, given $X = x$, is thus given by

$$\hat{F}(y|X = x) = \sum_{i=1}^{n} w_i(x)\ 1_{\{\ Y_i \leqslant y\}}\ , \tag{4}$$

where $1_{\{\ Y_i \leqslant y\}}$ is an indicator function. It is 1 if the condition is true and 0 otherwise. The estimation of the conditional distribution consists of three steps: (1) Grow k trees as in random forests, but for every leaf of every tree, take note of all observations in this leaf, not just their averages. (2) For a given $X = x$, drop x down all trees, compute the weight $w_i(x, \theta)$ of observation for every tree and then the weight $w_i(x)$ for every observation in the forest. (3) Compute the estimate of the distribution function for all y using Eq. (4).

For a continuous distribution function, the α-quantile $Q_\alpha(x)$ is defined such that the probability of Y being smaller than $Q_\alpha(x)$ is, for a given $X = x$, exactly equal to α. With the above estimation of the conditional distribution, the conditional quantiles $Q_\alpha(x)$ are derived using Eq. (5):

$$Q_\alpha(x) = \inf\{y :\ F(y|X = x) \geqslant \alpha\}. \tag{5}$$

Thus, the algorithm can derive the quantile of any α value in addition to the mean, as does standard random forest. For details of this algorithm, please see Ref. [29].

There are three important parameters in the model: number of variables used to train each tree (*mtry*), minimum number of terminal nodes (*nodesize*) and number of trees to be generated (*ntree*). We used the caret package [33] to optimize the *mtry* and *nodesize*, and then used optimal parameter values (Table S4 online) to construct final model for each soil property and depth. It is often not necessary to fine-tune the *ntree*. Its default value of 500 is usually sufficient to yield stable predictions.

Using the complete grid of environmental covariates as inputs, the trained models were applied over space to generate national gridded soil property maps with a resolution of 90 m at the depths 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm. The uncertainty of soil property predictions was simultaneously estimated at every pixel and depth, which was expressed as upper and lower limits of 90% prediction interval. The limits were identified using the 0.05 and 0.95 quantiles of empirical distribution. The prediction interval at the confidence level reports the range of values within which the true value is expected to occur 9 times out of 10 [34]. To facilitate comparison, we further calculated a ratio of the prediction interval to the median (i.e., 0.5 quantile), and used the ratio as an uncertainty index [35]. The bigger the ratio for a pixel, the higher the uncertainty of prediction at the pixel would be. In addition, for the data-transformed soil properties, their back-transformations were not performed directly on final predictions of a forest. For a pixel location, we input its environmental data into the trained forest to obtain the values in each leaf of each tree. We then did back-transformation, e.g., exponentiation, on the values at each leaf. From that we calculated mean and quantiles as outputs of the forest for this pixel. That is, the transformed values were used to build the trees, but once they were built we went back to the original scale. The transformed data is expected to lead to more balanced trees and more model stability compared to the original skewed data.

In order to analyze the controlling environmental factors on spatial variations of soil properties, we obtained the relative importance of covariates from the trained models. The relative importance was estimated based on the increase in mean square error (i.e., %IncMSE) when a covariate is randomly permuted. The bigger the increase, the more important is the covariate.

3.3. Performing high-performance parallel computing

The prediction grid covers the territory as a 90 m horizontal resolution raster, in an Albers Equal-area projection system with standard parallels 25° and 47°N and a central meridian of 105°E. Due to

large area and fine resolution, prediction computation was made on approximately 1.2 billion pixels for each soil property at a depth. There were 72 GB data of covariates as model inputs. This demands a large amount of computation resources, which is prohibitive for an ordinary computer. We constructed a high-performance parallel computing environment of national geographical space for implementing the soil predictions. It employed two Lenovo ThinkStation P700 workstations, each having two Intel Xeon E5-2643V3 CPUs with 12 cores and 64 GB RAM (random access memory). The territory was divided into 107 rectangle tiles for parallel computation, each 400 km × 400 km. The open source R programming language environment (http://www.r-project.org/index.html) was used with packages "ranger" [36] for model construction, "caret" [33] for model optimization, "snowfall" [37] for parallel computation, and "rgdal" [38] and "ggplot2" [39] for data processing and visualization.

3.4. Evaluation criteria

The 10-fold cross validation method was used to evaluate accuracy performance of the predictive mapping of each soil property and depth. Statistics including Model Efficiency Coefficient (MEC) [40], root mean square error (RMSE), and mean error (ME) were calculated for the evaluation. The predicted soil property maps of this study were compared with three previous soil map datasets. The first is the 250 m resolution SoilGrids250m developed by Ref. [15] with predictive mapping methods. The second is the 1 km resolution Soil Characteristics dataset made by Ref. [41] with a polygon linkage method (http://globalchange.bnu.edu.cn). The last is the Harmonized World Soil Database (HWSD) derived soil property maps using a soil type linkage method (http://www.fao.org/soils-portal/data-hub/soil-maps-and-databases/harmonized-world-soil-database-v12/en/). The HWSD was developed through merging regional and national soil data across the world [42]. The improvement of our predictions relative to a previous soil map was calculated based on the MEC and RMSE respectively using the following equations:

$$RI_{MEC} = \frac{MEC_{new} - MEC_{previous}}{MEC_{previous}}, \quad (6)$$

$$RI_{RMSE} = \frac{RMSE_{previous} - RMSE_{new}}{RMSE_{previous}}, \quad (7)$$

where RI_{MEC} and RI_{RMSE} are relative improvement with regard to MEC and RMSE respectively, MEC_{new} and $RMSE_{new}$ are accuracy statistics for our predictions, and $MEC_{previous}$ and $RMSE_{previous}$ are accuracy statistics for a previous soil map.

In addition, a prediction interval coverage probability (PICP) was calculated for each soil property and depth to evaluate prediction uncertainty. This is the proportion of observations that are included within the corresponding prediction interval [34]. If the uncertainty estimates have been reasonably defined, the PICP should be close to 0.90 for a 90% prediction interval.

4. Results and discussion

4.1. Predictive performance

Table 1 lists 10-fold cross validation results for the soil property predictions of our study at multiple depths. Model performance varied with specific soil properties. Soil pH was predicted with the best accuracy, MEC = 0.71–0.72 at the depths less than 15 cm. Over 70% of pH variation was explained, and there was good agreement between the predicted and the observed values. This is in line with several studies which reported the best prediction accuracy for pH among soil properties [7,15]. SOC content was predicted with good accuracy, MEC = 0.54–0.55 at the depths less than 15 cm, i.e., about 55% of SOC variation explained. This is substantially better than the SOC prediction accuracy reported by Ref. [7] for France, Ref. [9] for the United States and Ref. [8] for Chile. Soil thickness was predicted with moderate accuracy (MEC = 0.49), which was much better than the prediction accuracy (MEC = 0.11) reported by Ref. [7] in the French national soil predictions. Ref. [43] found it difficult to establish reliable predictive models for soil thickness in Australia-wide predictions. The TN, CEC, BD, TP, TK, and CF contents were predicted with moderate accuracy (MEC = 0.36–0.48) at the depths less than 15 cm, i.e., 36%–48% of soil property variations explained. The accuracy was comparable to that of the studies of Refs. [7–9].

Model performance also varied with depth. The prediction accuracy of SOC, TN, and BD decreased substantially with increasing depth while CEC, CF, and TK decreased slightly with increasing depth. Such decline in accuracy has been observed by previous SOC prediction studies [7,8,43]. A major reason is that most covariates mainly characterize surface environmental conditions and thus have relatively weaker relationships with the deeper soil layers. In contrast, the prediction accuracy of pH and TP content slightly increased with increasing depth. This may be partly because the two properties could be more stable at subsurface layers in a broad scale and thus respond more stably with regional covariates. This is similar to the result of Ref. [8] which showed an overall increase of prediction accuracy of pH with increasing depth across Chile, but most studies [7,43] reported an opposite pattern. In addition, almost all our predictions were overall unbiased with ME values close to zero.

4.2. Comparisons with the existing soil map datasets

Table 1 also lists 10-fold cross validation results for previous soil map datasets. Compared with them, our predictions achieved remarkable accuracy improvement at almost all depths. Specifically, relative to the SoilGrids250m [15], our prediction of pH had accuracy improvement of 11%–15% by MEC and 14%–18% by RMSE, and the predictions of other properties (SOC, CEC, BD, and CF) had accuracy improvement of 65%–482% by MEC and 8%–28% by RMSE. Relative to the Soil Characteristics dataset [41], our prediction of pH had accuracy improvement of 24%–35% by MEC and 18%–26% by RMSE, and the predictions of other properties (SOC, CEC, TN, TP, TK, BD, and CF) had accuracy improvement of over 124% by MEC and 9%–28% by RMSE. Relative to the HWSD [42], our predictions of soil properties had accuracy improvements of over 135% by MEC and 8%–43% by RMSE. In addition, the ME values indicate that the SoilGrids250m obviously over-estimated SOC content while the Soil Characteristics dataset and HWSD maps under-estimated SOC content. In addition, there were significant differences in spatial details between our predictions and previous soil maps. Taking soil BD at 0–5 cm depth as an example, Fig. 4 shows four BD maps excerpted from our predictions, SoilGrids250m, Soil Characteristics dataset, and HWSD, respectively, in an 85 km × 63 km window (108.51°–109.29°E and 35.42°–35.85°N) located in northern Shaanxi Province. Our map clearly shows BD spatial variation with local landscape patterns, and is much more detailed than other soil maps. Thus, our predictions better represent the spatial variation of soil properties across China than the previous soil datasets.

4.3. Environmental controls of spatial patterns of soil properties

Relative importance of the environmental covariates used in the soil spatial predictions is shown in Fig. 5. Although all covariates contributed to the predictions, their importance was different for

Table 1
Predictive performance of our soil property predictions, SoilGrids250m [15], Soil Characteristics [41] and HWSD [42] using 10-fold cross validation[a)].

	Depth (cm)	Our predictions			SoilGrids250m			Soil Characteristics			HWSD		
		MEC	RMSE	ME	MEC	RMSE	ME	MEC	RMSE	ME	MEC	RMSE	ME
pH	0–5	0.711	0.791	0.001	0.635	0.916	−0.038	0.572	0.966	−0.008	0.152	1.33	−0.179
	5–15	0.724	0.767	0.003	0.652	0.893	−0.066	0.582	0.945	−0.009	0.158	1.31	−0.215
	15–30	0.740	0.730	0.003	0.659	0.878	−0.141	0.567	0.953	−0.050	0.169	1.29	−0.334
	30–60	0.737	0.732	0.002	0.647	0.898	−0.194	0.546	0.983	−0.111	0.211	1.25	−0.188
	60–100	0.736	0.728	0.004	0.641	0.888	−0.184	0.550	0.977	−0.131	0.217	1.21	−0.245
	100–200	0.741	0.720	0.006	0.643	0.879	−0.143	0.573	0.944	−0.084	0.208	1.23	−0.271
SOC	0–5	0.551	14.68	0.446	0.124	20.50	7.94	0.223	20.06	−2.268	0.061	21.25	−6.954
	5–15	0.535	13.53	0.423	0.325	16.33	5.158	0.219	17.69	−2.531	0.070	19.11	−5.427
	15–30	0.442	12.68	0.410	0.227	14.94	2.910	0.151	15.72	−2.159	0.082	16.30	−1.061
	30–60	0.360	12.14	0.396	0.140	14.08	1.906	0.099	14.61	−2.759	0.074	14.65	−2.468
	60–100	0.286	13.59	0.408	0.089	15.36	1.378	0.060	15.60	−2.479	0.084	15.41	−0.717
	100–200	0.237	12.79	0.351	0.057	14.24	2.199	0.088	14.01	−1.369	0.101	13.93	0.440
TN	0–5	0.483	1.057	0.029	—	—	—	0.211	1.353	−0.126	—	—	—
	5–15	0.459	1.031	0.029	—	—	—	0.205	1.285	−0.188	—	—	—
	15–30	0.388	0.914	0.031	—	—	—	0.144	1.105	−0.137	—	—	—
	30–60	0.352	0.813	0.028	—	—	—	0.089	0.982	−0.193	—	—	—
	60–100	0.280	0.771	0.026	—	—	—	0.047	0.888	−0.169	—	—	—
	100–200	0.294	0.731	0.022	—	—	—	0.041	0.852	−0.024	—	—	—
TP	0–5	0.390	0.482	0.016	—	—	—	0.012	0.626	−0.063	—	—	—
	5–15	0.395	0.473	0.015	—	—	—	0.014	0.608	−0.063	—	—	—
	15–30	0.413	0.475	0.011	—	—	—	0.020	0.61	−0.036	—	—	—
	30–60	0.411	0.381	0.010	—	—	—	0.019	0.484	−0.035	—	—	—
	60–100	0.406	0.368	0.013	—	—	—	0.012	0.48	−0.041	—	—	—
	100–200	0.357	0.441	0.017	—	—	—	0.019	0.55	0.027	—	—	—
TK	0–5	0.383	5.474	0.074	—	—	—	0.044	6.83	0.764	—	—	—
	5–15	0.388	5.445	0.067	—	—	—	0.048	6.76	0.639	—	—	—
	15–30	0.384	5.484	0.061	—	—	—	0.047	6.85	0.860	—	—	—
	30–60	0.380	5.583	0.071	—	—	—	0.050	6.93	1.019	—	—	—
	60–100	0.376	5.655	0.079	—	—	—	0.051	7.01	0.935	—	—	—
	100–200	0.348	5.812	0.094	—	—	—	0.077	6.94	0.462	—	—	—
CEC	0–5	0.417	8.437	0.223	0.176	10.07	2.212	0.137	10.28	−3.427	0.048	10.82	−2.957
	5–15	0.426	8.046	0.213	0.176	9.68	−0.011	0.142	9.8	−3.485	0.048	10.36	−2.721
	15–30	0.419	7.667	0.211	0.157	9.30	−0.487	0.144	9.31	−3.459	0.048	9.83	−2.069
	30–60	0.392	7.524	0.233	0.133	9.03	−0.067	0.132	9.03	−3.640	0.064	9.32	−1.953
	60–100	0.348	7.986	0.264	0.115	9.28	0.214	0.122	9.29	−3.635	0.062	9.60	−1.609
	100–200	0.337	8.184	0.250	0.085	9.76	0.913	0.128	9.39	−3.883	0.080	9.63	−1.422
BD	0–5	0.483	0.147	−0.002	0.268	0.185	0.058	0.071	0.204	0.014	0.051	0.237	0.116
	5–15	0.479	0.138	−0.002	0.279	0.172	0.053	0.081	0.187	0.013	0.056	0.213	0.089
	15–30	0.457	0.132	−0.002	0.263	0.160	0.043	0.049	0.181	0.007	0.059	0.182	0.010
	30–60	0.403	0.128	−0.002	0.193	0.158	0.050	0.041	0.166	0.003	0.032	0.174	0.013
	60–100	0.303	0.126	−0.003	0.098	0.158	0.065	0.019	0.152	0.013	0.027	0.167	−0.006
	100–200	0.265	0.126	−0.003	0.057	0.159	0.071	0.002	0.15	0.005	0.024	0.169	−0.019
CF	0–5	0.361	10.29	0.437	0.062	12.51	4.513	0.034	12.68	3.946	0.003	12.88	2.486
	5–15	0.372	11.36	0.484	0.072	13.79	4.011	0.040	14.04	3.974	0.003	14.30	1.655
	15–30	0.351	14.50	0.551	0.085	17.25	2.123	0.034	17.67	2.515	0.002	17.97	−1.105
	30–60	0.331	17.62	0.593	0.112	20.30	−0.189	0.049	21.03	1.657	0.002	21.55	−4.108
	60–100	0.333	19.63	0.576	0.118	22.61	−2.193	0.034	23.68	0.064	0.002	24.03	−6.931
	100–200	0.224	21.05	0.900	0.086	22.85	−0.537	0.019	23.65	−3.493	0.001	23.86	−6.364
Thickness		0.485	57	0.431									

a) SOC: soil organic carbon; BD: bulk density; CEC: cation exchange capacity; CF: coarse fragment content; TN: total nitrogen; TP: total phosphorus; TK: total potassium; "—" represents the soil property not mapped or available from the datasets.

specific soil properties and depths, suggesting the complexity and non-stationarity of multi-factor interaction in the process of soil development.

For pH, the most important covariate was regolith thickness. Less important but still useful covariates were wind speed, MAP (mean annual precipitation), solar radiation, precipitation seasonality, elevation, and slope gradient. This may indicate that parent materials and climate conditions mainly controlled pH spatial pattern at national scale and terrain exerted its influence at local scale. Previous studies mainly reported MAP as the most dominant factor [44]. The climatic covariates were more important at shallow than deep depths.

For SOC content, the most important covariates were NDVI, solar radiation, MAP, and growing season LST, followed by MAAT (mean annual air temperature), Diurnal air temperature range, wind speed, band 5, band 7, elevation, slope gradient, and TWI (topographic wetness index). This indicates that bioclimatic conditions mainly controlled SOC spatial pattern. Solar radiation increased its importance with depth, but other climatic covariates were more important at shallow depths. NDVI mean was more important at shallow depths while NDVI standard deviation was important at almost all depths. Few studies have recognized solar radiation as an important predictor for SOC prediction, but many have reported NDVI, land use, MAP, and MAAT as important predictors [12].

The importance of covariates in TN prediction was similar to that in SOC prediction. For TP content, the most important covariates were solar radiation, wind speed and precipitation seasonality,

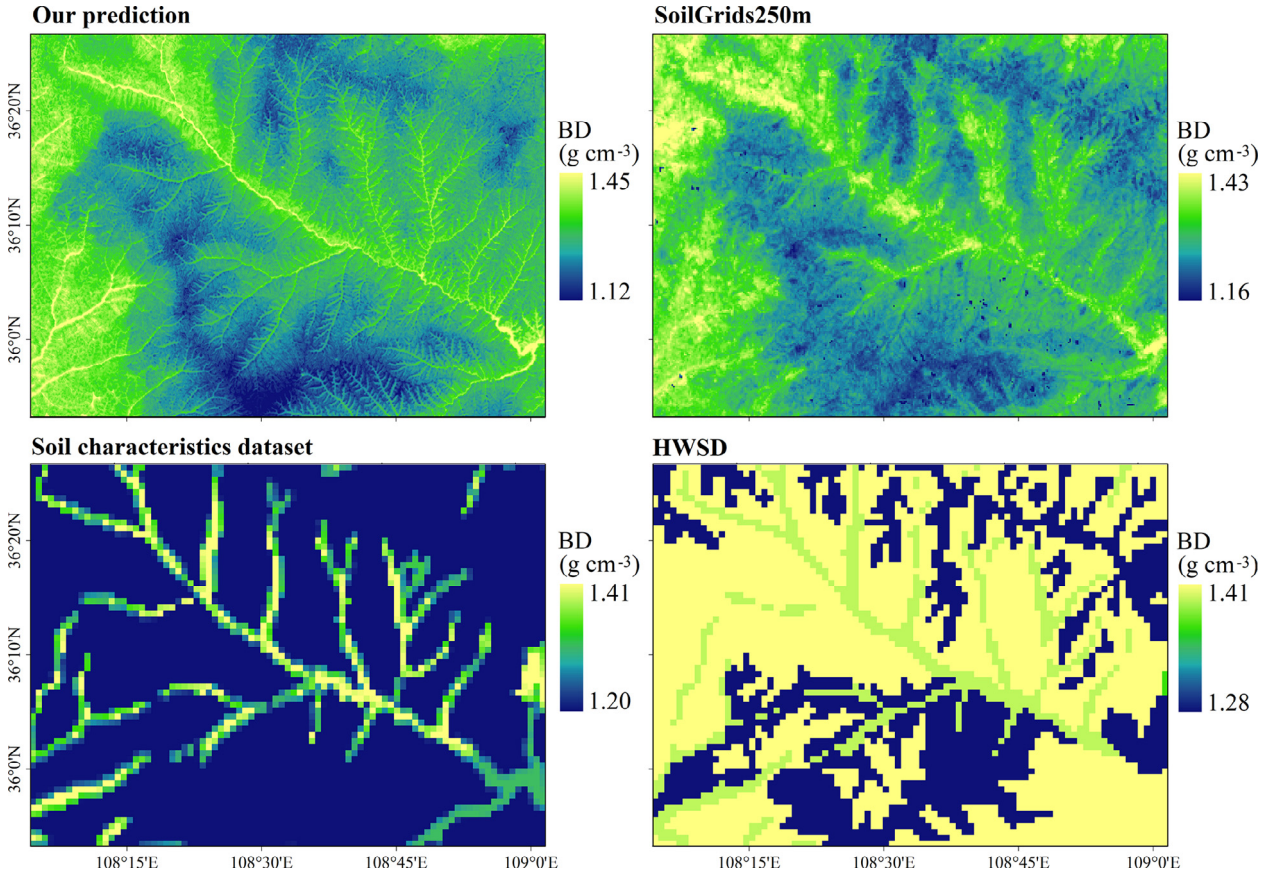


Fig. 4. Surface (0–5 cm) bulk density (BD) maps excerpted from our predictions, SoilGrids250m, Soil Characteristics dataset and HWSD, respectively, in an 85 km × 63 km window (108.51°–109.29°E and 35.42°–35.85°N) located in Shaanxi Province.

followed by TWI, NDVI standard deviation, elevation, regolith thickness, and annual air temperature range. This indicates that climate, terrain, and land uses were dominant factors. For TK content, the most important covariates were solar radiation, diurnal/annual air temperature range, and topographic exposure to wind, followed by wind speed and TWI, indicating that climate and parent materials were dominant factors. NDVI standard deviation, which reflects land use and cropping pattern information, was much more important for TN predictions, less important but useful for TP predictions, and not important for TK predictions. This is similar to the result of Ref. [45] in Renshou County, Sichuan Province.

For CEC, the most important covariates were solar radiation and NDVI standard deviation, followed by elevation, wind speed, band 3, NDVI mean, and regolith thickness. Among them, NDVI and band 3 are related to vegetation growth, land uses, and hence SOC accumulation process, while others are related to weathering intensity, erosion, deposition, and sorting process. Previous studies [46] mostly discussed the relations of CEC with SOC, clay content, and particle diameter.

For BD, the most important covariates were wind speed, solar radiation, elevation, MAP, NDVI standard deviation, daytime LST standard deviation, and June–July LST, followed by summer maximum air temperature, NDVI mean, precipitation seasonality, slope gradient, and TWI. This indicates that evaporation-precipitation balance, terrain, vegetation and land uses were dominant factors. NDVI, daytime June–July LST, summer maximum air temperature, and precipitation seasonality were more important at shallow than deep depths. Ref. [47] reported that land use and MAAT were important for topsoil BD prediction.

For CF content, the most important covariate was regolith thickness, which is associated with the processes of bedrock weathering, erosion and deposition. Less important but still useful covariates were topographic exposure to wind, wind speed, elevation, and slope gradient, which are related to wind, gravity, and water erosions. This indicates that geomorphic processes were a major controller for spatial pattern of CF content. Previous studies [48] reported significant relationships of CF content with slope gradient in local areas. For soil thickness, the most important covariates were regolith thickness and solar radiation, followed by daytime LST, band 7, wind speed, TWI, elevation, and maximum temperature of warmest month. Compared to CF content, moisture and thermal conditions became more important. This may indicate that both geomorphic and pedological processes are major controllers for the pattern of soil thickness.

4.4. The predicted soil spatial patterns and their local uncertainties

Fig. 6 shows the predicted maps of soil pH, SOC, TN, TP, TK, CEC, BD, and CF at 0–5 cm depth. Figs. S2–S9 (online) show their three-dimensional distributions. The spatial patterns of these soil properties had distinct characteristics and varied with depths, revealing high soil heterogeneity in three-dimensional geographical space.

The pH exhibits a gradually increasing trend from southeast to northwest. Alkaline soils (>7.0) are predicted to occur in the northwest and north with arid climate conditions, and acid soils (<6.5) in the south and mountains of Northeast China. Desert areas are predicted to be extremely alkaline, and mountainous and hilly areas in southeast are predicted to be extremely acid. The pH gradually increases with increasing depth. This is probably because the

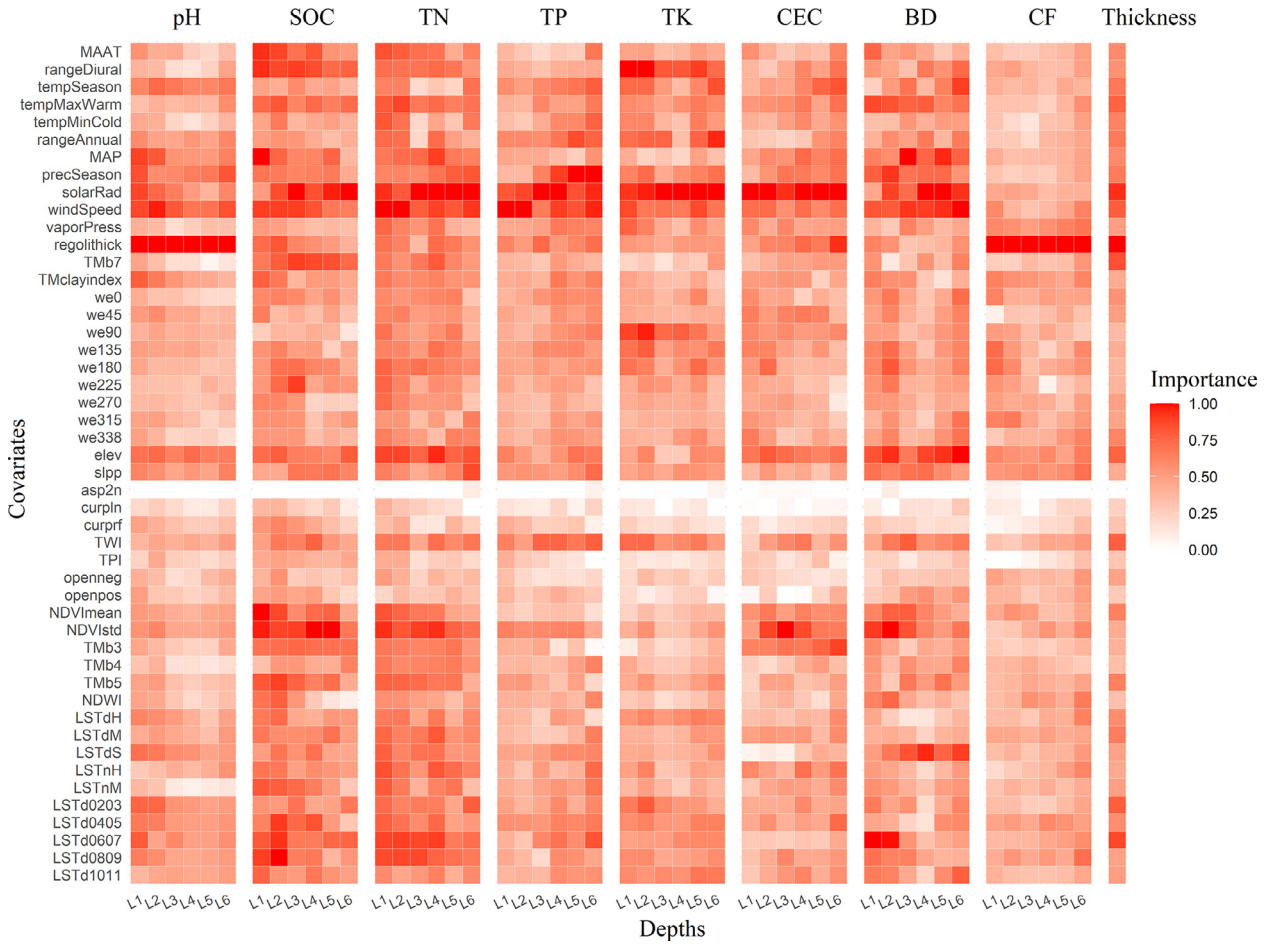


Fig. 5. Relative importance (%IncMSE) of environmental covariates in the spatial predictions of soil pH, soil organic carbon (SOC), total nitrogen (TN), total phosphorus (TP), total potassium (TK), cation exchange capacity (CEC), bulk density (BD), coarse fragments (CF) and soil thickness. L1, L2, L3, L4, L5 and L6 represent the depths 0–5, 5–15, 15–30, 30–60, 60–100 and 100–200 cm, respectively.

inputs of acid rains and nitrogen fertilizer have stronger influence at shallow than deep depths in the south and calcium carbonates accumulate at subsurface soils in the northwest and north.

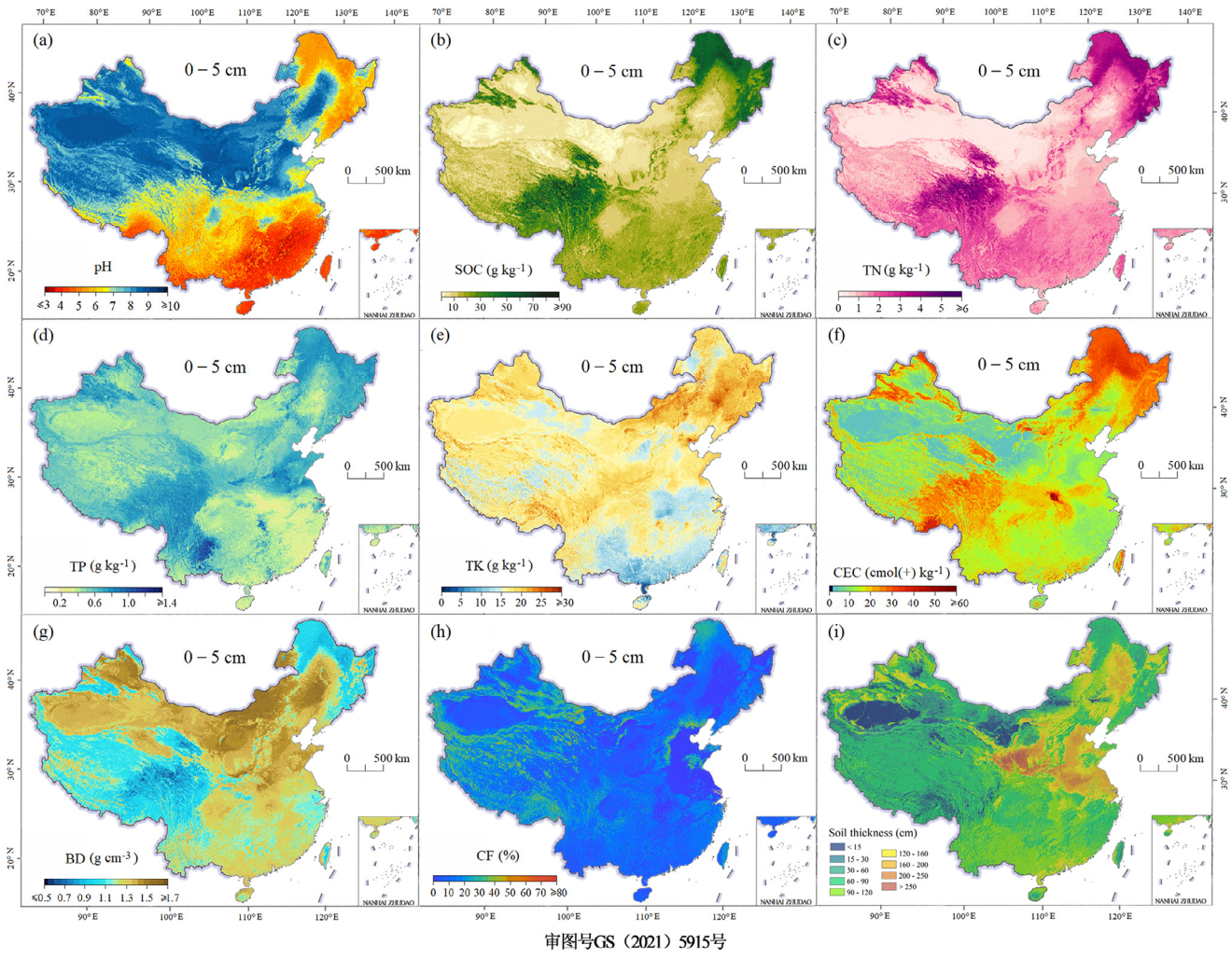
The SOC content is predicted to be highest in the eastern Qinghai-Tibet Plateau, Northeast China and Tianshan Mountains, and lowest in desert areas in the northwest. There is a decreasing trend from southeast to northwest, which is consistent with the influence of the southeast monsoon. In eastern China, the southern part dominated by paddy fields and forests has obviously higher predicted SOC content than the northern part dominated by drylands, especially for the depths less than 30 cm. SOC content rapidly decreases with increasing depth in most areas. The TN content exhibits similar patterns to SOC content, but the decreasing trend from the eastern Qinghai-Tibet Plateau to the east and south is much more gradual.

The TP content is predicted to be low in the south but high in other areas. Sedimentary rocks in Southwest China are rich in phosphorus, so soils derived on these contain relatively higher TP, while in south China soils are normally depleted with TP, as soils are highly weathered and leached [49]. Alpine mountains where large amount of organic matter accumulates are predicted to have relatively high TP content. The TP content is predicted to decrease with increasing depth in most areas, because plant roots absorb phosphorus from subsoil and then return it to the surface in the form of organic residues. Fertilization also increases phosphorus content in tillage layer, as we can see significantly higher content at shallow than deep depths in the North China Plain.

The TK content exhibits an increasing trend from south to north, and is predicted to be relatively high in mountainous areas. As it is mainly associated with weatherable minerals, pedologically younger soils normally have higher TK, while soils in tropical and subtropical monsoon areas have lower TK. The lowest content is predicted to occur in the three southernmost provinces (Hainan, Guangdong, and Guangxi), while the highest content is predicted to occur in Northeast China. TK content increases slightly with increasing depth, probably because mineral potassium in upper depths is easily released and leached.

The CEC exhibits an overall increasing trend from south to north and from west to east. Alpine areas (e.g., the eastern Qinghai-Tibet Plateau) are predicted to have relatively high CEC, mainly due to rich organic matter accumulation. The higher CEC of central China compared to its north and south mainly results from their differences in clay minerals. Relatively low CEC in the southeast is due to high air temperature and rainfall, leading to strong leaching loss of exchangeable substances. The alpine areas show faster decrease of CEC with increasing depth than other areas, due to significant decline of organic matter with depth.

The BD exhibits an overall decreasing trend from north to south. The highest predicted values at 0–5 cm depth occur in the middle part of Inner Mongolia, characterized by arid climate and low soil organic matter content, while the lowest predicted values at the depth occur in the eastern Qinghai-Tibet Plateau, characterized by alpine climate and high soil organic matter content. It increases with increasing depth.



审图号GS（2021）5915号

Fig. 6. The predicted maps of soil properties. (a) pH; (b) soil organic carbon (SOC); (c) total nitrogen (TN); (d) total phosphorus (TP); (e) total potassium (TK); (f) cation exchange capacity (CEC); (g) bulk density (BD); (h) coarse fragments (CF) at 0–5 cm depth; (i) soil thickness.

The CF content is predicted to be high in mountainous areas (e.g., the Qinghai-Tibet Plateau, Greater Khingan Range, and Lesser Khingan Range) and low in plains (e.g., the North China Plain, Northeast Plain, and Middle-Lower Yangtze Plain) and deserts. Most areas at 0–5 and 5–15 cm layers have low predicted CF content. It increases with increasing depth especially in mountainous areas.

Soil thickness is predicted to be biggest in the Loess Plateau and North China Plain, followed by the Northeast China Plain, lower Yangtze Plain and Pearl River Delta Plain, and smallest in deserts and high mountain ridges. The soils are predicted to be much thicker in the north than in the south, and also much thicker in the east than in the west (Fig. 6i).

Table S5 (online) shows that all PICP values are higher than 90%, indicating reliable uncertainty estimations for the predictions of soil properties and depths. It was found that different soil properties had distinct spatial patterns of prediction uncertainty but different depths of the same property had similar patterns. Fig. S10 (online) shows maps of uncertainty for soil pH, SOC, and BD predictions at depths 0–5 and 30–60 cm as an example. High prediction uncertainty of pH occurs in Southwest China, that of SOC in eastern Qinghai-Tibet Plateau and Greater and Lesser Khingan Ranges, and that of BD in almost all alpine areas. Most of these areas had sparse samples in complex soil landscapes.

4.5. Insights for soil prediction over large and complex areas

Large and complex areas are generally characterized by strong multi-factor interaction, nonlinearity, and nonstationary, leading to highly heterogeneous soils over space. Through this study, we recognized several important aspects of this relevant for making accurate and detailed soil prediction in such areas.

First, the predictive algorithm should be flexible and robust. Here "flexibility" means that the algorithm can model local soil-environment relationships, while "robustness" means that it can give reasonable predictions for new situations, even in the absence of sufficient information. Ensemble learning has been demonstrated to be helpful for improving robustness. Recent studies report that complex algorithms yield better predictive performance than simple ones [50]. However, according to the Occam's razor principle, one should not make a model more complex than necessary. It should be noted that highly complex models generally need large number of training data and can be sensitive to overfitting.

Second, environmental covariates play a critical role in being able to reveal soil spatial variation with a limited number of soil samples. Considering the diversity of soil landscapes in such areas, it would be better to characterize soil-forming environments from various angles, for example, remote sensing from visible, near

infrared and shortwave infrared bands. Detailed covariates can more adequately represent environmental conditions and thus may lead to more accurate predictions [51]. But, there are studies showing an opposite view that the performance of soil prediction models may be improved if covariates are aggregated to larger supports before they are used in the models [52]. This issue would be an interesting point for future research.

Finally, the estimation of spatial uncertainty of soil predictions is important in such areas. It can inform us of the reliability of soil predictions, especially for local areas with high heterogeneity or a lack of samples. This information is valuable for end-users to make an enlightened decision in applications of the predicted soil maps [20,53]. But, it should be acknowledged that the presentation and use of uncertainty maps is a field that still needs to be developed.

4.6. Limitations and further improvement

Our soil survey samples are sparse over space, especially in the western China. The on-going second Qinghai-Tibet Plateau soil survey will increase the number of soil samples in the west. Adding more samples may improve the mapping, but it would be better to add more samples in the feature space with high uncertainty than in the geographic space with high uncertainty.

Soil formation is a long process. The environmental covariates in this study mainly characterized current soil-forming environmental conditions, which in many cases are different from the historical environmental conditions under which pedogenesis really took place [54]. Moreover, current covariates may not adequately characterize the factors that shape soil spatial patterns at deep depths. It is an important issue to identify new covariates which can reflect historical process and soil differences.

Soil formation is a complex process. The ensemble tree-based model in this study is just an empirical simplification of soil formation mechanism. It may only partly and implicitly model the interaction and comprehensive effects of soil-forming factors. Moreover, machine learning is data-driven and depends completely on the data it has to make prediction, leading to a risk of generating unrealistic results in areas without samples in geographic and/or feature spaces. Knowledge based on the experience of soil survey experts could be a useful complement for the data-driven method. It would be an interesting topic to explore their integration in the future. In addition, it is necessary to develop new predictive mapping strategies to deal with the imbalance of samples distribution in geographic and feature spaces.

Applying standard cross-validation for which hyperparameters have been optimized using the same data may yield over-optimistic validation results. Ref. [55] argued that standard cross-validation does not suffice for adequate model evaluation and presented a nested cross-validation method. Also, it should be noted that the standardization of horizon data using the spline fittings was not error-free, but due to the lack of a "true" depth function (vertically intensive samples) of each soil profile, the standardization error could not be quantitatively estimated and taken into account.

4.7. Potential applications of high resolution National Soil Information Grids

The high resolution soil property maps developed in this study may find their wide applications in soil, agriculture, hydrology, ecology, climate, environment and forensic science. There are several major aspects.

First, soil monitoring and management. Our dataset can serve as a baseline against which to assess soil spatiotemporal changes and identify the underlying driving factors. It provides strong support for national and regional soil resource management and strategic decision-making.

Second, soil function and threat assessment. With our dataset, soil functions (e.g., nutrient storage, water infiltration, productivity, biodiversity) and threats (e.g., organic matter decline, acidification, erosion, pollution) can be quantitatively assessed in a spatial and detailed way to ensure a secured and healthy soil in the future.

Third, land surface processes modelling. The lack of detailed and accurate soil information has long been a bottleneck in the modelling of land surface processes. Our dataset is promising to fix the problem and improve the modelling of processes of carbon, water and energy in the Earth surface system.

Fourth, forensic investigation. Soil samples collected from suspects' clothing, footwear and vehicles have been considered as a kind of evidence in forensics. Identifying their places of origin, i.e., soil provenance, has been restricted by the lack of detailed soil spatial information. In this regard, our dataset can provide valuable support.

Fifth, civil engineering. Our dataset is useful for underground pipeline planning and road construction. Soil corrosivity, which is associated with soil pH, is an important concern in the planning of pipeline paths. Spatial distributions of soil physical properties are basic information for highway and railway constructions.

5. Conclusions

This study developed the first version of high resolution National Soil Information Grids of China with limited samples from a recent national soil survey, as an alternative to the existing out-of-date, spatially-coarse national soil maps. It was achieved in such a large area with complex soil landscapes through integrating predictive soil mapping paradigm with adaptive depth function fitting, quantile regression forest machine learning and detailed soil-forming environmental characterization in a high-performance parallel computing environment. It consists of 90 m resolution national gridded maps of a set of key soil properties at multiple depths, which clearly show regional patterns as well as substantial local details. Compared to previous soil map datasets, it is significantly more accurate and detailed and includes local uncertainty information, and can well represent soil spatial variations across the territory. The gridded soil property maps are a contribution to the GlobalSoilMap.net project, which can serve wide applications in soil management, agriculture production, hydrological modeling, ecological construction and climate change mitigation.

We also found that although all soil-forming factors contributed to the shaping of soil spatial patterns, their relative importance varies with specific soil properties and depths. This suggests the complexity and non-stationarity of comprehensive multi-factor interaction in the process of soil development at a national extent. The finding provides an insight for soil evolution modelling and decision making to ensure sustainable development in the future.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by the National Key Basic Research Special Foundation of China (2008FY110600 and 2014FY110200), the National Natural Science Foundation of China (41930754 and 42071072), the 2nd Comprehensive Scientific Survey of the Qinghai-Tibet Plateau (2019QZKK0306), and the Project of "One-Three-Five" Strategic Planning & Frontier Sciences of the Institute

of Soil Science, Chinese Academy of Sciences (ISSASIP1622). We thank the colleagues involved in the project of Chinese Soil Series Survey and Compilation of Chinese Soil Series and all soil surveyors and technical assistants in the survey and laboratory work. We also thank Dr. David G. Rossiter for his great work in improving language expression of the paper, Mr. Longquan Du and Zhenkun Liu for their wonderful work in making maps, and Dr. Kai Pan for his professional help in making the dataset accessible over the Internet. Maps in this article were reviewed by Ministry of Natural Resources of the People's Republic of China (GS(2021)5915).

Author contributions

Feng Liu and Gan-Lin Zhang designed the study, carried out predictive mapping, and wrote the manuscript; Huayong Wu and Decheng Li contributed to data compilation and analysis; Zhou Shi and A-Xing Zhu provided ideas to method development; Yuguo Zhao, Decheng Li, Feng Liu, Huayong Wu, Jinling Yang and Xiaodong Song evaluated the results.

Data availability

The national gridded soil property maps produced in this study are available for download at http://soil.geodata.cn/data/datadetails.html?dataguid=36810085119113 or http://doi.org/10.11666/00073.ver1.db.

Appendix A. Supplementary materials

Supplementary materials to this article can be found online at https://doi.org/10.1016/j.scib.2021.10.013.

References

[1] Food and Agriculture Organization of the United Nations, Intergovernmental Technical Panel on Soils. Status of the World's Soil Resources (SWSR)–Technical Summary. 2015.
[2] Zhang GL, Wu HY. From "Problems" to "Solutions": soil functions for realization of Sustainable Development Goals. Bull Chin Acad Sci 2018;33:124–34.
[3] China Soil Survey Office. Techniques of China soil survey. Beijing: China Agricultural Press; 1992.
[4] Shi XZ, Yu DS, Warner ED, et al. Soil database of 1:1000000 digital soil survey and reference system of the Chinese genetic soil classification system. Soil Surv Horizons 2004;45:129–36.
[5] Arrouays D, Grundy MG, Hartemink AE, et al. Globalsoilmap: toward a fine-resolution global grid of soil properties. Adv Agron 2014;125:93–134.
[6] Rossel RAV, Chen C, Grundy MJ, et al. The Australian three-dimensional soil grid: Australia's contribution to the GlobalSoilMap project. Soil Res 2015;53:845–64.
[7] Mulder VL, Lacoste M, Richer-de-Forges AC, et al. Globalsoilmap France: high-resolution spatial modelling the soils of France up to two-meter depth. Sci Total Environ 2016;573:1352–69.
[8] Padarian J, Minasny B, McBratney AB. Chile and the Chilean soil grid: a contribution to GlobalSoilMap. Geoderma Reg 2017;9:17–28.
[9] Ramcharan A, Hengl T, Nauman T, et al. Soil property and class maps of the conterminous United States at 100-meter spatial resolution. Soil Sci Soc Am J 2018;82:186–201.
[10] Liang Z, Chen S, Yang Y, et al. National digital soil map of organic matter in topsoil and its associated uncertainty in 1980's China. Geoderma 2019;335:47–56.
[11] Adhikari K, Kheir RB, Greve MB, et al. High-resolution 3-D mapping of soil texture in Denmark. Soil Sci Soc Am J 2013;77:860–76.
[12] Gomes LC, Faria RM, de Souza E, et al. Modelling and mapping soil organic carbon stocks in Brazil. Geoderma 2019;340:337–50.
[13] Hengl T, Heuvelink GBM, Kempen B, et al. Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions. PLoS One 2015;10:e0125814.
[14] Heuvelink GBM, Kros J, Reinds GJ, et al. Geostatistical prediction and simulation of European soil property maps. Geoderma Reg 2016;7:201–15.
[15] Hengl T, de Jesus JM, Heuvelink GBM, et al. Soilgrids250m: global gridded soil information based on machine learning. PLoS One 2017;12:e0169748.
[16] Grimm R, Behrens T. Uncertainty analysis of sample locations within digital soil mapping approaches. Geoderma 2010;155:154–63.
[17] Nauman TW, Duniway MC. Relative prediction intervals reveal larger uncertainty in 3D approaches to predictive digital soil mapping of soil properties with legacy data. Geoderma 2019;347:170–84.
[18] Ma Y, Minasny B, McBratney AB, et al. Predicting soil properties in 3D: depth as a covariate? Geoderma 2021;383:114794.
[19] Heuvelink G. Uncertainty quantification of GlobalSoilMap products. In: Arrouays D, McKenzie N, Hempel J, editors. GlobalSoilMap: basis of the global spatial soil information system. Leiden: CRC Press; 2014. p. 327–32.
[20] Lagacherie P, Arrouays D, Bourennane H, et al. How far can the uncertianty on a Digital Soil Map be known: a numerical experiment using pseudo values of clay content obtained from Vis-SWIR hyperspectral imagery. Geoderma 2019;337:1320–8.
[21] Gong ZT, Huang JR, Zhang GL. Soil geography of China. Beijing: Science Press; 2014.
[22] Cooperative Research Group on Chinese Soil Taxonomy. Keys to Chinese Soil Taxonomy. 3rd ed. Hefei: Press of University of Science and Technology of China; 2001.
[23] Zhang GL, Gong ZT. Soil survey laboratory methods. Beijing: Science Press; 2012. p. 8–23.
[24] Yan FP, Shangguan W, Zhang J, et al. Depth-to-bedrock map of China at a spatial resolution of 100 meters. Sci Data 2020;7:2.
[25] Gao B-C. NDWI—a normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens Environ 1996;58:257–66.
[26] Jenny H. Factors of Soil Formation: A System of Quantitative Pedology. New York: McGraw-Hill; 1941. p. 1–20.
[27] McBratney AB, Mendonça Santos ML, Minasny B. On digital soil mapping. Geoderma 2003;117:3–52.
[28] Bishop TFA, McBratney AB, Laslett GM. Modelling soil attribute depth functions with equal-area quadratic smoothing splines. Geoderma 1999;91:27–45.
[29] Meinshausen N. Quantile regression forests. J Mach Learn Res 2006;7:983–99.
[30] Vaysse K, Lagacherie P. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma 2017;291:55–64.
[31] Gyamerah SA, Ngare P, Ikpe D. Probabilistic forecasting of crop yields via quantile random forest and Epanechnikov Kernel function. Agric For Meteorol 2020;280:107808.
[32] Breiman L. Random forests. Mach Learn 2001;45:5–32.
[33] Kuhn M, Johnson K. Applied Predictive Modeling. New York: Springer; 2013.
[34] Malone BP, McBratney AB, Minasny B. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. Geoderma 2011;160:614–26.
[35] Poggio L, de Sousa LM, Batjes NH, et al. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil 2021;7:217–40.
[36] Wright MN, Ziegler A. ranger: a fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw 2017;77:1–17.
[37] Knaus J, Porzelius C, Binder H, et al. Easier parallel computing in R with snowfall and sfCluster. R J 2009;1:54–9.
[38] Bivand RS, Pebesma E, Gómez-Rubio V. Applied spatial data analysis with R. 2nd ed. New York: Springer; 2013.
[39] Wickham H. ggplot2: elegant graphics for data analysis. 2nd ed. New York: Springer; 2016.
[40] Krause P, Boyle DP, Base F. Comparison of different efficiency criteria for hydrological model assessment. Adv Geosci 2005;5:89–97.
[41] Shangguan W, Dai Y, Liu B, et al. A China data set of soil properties for land surface modeling. J Adv Model Earth Syst 2013;5:212–24.
[42] Food and Agriculture Organization of the United Nations, International Institute for Applied Systems Analysis, International Soil Reference and Information Centre, Institute of Soil Science Chinese Academy of Sciences, Joint Research Centre of the European Commission. Harmonized World Soil Database (version 1.2). 2012.
[43] Henderson BL, Bui EN, Moran CJ, et al. Australia-wide predictions of soil properties using decision trees. Geoderma 2005;124:383–98.
[44] Gardi C, Yigini Y. Continuous mapping of soil pH using digital soil mapping approach in Europe. Eurasian J Soil Sci 2012;2:64–8.
[45] Gao XS, Xiao Y, Deng LJ, et al. Spatial variability of soil total nitrogen, phosphorus and potassium in Renshou County of Sichuan Basin, China. J Integr Agric 2019;18:279–89.
[46] Shiri J, Keshavarzi A, Kisi O, et al. Modeling soil cation exchange capacity using soil parameters: assessing the heuristic models. Comput Electron Agric 2017;135:242–51.
[47] Taalab KP, Corstanje R, Creamer R, et al. Modelling soil bulk density at the landscape scale and its contributions to C stock uncertainty. Biogeosciences 2013;10:4691–704.
[48] Chen HS, Liu JW, Wang KL, et al. Spatial distribution of rock fragments on steep hillslopes in karst region of northwest Guangxi, China. Catena 2011;84:21–8.
[49] Xiong Y, Li QK. Soils of China. 2nd ed. Beijing: Science Publishing House; 1987.
[50] Keskin H, Grunwald S, Harris WG. Digital mapping of soil carbon fractions with machine learning. Geoderma 2019;339:40–58.

[51] Samuel-Rosa A, Heuvelink GBM, Vasques GM, et al. Do more detailed environmental covariates deliver more accurate soil maps? Geoderma 2015;243-244:214–27.
[52] Behrens T, Schmidt K, MacMillan RA, et al. Multiscale contextual spatial modelling with the Gaussian scale space. Geoderma 2018;310:128–37.
[53] Poggio L, Gimona A, Brewer MJ. Regional scale mapping of soil properties and their uncertainty with a large number of satellite-derived covariates. Geoderma 2013;209-210:1–14.
[54] Heuvelink GBM, Webster R. Modelling soil variation: past, present, and future. Geoderma 2001;100:269–301.
[55] Pejović M, Nikolić M, Heuvelink GBM, et al. Sparse regression interaction models for spatial prediction of soil properties in 3D. Comput Geosci 2018;118:1–13.

Feng Liu is an associate professor at the Institute of Soil Science, Chinese Academy of Sciences. His research interest focuses on the development of state-of-the-art predictive (digital) soil mapping techniques to reveal soil variations over time and space, and understand the relationships between soil and environments.

Gan-Lin Zhang is a professor at the Institute of Soil Science, Chinese Academy of Sciences and Nanjing Institute of Geography & Limnology, Chinese Academy of Sciences. His research fields include pedology, soil classification, soil mapping, soil resource assessment, and Earth critical zone.